Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Datová analýza diskuze o znečištění plasty na Twitteru

magisterská diplomová práce

Autor: Bc. Robert Jaworek

Vedoucí práce: Mgr. Vladimír Matlach

Olomouc

2018

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Datová analýza diskuze o znečištění plasty na Twitteru“ vypracoval samostatně a uvedl jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne 20. 08. 2018 Podpis

Na tomto místě bych rád poděkoval Mgr. Vladimíru Matlachovi za jeho vstřícné, velmi podnětné a trpělivé vedení mé magisterské diplomové práce. Děkuji také rodičům za jejich podporu během mého studia. A Péti za podporu vůbec.

Abstrakt

Název práce: Datová analýza diskuze o znečištění plasty na Twitteru

Autor práce: Bc. Robert Jaworek

Vedoucí práce: Mgr. Vladimír Matlach

Počet stran a znaků: 130 stran & 159 009 znaků

Počet příloh: 1

Abstrakt: Tato diplomová práce se zabývá dolováním textových dat z mikroblogovací sítě Twitter. Teoretická část obsahuje teoreticko-historický úvod do vědecké disciplíny Digital Humanities. Dále práce v rámci teoretické části obsahuje pojednání o teorii a historickém vývoji dolování textových dat, teorie sítí a dolování sociálních médií. Poté následují kapitoly věnované teorii Latentní Dirichletovy alokace a postojové analýzy. Teoretická část je ukončena představením enviromentálního problému znečištění plasty. V praktické části je popsáno krok za krokem dolování textových dat z mikroblogovací sítě Twitter a jejich následná analýza pomocí programovacího jazyka R a programu Cytoscape. Analýza se soustředí na objevení skrytých témat v korpusu tweetů a na vztahy mezi těmito tématy. Dále je provedeno dolování sentimentu z tweetů a identifikace entit a jevů, k nimž se vztahují. Dílčí výsledky analýzy jsou průběžně vizualizovány pomocí R a v programu Cytoscape. V závěru práce nabízíme shrnutí výsledků analýzy.

Klíčová slova: datová analýza, digital humanities, dolování sociálních médií, dolování textových dat, postojová analýza, teorie sítí, twitter, znečištění plasty

Abstract

Title: Data analysis of the discussion about plastic pollution on Twitter

Author: Bc. Robert Jaworek

Supervisor: Mgr. Vladimír Matlach

Number of pages and characters: 130 pages & 159 009 characters

Number of appendices: 1

Abstract: The subject of this Master thesis is mining of text mining on the Twitter microblogging site. The theoretical part contains the theoretical and historical introduction to the scientific field of Digital Humanities. Further, the thesis presents information on the theory and the historical development of text mining, network theory and social media mining. The following chapters are dedicated to the theory of Latent Dirichlet allocation and opinion analysis. The final portion of the theoretical part presents the topic of the environmental crisis caused by plastic pollution. The practical part describes step by step the process of text mining of the microblogging site Twitter and their subsequent analysis using the programming language R and Cytoscape program. The analysis focuses on discovering latent topics in the corpus of tweets and the exploration of relations between these topics. Mining is also applied to identify sentiment from tweets, as well as entities and phenomena to which they refer. Partial results of the analysis are continuously visualized with the use of R and Cytoscape. The thesis is concluded with a summary of the analysis results.

Keywords: data analysis, digital humanities, network theory, opinion mining, plastic pollution, social media mining, text mining, twitter

Fragments on the sea

Birds of prey above

All that lies beneath

No longer what it was, it changed

Why can't we relate

– Jack Johnson

Obsah

Úvod

Žijeme v době, která překypuje daty. Ta jsou generována nejrůznějšími měřícími přístroji teploměrem začínaje vesmírnými sondami konče. Data ale vytváříme neustále také my, lidé. Vezměme si např. komunikaci s našimi bližními a naším okolím. Téměř každý z nás používá denně mobilní telefon, e-mail nebo např. sociální média. Přičemž nejenom textová data z nich pocházející jsou průběžně ukládána do obsáhlých databází stejně jako knihy nebo např. novinové články. Z databází můžeme poté dolováním získávat cenné informace. Předmětem této práce je dolování dat textových, vycházející z tradice vědní disciplíny Digital Humanities. Konkrétně jsme se budeme věnovat dolování a analýze textových dat pocházejících z mikroblogovací sítě Twitter.

Po krátkém nástinu historického vývoje a povahy Digital Humanities se budeme věnovat charakteristice a historii dolování textových dat až po současnou podobu technik, jež se k tomuto účelu užívají. Zmíníme také teorii sítí, kterou budeme aplikovat v praktické části této práce převážně na vizualizaci textových dat pro nahlédnutí jejich povahy a vztahů mezi nimi. Vzhledem k tomu, že teorie sítí vychází z teorie grafů, předložíme zde rovněž základní pojmy tohoto oboru. Teoretická část bude dále pokračovat představením dolování sociálních médií, do nichž Twitter spadá. Zmíněn bude rovněž jejich vývoj a jejich současné kategorie.

Následně popíšeme způsoby komunikace na mikroblogovací síti Twitter a předložíme také pojednání o jejím vývoji. Další kapitoly představí postupně metody modelování skrytých témat a postojové analýzy, které budeme aplikovat v praktické části. Teoretická část této práce bude zakončena představením ilustrativního tématu znečištění plasty, tedy enviromentálního problému, jímž bude primárně vymezen zkoumaný dataset, na který budou aplikovány explikované metody.

V praktické části této práce budeme užitím programovacího jazyka R dolovat textová data z Twitteru a následně provedeme jejich analýzu. Užijeme při tom metod popsaných v teoretické části a převážně vlastního kódu, který jsme za účelem dolování a analýzy napsali. Budeme se věnovat objevení skrytých témat ve specializovaném korpusu zpráv z Twitteru (*tweety*) týkajících se znečištění plasty. Dílčí výsledky analýzy

budeme ilustrovat grafy zhotovenými v R a v programu Cytoscape. Dále provedeme automatickou analýzu sentimentu obsaženého v *tweetech* a zjistíme, jaké postoje jimi pisatelé vyjadřují. V závěru práce přineseme shrnutí výsledků naší analýzy.

1 Nástin vymezení Digital Humanities

We worry that the humanities are in a death spiral,
but they'll rise from the ashes…[1]

Dříve než se pokusíme o stručné vymezení Digital Humanities, pokládám za rozumné předložit definici humanitních věd vůbec. Skrze předmět jejich zájmu je spolu s Helen Small chápu takto: *„Humanitní vědy studují významotvorné praktiky lidské kultury, jejich minulost a přítomnost. Zaměřují se na jejich interpretaci a kritické zhodnocení, a to především z individuálního hlediska, avšak bez subjektivního omezování.“*[2] Předmětem svého zkoumání a převážně též svými metodami se tradičně odlišovaly od věd přírodních.

S příchodem třetí průmyslové revoluce, označované rovněž jako digitální, se však významný díl významotvorných praktik lidské kultury přesunul na web. Technologický vývoj umožnil humanitním vědám akvizici některých metod náležejících dříve jen vědám přírodním, a zároveň jim nabídl výpočetní kapacitu pro zpracování dříve nepředstavitelného množství dat. Minimálně bez druhého by nebyl vznik Digital Humanities možný.

První vlna humanitních věd, jež se zabývaly zpracováním a interpretací dat vytvářených lidmi, se vzedmula v pozdních čtyřicátých letech dvacátého století. Cílila primárně na sestavování korpusů a databází, jež by ulehčily práci vědcům.[3] Za nejvýznačnějšího pionýra v této oblasti je pokládán jezuita Roberto Busa, který v roce 1949 přišel s projektem počítačového zpracování kompletního díla Tomáše Akvinského. V letech 1974–1980 byl projekt ve spolupráci s firmou IBM realizován. Dílo bylo nejprve

[1] Domingos, 2015, s. 278.
[2] 2012, s. 23.
[3] Budrick, Drucker, Lunenfeld, Presner & Schnapp, 2012, s. 8.

přepsáno do děrných štítků a následně byla počítačem vytvořena úplná konkordance.[4] Dnes je celý Index Thomisticus volně dostupný na internetu.[5]

Tou dobou se však ještě pro podobné podniky běžně užíval termín „Humanities Computing". Až teprve v devadesátých letech jej Johanna Drucker, John Unsworth, Jerome McGann a další z *Institute for Advanced Technology in the Humanities* při *University of Virginia* nahradili termínem: „Digital Humanities".[6] S nástupem webu pak začaly vznikat studijní programy, výzkumná centra a akademické žurnály nově vzniklého oboru.[7] Databázové nástroje položily napříč celým světem základy prvních projektů Digital Humanities, které sjednocovaly především zájmy o textovou analýzu, katalogizaci a studium lingvistických prvků. Výzkumné otázky poháněla analýza strukturovaných dat[8], tj. dat, která již předem někdo roztřídil a přidělil jim význam. V pozdních devadesátých letech pak začaly první projekty na základě sběru digitálních dat předkládat např. vizualizace či komplexní síťové analýzy převážně textů.[9]

Příchod Digital Humanities s sebou přinesl nutnost reinterpretace humanitních věd, jejichž výstupy již nejsou pouze texy, ale stávají se jimi též obrazy, crossmediální korpusy či software.[10] Výzkum je nevyhnutelně čím dál více zprostředkováván digitálními technologiemi[11] a využívá metod pro vizualizaci a analýzu dat, síťovou analýzu či zpracování obrazu za účelem objevení nových úhlů pohledu na humanitní texty a nových způsobů, jak k nim přistupovat.[12] V tomto kontextu bývá často jmenována digitalizace tištěných knih nebo depozitářů galerií výtvarných umění. V současné době již ale vznikají také mnohem ambicióznější projekty jako kupříkladu Rome Reborn, jehož cílem je vytvoření trojrozměrných digitálních modelů reprezentujících pomocí virtuální reality vývoj starověkého Říma od pozdní doby bronzové až po ranný středověk.[13]

[4] Busa, 1980.
[5] http://www.corpusthomisticum.org/it/index.age, cit. 07. 03. 2018.
[6] Hayles, 2012, s. 43.
[7] Ibid., s. 42.
[8] Budrick, Drucker, Lunenfeld, Presner & Schnapp, 2012, s. 8.
[9] Ibid., s. 9.
[10] Ibid., s. 10.
[11] Berry, 2012, s. 1.
[12] Ibid., s. 161.
[13] http://romereborn.squarespace.com/, cit. 07. 03. 2018.

Výzkumná oblast Digital Humanities se však neomezuje pouze na digitalizování již vzniklých kulturních artefaktů, výraznou měrou se zabývá také digitálně generovanými obsahy nových médií, skrze něž čím dál více komunikujeme. Spolu s Levem Manovichem je lze definovat jako: *„[…] grafy, pohyblivé obrazy, zvuky, tvary a text, který je spočitatelný, tzn. jednoduše další soubor počítačových dat.“*[14] Pakliže nástup počítačů umožnil vznik Digital Humanities, nová média exponenciálně rozšiřují množství dat, s nimiž lze pracovat. Výzkumné aktivity zahrnují rovněž studium digitálních technologií po jejich kulturní stránce, možnosti jejich kreativního využití a jejich společenský impakt.[15]

Digitální povaha dat s sebou ale přináší pro humanitní vědce také úskalí v podobě nutnosti osvojení si dovedností a výzkumných metod náležejících tradičně výzkumníkům z oblasti přírodních věd. V počátcích tohoto nového oboru spolupracovali humanitní vědci s informatiky, programátory apod., případně se učili nezbytným dovednostem sami.[16] V současné době již ale univerzity napříč celým světem nabízejí stále častěji studijní programy, jejichž absolventi by měli disponovat jak humanitním vzděláním, tak znalostí programování.

Nemenším úskalím je vymezení Digital Humanities, o něž se zainteresovaní autoři pokoušejí. Přestože nacházejí průniky v nástrojích, metodách a předmětech svých výzkumů, předložit jednotnou definici svého oboru se jim nedaří. Gardiner a Musto se dokonce ptají, zdali není termín Digital Humanities redundantní.[17] Záměrně staví následně do kontrastu jeho definici z Wikipedie[18] s vymezením Budrickové, Druckerové, Lunenfelda, Presnera & Schnappa. Zatímco první z nich zdůrazňuje synergii komputačních, humanitních a sociálních věd, druhá vyzdvihuje trans-historický a trans-mediální přístup k poznání, jenž si klade otázku: co znamená býti člověkem v dnešním propojeném informačním věku?[19] Sami se přiklání k definici druhé. Nicméně debata

[14] 2001, s. 43.
[15] Schreibman, Siemens, Unsworth, 2016, s. xvii.
[16] Hayles, 2012, s. 59.
[17] 2015, s. 18.
[18] „Digital humanities,“ n.d.
[19] 2012, s. 82.

o vymezení Digital Humanities neustále probíhá, a protože pro účely této práce postačuje její výše předložený nástin, tak ji v tomto bodě opustíme.

Již od svých prvopočátků se tento obor zaobíral texty, kterým věnuje značnou pozornost i dnes; a jejichž zpracování je jedním z hlavních témat této práce. V následující kapitole proto představíme metodu, díky které jsme schopni objevit a extrahovat informace z textů, tj. dolování textových dat.

2 Dolování textových dat

It is estimated that
80 % of the world's data is unstructured...[20]

Dolování textových dat (Text Mining) bývá zpravidla pomocí počítače aplikováno na korpus textů za účelem objevování a extrahování nových, předem neznámých informací.[21] Zastřešuje širokou škálu nástrojů a technik, ke kterým se řadí dolování dat, strojové učení, zpracování přirozeného jazyka, umělá inteligence, shluková analýza, vytěžování vědomostí, textová analýza, komputační a korpusová lingvistika, analýza sentimentu apod.[22]

Odhaduje se, že 80 % veškerých digitálních dat je nestrukturovaných, přičemž jejich největší část tvoří právě text; dalšími jsou fotografie a video. Oproti strukturovaným datům – která jsou většinou numerická a předem roztříděná dle významu – vyžadují ta nestrukturovaná nejprve pečlivou úpravu. Ústřední roli při aplikaci dolování textových dat proto hraje sběr, čištění a transformace dat do takové podoby, z níž jsme schopni vytěžit co nejvíce poznatků. Nestrukturovaná data nám totiž mohou odhalit informace, které bychom jinak nezískali. Ať už jsou to např. emoce obsažené v textu, názory spotřebitelů na nějaký produkt nebo dokonce předzvěsti politických revolucí.[23]

Před započetím samotného dolování textových dat musí být vytvořen korpus textových dokumentů, jehož velikost je omezena pouze pamětí a výpočetní kapacitou počítače. Vzhledem k veliké rozsáhlosti takových korpusů jsou manuální pokusy usouvztažňovat data v nich obsažená, mapovat komplexní vztahy mezi nimi či identifikovat skryté tendence velmi náročné. Aplikace automatických metod jako je dolování textových dat ale výrazně zvyšuje rychlost a efektivitu práce s nimi.[24]

[20] Schneider, 2017, cit. 10. 03. 2018.
[21] Hearst, 2003, cit. 18. 03. 2018.
[22] Lin, 2012, s. 301.
[23] Brandom, 2014, cit. 12. 03. 2018.
[24] Sanger, & Feldman, 2002, s. 2.

2.1 Historie dolování textových dat

Pro pochopení významu současné podoby dolování textových dat je dobré nahlédnout jej v kontextu technologií, které umožnily jeho nástup. Technologický vývoj byl krom jiného popoháněn rychle se rozrůstajícím množstvím textových informací, jež bylo potřeba zpracovat. Jednou z výzev bylo, jak přistupovat ke konkrétním informacím ukrytým v obrovském množství textových dat např. v knihovnách či archivech. Dnes se k tomu běžně používají dvě navzájem se doplňující metody: vyhledávání a extrakce informací.[25] Avšak dříve, než byly vyvinuty, zabývali se výzkumníci jinými otázkami: zaprvé bylo potřeba přijít na to, jak sumarizovat nejrůznější textové dokumenty (vystihnout ve zkratce jejich obsah), aby se s nimi následně dalo dobře pracovat; když byla tato otázka uspokojivě vyřešena, vyvstala přirozeně potřeba klasifikovat nově vzniklou myriádu sumarizací. Řešení vzešla mezi disciplínami knihovnictví, informační vědy a zpracování přirozeného jazyka.[26]

Jedním z prvních systémů sumarizace a klasifikace textů byl knihovní katalog, který v roce 1674 vymyslel Thomas Hyde pro *Bodleyovu knihovnu* při *Oxfordské univerzitě*. Melvil Dewey pak roku 1876 představil kartotéční lístky. Dalším krokem ve vývoji zpracování textu bylo jeho shrnutí pro tvorbu abstraktů. Ačkoli na tomto problému začaly pracovat *Institution of Electrical Engineers* a *Physical Society of London* již v roce 1898, vyřešen byl až H. P. Luhnem na konci padesátých let minulého století. Luhn analyzoval pomocí počítače IBM 701 frekvence slov, z nichž odvozoval jejich relativní míry významnosti. Poté pro každou větu sečetl všechna relativně významná slova a výsledek zkombinoval s lineární vzdáleností mezi slovy, čímž vytvořil měřítko důležitosti věty. Abstrakty dokumentů byly tvořeny těmi nejvýznamnějšími.[27]

Nástup počítačů umožnil elektronické ukládání textu a jeho automatické vyhledávání. Osobní počítače zase vedly k vývoji uživatelských rozhraní s hypertextovými odkazy, s nimiž přišel v roce 1965 při práci na projektu Xanadu Theodor H. Nelson. Poskytovaly téměř okamžitý přístup k celým textům, a ne pouze jejich abstraktům. Konečně Shannonova teorie informace z roku 1948 umožnila vznik

[25] Anglicky: Information Retrieval & Information Extraction.
[26] Miner, Elder, Fast, Hill, Nisbet, & Delen, 2012, s. 4–5.
[27] Ibid., s. 5–6.

informačního toku na internetu stejně jako vznik bibliometrie či vyhledávacích systémů. Budoucí vývoj dolování textových dat byl v těchto raných aplikacích informační vědy pěvně zakořeněn.[28]

Mezi lety 1960 až 1985 dominoval lingvistice, psychologii, umělé inteligenci a zpracování přirozeného jazyka racionalistický přístup hlásající, že výrazná část lidského vědění není odvozena smysly, ale je naopak stanovena geneticky. V lingvistice byl tento přístup zpopularizován Noamem Chomským. V oblasti umělé inteligence lze podle Manninga a Schützeho týž přístup vnímat jako snahu vytvoření inteligentních systémů tím, že se jim nejprve dodá veliké množství znalostí a usuzovací mechanismy, aby poté kopírovaly proces učení se lidského mozku. Oproti tomu empirický přístup, jenž dominoval výše zmíněným oblastem mezi lety 1920 až 1960 předpokládá, že mozek dítěte se začíná učit asociacemi, rozpoznáváním vzorů a zobecňováním, které aplikuje na bohaté senzorické vjemy, jež má k dispozici pro osvojení si přirozeného jazyka.[29]

Rané aktivity dolování textových dat v oblastech knihovnictví, informační vědy a zpracování přirozeného jazyka soustředily svou pozornost primárně na rozličné formy metod vyhledávání informací a jejich sumarizaci v podobě indexů, abstraktů nebo seskupení dokumentů. V pozdějších letech se výzkum přesunul k procesu extrahování informací[30], který se zpravidla skládá ze série kroků navržených k extrakci termínů a jejich atributů, faktů a obsáhlejších událostí.[31]

Pro tyto účely se používají např. metody detekce témat, která jsme schopni objevit napříč velikým množstvím textových dokumentů. Získáme-li takto obecnou představu o tom, co je jejich obsahem, můžeme k jejich hlubšímu poznání dále využít metod detekce pojmenovaných entit. Ilustrovat si to můžeme např. pomocí následující situace: představme si, že máme k dispozici tisíce zákaznických e-mailů, ale nevíme, o čem pojednávají. Rozpoznáme-li v nich kupříkladu téma týkající se produktů, můžeme v nich dále pomocí k tomu určených metod veškeré zmíněné produkty odhalit. Jako další krok se v takovém případě nabízí zjistit, jaké emoce či hodnocení k těmto

[28] Ibid., s. 6–7.
[29] 1999, s. 6–7.
[30] Miner, Elder, Fast, Hill, & Nisbet, 2012, s. 9.
[31] Sanger, & Feldman, 2002, s. 96.

produktům zákaznící vyjadřují. Pro tento účel byly vytvořeny metody detekce sentimentu a postojové analýzy.

Miner, Elder, Fast, Hill a Nisbet nabízejí obecné vymezení dosavadních vývojových fází dolování textových dat[32]:

I. 1674–1960: sumarizace textu
II. 1960–1990: extrahování informací (automatická tvorba abstraktů)
III. 1990–2010: objevování skrytých struktur v textech (shluková analýza)
IV. 2010–současnost: hluboké zkoumání textu (detekce idiomů, sarkazmů apod.)

Moderní analýzy dolování textových dat již nemají za cíl pouze deskripci jazykového materiálu, nýbrž se snaží o predikci jazykových jevů. Příkladem může být překladač od společnosti Google[33]: rozpoznat věty pasné např. anglicky a německy a přeložit je mezi sebou dokázal, ač stylisticky nekvalitně, již před několika lety. Avšak teprve díky nedávné aplikaci neuronových sítí zlepšil kvalitu svých výstupů (predikci jazykových jevů) na takovou úroveň[34], že začíná být užíván pro překlady technických návodů apod. i některými překladatelskými agenturami.

V současnosti posledním vývojovým stupněm dolování textových i netextových dat je strojové učení, jehož cílem je zjištění struktury dat a jejich vlastností. Pomáhá nám např. při určování autorství textu, detekci spamu či strojovém překladu. Mezi jeho nejznámější algoritmy patří: metoda podpůrných vektorů (SVM), Naivní Bayes, K-nejbližších sousedů (k-NN) či Latentní Dirichletova alokace (LDA), kterou využívá pro klasifikaci dokumentů podle témat např. populární platforma pro šíření digitálních publikací Issuu. V centru aktuálního výzkumu a diskuzí pak stojí hlavně neuronové sítě. Tento výpočetní model hlubokého strojového učení, jehož architektura je inspirována lidským mozkem, již dnes čím dál více při řešení úloh nahrazuje výše zmíněné algoritmy.

Nyní, když jsme představili dolování textových dat a ve zkratce nastínili jeho historický vývoj, je záhodné, abychom vymezili data, s nimiž pracuje. Následující řádky

[32] 2012, s. 11–12.
[33] https://translate.google.cz/?hl=cs&tab=wT
[34] Le, & Schuster, 2016, cit. 12. 03. 2018.

proto postupně pojednávají o textu, dokumentu a korpusu jakožto základních seskupeních textových dat.

2.2 Text

Text může mít mnoho podob v závislosti na tom, kdo je jeho původcem a příjemcem, za jakým účelem byl vytvořen, ale také v závislosti na tom, skrze jaké médium je šířen a jak je konzumován. Abychom se v něm dokázali orientovat a aplikovat na něj metody dolování textových dat je proto potřeba jej jasně vymezit. Spolu s Františkem Štíchou v této práci nahlížím text jako: *„[…] spojitý jazykový útvar, který se uplatňuje v komunikaci jako komplexní jednotka sdělování.“*[35] A jehož vlastnosti jsou[36]:

(a) komplexnost, ucelenost, celistvost: tzn. text je tvořen určitým počtem textových jednotek (vět či souvětí);
(b) integrovanost, vnitřní organizovanost, obsahová jednotka: jednotky skládají dohromady uspořádaný celek, jenž zprostředkovává účastníkům komunikace ucelenou informaci;
(c) uzavřenost, ohraničenost: ustálené formule na začátku a konci určitých textových typů;
(d) spojitost a soudružnost: textové jednotky jsou mezi sebou propojeny specifickými způsoby a prostředky;
(e) funkčnost, informativnost: autor vytváří text s určitým záměrem, chce sdělit adresátovi určitou informaci a dosáhnout v komunikaci určitého cíle;
(f) zapojení textu do komunikační situace;
(g) zapojení textu do kontextu jiných textů, na něž navazuje a přejímá z nich určitá kompoziční schémata.

Textem zde chápu text tištěný, pro nějž jsou výše popsané vlastnosti typické. Digitální texty na internetu dnes ale často vykazují rovněž rysy, které bývají zpravidla přisuzovány projevům mluveným. Výrazně lze tento jev pozorovat v internetových diskuzích na sociálních médiích, jež používají jak běžní občané k všednodennímu

[35] 2013, s. 887.
[36] Ibid.

dorozumívání, tak firmy jako oficiální komunikační kanál. Účastníci komunikačních situací se zde často vyjadřují neuspořádaně a neurčitě, užívají významově vyprázdněné vycpávky nebo naopak připojují vsuvky či dodatky. Dále také používají emotikony – specifické grafické ikony sloužící k nahrazení mimiky, gestiky, proxemiky apod. To vše dělá digitální texty na jednu stranu významově značně bohaté, na druhou stranu to ale snižuje jejich čitelnost počítačem.

2.3 Dokument & Korpus

Korpusem rozumím spolu s Františkem Čermákem veliký soubor elektronických textů, jenž představuje relativně objektivní model jazykové empirie a měl by být při svém zpracování ve výchozím smyslu zcela zbaven jakékoliv apriorní teorie a jiných vlivů, které jazyk nutně deformují. Tím, že korpus zobrazuje jazykové jevy v jejich přirozeném kontextu, umožňuje vytvářet na reálných jazykových datech podložený jazykový výzkum, který by byl bez užití počítače nemyslitelný. Jednou z jeho dalších předností je schopnost vypovídat o frekvenci jevů a jejich typickém užití.[37]

Korpus může být jak statický, tak dynamický. V prvním případě se počet dokumentů v něm obsažených nemění jako je tomu například u českého korpusu SYN2015.[38] V případě druhém bývají buď stávající dokumenty v průběhu analýzy aktualizovány, případně je korpus dále rozšiřován o dokumenty nové[39] jako je tomu například u Korpusu současné americké angličtiny.[40]

Korpus se dále dělí na dokumenty, které můžeme definovat jako jednotky diskrétních textových dat, kterými jsou např. knihy či novinové články. V těch se dále snažíme nacházet komponenty, např. kapitoly nebo věty, s co největší informační hodnotou. K počátečnímu rozpoznání potenciálně důležitých komponentů dokumentu

[37] 2017.
[38] Křen et al., 2015, cit. 12. 03. 2018.
[39] Sanger, & Feldman, 2002, s. 2
[40] https://corpus.byu.edu/coca/, cit. 13. 03. 2018.

často slouží typografické prvky jako interpunkce, kapitálky, číslice a speciální znaky: textové proluky, CR[41], podtržení, hvězdičky, tabulky, sloupce apod.[42]

Nejméně přehledné jsou slabě strukturované dokumenty jako např. soukromé zprávy ze sociálních sítí. Oproti tomu dokumenty s konzistentním formátováním jako např. novinové články nebo webové stránky jsou označovány za semistrukturované. Ke zjednodušené reprezentaci dokumentů se užívají modely tvořené podmnožinami jejich vlastností, jež nám usnadňují práci s nimi. Sanger a Feldman takový model ilustrují korpusem 15 000 dokumentů zpravodajské agentury Reuters, který obsahuje více než 25 000 netriviálních slovních kmenů. Podobně vysoká vícerozměrnost je pro korpusy přirozeného jazyka typická.[43]

2.4 Dolování textových dat v praxi

Díky obrovskému množství vědomostí obsažených v textech a tomu, že se při komunikaci spoléháme primárně na textová data, lze si představit aplikaci dolování textových dat takřka v každém oboru. Můžeme rozlišit dva základní druhy dolování textových dat. První z nich nahrazuje manuální práci s texty, a tudíž zvyšuje její produktivitu jako např. automatické třídění e-mailů detekující spam. Nedisponuje však ničím, co by člověk bez užití počítače nezvládl. Oproti tomu druhý způsob umožňuje odhalit informace a syntetizovat je do vzorců, jejichž odhalení by bylo pro lidi časově příliš náročné. Například inteligentní analyzátor biomedicínské literatury by mohl objevit řetězec asociací genů a chorob napříč rozličnými vědeckými články; a tím nabídnout potenciální způsob cílení léků na konkrétní geny při léčbě konkrétní choroby. Podobný výzkum by byl pro lidi bez využití počítačů velmi náročný nejen časově, ale také např. tím, že by museli ovládat veškeré světové jazyky, v nichž byly a jsou relevantní vědecké články psány.[44]

[41] Carriage return – v počítačové terminologii název pro speciální netisknutelný znak, který posune kurzor na začátek stávajícího řádku.
[42] Sanger, & Feldman, 2002, s. 3.
[43] Ibid., s. 4.
[44] Zhai, & Massung, 2016, s. 242–243.

V praxi lze dolování textových dat využit k identifikaci nelegálních příspěvků na sociálních médiích, k měření významnosti a dosahu marketingových kampaní nebo k modelování a predikci chování publik.[45] Textová data vytěžená ze sociálních médií mohou ale také posloužit k optimalizaci rozhodování. Zhai a Massung se zamýšlejí nad inteligentním systémem, který by shromažďoval veškerá textová data určitého typu, řekněme zprávy z mikroblogovací sítě Twitter. Analýza sentimentu obsaženém v tzv. tweetech by dle jejich názoru mohla kupříkladu pomoci pochopit reakce společnosti na politické dění. Analýza tweetů v reálném čase by mohla být prospěšná také pro rychlejší poskytování humanitární pomoci oblastem postiženým přírodními katastrofami.[46] Oproti klasickým médiím jako je tisk či rozhlas je totiž výhodou těch sociálních, že se jimi informace šíří mnohem rychleji.

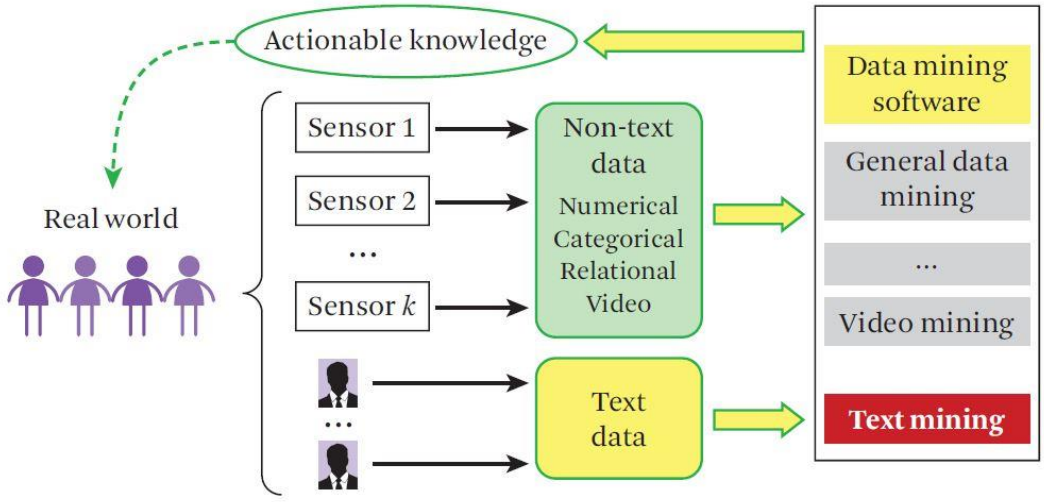
2.5 Textová versus netextová data

Pro účely dolování textových dat je užitečné, soudí Zhai a Massung, pohlížet na ta textová – vytvářená bezprostředně lidmi – jako na data pocházející ze subjektivních senzorů. V rámci této perspektivy bychom mohli přirovnat lidi k senzorům fyzikálním jako je např. teploměr. Každý senzor snímá určité aspekty okolního světa a následně o nich informuje v různých datových podobách. Zatímco teploměr snímá teplotu reálného světa a informuje o ní ve formě stupňů Celsia či ve Fahrenheitech, lidé sledují okolní svět každý svým vlastním pohledem a o tom, co vnímají, referují řečí nebo pomocí textu.[47]

Analogie člověka jako subjektivního senzoru je prospěšná především při výzkumu dat pocházejících z internetu, konkrétně např. ze sociálních sítí. Většinou jsou totiž texty extrahovány spolu s daty netextovými (geolokační údaje, typ zdrojového zařízení apod.). Skloubením analýzy obou typů dat je možno dosáhnout komplexnějšího pohledu na zkoumaný problém, než kdybychom se zabývali pouze textem a numerickými, kategorickými či např. multimediálními daty odděleně.[48] Ostatně žádná data neexistují

[45] Kwartler, 2017, s. 2.
[46] 2016, s. 243–244.
[47] Ibid., s. 244.
[48] Ibid., s. 244.

sama o sobě, ale jsou součástí větších celků, pro jejichž pochopení je vhodné nahlížet je komplexně: např. pomocí teorie sítí, kterou si představíme níže.



Obrázek č. 1: Syntéza analýzy textových a netextových dat

Zdroj: Zhai & Massung 2016. s. 245.

3 Teorie sítí

Síťové analýzy hrají důležitou roli v mnoha průmyslových odvětvích a vědních disciplínách – jak přírodních, tak humanitních i sociálních. Pro pochopení komplexních systému je totiž nezbytné předchozí porozumění sítí stojící v jejich pozadí. Každý z nás je součástí řady systémů, jsme jimi obklopeni den co den. Sítí zde uvažujeme množinu entit (vrcholů), jenž jsou navzájem propojeny určitými vztahy (hranami). Už rodina či síť našich přátel, které jsou poměrně malé, mohou být složité. Avšak např. komunikační infrastruktura zahrnující miliardy mobilních telefonů, počítačů a satelitů na oběžné dráze je daleko složitější. Podobně lidská schopnost uvažovat a rozumět okolnímu světu vyžaduje interakci miliard neuronů v našem mozku.[49]

Lewis nabízí definici teorie sítí jako vědy zkoumající teoretické základy síťových struktur a jejich dynamického chování spolu s aplikacemi v oblastech analýzy sociálních sítí či na poli syntetických emergentních systému jako je internet či elektrická síť a na poli přírodních a humanitních věd.[50]

Počátky teorie sítí sahají do roku 1736, kdy švýcarský matematik Leonhard Euler položil základy teorie grafů, z nichž teorie sítí těží. Obyvatelé města Königsberg (dnešní Kaliningrad), které leží při ústí řeky Pregola, a jehož části byly tehdy propojeny sedmi mosty, dumali nad otázkou, zdali je možné projít celé město, aniž by přešli kterýkoli z mostů více než jednou. Krom toho, že Euler problém vyřešil, použil poprvé termíny – graf, vrchol, hrana, stupeň vrcholu a tah – které jsou jádrem jak teorie grafů, tak dnešní teorie sítí.[51]

Během pozdních šedesátých a sedmdesátých let byla teorie používána také sociálními vědci k modelování sociálních sítí a ke studiu lidského chování v rámci skupin. Jednou z postulovaných otázek bylo např. jak je chování jedinců ovlivněno topologií jejich sociální sítě a naopak. Roku 1967 provedl Stanley Milgram slavný experiment, při němž ukázal, že vzdálenost mezi dvěma náhodně vybranými lidmi z celé populace USA je dána v průměru šesti prostředníky. Participanti ze států Kansas

[49] Barabási, 2016, cit. 12. 03. 2018
[50] 2008, s. 7.
[51] Estrada, & Knight, 2015, s. 6–7.

a Nebraska byli požádáni, aby poslali dopis neznámé cílové osobě v Bostonu, Massachusetts nebo v Cambridge. Každý z dobrovolníků vždy poslal dopis známému, jenž byl blíže cíli než on sám. Mnoho dopisů sice nikdy nedošlo, ale ty, které ano, urazily v průměru vzdálenost 5,2 prostředníka. Zrodil se tzv. fenomén malého světa.[52]

Moderní podoba teorie sítí se zrodila v pozdních devadesátých letech, když se začali sítěmi zabývat vědci napříč dalšími vědními disciplínami včetně biologie či fyziky. Pro inženýra komunikace je síť počítačová, pro sociologa diagram reprezentující sociální interakce mezi lidmi, fyzikové zase přemýšlejí o sítích jako modelech fázového přechodu nebo magnetismu. Uplatnění nachází teorie sítí v marketingu, např. při tvorbě virálních marketingových kampaní, a stejně tak užitečná se může ukázat i při zkoumání přirozeného jazyka, chceme-li např. určit literární žánry textů.[53]

Vzhledem k jejich velikosti a spletitosti není sítím lehké porozumět. Už Euler však ukázal, že některé výzkumné problémy se stanou jednoduššími, pokud je znázorníme formou grafu. Chceme-li totiž pochopit jakýkoli komplexní systém, musíme nejprve zjistit, jakým způsobem jsou propojené jeho komponenty; tj. potřebujeme vytvořit síťový graf.[54] Proto se nyní podíváme na některé ze základních pojmů z teorie grafů, s nimiž budeme dále pracovat.

3.1 Graf

Grafy reprezentují množiny prvků a vztahy mezi nimi. Vezměte si např. veškeré vaše přátele: pokud byste je zakreslili na papír v podobě koleček, a spojili byste čarou každé dva z nich, kteří se navzájem znají, získali byste přehled o tom, jak hustě je síť vašich přátel propojená. Kdo má přátel nejvíce a stojí tudíž v jejím centru, a kdo je spíše na okraji.

Formálně uvažujeme graf $G = (V, E)$ jako abstraktní objekt tvořený množinou vrcholů V a množinou hran E, které spojují jednotlivé vrcholy a reprezentují vztahy mezi

[52] Lewis, 2008, s. 4.
[53] Ibid., s. 4.–6.
[54] Barabási, 2016. cit. 12. 03. 2018.

nimi. Množina vrcholů *V* grafu *G* se značí *V(G)*, množina hran *E* grafu *G* pak *E(G)*. Kardinalita (počet prvků množiny) *V* je obvykle značena jako *n*, kardinalita *E* jako *m*. Dva vrcholy spojené hranou se nazývají vrcholy sousední.

Grafy mohou být jak neorientované, tak orientované. U neorientovaných grafů je uspořádání sousedních vrcholů nepodstatné, protože nezáleží na směru interakcí mezi nimi. Typicky si neorientovaný graf můžeme představit např. jako skupinu tří přátel: Adama (A), Barbory (B) a Cedrika (C). Pokud je Adam přítelem Barbory, je Barbora zároveň přítelkyní Adama. Stejně tak pokud je A přítelem C, je C přítelem A, a konečně, je-li B přítelkyní C, je C přítelem B. Ve srovnání s tím u orientovaných grafů na směru interakcí mezi vrcholy záleží. Jako ilustrace nám může posloužit např. graf fanoušků (A, B, C) rockové skupiny The Rolling Stones. Jak A, tak B i C poslouchají její skladby, naopak však tento vztah neplatí.

Neorientovaná hrana spojující vrcholy $u, v \in V$ je značena $\{u, v\}$. Oproti tomu u orientovaných grafů má každá orientovaná hrana (v podobě šipky) svůj výchozí bod a cíl. Hrana s počátkem $u \in V$ a cílovým $v \in V$ je reprezentována uspořádanou dvojicí (u, v). V orientovaném grafu $G = (V, E)$ existuje skrytý neorientovaný graf, pakliže má množinu vrcholů V mající neorientovanou hranu mezi vrcholy $u, v \in V$ pokud (u, v) nebo (v, u) leží v E. Hrana, jejíž cílové vrcholy jsou identické (tzn. spojuje vrchol se sebou samým) se nazývá smyčkou. Jestliže graf neobsahuje žádné smyčky, jedná se o graf jednoduchý.[55]

Nyní, když jsme si představili pojem grafu, podíváme se blíže na některé jeho vlastnosti a specifické komponenty. Často nám totiž pomohou k lepšímu pochopení sítí, které pomocí grafů znázorňujeme, a přinesou nám o nich více informací. Veškeré vlastnosti a komponenty, jež budou následovat, vysvětlíme formálně a zároveň je budeme ilustrovat na příkladech. Aplikovat je pak budeme v praktické části této práce.

[55] Viz např. Brandes, & Erlebach, 2005.

3.2 Vážený graf

Jsou situace, kdy můžeme vztahy mezi vrcholy grafu ohodnotit; pokud tak učiníme, hovoříme o grafu váženém. Přidržíme-li se našeho příkladu se skupinou přátel, uvažujme následující situaci: Adam a Barbora žijí v Olomouci, ale Cedrik se odstěhoval za prací do Vídně a je časově velmi vytížený. Se svými olomouckými přáteli se proto vídá jenom párkrát do roka a ani si s nimi příliš často nepíše, nevolá. Oproti tomu Adam potkává Barboru několikrát do týdne. Frekvence interakcí mezi A a B je tak daleko vyšší než frekvence interakcí mezi A a C a než mezi B a C; hraně spojující A s C v takovém případě přiřadíme větší váhu než hranám zbylým. Nicméně vahami může být cokoli, co si jen zvolíme – vždy záleží na tom, co graf znázorňuje.

Formálně hovoříme o váženém grafu $G = (V, E)$, pakliže přiřadíme jeho hranám či vrcholům numerické hodnoty (váhy). Váhy hran mohou být reprezentovány funkcí jako např. $\omega: E \rightarrow \mathbb{R}$, která každé hraně $e \in E$ přiřazuje váhu $\omega(e)$.

3.3 Stupeň

Přiřadíme-li konkrétní hodnoty vrcholům orientovaného grafu, hovoříme o tzv. stupních. Jak Adam, tak Barbora i Cedrik poslouchají The Rolling Stones, Adam a Cedrik jsou však také fanoušky kanadské zpěvačky Feist. Ta má proto v takovém případě stupeň o hodnotě dvě, zatímco The Rolling Stones mají stupeň tři.

Formálně uvažujeme stupeň vrcholu v orientovaného grafu $G = (V, E)$, značený $d(v)$, jako počet hran v E, pro něž je v konečným vrcholem. Množina hran, pro které je v konečným vrcholem je značena $T(v)$. Množina sousedů v je značena $N(v)$. Pro orientovaný graf $G = (V, E)$ je výstupní stupeň $v \in V$, značený $d^+(v)$, počet hran ležících v E a vycházejících z v. Oproti tomu vstupní stupeň $v \in V$, značený $d^-(v)$, je počet hran končících ve v.

U vážených grafů se tyto pojmy zobecňují součtem nad hranami. Množina hran vycházejících z v je značena $T^+(v)$, množina hran směřujících do v pak $T^-(v)$. Množina cílů hran v $T^+(v)$ je značena $N^+(v)$, množina zdrojů hran v $T^-(v)$ je značena $N^-(v)$.

3.4 Podgraf

Ačkoli jsou Cedrik, Adam i Barbora přátelé, přeci jenom pobývá Cedrik od zbylých dvou několik stovek kilometrů daleko. Pokud bychom se na tuto trojici rozhodli nahlížet podle toho, kde bydlí, Adam s Barborou by vytvořili podmnožinu přátel v Olomouci. Stejně tak odebereme-li z původního grafu některé jeho vrcholy a všechny hrany do nich zasahující, vznikne podgraf.

Graf $G' = (V', E')$ je podgrafem grafu $G = (V, E)$, jestliže $V' \subseteq V$ a $E' \subseteq E$. Pokud E' obsahuje všechny hrany e ϵ E, jež zároveň spojují vrcholy V, jedná se o indukovaný podgraf. Indukovaný podgraf $G' = (V', E')$ s vrcholy $V' \subseteq V$ se značí $G[V']$. Oproti tomu indukovaný podgraf $G' = (V', E')$ s hranami $E' \subseteq E$, označovaný jako $G[E']$, obsahuje V' – množinu všech vrcholů V, které jsou konečnými vrcholy alespoň jedné hrany v E'.

3.5 Sled, tah, cesta a cyklus

Sled z x_0 do x_k v grafu $G = (V, E)$ je posloupností vrcholů a hran $x_0, e_1, x_1, e_2, x_2, ..., x_{k-1}, e_k, x_k$, kde $e_i = \{x_{i-1}, x_i\}$ v případě neorientovaného grafu, a $ei = (x_{i-1}, x_i)$ v případě grafu orientovaného. Délka sledu je definována součtem délek jeho hran. Představit si jej můžeme např. jako trasu, kterou urazí Barbora z Olomouce přes Brno, pojede-li navštívit Cedrika do Vídně a zpět. Cestou zpátky se ale v Olomouci jenom rychle osprchuje, a bude pokračovat dál na chatu na Dolní Moravu. Pakliže $e_i \neq e_j$ pro $i \neq j$ (žádná hrana ve sledu se neopakuje), hovoříme o tahu. Dále tah, v němž se neopakuje žádný vrchol nazýváme cestou a tah, pro který platí $x_0 = x_k$, je nazýván cyklem.

3.6 Komponenty grafu

Představme si, že naše známá trojice přátel vyrazila na koncert. Když bychom znázornili formou grafu všechny skupiny přátel, které si toho večera na daný koncert zašly, dostali bychom nesouvislý graf – party přátel se navzájem vůbec neznají – přesně o tolika komponentách, kolik bylo na koncertu skupin kamarádů.

Neorientovaný graf $G = (V, E)$ je souvislý, pokud může být každý jeho vrchol dosažen z každého jiného vrcholu, tzn. pokud existuje pro každý vrchol hrana spojující jej s všemi ostatními vrcholy. Graf pouze o jednom vrcholu je rovněž považován za spojený. Pro neorientovaný graf $G = (V, E)$ je jeho souvislou komponentou G indukovaný podgraf $G' = (V', E')$, který je maximálně souvislý (tj. neobsahuje žádný indukovaný podgraf $G'' = (V'', E'')$ s $V'' \subseteq V'$).

Orientovaný graf $G = (V, E)$ je silně souvislý, pokud v něm pro každou dvojici vrcholů existuje sled. Silně spojenou komponentou orientovaného grafu G je indukovaný podgraf, jenž je silně a maximálně souvislý. Pokud je k orientovanému grafu připojen jeho základní neorientovaný graf, jedná se o slabě souvislý orientovaný graf.

3.7 Míry centrality

Míry centrality kvantifikují a popisují vztahy mezi jednotlivými vrcholy nebo hranami grafu/sítě (např. sociální) a pomáhají určit ty významnější z nich (např. lidé či slova); jsou závislé na struktuře grafu.

3.7.1 Stupeň

Stupeň jsme detailně popsali výše, zde proto jenom zmíníme, že je nejjednodušší mírou centrality. Pro neorientovaný graf je stupeň centrality $c_D(v)$ vrcholu v definován jako stupeň $d(v)$ vrcholu v. U orientovaných grafů lze rozlišit dva typy: vstupní stupeň centrality $c_{iD}(v) = d^-(v)$ a výstupní stupeň centrality $c_{oD}(v) = d^+(v)$. Užitečný je stupeň centrality např. při zjišťování frekvence výskytu a souvýskytu slov v rámci korpusu. Vzhledem k tomu, že centralita vrcholu je dána pouze jeho sousedy, je stupeň lokální mírou.

3.7.2 Blízkost

Při analýze sociálních sítí se centralita blízkosti užívá k měření důležitosti jedinců v sítí. Popisuje provázanost jednotlivých vrcholů grafu, přičemž čím blíže je vrchol vrcholům

ostatním, tím vyšší je jeho ukazatel centrality blízkosti. Např. ze skupiny dvaceti lidí, kdy všichni nejsou nutně přátelé, bude mít nejvyšší ukazatel centrality blízkosti ten, k jehož zprostředkovanému oslovení bude potřeba od všech zbylých devatenácti jedinců použití nejméně prostředníků. Suma vzdáleností vrcholu $u \in V$ ku ostatním vrcholům grafu $G = (V, E)$ se značí jako celková vzdálenost $\sum_{v \in V} d(u, v)$. Nejčastěji používanou definicí centrality blízkosti je: $c_C(u) = \frac{1}{\sum_{v \in V} d(u,v)}$.

3.7.3 Mezilehlost

Mezilehlost můžeme určit jak pro vrcholy grafu, tak pro jeho hrany. Centralita mezilehlosti vrcholů určuje jejich významnost v rámci sítě podle množství hran mezi různými vrcholy, jež jimi procházejí. Představme si např. zájezd deseti Čechů do Jižního Tyrolska, z nichž pouze jeden umí německy. Při rozpravách s místními by německy mluvící Čech působil jako tlumočník: zprostředkovával by komunikaci mezi devíti svými krajany a všemi Italy. Měl by tudíž nejvyšší ukazatel centrality mezilehlosti vrcholů. Formálně centralitu mezilehlosti vrcholu $v \in V$ definujeme jako: $c_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, kde σ_{st} je počtem nejkratších cest mezi vrcholy s a t, a $\sigma_{st}(v)$ je počet těchto cest procházejících vrcholem v.

Oproti tomu centralita mezilehlosti hran je definována jako suma nejkratších tahů, které procházejí hranami v grafu. Nejvyšší ukazatel centrality mezilehlosti hran mají hrany, jež spojují části grafu takovým způsobem, že by jejich absence výrazně ovlivnila či přímo znemožnila komunikaci mezi některými skupinami vrcholů.[56]

Teď už známe některé z vlastností a komponentů, jimiž mohou grafy disponovat. Víme zároveň, že nám díky nim mohou poodhalit podoby vztahů, které pojí jejich vrcholy. Těmi mohou být např. lidé v rámci společenství či skupin, ale též např. jednotlivá slova vyskytující se v korpusu. A protože budeme v praktické části dolovat textová data ze sociálních médií, představíme si na následujících řádcích stručně jak historii a kategorie sociálních médií, tak samotné dolování dat z nich pocházejících.

[56] Lu, & Zhang, 2013, cit. 26. 04. 2018.

4 Dolování sociálních médií

Dolování sociálních médií analyzuje data vytvářená lidmi spolu s jejich společenskými vazbami za účelem objevení, měření a modelování schémat stojících v jejich pozadí.[57] Nástup sociálních médií značně proměnil způsob, kterým mezi sebou komunikujeme. Výrazná část mezilidské komunikace se dnes odehrává na webu, protože to s sebou přináší řadu výhod. Se svými blízkými můžeme sdílet nejenom textové, ale i audiovizuální či multimediální komunikáty, a to téměř z kteréhokoli zákoutí Země a v kteroukoli denní nebo noční hodinu. Každým okamžikem na světě skrze sociální média komunikují miliardy uživatelů. Informace jsou shromažďovány, upravovány a publikovány jednotlivci a zároveň sdíleny tisíci dalšími, kteří na ně spontánně reagují. To vše nabízí možnost zkoumat mezilidskou komunikaci v obrovském měřítku (ke 3. dubnu 2018 měla sociální síť Facebook více než 2,1 miliardy měsíčně aktivních uživatelů[58]).

Obecně můžeme na sociální média nahlížet jako na svět sociálních atomů (jednotlivců), entit (obsahů, webových stránek, sítí apod.) a interakcí mezi jednotlivci a entitami, jež jsou řízeny společenskými normami. Vzhledem k povaze dat pocházejících ze sociálních médií, která jsou značně nestrukturovaná a zpravidla zahlcená množstvím nepotřebných informací, vyžaduje jejich těžení kombinaci technik pocházejících z dolování textových i netextových dat, lingvistiky, strojového učení, analýzy sociálních sítí či statistiky.[59]

Především však dolování sociálních médií vychází z tradice analýzy sociálních sítí, která se zrodila již na začátku dvacátého století v sociologických pracích Émila Durkheima. Tehdy byla data potřebná pro výzkum sbírána manuálně a experimenty často cílily na příslušníky střední nebo vyšší společenské třídy, aby se zvýšila pravděpodobnost jejich logického průběhu. Výsledky těchto experimentů byly přijímány

[57] Zafarani, Abbasi, & Liu, 2014, s. 2.
[58] „Top 20 Facebook Statistics," 2018, cit. 01. 05. 2018.
[59] Zafarani, Abbasi, & Liu, 2014, s. 2.

z kvalitativního, avšak ne kvantitativního hlediska. Oproti tomu dnes máme k dispozici digitální data umožňujícími nejenom kvalitativní, ale také kvantitativní výzkum.[60]

Dolování sociálních médií zpravidla zahrnuje dvě základní skupiny analýz[61]:

(a) Analýza vazeb a strukturální analýza – důraz je při nich kladen na zkoumání vazeb a jejich chování v rámci sítě za účelem určení důležitých uzlů, komunit a vazeb mezi nimi.
(b) Obsahová analýza – zkoumá obrovské množství multimediálních dat umístěných na sociálních platformách pro sdílení obsahu jako je Flickr sloužící ke sdílení fotografií, YouTube, jež slouží primárně ke sdílení videí, nebo např. Twitter, na kterém nalezneme nepřeberné množství textu šířené uživateli.

Abychom lépe pochopili, co přesně sociální média jsou, a proč je užitečné dolovat z nich data, podíváme se v následující kapitole na jejich historický vývoj a jejich současné kategorie. Od těch jsou totiž odvislé jak způsoby komunikace, které svým uživatelům sociální média nabízí, tak typy dat, jež z nich můžeme dolovat.

4.1 Historie a kategorie sociálních médií

Definovat můžeme sociální média spolu s Kaplanem a Haenleinem jako skupinu internetových aplikací, jež jsou vybudovány a ideologicky vycházejí z technologických základů webu 2.0. Tento termín byl poprvé použit v roce 2004 k popsání nového způsobu, kterým začali softwaroví vývojáři a uživatelé využívat World Wide Web. Obsah a aplikace již nebyly nadále vytvářeny a publikovány pouze jedinci, nýbrž byly upravovány navzájem spolupracujícími uživateli. Zatímco internetové aplikace jako webové stránky nebo Encyclopedia Britannica Online spolu se základní myšlenkou publikování obsahu patří do éry webu 1.0, participativní a kooperativní projekty typu blogů, vlogů či Wikipedie charakterizují současný web 2.0. Rozličné formy veřejně

[60] Aggarwal, 2011, s. 3–4.
[61] Ibid., s. 5–6.

dostupného mediálního obsahu, jenž je vytvářen a konzumován uživateli těchto aplikací, shrnuje termín User Generated Content.[62]

Pojďme se však ještě krátce podívat na úplné počátky sociálních médií. Roku 1978 byl vytvořen Bulletin Board System, který jako první umožnil uživatelům vzájemnou komunikaci na internetu. O rok později představili Tom Truscott a Jim Ellis diskuzní systém Usenet umožňující uživatelům na celém světě veřejně publikovat zprávy a články. První sociální síť – Geocities – byla vytvořena v roce 1994 a umožňovala svým uživatelům tvorbu vlastních domovských stránek. O pár let později vznikl AOL Instant Messenger, jenž jako první umožňoval komunikaci v reálném čase; jeho uživatelé si zároveň mohli utvářet profily a sdílet tak s ostatními své zájmy apod. Rostoucí dostupnost internetu, přenosných počítačů a mobilních telefonů spolu s oblibou podobných konceptů vedla ke vzniku sociálních médií, jak je známe dnes.[63]

Sociální média zahrnují v dnešní době kategorie jako např.:

- sociální sítě: Facebook, LinkedIn, VKontakte, WeChat
- mikroblogy: Twitter, Tumblr, Sina Weibo
- média shromažďující zprávy: The Old Reader
- platformy pro sdílení fotografií: Instagram, Flickr
- platformy pro sdílení videí: YouTube, Vimeo, Youku Tudou
- instant messaging: Skype, Facebook Messenger, WhatsApp, Telegram
- online seznamky: Tinder, Grindr, Badoo, Momo

V současnosti nejoblíbenějším sociálním médiem je americký Facebook, který svým uživatelům nabízí primárně možnost propojit se na internetu se svými přáteli, tzn. stát se přáteli na Facebooku. Uživatelé zde mohou navzájem komunikovat vícero způsoby: od pouhého pasivního sledování příspěvků svých známých, přes jejich hodnocení (pomocí tzv. „lajků“ a emotikon), až po aktivní zapojování se do veřejných diskuzí, komunikaci skrze soukromé zprávy a šíření vlastního textového či audiovizuálního obsahu (též formou živého vysílání).

[62] 2010.
[63] Ibid.

Krom běžných uživatelů využívají tuto sociální síť hojně také soukromé firmy ke komunikaci se zákazníky a ke své propagaci. Pro tyto účely mohou např. pořádat tzv. „facebookové události“, na nichž uživatelé přislíbí svou účast online a většinou se jí mohou následně zúčastnit též v reálném světě. Dále se mohou uživatelé přihlásit k odběru novinek profilů svých oblíbených firem nebo celebrit. Případně mohou založit online zájmové skupiny sdružující např. spolužáky či profesní komunity. Na základě uživatelských dat[64] nabízí Facebook firmám cílené šíření reklam.

Platformy pro sdílení fotografií či videí svým uživatelům zpravidla poskytují méně způsobů komunikace než sociální sítě. I přesto je ale jejich škála poměrně široká: mohou komentovat fotografie či videa, diskutovat spolu pomocí soukromých zpráv, ale též navazovat přátelství nebo odebírat novinky svých oblíbených profilů. Kromě toho si ti, kteří zde publikují často, jejichž příspěvky jsou nadprůměrně sledované a jejichž profily mají mnoho odběratelů, mohou vydělávat peníze. Viz např. Casey Neistat a tzv. „youtubering“, kdy lidé pomocí filmu mapují své životy, případně se vyjadřují ke společensky aktuálním tématům. Firmy podobným tvůrcům platí jak za tzv. „product placement“ přímo v jejich videích, tak např. za to, že se uživateli před samotným spuštěním videa přehraje nejprve reklama na konkrétní produkt. Oblíbené jsou též vzdělávací kanály (např. Kurzgesagt – In a Nutshell).

Jsou však země, ve kterých se k některým z globálně nejoblíbenějších sociálních médií nepřipojíte. Čína aktuálně blokuje např. Facebook, Twitter, Instagram, YouTube, Snapchat či Tinder.[65] Tamní uživatelé mají nicméně k dispozici lokální alternativy jako třeba WeChat. V Rusku se zase těší větší oblibě než americký Facebook, přestože je zde povolený, jeho ruská varianta VKontakte.

Jak vidíme, od svých počátků si prošla sociální média velikým vývojem. V dnešní době existuje celá řada jejich kategorií a způsobů, jimiž na nich lidé mohou komunikovat. Někteří se dokonce vytvářením obsahu pro sociální média živí. Vzhledem k jejich množství a rozmanitosti, mohou být data pocházející ze sociálních médií

[64] Soukromí uživatelů Facebooku a způsoby, jakými lze nakládat s jejich osobními daty, začaly být výrazně diskutovány v souvislosti se skandálem týkajícím se zneužití uživatelských dat firmou Cambridge Analytica. Ta údajně data poskytla politickým subjektům, které díky nim měly ovlivnit americké prezidentské volby 2016.
[65] „List of Websites and Apps Blocked in China,” 2018.

cenným zdrojem informací o jedincích, ale také o celých společnostech. Pojďme se proto nyní na jedno z nich – Twitter – podívat blíže.

5 Twitter

Aktuálně je Twitter[66] nejznámější mikroblogovací sítí na světě. Ty se oproti ostatním sociálním médiím liší v tom, že z části vycházejí z klasických internetových blogů, tj. stránek, na nichž lidé nebo firmy publikují krátké textové útvary. Může se jednat osobní deníkové zápisky či např. cestopisy, ale též o firemní zpravodajství. Mikroblogovací sítě oproti blogům umožňují sdílení menších obsahů – ať už se jedná o krátké věty, obrázky nebo odkazy na videa.[67] Za dobu své existence si Twitter prošel řadou změn. Zprvu šlo pouze o mobilní službu pro SMS (Short Message Service), dnes se na něj mohou lidé připojit jak na internetu, tak skrze mobilní aplikace. A protože chápat jeho vývoj je důležité pro porozumění některým jeho současným vlastnostem, podíváme se nyní na cestu, kterou od svého založení urazil.

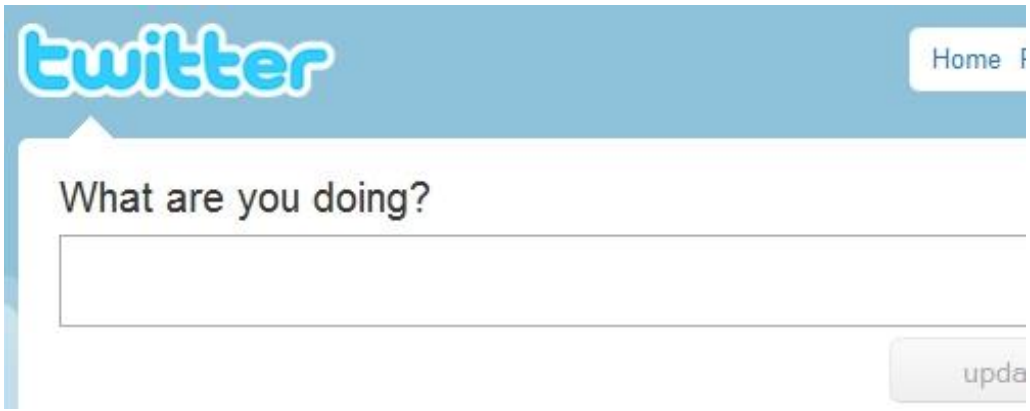
Twitter, jehož původní název zněl Twttr, založili v roce 2006 v San Franciscu Jack Dorsey spolu s Noahem Glassem, Bizem Stonem a Evanem Williamsem. Zprvu služba poskytovala možnost propojení malých skupin uživatelů skrze mobilní telefony, resp. textové zprávy SMS. Ty byly a jsou omezeny rozsahově na 160 znaků (po překročení limitu se zpráva rozdělí na dvě), což určovalo také délku sdělení na Twitteru: 20 znaků bylo vyhrazeno přezdívce uživatele, 140 znaků samotné zprávě. Potřebná stručnost vedla následně rovněž ke zkracování URL adres (Uniform Resource Locator), které slouží ke specifikaci umístění zdrojů informací na internetu, a jsou při komunikaci skrze sociální média hojně využívány.[68]

V průběhu svého vývoje se postupně Twitter přesunul na web. Uživatelé se na něj dnes mohou přihlásit nejen s pomocí stolního či přenosného počítače, ale také skrze mobilní aplikace určené chytrým telefonům nebo např. skrze herní konzoli Xbox. Ruku v ruce s technologickými změnami se proměňovala rovněž podoba samotných zpráv. Kromě těch soukromých, určených konkrétním uživatelům, je na světě denně na Twitteru publikováno na 500 miliónů tzv. *tweetů*[69] – zpráv, které uživatelé zveřejňují

[66] www.twitter.com, cit. 01. 05. 2018.
[67] Kaplan, & Haenlein, 2011.
[68] Twitter vyvinul svůj vlastní zkracovač URL adres: t.co.
[69] „Twitter Usage Statistics," 2018, cit. 01. 05. 2018.

na svých komukoli přístupných profilech. Původní limit 140 znaků byl začátkem listopadu 2017 navýšen na 280.

Během prvního období své existence se Twitter svých uživatelů před publikováním *tweetu* tázal: „Co zrovna děláš?“ Zprávy převážně zprostředkovávaly názory a pocity svých autorů, a byly proto nahlíženy výzkumníky jako banální. Někteří dokonce odhadovali, že v roce 2009 tvořilo „blábolení“ 40 % veškerého obsahu Twitteru.[70]

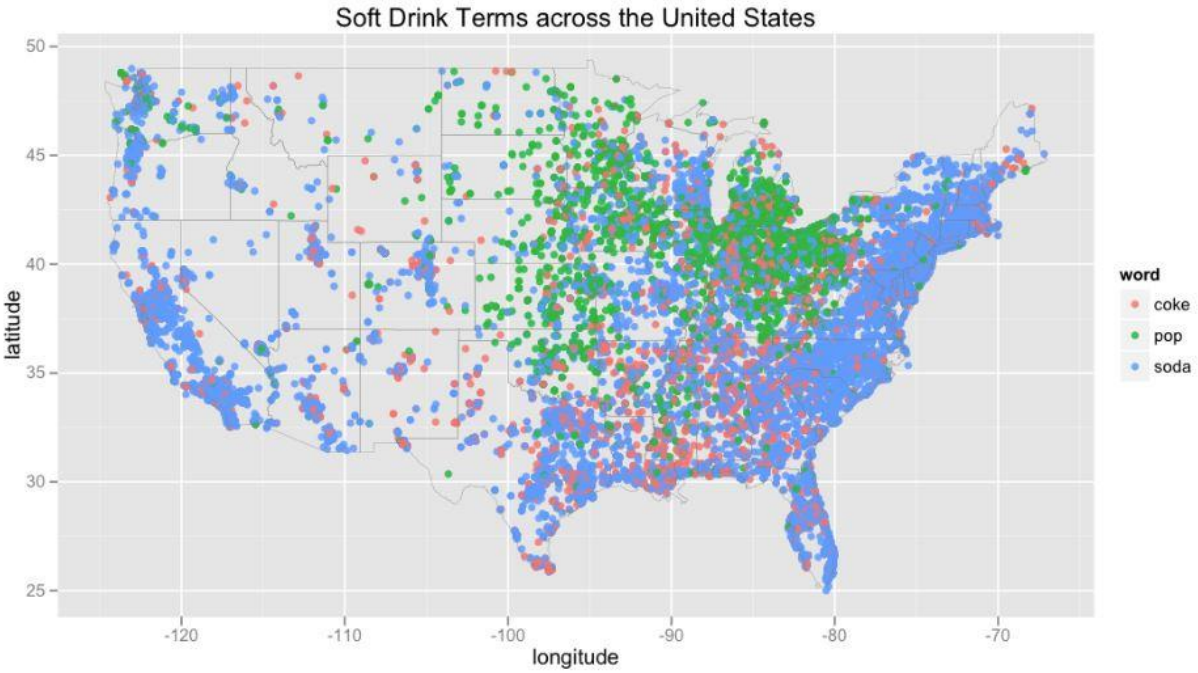


Obrázek č. 2: Twitter – Co zrovna děláš?

Zdroj: http://creation.co/knowledge/twitter-for-beginners-5-steps/, *cit. 15. 03. 2018.*

Často zmiňovanou kategorií blábolení byly *tweety* referující o jídle, které na první pohled nikdo nepovažoval za jakkoli hodnotné. Datoví analytici jako např. Edwin Chen to ale dle Brunse a Moeho změnili.[71] Ve svém výzkumu z roku 2012 se Chen zaměřil na způsob, jakým Američané *tweetují* o limonádách. Za pomoci zpráv obsahujících krom textu také geolokační údaje porovnal, v jakých částech USA užívají lidé výrazy „soda“, „pop“ a „coke“. Výsledná vizualizace dat rozdělila USA dle zkoumaného kritéria jasně na několik oblastí.[72] Všednodenní zprávy ze sociálních sítí získaly díky podobným studiím v očích odborné veřejnosti na relevanci.

[70] Kelly, 2009, cit. 15. 03. 2018.
[71] Bruns, & Moe, 2013, s. xiv.
[72] Chen, 2012, cit. 15. 03. 2018.

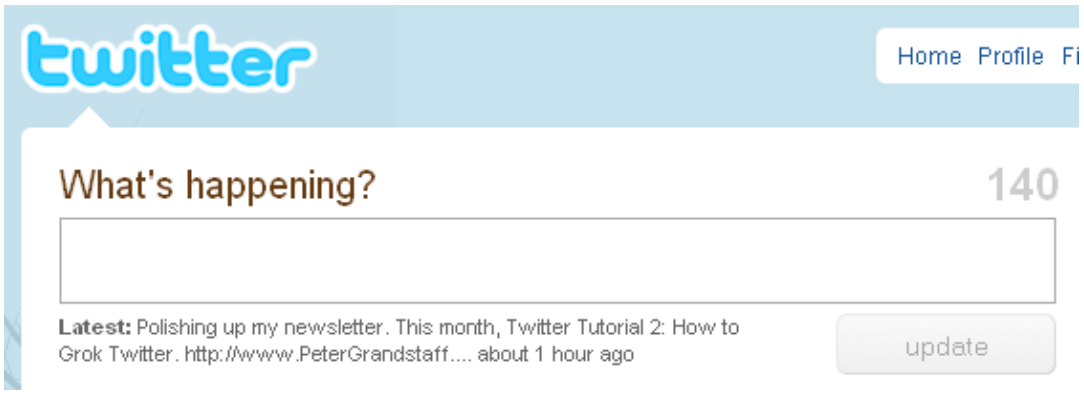


Obrázek č. 3: Tweetovaní o limonádách v USA

Zdroj: http://blog.echen.me/2012/07/06/soda-vs-pop-with-twitter/, *cit. 15. 03. 2018.*

K významné změně však došlo ještě závěrem téhož roku, kdy uživatelé na Twitteru převážně blábolili. Od listopadu 2009 již nebyli při publikování *tweetu* dotazováni na to, co zrovna dělají – otázka nově zněla: „Co se děje?" Tento přechod lze interpretovat jako interní posun ve směřování služby, stejně jako pobídnutí svých uživatelů ke sdílení hodnotnějších informací.[73] S velikou oblibou se začal postupně Twitter používat při konferencích, kdy organizátoři spolu s účastníky referují o jejich průběhu; při volbách, kdy jej využívají jak volební štáby jednotlivých kandidátů, tak jejich příznivci; či např. během přírodních katastrof a teroristických útoků, kdy oběti informují veřejnost o jejich průběhu nebo žádají o pomoc.

[73] Rogers, 2013, s. xvi.



Obrázek č. 4: Twitter – Co se zrovna děje?

Zdroj: https://www.petergrandstaff.com/marketing/twitter-tutorial-2, *cit. 15. 03. 2018.*

Významnou úlohu sehrává Twitter, spolu s dalšími sociálními médii, rovněž během revolucí, jak ukázalo např. arabské jaro. Protestující v Egyptě, Tunisku či Libyi je využívali nejenom pro komunikaci mezi sebou navzájem, ale obraceli se k nim také jako ke zdroji informací, když ztratili důvěru v prorežimní novináře a televizní komentátory. Na druhou stranu však nedemokratické režimy často sociální média nebo přímo celý internet na svém území (nejenom) při vzpourách blokovaly. Neopomenutelnou úlohu dále hrála sociální média při komunikaci utlačovaných obyvatel s vnějším světem. Každý, kdo vlastnil mobilní telefon s připojením k internetu, mohl např. *tweetovat* o aktuálním stavu ve svém městě, ve své ulici a připojit se tak k řadě dalších řadových občanů, kteří zásobovali svými texty a fotografiemi západní média.[74] Podobné situace jsme nyní ostatně svědky v rámci syrské občanské války.

Vzhledem k obrovskému a rozmanitému množství převážně textových dat – od rekapitulací jídelního lístku přes sdílení obsahu konferencí až po revoluce na blízkém východě – se stal Twitter hojně používaným zdrojem informací pro sociologické či lingvistické studie. Jeho další výhodou je relativně snadná dostupnost dat potřebných pro tvorbu výzkumného korpusu. Právě z těchto důvodů jsem zvolil *tweety* jako výchozí materiál pro svou práci.

[74] Viz např. Beaumont, 2011, cit. 15. 03. 2018.

5.1 Komunikační vrstvy na Twitteru

Nejsou to ale jenom témata a jejich důležitost, jež mohou na Twitteru variovat. Samotná komunikace se zde odehrává v několika navzájem se prolínajících vrstvách. Ty jsou určeny jak technickými specifiky mikroblogovací sítě, tak jejím uživatelským rozhraním a uživatelskými konvencemi od něj odvislých. Bruns a Moe rozpoznávají tři základní komunikační vrstvy: mikro vrstvu interpersonální komunikace, meso vrstvu v síti sledujících a makro vrstvu na úrovni hashtagů.[75]

5.1.1 Meso vrstva

Informační tok se na Twitteru realizuje především díky možnosti sledovat účty lidí, jež známe nebo kteří nás zajímají. Pakliže nemá uživatel nastavený svůj účet jako soukromý, může jej sledovat kdokoli. Primárně se ale odběrateli zobrazují pouze veškeré *tweety* publikované těmi uživateli, které sleduje. Dále pak záleží pouze na něm, nakolik důkladně zprávy generované sítí svých účtů pročítá. Právě tato výchozí úroveň šíření *tweetů* tvoří meso vrstvu. Proces odebírání *tweetů* přitom nemusí být oboustranný, tj. mezi počtem uživatelů, které jedinec sleduje, a jimiž je sledován, je většinou disproporce.[76]

Na skupinu odběratelů lze pro každý účet nahlížet jako na jeho tzv. osobní publikum. Průměrný uživatel publikující příspěvek na svém účtu je podle Brunse a Moea v podobné situaci jako např. při rodinném setkání, když má před všemi shromážděnými přednést osobní prohlášení. Tato analogie nicméně selhává u účtů (popkulturních celebrit, politiků apod.) majících velikou základnu sledujících, které uživatel nemůže reálně všechny osobně znát.[77]

[75] 2013, s. 16.
[76] Bruns, & Moe, 2013, s. 16–17.
[77] Ibid., s. 17.

5.1.2 Makro vrstva

Zprávy spadající do meso vrstvy obsahují často krom samotného textu či fotek/obrázků ještě tzv. hashtagy. Hashtagy jsou klíčová slova definovaná uživateli, jimž předchází symbol „#“, a která slouží zpravidla k označení *tweetů* vztahujících se k určitému tématu. Takto označené zprávy mají potenciál dosáhnout mnohem širšího publika, než je pouze síť odběratelů daného účtu. Uživatelé mohou totiž vyhledávat veškeré zprávy obsahující daný hashtag nezávisle na tom, zdali sledují účty, z nichž tyto zprávy pocházejí. Hashtagy se proto využívají také pro označování událostí jako jsou volby nebo přírodní katastrofy, čímž napomáhají koordinovat přenos informací a orientaci v nich.[78] S pomocí hashtagů Twitter rovněž v uživatelském rozhraní zobrazuje aktuální trendy (jako např. #Cannes2018 či #TuesdayThoughts).

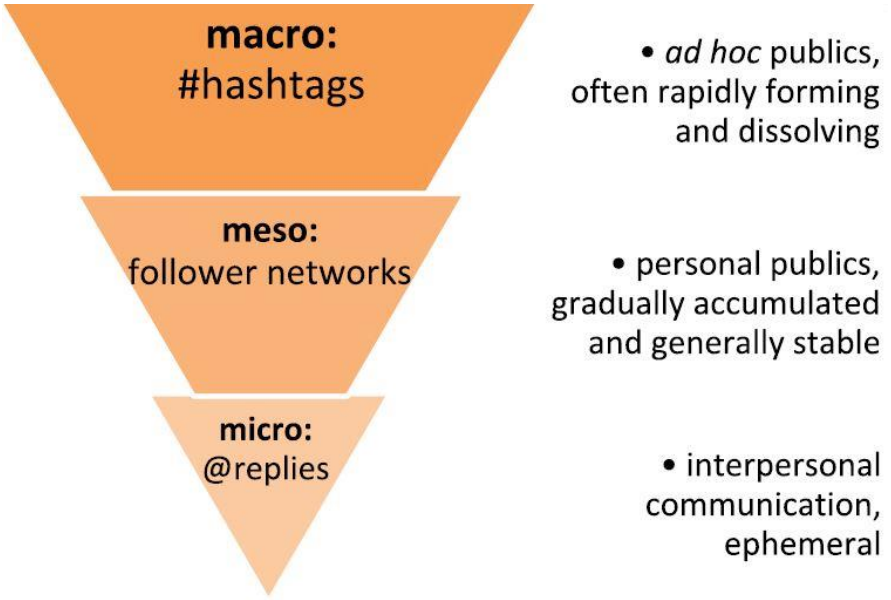
Vložením hashtagu do své zprávy uživatel dává zároveň najevo, že chce být součástí širšího komunikačního procesu s každým, kdo projeví zájem o stejné téma. Obzvlášť rychle se hashtagy hromadí a generují veliká *ad hoc* publika při referování o výše zmiňovaných přírodních katastrofách, především ve spojitosti s jejich aktuálním vývojem a při dalších mimořádných zprávách. Takto vzniklá publika mohou být na jednu stranu mnohem dynamičtější než sítě sledujících v meso vrstvě, na druhou stranu se však rovněž mohou přetavit ve stabilní a dlouho existující komunity uživatelů. Bruns a Moe se krom toho domnívají, že někteří uživatelé mohou přiložit hastagy do svých *tweetů*, čistě pro zvýšení jejich dosahu a bez jakékoli tematické spojitosti.[79]

Další skupinou jsou hashtagy vztahující se namísto aktuálních situací k náladám, emocím a prožitkům (např. #fail pro neúspěch, #headdesk pro frustraci či zklamání, #happy pro radost). Vzhledem k tomu, že jsou užívány převážně ke zvýraznění tweetů a nemají přílišnou informační hodnotu, bývají často řazeny do meso vrstvy.[80]

[78] Ibid., s 17–18.
[79] Ibid., s. 18.
[80] Ibid.

5.1.3 Mikro vrstva

Začlení-li uživatel do svého *tweetu* zmínku o uživateli jiném (např. @hadleywickham), dává tím najevo, že s ním chce navázat konverzaci. Jeden druhého přitom nutně nemusí mít ve své síti odběratelů, ani jej nemusí sledovat. Twitter ho na zmínku upozorní automaticky. Pokud adresát @zmínku opětuje a naopak, vznikne cyklus @odpovědí. Takovéto konverzace tvoří mikro vrstvu komunikace na Twitteru a mohou být připodobněné k off-line rozhovorům s několika pasivními přihlížejícími.[81] Ne všechny @zmínky ale představují snahu navázat konverzaci. Zvláště pokud jimi uživatel cílí na účet patřící celebritě nebo politikovi, může být jeho záměrem doslova pouze zmínka jako např. „Fandím @IvanCzechPir8“.[82]



Obrázek č. 5: Komunikační vrstvy na Twitteru

Zdroj: Bruns, & Moe, 2013, s. 20.

[81] Aby uživatel zajistil viditelnost @odpovědi také ostatním, kterým není primárně určena, musí použít následující syntax: .@uživatel. Symbol „.“ či jakýkoli jiný zde zajišťuje vizibilitu.
[82] Bruns, & Moe, 2013, s. 19–20.

5.1.4 Komunikace napříč vrstvami

Jak je patrné z výše předloženého popisu komunikačních vrstev, nikdy neexistují odděleně. Meso vrstva při tom slouží jako základní rovina, které se žádný aktivní účastník komunikace na Twitteru nemůže vyhnout. (Výjimkou jsou pouze soukromé zprávy.) Stejně tak každý, kdo má veřejný účet a publikuje *tweety*, musí brát v potaz, že ty mohou být přečteny kýmkoli s přístupem k Twitteru bez ohledu na to, jestli je daný čtenář odběratelem uživatelova účtu nebo ne.

Nejvýraznějším a hojně používaným způsobem přesouvání informací napříč jednotlivými vrstvami jsou však tzv. *retweety* – jedná se o přímé převzetí a opětovné publikování zpráv ostatních uživatelů na svém vlastním účtu. *Retweet* lze provést dvěma způsoby:

a) manuálně formou „RT @uživatel [původní zpráva]“;
b) nebo přímým stiskem ikony *retweet* pod původní zprávou.

Pokud je *tweet* přeposlán manuálně, může do něj uživatel vložit nejenom @zmínku, ale také hashtag, čímž se původní informace přesouvá z makro úrovně přes meso až po mikro úroveň. Na druhou stranu *retweety* pocházející z @odpovědí doplněné hashtagem přenášejí informaci v opačném směru: z mirko do meso případně dále do makro úrovně.[83]

Oproti tomu je-li *tweet* přeposlán pouze stisknutím ikony k tomu určené, informace se nepřesouvá vertikálně napříč vrstvami, nýbrž horizontálně mezi osobním publikem původního účtu a mezi meso vrstvou odběratelů účtu nového. Právě díky tomuto procesu je Twitter velmi efektivní při šíření aktuálních zpráv. S každým *retweetem* totiž původní sdělení zasahuje novou a od té původní odlišnou skupinu příjemců.[84]

[83] Ibid., s. 20–21.
[84] Ibid., s. 22.

6 Modelování skrytých témat

Procházení textového korpusu – zvláště, pokud je nám jeho charakter předem neznámý – je nelehký úkol. Velmi užitečné je v takové situaci objevení skrytých témat, jimiž je protkán, a která nám mohou napovědět, na které otázky pravděpodobně v korpusu nalezneme odpovědi.

Nicméně i pokud máme apriorní představu o tematickém charakteru korpusu, může nás zajímat, jaká další, *apriorně* neznámá témata, se s těmi, jež známe, pojí. Vzhledem k obrovskému množství nestrukturovaných textových dat vznikla přirozeně potřeba vyvinutí automatických metod pro organizaci a správu dokumentů a objevování jejich obsahů.

K tomuto účelu byly vytvořeny a jsou dnes běžně používány tzv. „topic models". Jedná se o převážně pravděpodobnostní modely pro objevování skryté sémantické struktury v korpusech založené na hierarchické Bayesově analýze. Aplikovány bývají např. na e-maily, vědecké abstrakty či na novinové archivy,[85] pokud je potřebujeme utřídit nebo v nich nalézt konkrétní texty. Vycházejí z klasických metod zpracování přirozeného jazyka, jakými jsou např. unigramový model a jeho směsice (mixture of unigram models) či latentní sémantická analýza (LSA).[86]

Oproti unigramovému modelu a jeho směsici se však topic models liší tím, že jsou modely smíšenými (mixed-membership models). U unigramového modelu se četnost každého slova – která je ekvivalentní četnosti jednoho tématu – vyvozuje z celkové distribuce slov pro celý korpus. Představit si to můžete např. tak, že byste ze všech knih vaší domácí knihovny vytrhali veškeré stránky, a poté z nich rozpoznávali témata, která vaše knihovna obsahuje. U směsice unigramových modelů je pak téma vyvozováno zvlášť pro každý dokument a všechna slova v konkrétním dokumentu mohou být vyvozena z distribuce jednoho daného tématu. V takovém případě by každá kniha pojednávala přesně o jednom tématu. Oproti tomu smíšené modely nepředpokládají, že každému dokumentu náleží pouze jedno téma, ale že jeden dokument může zároveň

[85] Blei, & Lafferty, 2009, s. 71.
[86] Grün, & Hornik, 2011, s. 1.

náležet několika různým tématům; a že jejich rozložení se napříč dokumenty různí.[87] Např. kniha Měření světa německého autora Daniela Kehlmana by pak nepojednávala pouze o přírodovědci Alexanderu von Humboldtovi, ale také o fyzikovi Carlovi Friedrichovi Gaussovi a o Amazonském pralese.

Jeden z prvních smíšených modelů – Pravděpodobnostní latentní sémantická analýza (PLSA), též nazývaný Pravděpodobnostní indexování latentní sémantiky (PLSI) – navrhnul v roce 1999 Thomas Hofmann. Předpokládal, že provázanost mezi slovy v dokumentu může být vysvětlena s pomocí skrytých témat v něm obsažených. V roce 2003 pak představili David M. Blei, Andrew Y. Ng a Michael I. Jordan model Latentní Dirichletovy alokace (LDA), který z PLSA vychází.[88] Jeho hlavní výhodou oproti PLSA je schopnost odhalit latentní témata u dokumentů bez použití trénovacího datasetu.

6.1 Latentní Dirichletova alokace

LDA vychází z myšlenky, že dokumenty jsou reprezentovány jako náhodné směsi slov nad latentními tématy, kde každé téma je charakterizováno svou distribucí nad fixní slovní zásobou. Každému dokumentu přitom může náležet vícero témat, přičemž proporcionální rozdělení se různí. Model dále předpokládá, že slova jsou tématy generována, a že ta jsou v rámci konkrétního dokumentu nekonečně měnitelná.[89]

Možnost přiřadit jednomu dokumentu více témat je přínosná, protože jej tímto způsobem lépe charakterizuje. Vezměme si např. text pojednávající o lingvistice. Pakliže bychom jej zkoumali modelem omezujícím se pouze na jedno téma, jenom stěží bychom zjistili důkladně, čemu se jazykověda věnuje a jaká je její podoba alespoň s některými jejími odstíny. LDA převádí tuto intuici do variabilního modelu dokumentů. Skryté variabilní modely jsou strukturovanými distribucemi, v nichž zkoumaná data interagují se skrytími náhodnými proměnnými. Při práci se skrytým variabilním modelem

[87] Ibid.
[88] Ibid., s. 1–2.
[89] Blei, Ng, & Jordan, 2003, s. 993–996.; Blei, & Lafferty, 2009, s. 73.

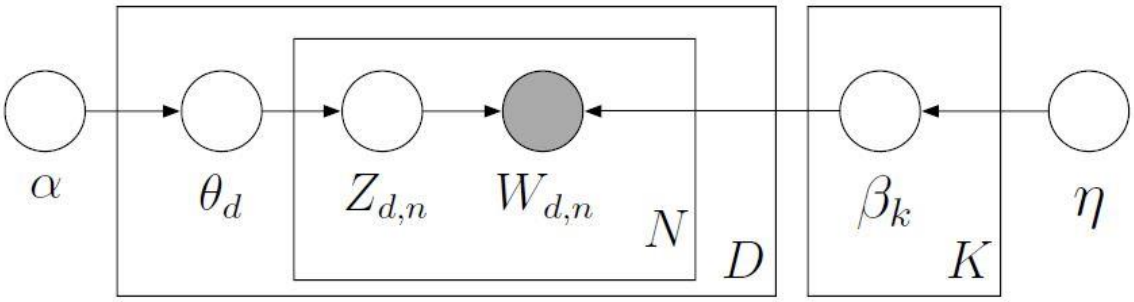
předpokládáme, že data obsahují skrytou strukturu, která je následně odhalena prostřednictvím posteriorní pravděpodobnostní inference.[90]

V modelu LDA jsou zkoumanými daty slova každého jednotlivého dokumentu a skryté proměnné reprezentují latentní lokální strukturu, tj. konkrétní témata a to, jakým způsobem se v každém dokumentu projevují. Vzhledem ke korpusu určuje posteriorní distribuci skrytých proměnných (témat), jež je dána zkoumanými dokumenty, skrytá lokální dekompozice korpusu, tj. to, jakým způsobem se korpus rozloží na menší tematické celky. Interakce mezi pozorovanými dokumenty a strukturou latentních témat se projevuje pravděpodobnostním generativním procesem spojeným s LDA; imaginárním náhodným procesem, o němž se předpokládá, že generoval zkoumaná data, resp. korpus následujícím způsobem[91]:

Nechť K je specifický počet témat, V je velikost slovníku, $\vec{\alpha}$ pozitivní K-vektor a η skalár. $\mathrm{Dir}_V(\vec{\alpha})$ denotuje V-dimenzionální Dirichletovu distribuci s vektorovým parametrem $\vec{\alpha}$, $\mathrm{Dir}_K(\eta)$ denotuje K-dimenzionální symetrickou Dirichletovu distribuci se skalárním parametrem η.

1. Pro každé téma,
   (a) vyber rozdělení nad slovy $\vec{\beta}_k \sim \mathrm{Dir}_V(\eta)$.
2. Pro každý dokument,
   (a) vyber vektor proporcí témat $\theta_d \sim \mathrm{Dir}_V(\vec{\alpha})$.
   (b) Pro každé slovo[92],

[90] Blei, & Lafferty, 2009, s. 73.
[91] Ibid.
[92] Viz grafická reprezentace modelu LDA na následující straně.



Obrázek č. 6: Grafická reprezentace modelu LDA

Uzly zde denotují náhodné proměnné; hrany denotují závislost mezi nimi.

Tmavý uzel denotuje zkoumané náhodné proměnné; světlé uzly denotují skryté náhodné proměnné.

Zdroj: Blei, & Lafferty, 2009. s. 74.

7 Postojová analýza

Po odkrytí témat vyskytujících se v korpusu se nabízí jako jedna z dalších možností analýzy zjistit, jaké postoje k nim autor/autoři textů zaujímají a s jakými emocemi se pojí. Právě to je cílem postojové analýzy.

Postoje a emoce sehrávají v lidské psychice velmi důležitou úlohu, protože značnou měrou ovlivňují naše chování. Naše chování je však ovlivněno také emocemi, názory a chováním našich přátel či naším širším okolím. Na základě lidského chování a postojů se rozhodují také státní instituce a soukromé firmy, aby mohly optimalizovat své služby. Aplikace nachází postojová analýza téměř ve všech průmyslových odvětvích zdravotnictvím začínaje politickým konzultantstvím konče. Dříve se výzkumy veřejného mínění prováděly pomocí manuálních dotazníkových šetření. Dnes se k tomuto účelu využívají speciálně vyvinuté komputační metody, jejichž vznik se datuje na počátek tohoto tisíciletí a byl značně stimulován textovými daty pocházejícími ze sociálních médií (např. tweety či on-line diskuze).

Rozdíl mezi emocemi a postoji nemusí být vždy na první pohled patrný, protože se navzájem často doplňují. Např. věta: „Stav české politiky mě znepokojuje," vyjadřuje emoce, kdežto věta: „Česká politika je na tom špatně," sděluje názor. Zatímco na první větu se přirozeně nabízí odpověď: „Cítím to stejně," na druhou bychom spíše odvětili: „Souhlasím/nesouhlasím s tebou." Nicméně obě věty jsou navzájem sémanticky spjaté, protože nálada popsaná v první větě je pravděpodobně způsobená stanoviskem v druhé. Stejně tak ale můžeme říci, že nálada vyjádřená ve větě první implikuje názor na českou politiku ve větě následující. V některých případech (např. „Myslím, že pojedu vlakem.") však názory žádné emoce implikovat nemusí. Postojová analýza se nicméně primárně soustředí na stanoviska vyjadřující pozitivní či negativní nálady.[93]

Sdělení vyjadřující pozitivní nebo negativní emoce jsou většinou subjektivní. Ovšem také věty objektivní, konstatující fakta, mohou implikovat emoce svých autorů tím, že popisují jimi chtěné či nechtěné skutečnosti. Kupříkladu z věty: „Boty, které jsem si koupil teprve před měsícem, mám již děravé," můžeme díky naší obecné znalosti světa

[93] Liu, 2015, s. 2.

poměrně spolehlivě usoudit, že autor pociťuje v souvislosti se svými botami negativní náladu.[94]

Při klasifikaci emocí a názorů je nutné brát v potaz také negaci, která může obracet nejenom význam slova (např. „Chci ne vodu, ale minerálku."), ale i celé věty (např. „Ano, opravdu mi tenhle výtečný kečup nechutná.").

Vzhledem k tomu, že klíčovou funkcí sociálních médií je možnost vyjádřit skrze ně svůj názor, stojí postojová analýza v centru jejich výzkumu i vývoje. Efektivní je v této oblasti hlavně proto, že lidé diskutují o tématech či událostech a vyjadřují k nim svá stanoviska, pokud jsou pro ně důležitá. Každá služba přitom svým uživatelům nabízí různé způsoby komunikace, které ovlivňují styl, jakým se vyjadřují. Vezměme si např. tweety – kvůli tomu, že jsou rozsahem omezeny pouze na 280 znaků, vyjadřují se jejich pisatelé zpravidla přímo k věci a neformálně. Z těchto důvodů se na nich podle Liu postojová analýza provádí snáze, než na delších a formálnějších textech, a její výstupy bývají přesnější.[95]

Při samotné analýze se rozlišují dva hlavní komponenty názoru/hodnocení: cíl hodnocení a hodnotící výraz spolu s jeho sentimentem. Ten bývá nejčastěji vyjádřen jako negativní, pozitivní, případně neutrální; pomocí numerického hodnocení; nebo může být kategorizován do předem určených emocí jako např.: vztek, znechucení, obavy, smutek, očekávání, překvapení, radost a důvěra. Dalšími komponenty názoru/hodnocení jsou jeho zdroj (autor) a čas, v němž bylo realizováno.[96]

Analýzu můžeme provádět na úrovni[97]:

a) dokumentu – kdy implicitně předpokládáme, že každý dokument vyjadřuje názor pouze na jednu entitu (např. produkt);
b) věty;
c) a na úrovni aspektu.

[94] Ibid., s. 3–4.
[95] 2015, s. 17.
[96] Ibid, s. 18.
[97] Ibid., s. 9.

Analýza na úrovni dokumentu, ani na úrovni věty většinou neodhalí přesný cíl hodnocení, jako nejefektivnější se proto jeví analýza na úrovni aspektu. Víme-li například, že věta: „Mám rád knihy nakladatelství Paseka," vyjadřuje celkově pozitivní emoce, pořád netušíme, k čemu přesně se vážou – ke knihám konkrétního českého nakladatelství. Víme jen, že např. věta č. 25 je pozitivní, ale nevíme už, co je jejím obsahem. Ne vždy je však lehké určit celkový emotivní náboj věty či názor v ní obsažený, jako např. zde: „V naší špatné ekonomice se nakladatelství Paseka daří dobře." Zatímco o ekonomice autor vyjadřuje negativní názor, o nakladatelství pozitivní. Analýza na úrovni aspektu se proto soustředí přímo na cíl hodnocení a hodnotící výraz spolu s jeho sentimentem.[98]

7.1 Cíl hodnocení a entita

Cíl hodnocení můžeme definovat jako entitu nebo její část, ke které se hodnocení (názor či emoce) vztahuje. Entita je pak např. tématem, produktem, událostí apod. a definovat ji lze jako pár *e: (T, W)*, kde *T* je hierarchií jejích částí a *W* je množinou jejích vlastností. Přičemž každá část disponuje rovněž svou vlastní množinou vlastností.[99]

Vezměme si např. text: „Před týdnem jsem si koupil knihu Literární spolek Laury Sněžné. Miluji ji! Má moc pěknou grafickou úpravu a obálku. Má přítelkyně však říká, že kvůli zvolenému fontu je špatně čitelná a má moc stran." Entitou je zde kniha Literární spolek Laury Sněžné, jejímiž vlastnostmi jsou grafická úprava a čitelnost, a jejímiž částmi jsou obálka a strany.

Výše zmíněná definice popisuje hierarchii částí entity založenou na vztazích mezi nimi, která by se dala znázornit stromovým grafem. Kořenovým uzlem by byl v takovém případě název entity (Literární spolek Laury Sněžné) a všechny ostatní uzly by byly jejími částmi. Názor by se přitom mohl vztahovat ke kterémukoli uzlu či jeho vlastnosti. Druhá věta z ilustračního textu tak vyjadřuje pozitivní emoce o celé knize, kdežto třetí věta značí pozitivní názor o její vlastnosti (grafická úprava) a části (obálka).

[98] Ibid.
[99] Ibid., s. 19–20.

7.2 Sentiment

Sentiment je skrytý pocit, postoj, hodnocení či emoce asociovaná s názorem. Liu nabízí jeho reprezentaci jako trojici *(y, o, i)* kde *y* je typem sentimentu, *o* jeho orientací a *i* intenzitou. Typy sentimentu mohou být různé, od lingvisticky přes psychologicky založené až po speciálně vytvořené pro spotřebitelský výzkum.[100] Vzhledem k tomu, že v této práci nebudu klasifikaci sentimentu z hlediska typů provádět, vynechám nyní jejich podrobnější výklad.

Orientace sentimentu může být pozitivní, negativní nebo neutrální, přičemž neutrální značí obvykle buď jeho absenci nebo žádný sentiment. Intenzitu svých emocí a názorů většinou lidé v textech vyjadřují dvěma způsoby. Tím prvním je užití výrazů (slov či frází), které mají hodnotící povahu. Např. „dobrý" značí menší hodnotu než „skvělý" a stejně tak „špatný" implikuje slabší sentiment než „strašlivý". Druhým způsob představují intenzifikační částice, jež stupňují intenzitu sentimentu. Jedná se o slova jako např. „opravdu", „extrémně" nebo „šíleně".[101]

Výše předložený popis sentimentu vychází více z komputačního než z lingvistických přístupů, které se vztahují ke všem hlavním oblastem popisu jazyka: od gramatiky přes lexikon až po pragmatiku.[102] Jeho cílem není zcela objasnit role konkrétních lingvistických aspektů vyjadřujících emoce, nýbrž nabídnout jeden z možných základních přístupů k analýze sentimentu, jenž postačuje záměru této práce.

V předchozích kapitolách jsme si představili metodu dolování textových dat vycházející z vědecké tradice Digital Humanities. Podívali jsme se na to, jakým způsobem lze pomocí počítače objevit témata skrytá v textech. A nahlédli jsme také do oblasti sociálních médií, skrze něž v současnosti komunikují lidé na celém světě, čímž vytvářejí obrovské množství dat využívaných např. k vytěžování postojů. A protože porozumět lidské komunikaci je složitý úkol, osvojili jsme si spolu Eulerem některé ze základních konceptů z teorie sítí. Znázorní-li se totiž některé výzkumné problémy formou grafu, stanou se rázem jasnějšími. Abychom však nezůstali pouze u teorie,

[100] 2015, s. 20.
[101] Ibid, s. 21.
[102] Veselovská, 2017, s. 7.

ukážeme si nyní v praktické části této práce aplikace výše popsaných metod při zkoumání komunikace na mikroblogovací síti Twitter. Záměrem následující analýzy bude nahlédnout, jakým způsobem zde diskutují lidé o enviromentálním problému znečištění plasty. Cílem pak bude odhalit konkrétně o jakých tématech zde lidé v souvislosti s tímto problémem píší a jaké k němu zaujímají postoje. Avšak ještě před započetím samotné analýzy o znečištění plasty krátce pojednáme.

8 Znečištění plasty

Znečištění plasty je jedním z globálních enviromentálních problémů. Přestože ve své současné podobě existují plasty teprve posledních 60 až 70 let, změnily téměř všechna průmyslová odvětví: textilním počínaje, vesmírným inženýrstvím konče. Snídáme jogurty z plastových kelímků, nosíme oblečení s příměsí elastanu, pijeme vodu z plastových lahví, před spaním polykáme pilulky z platových krabiček. Plasty jsou s námi 24 hodiny denně, doprovázejí nás od narození až do smrti. Jednou z jejich největších výhod je totiž to, že jsou navrženy, aby vydržely velmi dlouho. Téměř všechny plasty, jež byly kdy vytvořeny, stále existují.[103]

Odhaduje se, že ku dnešnímu dni bylo vyrobeno více než 8 miliard tun plastů a pokud bude současný trend pokračovat, v roce 2050 jich bude na skládkách či ve volné přírodě na 12 bilionů tun.[104] Nejvíce se jich spotřebuje na obaly, které jsou z drtivé většiny pouze na jedno použití. Žádný z běžně používaných plastů přitom není biologicky rozložitelný, a proto na ně můžeme narazit téměř kdekoli: plastový odpad byl nalezen na pobřežích těch nejodlehlejších ostrovů i na dně všech největších oceánských pánví.[105] Je dokonce natolik všudypřítomný, že byl některými vědci navržen jako ukazatel antropocénu.[106]

Ve světových oceánech existuje celkem pět plastových skvrn, které díky mořským proudům absorbují veškerý plovoucí odpad. Vědci se domnívají, že největší z nich – Velká tichomořská plastová skvrna nacházející se mezi Kalifornií a Havají – obsahuje na 1.8 trilionů plastových částic a exponenciálně roste.[107] Tato situace představuje problém jak pro živočichy žijící v mořích a oceánech, tak pro lidi. Na pobřežích např. Španělska[108] či Norska[109] byly nalezeny velryby, které měly ve svých útrobách

[103] „Seven charts that explain the plastic pollution problem“, 2017, cit. 20. 04. 2018.
[104] Geyer, Jambeck, & Law, 2017, cit. 20. 04. 2018.
[105] Barnes, Galagni, Thompson, & Barlaz, 2009, cit. 20. 04. 2018.
[106] Zalasiewicz, et al., 2016, cit. 20. 04. 2018.
[107] Lebreton, et al., 2018, cit. 20. 04. 2018.
[108] „Whale found dead“, 2018, cit. 13. 04. 2018.
[109] Horton, 2017, cit. 20. 04. 2018.

až 29 kilogramů plastů. Plastové mikročástice byly ale objeveny také v potravinách jako jsou balená voda[110], pivo[111] a med[112].

Pokud plasty nekončí ve volné přírodě či na skládkách, jsou buď tepelně zničeny, nebo recyklovány. Odhaduje se, že např. v Evropě je jich ročně recyklováno přibližně 30 %. K recyklaci plastů se při výrobě svých produktů uchyluje stále více soukromých firem jako např. americká outdoorová firma Patagonia nebo německý Adidas, který ve spolupráci s projektem Parley vyvinul řadu sportovního oblečení a bot vyrobených z plastového odpadu vyloveného z moře. Podobné projekty mají za cíl zvýšit všeobecné povědomí o výše nastíněném problému a nabídnout ekologicky šetrné alternativy spotřebního zboží. Dále kupříkladu Tchaj-wan plánuje do roku 2030 zcela zakázat prodej jednorázových plastových výrobků, konkrétně: sáčků, kelímků a slámek.[113] A v neposlední řadě nizozemská nezisková organizace The Ocean Cleanup vyvíjí systém, který by dokázal lapit a odebrat plasty z oceánů jednou pro vždy.

Závažnost problému znečištění plasty a rostoucí snaha některých ať už státních či soukromých subjektů upozornit na něj veřejnost, případně nabídnout alespoň částečné řešení byly impulsem pro následující datovou analýzu.

[110] Mason, et al., 2018, cit. 20. 04. 2018.
[111] Kosuth, Mason, & Wattenberg, 2018, cit. 20. 04. 2018.
[112] Liebezeit, & Liebezeit, 2013, cit. 21. 04. 2018.
[113] „Taiwan to ban disposable plastic", 2018, cit. 21. 04. 2018.

9 Diskuze o znečištění plasty na Twitteru

Cleaning and transforming data may not bet the fun part [...],
but you'll probably spend more of your life than you care to doing it.[114]

V předchozích kapitolách jsme popsali metody pro dolování, extrakci, zpracování a vizualizaci textových dat, které nyní upotřebíme v praktické části této práce. Budeme je aplikovat na dataset anglicky psaných *tweetů* dotýkajících se explicitně tématu znečištění plasty za účelem odhalení dalších, předem neznámých, témat, s nimiž je pisatelé spojují. K těmto účelům použijeme programovací jazyk R[115], který představíme na dalších řádcích. Pro síťovou vizualizaci dat využijeme program Cytoscape[116], jímž se budeme dále rovněž blíže zabývat. Poté, co detekujeme a zobrazíme prostřednictvím grafů témata obsažená v datasetu, budeme se věnovat odhalení sentimentu obsaženém v *tweetech* a postojům pisatelů, jež zaujímají vůči detekovaným tématům a entitám, které jsou jejich součástí.

9.1 Programovací jazyk R

Tato datová analýza byla provedena v programovacím jazyce R určeném primárně pro statistické výpočty a jejich vizualizace. Jednou z výhod R je veliké množství dostupných ať už statistických nebo grafických metod, které mohou být dále rozšiřovány pomocí tzv. „knihoven" obsahujících nové funkce. Knihovny může vytvořit kdokoli disponující potřebnými dovednostmi a dát je poté k dispozici celosvětové komunitě uživatelů jazyka R. Ta zahrnuje jak programátory, datové vědce a analytiky, tak statistiky či např. bioinformatiky a v neposlední řadě rovněž lingvisty. Výhodou této komunity je množství publikací, návodů a online fór usnadňujících programování funkcí – v našem případě pro zpracování *tweetů* – a manipulaci s daty – tj. jejich zpracování

[114] Cotton, R. 2013. s. 191.
[115] www.r-project.org, cit. 22. 04. 2018.
[116] www.cytoscape.org, cit. 22. 04. 2018.

a vizualizaci. Před samotným programováním a manipulací s daty je ale nejdřív musíme získat. Popíšeme si proto nyní, jak se skrze R připojit k Twitteru a stáhnout z něj dataset *tweetů*.

9.2 Knihovna jazyka R: rtweet & Twitter API

Pro získání dat z Twitteru a základní manipulaci s nimi jsme použili knihovnu rtweet[117], který k tomuto účelu vytvořil Michael W. Kearney. Knihovna rtweet se připojuje k rozhraní pro programování aplikací (Application Programming Interface; dále API) Twitteru pomocí přístupových údajů, které získá zdarma kdokoli, kdo o ně na jeho internetových stránkách pro vývojáře[118] zažádá.

API je souborem funkcí a procedur, které slouží uživateli k tomu, aby při komunikaci s aplikací či databází dostal požadovanou odpověď na svůj příkaz. Představit si API můžeme např. jako číšníka v restauraci[119], jenž standardizuje komunikaci hostů s kuchaři. Když si vybereme z jídelního lístku, potřebujeme, aby se naše objednávka dostala do kuchyně, a aby nám bylo oplátkou na stůl doneseno jídlo. Požádáme tedy číšníka, který objednávku předá kuchaři a o chvíli později nám jídlo přinese. Stejně tak nám API Twitteru dodá námi požadované *tweety*.

Přístup k datům na Twitteru pomocí API je rozdělen do tří kategorií: standartní, prémiový a podnikatelský. Standartní je zcela zdarma, další dvě kategorie jsou zpoplatněny. Kategorie se liší převážně v tom, jak staré *tweety* lze získat, dále např. nabídkou analytických nástrojů ze strany Twitteru. Data analyzovaná v této práci byla získána skrze přístup standartní, který od 1. ledna 2018 umožňuje pouze stahování *tweetů* zveřejněných nejpozději před deseti dny od data stahování. Pro naše účely je standartní přístup dostačující. Pojďme se dále podívat, jaká data jsme pro naši analýzu získali.

[117] Kearney, M. W. (2017). rtweet: Collecting Twitter Data. R verze knihovny 0.6.0. Získáno z https://cran.r-project.org/package=rtweet, cit. 22. 04. 2018.
[118] apps.twitter.com, cit. 22. 04. 2018.
[119] MuleSoft Videos, 2016, cit. 22. 04. 2018.

9.3 Metodika sběru dat

Vzhledem k malému množství zpráv publikovaných na Twitteru, jež se explicitně vztahují k tématu znečištění plasty a jsou psány v českém jazyce, jsme se rozhodli pracovat se zprávami psanými v jazyce anglickém. Pro potřeby této analýzy bylo pomocí funkce search_tweets[120] staženo celkem 64 299 *tweetů* obsahujících výraz „plastic pollution", jež byly publikovány v časovém rozmezí od 13. 04. 2018 19:01:20 do 23. 04. 2018 18:24:05 letního českého času.

Výsledný dataset byl převeden do základního datového formátu R: data frame suplujícího tabulku, abychom si byli jistí, v jakém formátu data máme a mohli s nimi dále přehledně pracovat. Každý *tweet* je tvořen 42 částmi a vlastnostmi jako jsou např. jméno uživatele, plný text zprávy nebo informace o tom, zdali se jedná o *retweet*, či nikoli. Ne všechny *tweety* však obsahují všechny části a vlastnosti. Kupříkladu název místa, z něhož byl konkrétní *tweet* odeslán, lze zjistit jen velmi zřídka, protože tuto informaci přidává manuálně ke *tweetu* jen málo uživatelů. Absence údaje je v takovém případě značena neznámou hodnotou, zkracovanou jako NA (not available).

Díky funkcím knihovny rtweet lze z datasetu jednoduše extrahovat jednotlivé části a vlastnosti *tweetů* jako např. hashtagy či informace, zdali se jedná o *retweet*, či nikoli. Pro následující manipulaci s daty je však nezbytné užití dalších ať už předpřipravených nebo vlastních funkcí. To platí zejména pro zpracování *tweetů* a vytvoření korpusu, jemuž se budeme věnovat na následujících řádcích.

9.4 Vytvoření korpusu tweetů

Korpus *tweetů* vytváříme, abychom mohli přehledně pracovat s *tweety* beze všech jejich, pro naše účely, nadbytečných částí a vlastností; tj. abychom pracovali pouze s texty, resp. slovy. Prvním krokem k vytvoření korpusu je proto oproštění *tweetů* od všech pro nás nadbytečných částí a vlastností. Pokud bychom to provedli pomocí základní funkce knihovny rtweet, získali bychom sice veškeré texty, avšak část z nich – texty pocházející

[120] Kearney, M. W. (2017). rtweet: Collecting Twitter Data. R verze knihovny 0.6.0. Získáno z https://cran.r-project.org/package=rtweet, cit. 22. 04. 2018.

z *retweetů* – by nebyla ve své úplné délce. Provedení analýzy na neúplných textech by následně mohlo vést ke zkresleným závěrům. Z tohoto důvodu jsme vytvořili vlastní funkci replaceRetweetedText, jejímž vstupem je původní tabulka a výstupem tabulka pozměněná. Funkce projíždí tu původní řádek po řádku a pokaždé, když narazí na *retweet*, nahradí jeho plným textem zkrácený text zprávy původní. Teprve z takto pozměněné tabulky jsme pomocí základní funkce knihovny rtweet oddělili texty pro další zpracování.

Texty *tweetů*, stejně jako zprávy publikované v jiných sociálních médiích, obsahují často hypertextové odkazy, emotikony, číslice a matematické znaky (např. +, -). Vzhledem k tomu, že jsme do korpusu chtěli zahrnout pouze slova, veškeré výše zmíněné grafické symboly spolu s interpunkčními znaménky musely být odstraněny. Pro tento účel jsme vytvořili funkci cleanTweets, jejímž výstupem jsou texty v požadované formě.

Následně jsme veškeré texty tokenizovali prostřednictvím k tomu napsané funkce tokenizeListsOfTweets. Ze vzniklého seznamu tokenizovaných textů bylo dále pro naše potřeby nutné odstranit všechna stop slova. Jedná se o slova, která se v korpusu vyskytují příliš často a zároveň vypovídají jen velmi málo o jeho sémantickém obsahu. Pro naši analýzu jsme použili seznam anglických stop slov z knihovny tm[121] určené pro dolování textových dat. Základních 174 stop slov jsme doplnili o 6 dalších na základě vlastních zkušeností. Výsledný seznam StopWords_pp_included (ve formátu CSV) obsahoval celkem 180 stop slov, do nichž jsme zahrnuli také „plastic“ a „pollution“. A sice proto, že se tato slova vyskytovala nutně v každém *tweetu* zkoumaného datasetu, a tudíž měla pro další analýzu z hlediska objevení skrytých témat pouze malou informační hodnotu. Ze seznamu tokenizovaných textů jsme stop slova odstranili pomocí vlastní funkce removeStopWordsFromListOfTokenizedTweets.

Poté jsme za užití funkce Corpus[122] převedli výsledný seznam do datové struktury korpusu, kde každý jeden tokenizovány text představoval jeden dokument. Korpus jsme

[121] Ingo Feinerer & Kurt Hornik (2017). tm: Text Mining Package. R verze knihovny 0.7-3. Získáno z https://CRAN.R-project.org/package=tm, cit. 22. 04. 2018.
[122] Ibid.

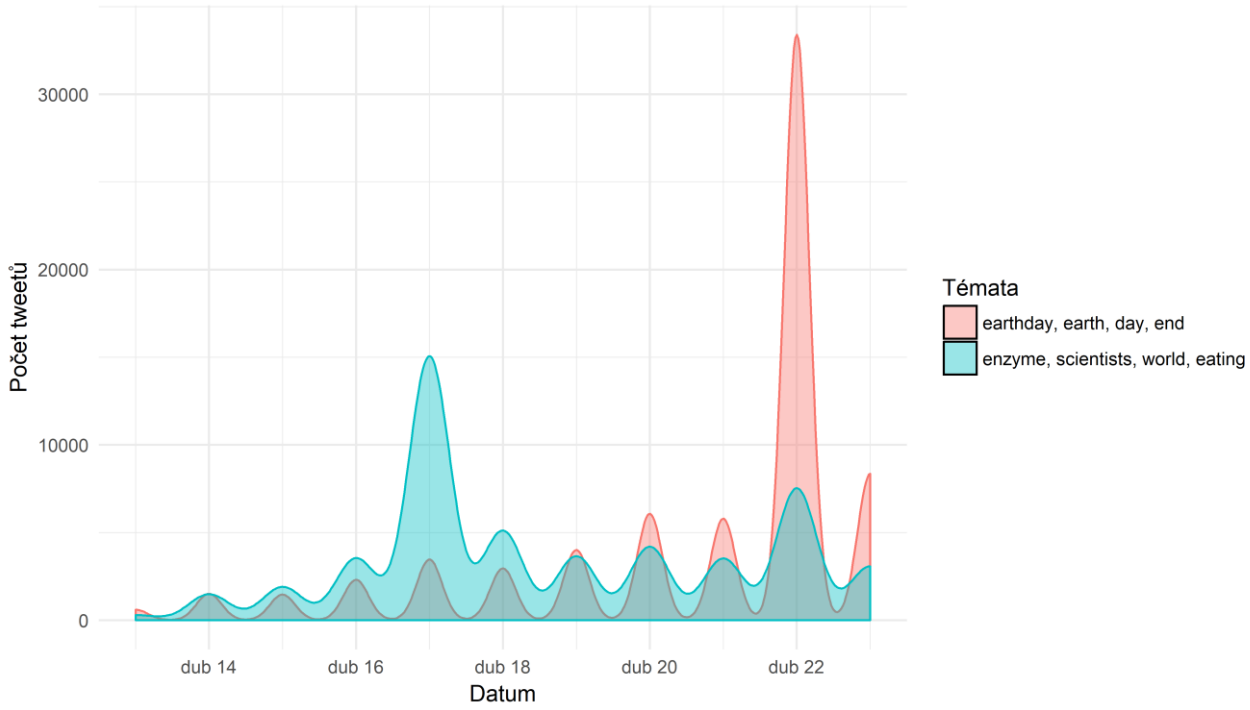
dále pomocí funkcí TermDocumentMatrix[123] a as.DocumentTermMatrix[124] převedli na matici, která je požadovaným vstupem námi zvolené implementace algoritmu Latentní Dirichletovy alokace.

9.5 Detekce tematických celků pomocí Latentní Dirichletovy alokace

Pro objevení skrytých témat z korpusu jsme použili nejprve knihovnu topicmodels[125] od Bettiny Grünové a Kurta Hornika. Konkrétně jsme pro větší přesnost při určování témat zvolili metodu Latentní Dirichletovy alokace založené na Gibbsově vzorkování.

Kvůli komputačním nárokům algoritmu a časové náročnosti výpočtů jsme se rozhodli použít Latentní Dirichletovu alokaci (LDA) pouze pro detekci dvou nejvýraznějších témat zastoupených v našem korpusu. Pro dobrou přehlednost výstupu jsme stanovili počet termínů charakterizujících každé téma na 4. Vzhledem k tomu, že každý *tweet* obsahuje údaj o tom, kdy byl publikován, lze zastoupení témat sledovat i v průběhu času. Výsledkem takového sledování je graf č. 1.

[123] Ibid.
[124] Ibid.
[125] Grün, B. & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. R vezre knihovny 0.2-7. Získáno z https://cran.r-project.org/package=topicmodels, cit. 22. 04. 2018.



Graf č. 1: Dvě nejvýraznější témata obsažená v korpusu a jejich vývoj v čase

V grafu č. 1 vidíme dvě nejvýraznější témata: každé z nich je znázorněno křivkou a odpovídající barevnou plochou pod ní. Přestože jsme předpokládali jejich částečnou homogennost, skutečnost, že se termíny v nich obsažené nepřekrývají svědčí o opaku. Kromě toho můžeme určit, že největší koncentrace červeně znázorněného tématu nastala 22. dubna, kdežto modře vyvedené téma kumuluje nejvíce dubna 17. a následně jeho koncentrace stoupá o pět dní později spolu s tématem prvním.

Už samotné termíny nám napoví, čeho se jednotlivá témata týkají. První z nich (pořadí je určeno jejich frekvencí) obsahuje termíny: „earthday“, „earth“, „day“ a „end“. Lze se tedy domnívat, že 22. dubna byl Den Země – nahlédnutí do kalendáře nám naši domněnku potvrdí. Na druhou stranu nám ale neodhalí, co má společného Den Země se znečištění plasty.

Pro tento účel jsme napsali funkci findTweetsContainingSpecificTerms, do níž jsme zadali dva první termíny[126] („earthday“ a „earth“), a nechali si poté vypsat veškeré unikátní *tweety*, jež je oba zároveň obsahují. Následně jsme provedli prostý náhodný výběr tří z nich, které nám poskytly vodítko k interpretaci vztahu mezi Dnem Země a znečištěním plasty. Dozvídáme se z nich, že tématem Dne Země 2018 je ukončení problému znečištění plasty, viz tabulka č. 1.

"Earth Day is focused on ending plastic pollution – a timely topic, considering the giant floating mass of plastic in the Ocean. This is a shocking story! To gain perspective. Watch: A Plastic Ocean Official Trailer https://t.co/NFuAJH6tnS via @YouTube. #earthday #planet #earth https://t.co/r1iqiHbsCX"
"Today we bring our attention to #earthday \nThis year the topic on Earth Day is a world free of plastic pollution!\nLet's all take care of our home \xf0\u009f\u0092\u09a\n#earthday2018 #home #plasticfree #takecare #earth #greenearth #plasticpollution #lovetheplanet #ourplanet #ourhome #seva #volunteer https://t.co/O7l1yBNoiS"
"Today is #earthday end plastic pollution. Our #earth close up find ID14515993 #australianviews… https://t.co/q0QDMCBTm0"

Tabulka č. 1: *Tweety* obsahující termíny „earthday“ a „earth“

[126] Vstupem funkce jsme zvolili záměrně pouze dva, a ne všechny termíny, protože jsme si nemohli být jistí, zda některé *tweety* daného tématu vůbec zároveň všechny čtyři termíny obsahují. Funkce je citlivá na velká a malá písmena, a zároveň nerozlišuje mezi termíny a hashtagy: vyhledá jak zprávy obsahující „earthday“, tak ty, jež obsahují „#earthday“. Tato její vlastnost však nebrání účelu, ke kterému jsme ji použili.

Odvození konkrétního obsahu druhého tématu jsme provedli stejným způsobem. Termíny: „enzyme“, „scientists“, „world“ a „eating“ napovídají, že je zde referováno o vědě, světě a jídle. Na první pohled ale není jasné, jaký vztah je pojí. Nechali jsme si proto vypsat veškeré unikátní *tweety* obsahující zároveň termíny „enzyme“ a „scientists“. Na základě nahlédnutí tří z nich náhodně vybraných jsme byli schopni odvodit následující závěr: *tweety* informují o enzymu, který je údajně schopný pojídat plast, a vědci v něm spatřují pomoc při boji proti znečištění plasty, viz tabulka č. 2.

"Plastic-eating enzyme could help fight pollution, scientists say | CBC News https://t.co/IrGzI5GdmN."
"Normally it takes 400 years for a plastic bottle to decompose but now scientists accidentally created an enzyme that eats plastic. YAY! https://t.co/NcKjIyFH6m"
"But hope for an end to plastic pollution may be on the horizon: scientists recently stumbled upon a new enzyme that degrades and digests petroleum-derived products\nhttps://t.co/yNCWoY6wdn."

Tabulka č. 2: *Tweety* obsahující termíny „enzyme“ a „scientists“

Třetí *tweet* ze vzorku poskytuje více informací o výše zmíněném enzymu. Jeho autor v něm spojuje objevení enzymu s potenciálním koncem znečištění plasty, neboť tento enzym údajně dokáže rozložit produkty zhotovené z petrolejových derivátů. V tuto chvíli již máme představu o vzájemném vztahu tří ze čtyř termínů druhého tématu. V dalším kroku jsme se rozhodli osvětlit význam termínu „world“, nechali jsme si proto vypsat všechny *tweety* obsahující termíny „world“ a „scientists“ a pomocí prostého náhodného výběru jsme z nich tři vydělili, viz tabulka č. 3.

"Understanding how bacteria can help degrade plastic could be key to addressing the world's plastic pollution problem this #EarthDay2018. These Korean scientists are help ing advance the field: https://t.co/ZMDZfJ3roF https://t.co/p3GxZmDBlJ"
"In a world first, scientists are testing plastic-eating enzyme in bid to fight pollution https://t.co/k8p5cIMezS https://t.co/AHCWhwUWVg*."*
"According to CNN, scientists are working on further developing a \"plastic eating enzyme\" which they hope will curb the world's pollution problem. #2150CE"

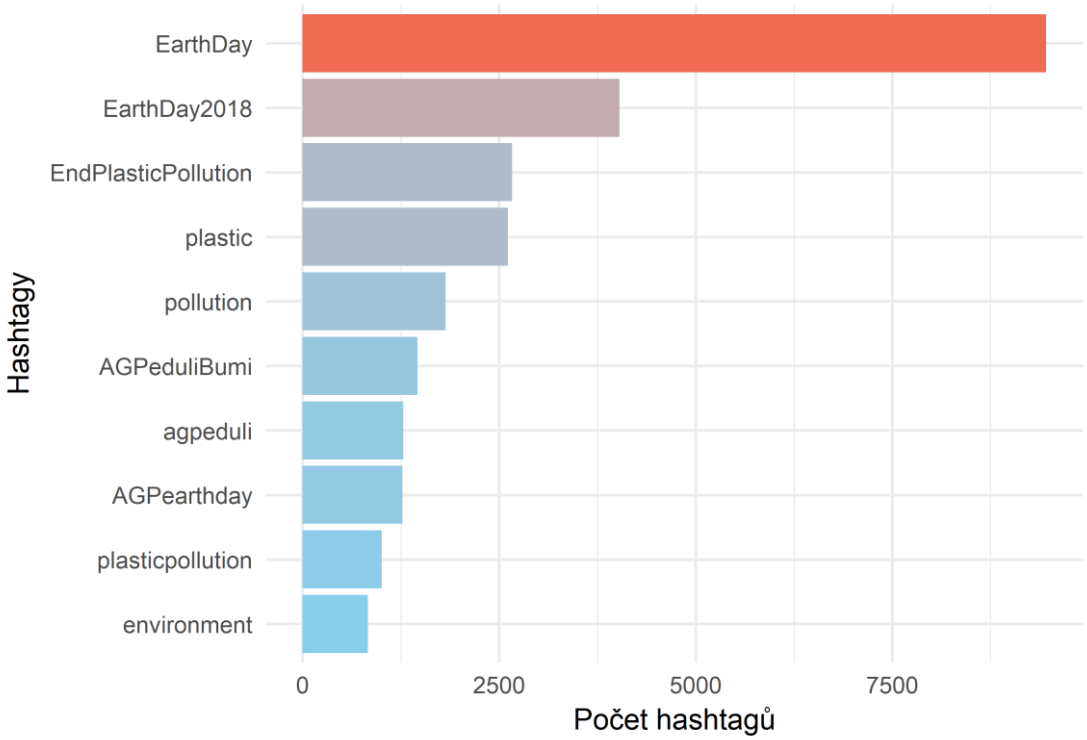
Tabulka č. 3: *Tweety* obsahující termíny „world" a „scientists"

Tweety z tabulky č. 3 nám říkají, že: a) korejští vědci se snaží porozumět tomu, jak dokáže bakterie pojídat plast; b) vědci poprvé na světě testují plasty pojídající enzym; c) dle CNN pracují vědci na vývoji enzymu schopného pojídat plast. V prvním *tweetu* je referováno o světovém problému znečištění plasty, v druhém o tom, že objev enzymu pojídajícího plasty byl učiněn poprvé na světě, a třetí *tweet* referuje o světovém problému znečištění. Na základě nahlédnutí tohoto vzorku můžeme říci, že termín „world" se v rámci druhého nejvýraznějšího tématu vztahuje buď ke světovému problému znečištění plasty, případně k objevu enzymu schopného plasty pojídat.

Nyní již máme představu o dvou nejvýraznějších tématech vyskytujících se v našem korpusu. Jsou jimi Den Země a vědecký objev enzymu schopného údajně pojídat plast. Zjistili jsme, že tématem Dne Země 2018 je boj se světovým problémem znečištění plasty, a že výše zmíněný enzym by mohl být v tomto snažení nápomocný. Dále se podíváme, jak se detekovaná témata projevují pouze na makroúrovni v podobě hashtagů.

9.6 Detekce tematických celků na makroúrovni

Abychom zjistili, jak se detekovaná témata projevují na makroúrovni – jež je tvořena klíčovými slovy, tzv. hashtagy, definovanými uživateli – vytvořili jsme frekvenční tabulku veškerých hashtagů ze všech kompletních *tweetů* našeho korpusu. Zároveň jsme neomezili naši otázku pouze na ty, které se týkají vědeckého objevu a Dne Země, tj. neodebrali jsme z nich stop slova. Z frekvenční tabulky jsme následně vydělili z ilustrativních důvodů prvních deset nejčastějších hashtagů a znázornili je pro dobrou přehlednost grafem č. 2.



Graf č. 2: 10 nejfrekventovanějších hashtagů

V grafu č. 2 vidíme, že dva nejfrekventovanější hashtagy – „#EarthDay“ a „#EarthDay2018“ – se vztahují ke Dni Země. V následujícím kroku jsme se pomocí vlastní funkce findTweetsContainingSpecificTerm, která vyhledá všechny *tweety* obsahující

zvolený hashtag či termín[127], rozhodli podívat, zdali souvisí se Dnem Země též třetí nejfrekventovanější hashtag: „#EndPlasticPollution“. Vyhledali jsme proto veškeré unikátní *tweety* obsahující termín „EndPlasticPollution“ a poté opět provedli prostý náhodný výběr tří z nich. Zjistili jsme, že i jej obsahují *tweety* pojednávající o Dni Země, viz tabulka č. 4.

"Raising Awareness of Plastic Waste: BNF Brings Together Superstore on the #EarthDay2018 to #EndPlasticPollution https://t.co/z1iWOqEAFi https://t.co/AFYtedg5vy"
"Celebrating Earth Day 2018! #Symphony Takes Over NYC and D.C. To End Plastic Pollution! #EarthDay2018 #D2W #NYC #DC #Plastic #LessPlastic #EndPlasticPollution https://t.co/yuXqkAe3el"
"Solving the plastic pollution of our oceans starts at home! Come find out how to reduce your plastic footprint. #EndPlasticPollution https://t.co/hbVoUuQcqc"

Tabulka č. 4: *Tweety* obsahující hashtag „#EndPlasticPollution“

V tabulce č. 4 vidíme, že dva ze tří náhodně vybraných *tweetů* obsahují jak hashtag „#EndPlasticPollution“, tak „#EarthDay2018“. Můžeme tedy předpokládat převahu tohoto tématu nad tématem vědeckého objevu enzymu schopného pojídat plast jak na meso úrovni (celý korpus), tak pouze samostatně na makro úrovni. U textů obsahujících hashtagy „#plastic“, „#pollution“ a „#plasticpollution“ předpokládáme, že se budou vyskytovat napříč celým korpusem nezávisle na tématech, neboť právě

[127] Funkce je citlivá na velká a malá písmena, a zároveň nerozlišuje mezi termíny a hashtagy: vyhledá jak zprávy obsahující „EndPlasticPollution“, tak ty, jež obsahují „#EndPlasticPollution“. Spojení několika slov dohromady bez užití mezer je však typické pro hashtagy – proto lze předpokládat, že právě ty zde budou převažovat.

znečištění plasty je to, co mají úplně všechny zprávy společné – při stahování *tweetů* jsme zadali jako jediné kritérium jejich výběru výraz „plastic pollution“.

Na rozdíl od hashtagů „#plastic“, „#pollution“ a „#plasticpollution“ abychom byli schopni interpretovat hashtagy „#AGPeduliBumi“, „#agpeduli“ a „#AGPearthday“, musíme se nejprve podívat na *tweety*, které je obsahují. Vyhledali jsme proto z korpusu všechny unikátní *tweety* zahrnující „AGPeduliBumi“ a následně provedli prostý náhodný výběr. Přesný význam tohoto termínu/hashtagu jsme však nedokázali zjistit, neboť není v našem vzorku jasně explikován. Často se ale ve *tweetech* vyskytuje spolu s nejfrekventovanějším hashtagem „#EarthDay“, jak ilustruje tabulka č. 5.

"@pandu98dj - End the plastic pollution \n#agpeduli \n#AGPeduliBumi #AGPearthday\n#EarthDay… https://t.co/4bohmSfKLm"
"No Plastic Pollution \n#agpeduli \n#AGPeduliBumi \n#AGPearthday\n#EarthDay \n#EndPlasticPollution \n#lessplastic \n#earthday2018 \n@tamblingwildlife \n@agpeduli \n#AGPeduliCitarum \n#citarumlestari \n#citarumharum\n@tamblingID\n@citarumlestari https://t.co/YJC9doML7U"
"#agpeduli #AGPeduliBumi #AGPearthday #EarthDay #EndPlasticPollution \nlets end plastic pollution https://t.co/QmFQTbta4H"

Tabulka č. 5: *Tweety* obsahující hashtag „#AGPeduliBumi“

Nechali jsme si proto v dalším kroku vypsat veškeré *tweety* obsahující tyto dva termíny/hashtagy („AGPeduliBumi“ a „EarthDay“) zároveň, celkem je jich 1533. Oproti tomu v celém korpusu neexistuje žádný *tweet*, jenž by spolu s hashtagem „#AGPeduliBumi“ či termínem „AGPeduliBumi“ zahrnoval též termíny „enzyme“

nebo „scientists“. Lze tedy usuzovat, že hashtagy „#AGPeduliBumi“, „#agpeduli“ a „#AGPearthday“ se v rámci našeho korpusu vztahují spíše k tématu Dne Země než k vědeckému objevu.

Konečně poslední z deseti nejfrekventovanějších hashtagů „#environment“ se ve *tweetech* vyskytuje jak vedle hashtagů vztahujících se ke Dni Země, tak vedle těch, které referují o vědcích a jejich objevu enzymu schopného pojídat plast. Náš vzorek ale obsahuje rovněž *tweet*, který hovoří o filmu o znečištění plasty. Nicméně o hashtagu „#environment“ lze usuzovat, že je společný oběma hlavním tématům korpusu, viz tabulka č. 6.

"#EarthDay coming up this Sunday, April 22. See what we can do for our homeland. https://t.co/fuOJ9MWVno *\xf0\u009f\u008c\xf0\u009f\u008c\u008d#environment #greenenergy #plastic #pollution #ocean"*
"An Enzyme That Dissolves Plastic Pollution: CityLab https://t.co/Aa27riHYvo #environment More: https://t.co/GO4ceKll1g"
"This beautiful, haunting film highlights the horrors of plastic pollution https://t.co/4qCE7rFvk5 #charity #environment"

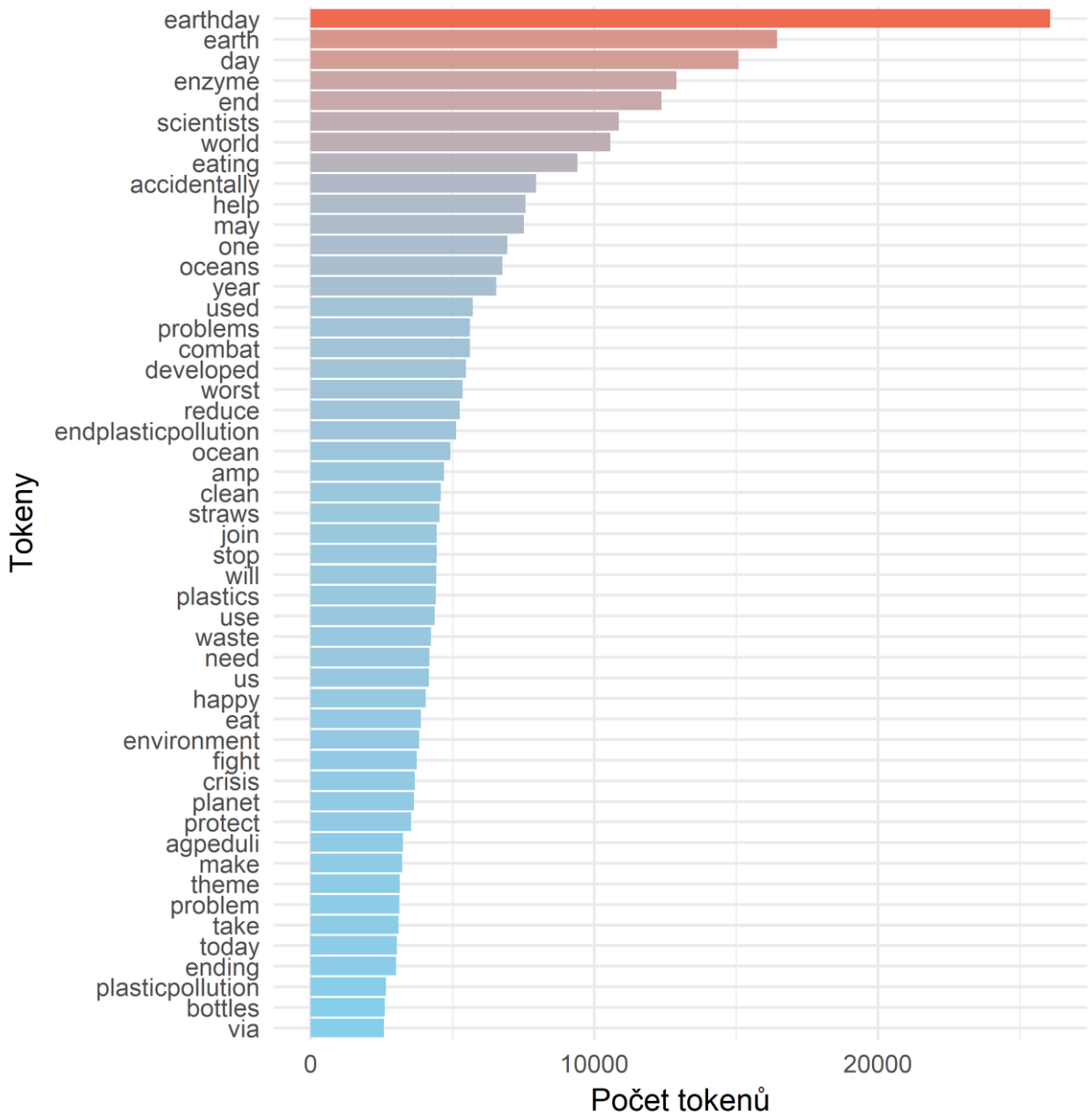
Tabulka č. 6: *Tweety* obsahující hashtag „#environment“

Nyní již máme základní představu o distribuci dvou nejvýraznějších témat. Víme zároveň, že jejich rozložení v meso a makro vrstvě korespondují, a že převažují *tweety* týkající se Dne Země. V následující kapitole se proto na projevy témat v rámci celého korpusu podíváme důkladněji.

9.7 Detekce tematických celků pomocí frekvenčních tabulek

Kvůli komputačním nárokům algoritmu a časové náročnosti výpočtů jsme se rozhodli použít Latentní Dirichletovu Alokaci pouze pro detekci dvou nejvýraznějších témat zastoupených v našem korpusu. Pro zjištění, zdali korpus obsahuje rovněž další shluky tematických celků, jsme zvolili jednodušší a výpočetně méně náročné metody.

Za účelem objevení některých tematických celků jsme se rozhodli nahlédnout distribuce nejfrekventovanějších tokenů a bigramů, které nám pro naše účely poskytly vodítko. Nejprve jsme opět veškeré *tweety* v jejich plné délce zbavili hypertextových odkazů a znaků. V dalším kroku jsme je spojili do jednoho velkého dokumentu a tokenizovali pomocí pro tento účel napsané funkce `tokenizeText`. Následně jsme z celého dokumentu odstranili stop slova (mezi něž jsme zahrnuli i „plasitc“ a „pollution“), zhotovili z něj frekvenční tabulku a pro ilustraci znázornili prvních 50 nejčastějších tokenů následujícím grafem č. 3.



Graf č. 3: 50 nejfrekventovanějších tokenů v korpusu

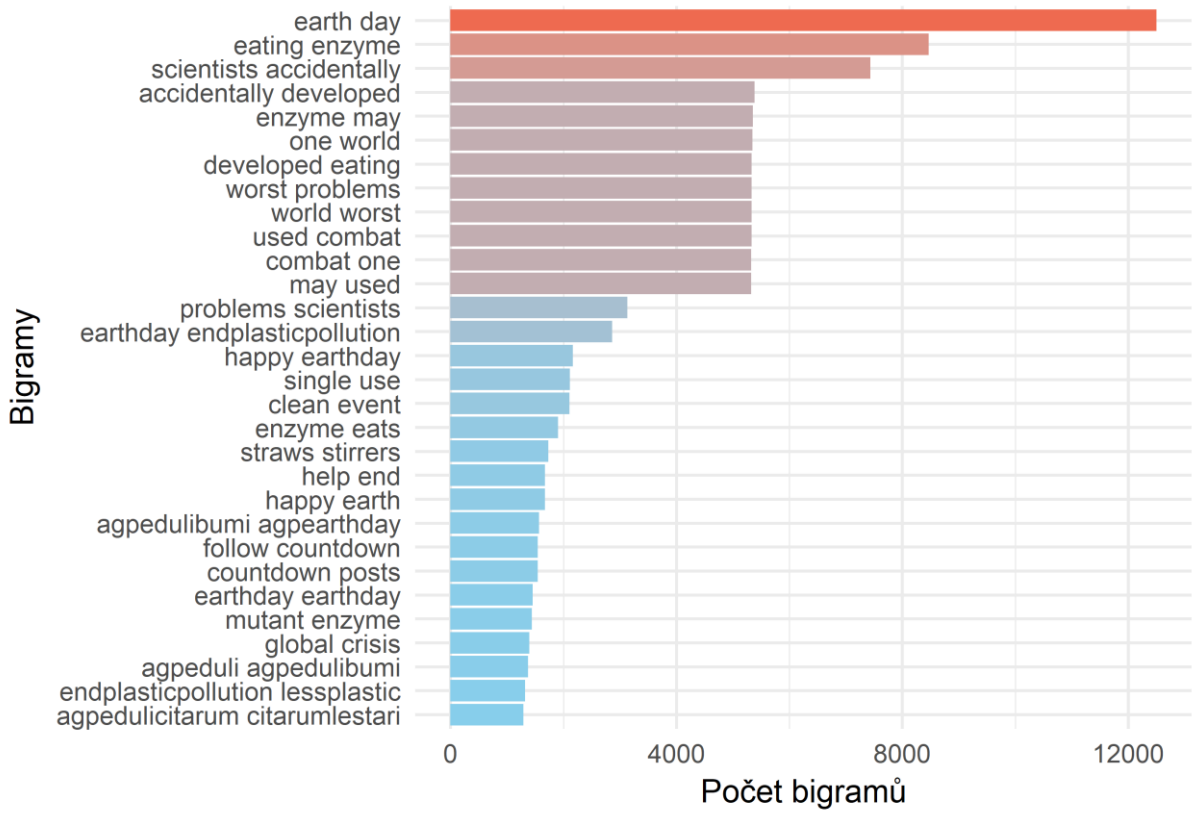
Vidíme, že nejfrekventovanější tokeny jsou: „earthday“, „earth“, „day“, „enzyme“, „end“ a „scientists“, což odpovídá rozložení dvou hlavních témat v korpusu. Významy pěti z nich jsme rozkryli již dříve, jenom token „end“ zde působí na první pohled překvapivě. Povšimneme-li si ale jiného – „endplasticpollution“ – můžeme odvodit, že token „end“ náleží slovnímu spojení „end plastic pollution“. Jeho výskyt v grafu je dán absencí dvou zbylých termínů tohoto slovního spojení, tj. „plastic“ a „pollution“,

které jsme zařadili do stop slov. Tento závěr nám potvrzuje nahlédnutí tří náhodně vybraných *tweetů* zahrnujících token „end“, které zobrazuje tabulka č. 7.

"On Earth Day, Bengalureans vow to end plastic pollution https://t.co/JouqCkSC1X"
"Happy Earth Day! Let's continue the fight to end plastic pollution. \xf0\u009f\u008c\u008f https://t.co/KqUrF5ohvH"
"Let’s take care of this amazing planet we call home. Together we can end plastic pollution, help save endangered species and reduce our ecological footprint. Find out how you can help here: https://t.co/kzYx03YaBb\n\n#EarthDay2018 #EndPlasticPollution #TakeActionToday https://t.co/j2DqtJOBiT"

Tabulka č. 7: *Tweety* zahrnující termín „end“

Dále v grafu č. 3 vidíme např. termíny jako „bottles“, „straws“, „waste“, „problems“, „crisis“, „combat“ nebo „fight“. Některé z nich referují k předmětům vyráběným zpravidla z plastu: k lahvím či slámkám, druhá část termínů značí problémy, krizi a boj – tyto termíny mají negativní konotace. V tuto chvíli ještě nemůžeme usuzovat, zda se výše zmíněné termíny v *tweetech* vyskytují společně, či nikoli. Pro ověření této domněnky jsme se podívali na 30 nejfrekventovanějších bigramů a pro dobrou přehlednost výsledků jsme je znázornili grafem č. 4.



Graf č. 4: 30 nejfrekventovanějších bigramů

V grafu č. 4 si můžeme povšimnout, že termín „problems“ se nejčastěji vyskytuje vedle termínů „worst“, „wordl“ a také „scientists“. Pomocí vlastní funkce findTweetsContaining_3_SpecificTerms jsme zjistili, že v korpusu existuje celkem 5330 *tweetů* obsahujících „problems“, „worst“ a „world“, přičemž 149 z nich je unikátních. Po nahlédnutí vzorku tří takovýchto náhodně vybraných *tweetů* lze dojít k závěru, že referují o jednom z největších světových problémů – znečištění plasty, viz tabulka č. 8.

"Scientists have accidentally developed a plastic-eating enzyme that may be used to combat one of the world's worst pollution problems https://t.co/T1ubCu91Tw https://t.co/gkgaGsxUGR"

"RT CNN \"Scientists have accidentally developed a plastic-eating enzyme that may be used to combat one of the world's worst pollution problems. https://t.co/yWhJHrdMhp https://t.co/JKDoqEehXK\""
"Scientists have accidentally developed a plastic-eating enzyme that may be used to combat one of the world's worst pollution problems. https://t.co/FtAhVQCXeb"

Tabulka č. 8: *Tweety* obsahující termíny "world", "worst", "problems"

Kromě toho jsme pročtením všech 149 unikátních zpráv zjistili, že všechny hovoří o vědeckém objevu enzymu schopného pojídat plast, který by tento problém mohl vyřešit.

Jestliže je znečištění plasty považováno pisateli za jeden z největších světových problémů, lze se domnívat, že bigram „global crisis" bude rovněž hodnocením situace znečištění. Celkem jej obsahuje osm unikátních *tweetů*, které reprezentujeme trojicí z nich náhodně vybraných, viz tabulka č. 9. Dva ze tří *tweetů* této tabulky opravdu hovoří o znečištění plasty jako o globální krizi, přičemž pisatel jednoho z nich se dále domnívá, že znečištění plasty potřebuje svou verzi Pařížské dohody o změně klimatu.

"90 % of sea birds have been found to have plastic in their bellies. This is a global crisis. https://t.co/CFD8jQjICY https://t.co/qucKRm66KJ."
"The global crisis of plastic pollution https://t.co/bPuQ2CXOGr via @newrepublic"

"Plastic pollution is a global crisis in desperate need of its own version of the Paris accord. https://t.co/CFD8jQjICY https://t.co/tXXyjibAFU"

Tabulka č. 9: *Tweety* obsahující bigram „global crisis“

Konečně první *tweet* tabulky č. 9 tvrdí, že u 90 % mořských ptáků byly zjištěny plasty v žaludku. Podívali jsme se tedy kolik zpráv zahrnujících společně termíny „sea“ a „birds“ náš korpus obsahuje: 233, z toho 21 unikátních. Prostým náhodným výběrem jsme z nich vydělili tři a znázornili je tabulkou č. 10.

"An autopsy of a young male sperm whale was found to have 65 pounds of plastic trash crammed into the dead whale's stomach and intestines. Each year an estimated 1 million sea birds, 100,000 sea mammals and countless numbers of fish are killed due to #plastic #pollution. https://t.co/JmbpWaxK7h"
"\"Every day approximately eight million pieces of plastic pollution find their way into our oceans, causing 100,000 marine mammals and turtles and one million sea birds are killed by marine plastic pollution annually.\" #saveourbeaches https://t.co/itOOkmVvAi"
"#Plastic Kills not only fish & birds, but also kills nearly 100,000 sea mammals every year #Pollution #Ocean #Nature https://t.co/JW7NvfJq68"

Tabulka č. 10: *Tweety* obsahující termíny “sea” a “birds”

Dle prvního a druhého *tweetu* z tabulky č. 10 zemře ročně v důsledku znečištění plasty na jeden milion mořských ptáků a 100 000 mořských savců, třetí referuje pouze o mořských savcích. Abychom nahlédli tato čísla konkrétněji, provedli jsme prostý náhodný výběr tří *tweetů* ze všech unikátních hovořících o velrybách. Výsledek ilustruje tabulka č. 11.

"This sculpture at @TMMC is made up of ocean trash from inside a whale. The 51-foot-long sperm whale washed ashore Point Reyes National Seashore with 450lbs of ocean garbage inside its stomach. @TMMC hopes to bring awareness to the plastic pollution problem. @KPIXtv https://t.co/S8JB1w9MCL."
"Plastic pollution killed sperm whale found dead on Spanish beach https://t.co/KcLVlhsTEx"
"A sperm whale has died after swallowing 29 kilos of plastic waste https://t.co/rH1vknmBsK #plasticpollution"

Tabulka č. 11: *Tweety* obsahující termín „whale“

Dva z výše uvedených *tweetů* referujících o velrybách hovoří o jejich smrti způsobené plasty, které měly ve svých útrobách. Znečištění plasty tak nabírá konkrétních obrysů – dozvídáme se o jeho fatálních následcích pro oceánskou faunu. První *tweet* tabulky č. 11 oproti tomu referuje o jednom ze způsobů, jakým lidé následně na tento problém reagují a snaží se na něj upozornit ostatní – vyrábějí z plastového smetí sochy. *Tweetů* obsahujících termín „sculpture“ je v korpusu celkem pouze 40, z nichž 4 jsou unikátní. Velikostí se tedy jedná o marginální téma. Je pro nás ale důležité, neboť nás informuje o podobě reakce umělců na problém. Pro jeho bližší nahlédnutí

jsme provedli prostý náhodný výběr tří *tweetů* ze 4 unikátních a výsledek znázornili tabulkou č. 12.

"World Earth Day… End plastic Pollution … Save Earth… My soap sculpture Art…\n\n#bijucgart… https://t.co/izN6QPqomA"
"This sculpture at @TMMC is made up of ocean trash from inside a whale. The 51-foot-long sperm whale washed ashore Point Reyes National Seashore with 450lbs of ocean garbage inside its stomach. @TMMC hopes to bring awareness to the plastic pollution problem. @KPIXtv https://t.co/S8JB1w9MCL"
"This isn't a real whale, it's a sculpture made out of discarded plastic by Greenpeace, I believe, to showcase the effects of marine pollution on aqautic life https://t.co/fmEUBn2OsJ"

Tabulka č. 12: *Tweety* obsahující termín „sculpture“

První *tweet* z tabulky č. 12 referuje o explicitně blíže nespecifikované soše z mýdla, druhý o soše vyrobené z plastového smetí nalezeného uvnitř mrtvé velryby, třetí pak o soše z plastového odpadu zhotovené enviromentální organizací Greenpeace.

Jedněmi z padesáti nejfrekventovanějších tokenů jsou dále „bottles“ a „straws“ – většinou plastové výrobky na jedno použití. Kromě toho je šestnáctým nejčastějším bigramem „single use“. V dalším kroku jsme se proto podívali, zda se v rámci korpusu vyskytují termíny „bottles“, „straws“ a bigram „single use“ ve *tweetech* společně, a pokud ano, co z nich můžeme vyčíst. Celkem je v něm 51 takovýchto unikátních *tweetů* – prostým náhodným výběrem jsme z nich 3 nahlédli blíže, viz tabulka č. 13.

"We can do our part to end plastic pollution: say no to plastic straws, ditch the plastic water bottles, and try to cut down on single use plastic. Let's help keep the local waterways and environments of St. Petersburg… (more:https://t.co/szqUXQsWvG) https://t.co/KtSrdAAUwq"
"We can do our part to end plastic pollution: say no to plastic straws, ditch the plastic water bottles, and try to cut down on single use plastic. Let's help keep the local waterways and environments of Muskoka clean… (more:https://t.co/mrTZu1RxfI) https://t.co/dRWvf6Uciu"
"They can have my plastic straws when they ban them! Which is of course, WHY they should definitely ban them. Same with my beloved single use water bottles and plastic latte cups. And my bags! I need coerced behavior modification, dammit!! \n\nhttps://t.co/MxZ5SqkAq3"

Tabulka č. 13: *Tweety* obsahující termíny "bottles", "straws" a bigram "single use"

První dva *tweety* z tabulky č. 13 vybízí čtenáře ke zřeknutí se plastových slámek, lahví na vodu a k pokusu o omezení užívání jednorázových plastů. Kromě toho jej také pobízí k udržování čistoty vodních kanálů a okolí Petrohradu a Muskoky: je zde patrný apel na lokálně situované čtenáře. Třetí *tweet* tabulky č. 14 představuje osobní zpověď pisatele chtějícího omezit svou spotřebu plastových výrobků na jedno použití.

Tweety obsahující termín „straws" vyzývaly čtenáře, aby je přestal používat. Můžeme se proto domnívat, že ty obsahující bigram „straws stirrers" budou apelovat na čtenáře, aby se vzdal též plastových míchadel, které se někdy podávají např. ke kávě či zmrzlině. Vyhledali jsme proto všechny unikátní zprávy obsahující termíny

„straws“ a „stirrers“.[128] Celkem je jich 75 unikátních, tabulka č. 14 zobrazuje tři z nich náhodně vybrané.

"Cities and nations are looking at banning #plastic straws and stirrers in hopes of addressing the world's plastic pollution problem. https://t.co/cpofi51Pnd @analysisnytimes https://t.co/sluieXkcHh"
"Britain plans to ban plastic straws, cotton swabs, and coffee stirrers\n\nBy Alex Bellury\n\nhttps://t.co/EpHrpGLnLs\n\n#Pollution #Nature #Environment #Europe"
"The Government says it is prepared to ban the sale of plastic straws, cotton buds and drinks stirrers in England in a bid to cut pollution. Prime Minister Theresa May said… https://t.co/Mfy3romHWR"

Tabulka č. 14: *Tweety* obsahující termíny „straws“ a „stirrers“

Tabulka č. 14 ukazuje *tweety* zmiňující plastové slámky a míchadla v souvislosti s jejich zákazem: ať už obecně na úrovni měst či národů nebo konkrétně – v případě Velké Británie.

Jak vidíme, pouze na základě padesáti nejfrekventovanějších tokenů a třiceti nejfrekventovanějších bigramů jsme byli schopni rozkrýt tři různé typy reakcí – ať už na úrovni jedinců, komunit, či státních celků – na problém znečištění plasty. Ten dostal zároveň jasnější obrysy, když jsme zjistili, jakým způsobem ovlivňuje život

[128] Nevyhledávali jsme v tomto případě bigram, protože vstupem funkce pro vyhledávání zpráv jsou kompletní *tweety*, a ne tokenizovány text. Pravděpodobnost, že termíny „straws“ a „stirrers“ budou následovat hned po sobě a nebudou odděleny ani čárkou je velmi malá.

mořských živočichů. V další části naší analýzy se pokusíme nahlédnout vztahy jednotlivých bigramů ještě detailněji za užití síťového grafu.

9.8 Detekce tematických celků pomocí síťového grafu

V předchozí části analýzy jsme nejprve nahlédli pomocí Latentní Dirichletovy alokace distribuci dvou nejvýraznějších témat našeho korpusu – Dne Země & enzymu schopného pojídat plast – abychom si udělali základní představu o jeho obsahu. Následně jsme se z důvodů komputační a časové náročnosti algoritmu LDA rozhodli pro detekci dalších témat užít frekvenční tabulky tokenů a bigramů vyskytujících se v korpusu nejčastěji.

Vždy jsme přitom vycházeli z korpusu, z něhož jsme nejdříve odebrali stop slova zahrnující také „plastic“ a „pollution“. Věděli jsme totiž, že veškeré *tweety* nutně oba výše zmíněné termíny obsahují – bylo to kritériem jejich zařazení do korpusu. Při tvorbě frekvenčních tabulek tokenů a bigramů jsme proto vycházeli z předpokladu, že kdybychom je předem neodebrali, tak by: a) ve frekvenční tabulce tokenů zaujaly termíny „plastic“ a „pollution“ přední místa; b) vyskytovaly by se často napříč frekvenční tabulkou bigramů. V obou případech by pro nás jejich zahrnutí, a tudíž výskyt ve frekvenčních tabulkách, měly pouze malou informační hodnotu.

U síťového grafu je tomu ale naopak. Znázorněním bigramů jako sítě slov získáme přehled o tom, jak jsou spolu jednotlivá slova navzájem propojena; zdali některá z nich sehrávají roli prostředníků mezi slovy jinými či jejich shluky. Pro ilustraci nám postačí následující příklad. Uvažujme cestu[129] A → B → C → D, kde písmena A, B, C, D představují rozdílná slova. Můžeme se např. ptát: o kolik slov jsou od sebe vzdálena slova A a D? Pokud bychom např. B předem z cesty odstranili, dostali bychom zcela jiný výsledek. Z tohoto důvodu jsme se při tvorbě síťového grafu rozhodli vycházet z korpusu zahrnujícího rovněž termíny „plastic“ a „pollution“. Podobný postup pro vizualizaci bigramů nabízí také např. Silge a Robinson.[130]

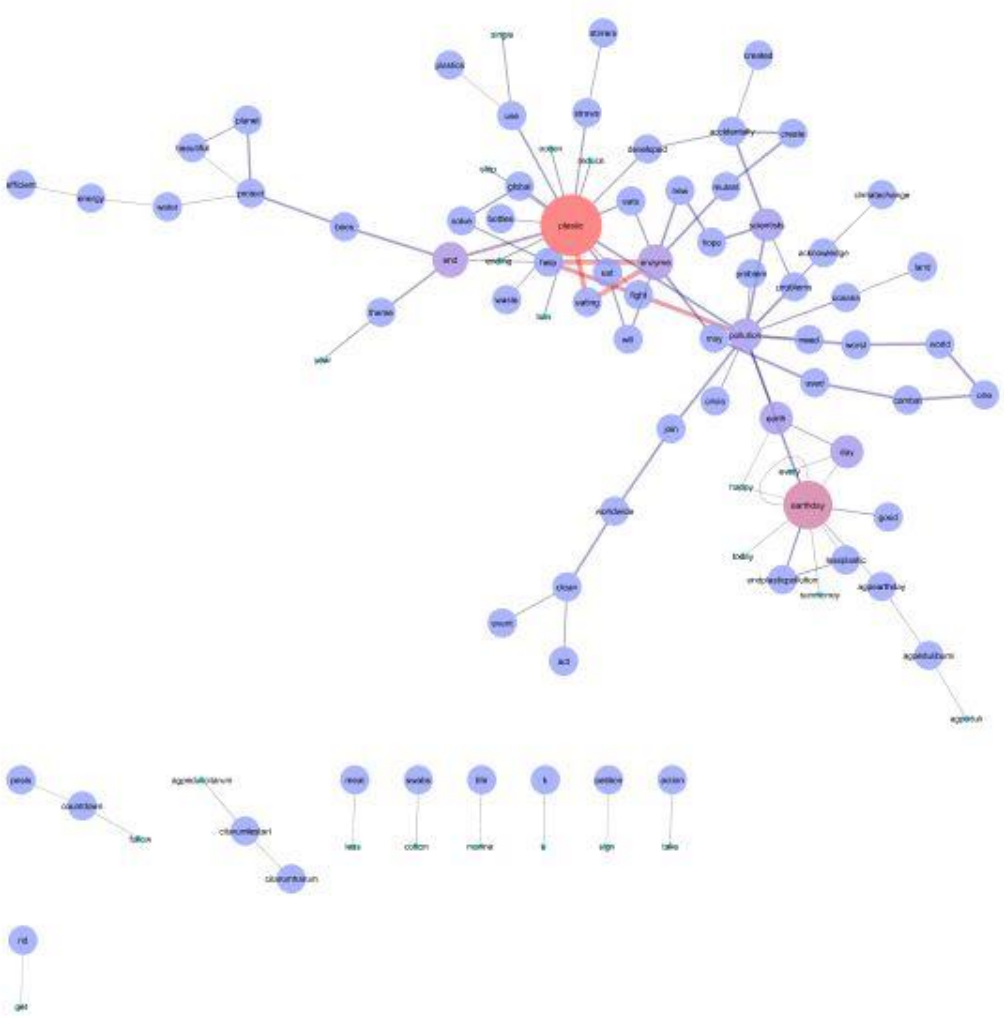
[129] Viz kapitola Teorie sítí: Sled, tah, cesta a cyklus.
[130] Silge, & Robinson, 2017, s. 54.

9.8.1 Program Cytoscape

Následující síťový graf byl zhotoven v grafickém programu Cytoscape určeném primárně k vizualizaci sítí molekulárních interakcí a dalších biologických dat. Vzhledem ke svým vlastnostem a přívětivému uživatelskému rozhraní je nicméně tento program vhodný také k vizualizaci dat textových. Pro dobrou čitelnost síťového grafu a orientaci v něm jsme se rozhodli znázornit pouze prvních sto nejfrekventovanějších bigramů a vztahy mezi nimi.

Rozložení síťového grafu jsme nechali vygenerovat programem Cytoscape na základě automatické analýzy sítě, tj. tabulky sto nejfrekventovanějších bigramů. Vrcholům jsme přiřadili barvu – od tmavší po světlejší – a velikost – od menší po větší – podle toho, jak vysoký je jejich vstupní stupeň. Hranám jsme přiřadili barvu – od tmavší po světlejší – a velikost – od tenčí po hrubší – na základě jejich mezilehlosti. Výsledný síťový graf můžeme vidět níže.

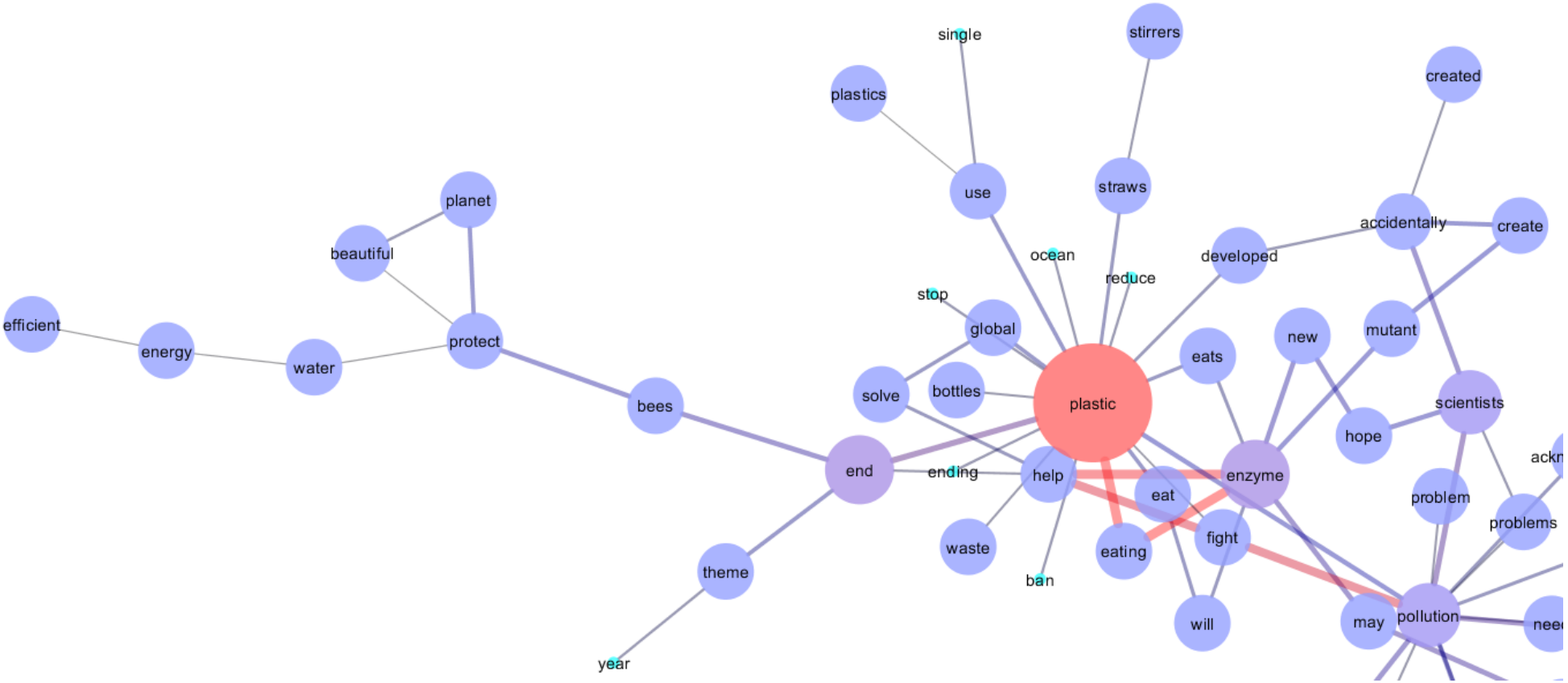
9.8.2 Detekce tematických celků pomocí síťového grafu



Graf č. 5: Síťový graf 100 nejfrekventovanějších bigramů

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.
Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší.

Jak vidíme, znázorněním sto nejfrekventovanějších bigramů nám vznikl orientovaný nesouvislý síťový graf, v němž můžeme rozlišit deset komponent. Největších z nich můžeme dále rozdělit na dvě skupiny vrcholů podle toho, ke kterému ze dvou největších a nejsvětlejších vrcholů se shlukují. Abychom grafu porozuměli, podíváme se nyní na jeho největší komponentu po částech blíže. Její levé části dominuje vrchol o termínu „plastic“, zatímco té pravé vrchol o termínu „earthday“. Nejprve se zaměříme na část levou.



Graf č. 6: První část hlavní komponenty síťového grafu

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.

Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší

Vrcholem s největším vstupním stupněm celého grafu č. 6 je „plastic“, následují jej: „end“, „enzyme“, „scientists“ a „pollution“. Tento fakt odpovídá rozložení hlavních témat v korpusu – bigram „plastic pollution“ se nutně vyskytuje ve všech *tweetech*. Větší vstupní stupeň centrality než „pollution“ má „plastic“, protože do něj směřuje více hran. Pojďme se nyní podívat, jaké termíny se na „plastic“ vážou.

Na první pohled lze spatřit cestu vycházející z vrcholu „efficient“ a směřující do „plastic“. Postupně jsme proto nahlédli do *tweetů*, které tento termín obsahují spolu s „energy“ a „water“. Celkem je jich v našem korpusu pět unikátních, za účelem jejich interpretace jsme z nich tři náhodně vybrali a zobrazili je v následující tabulce č. 15.

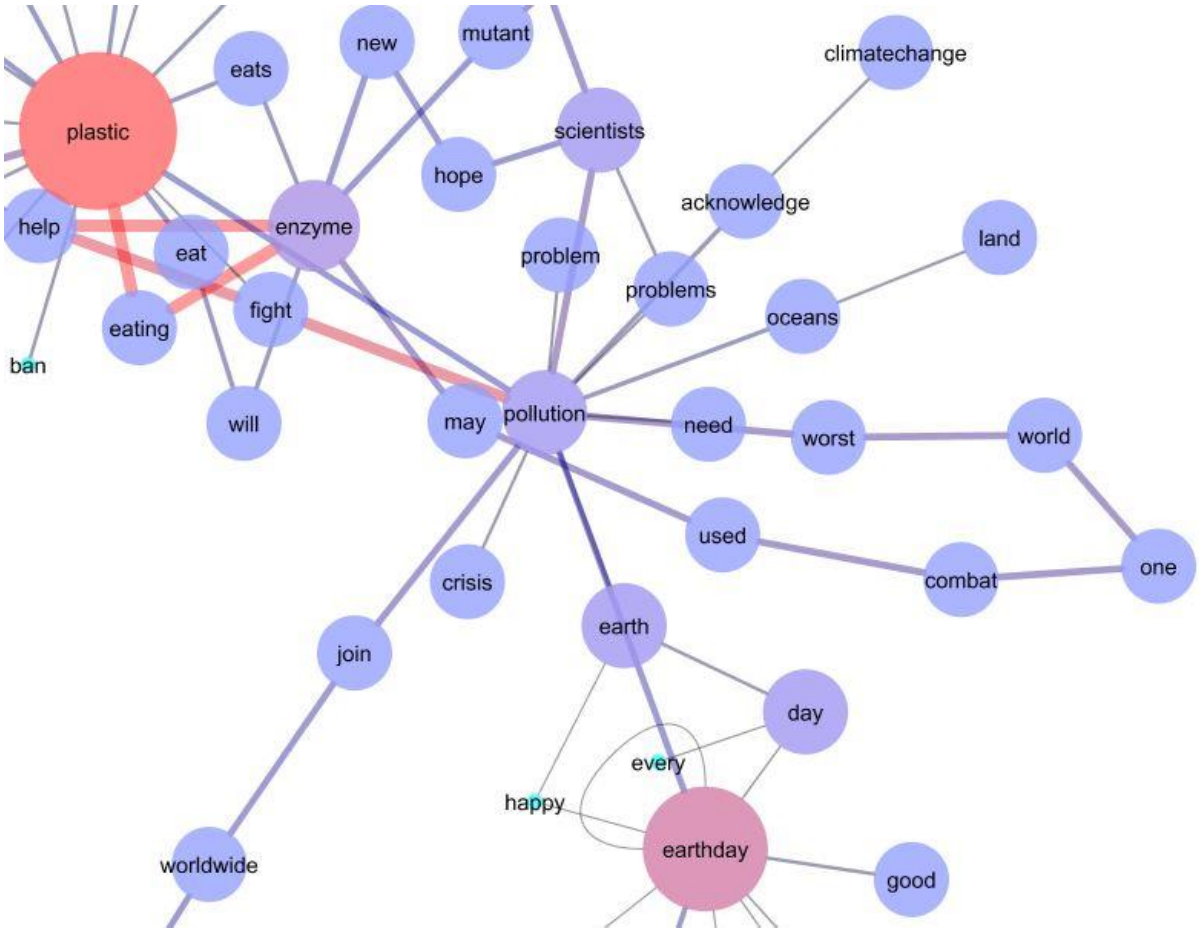
"Happy #EarthDay \n\nA few tips to protect our beautiful planet:\n\xfo\u009f\u008c\u008fProtect our water\n\xfo\u009f\u008c\u008fBe energy efficient\n\xfo\u009f\u008c\u008fEat less meat\n\xfo\u009f\u008c\u008fEat local, organic produce\n\xfo\u009f\u008c\u008fProtect the bees\n\xfo\u009f\u008c\u008fEnd plastic pollution\n\xfo\u009f\u008c\u008fAcknowledge #ClimateChange \n\xfo\u009f\u008c\u008fMost important, get rid of toxic @EPAScottPruitt https://t.co/DJI555s1Mv"
"Happy #EarthDay \n\nA few tips to protect our beautiful planet:\n\xfo\u009f\u008c\u008fProtect our water\n\xfo\u009f\u008c\u008fBe energy efficient\n\xfo\u009f\u008c\u008fEat less meat\n\xfo\u009f\u008c\u008fEat local, organic produce\n\xfo\u009f\u008c\u008fProtect the bees\n\xfo\u009f\u008c\u008fEnd plastic pollution\n\xfo\u009f\u008c\u008fAcknowledge #ClimateChange"

"Happy #EarthDay \n\nA few tips to protect our beautiful planet:\n\xf0\u009f\u008c\u008fProtect our water\n\xf0\u009f\u008c\u008fBe energy efficient\n\xf0\u009f\u008c\u008fEat less meat\n\xf0\u009f\u008c\u008fEat local, organic produce\n\xf0\u009f\u008c\u008fProtect the bees\n\xf0\u009f\u008c\u008fEnd plastic pollution https://t.co/EHeAjFsu2n"

Tabulka č. 15: *Tweety* obsahující termíny „efficient“, „energy“ a „water“

Všechny tři *tweety* z tabulky č. 15 jsou si velmi podobné a obsahují stejné poselství: vybízejí čtenáře, aby chránil vodní plochy a toky, využíval efektivně energii, jedl méně masa a jedl lokální a organické produkty, chránil včely a ukončil znečištění plasty. Interpretovat si to můžeme tak, že čím efektivnější bude výroba energie, tím čistší bude vodní ekosystém na naší planetě. Konkrétní nápady realizací těchto myšlenek ale v *tweetech* chybí. V dalším kroku jsme rozhodli přijít na kloub spojení termínů „protect“ a „bees“. Obsahují je pouze čtyři *tweety* téměř totožné s těmi z tabulky č. 15, a nepřinášejí žádné nové informace, proto je zde neuvádíme.

Významy termínů příslušících vrcholům kumulujících se okolo „plastic“ jsme nahlédli již v předchozích částech analýzy. Pojďme se proto nyní v našem síťovém grafu přesunout směrem doprava.



Graf č. 7: Druhá část hlavní komponenty síťového grafu

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.

Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší.

Na další části síťového grafu vidíme, že termíny „plastic“ a „earthday“ jsou spolu propojeny přes termín „pollution“, což nám opět potvrzuje, že tématem letošního Dne Země bylo znečištění plasty. Oproti přechozím grafům zde vidíme nové termíny „climatechange“ a „acknowledge“. Vrcholy, které je reprezentují jsou spojeny tahem s termínem „pollution“. Podívali jsme se tedy na *tweety*, v jakých se vyskytují. Následující tabulka č. 16 zobrazuje tři náhodně vybrané *tweety* obsahující termín „climatechange“, který se ve všech vyskytuje v podobě hashtagu „#climatechange“.

"#Iceland #Norway #Japan #Faroeislands #Denmark stop the #Whale hunts in this time of #Oceans #Marinelife at peril due to #pollution #climatechange #plastic #seismictests etc. etc. #Respect #PlanetEarth #BluePlanet #Greenpeace https://t.co/5iR30KFlh4"
"Solve it! 'Plastic pollution, unlike #climatechange, is an environmental cause without a vociferous band of sceptics denying the scientific consensus that action is needed urgently' @FT https://t.co/knOyovlpLl."
"What is plastic pollution, and why is it so harmful to the environment?\nhttps://t.co/nb9jeqOkzU #climatechange #climateaction \n#environment \n#energy https://t.co/eTVRy7bk1q"

Tabulka č. 16: *Tweety* obsahující termín „climatechange", resp. hashtag „#climatechange"

V prvním *tweetu* tabulky č. 16 se můžeme dočíst, že státy Island, Norsko, Japonsko a Dánsko spolu s Faerskými ostrovy zastavily lov velryb kvůli nebezpečí vzniklému údajně z důvodů znečištění plasty, klimatické změny apod. Pouze na základě našeho korpusu tyto informace nedokážeme ověřit, přesto je pro nás tato informace o reakcí rybářského průmyslu na problém znečištění plasty hodnotná. Druhý *tweet* z tabulky č. 16 zase tvrdí, že na rozdíl od změny klimatu, nemá problém znečištění plasty své kritiky. Konečně poslední z výše uvedené trojice *tweetů* si klade otázku: co je znečištění plasty, a proč je škodlivé pro životní prostředí? Odpověď je pravděpodobně skryta pod jedním z přiložených hypertextových odkazů. Stejně tak vztah změny klimatu a znečištění plasty pro nás zůstává po nahlédnutí našich tří náhodně vybraných *tweetů* neznámý.

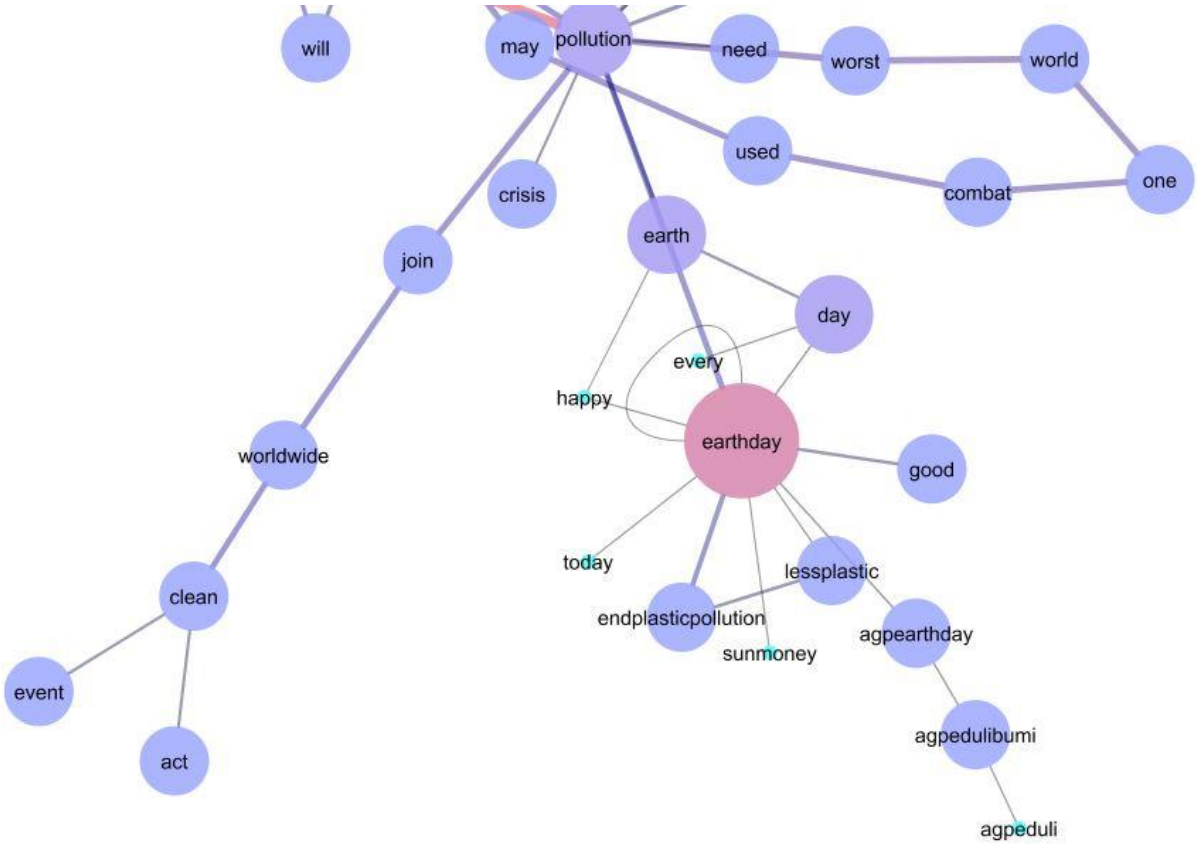
V následujícím kroku jsme se rozhodli pokračovat v tahu „climatechange“, „acknowledge“, „pollution“ a rozkrýt význam termínu „acknowledge“. Ten se vyskytuje v *tweetech* apelujících na čtenáře, aby seznámili veřejnost se znečištění plasty, jak níže ilustruje tabulka č. 17.

"Happy #EarthDay! This year, we acknowledge the growing problem of plastic pollution and the need to fundamentally change human attitude and behavior about plastics. Be the change you want to see in the world!v https://t.co/JDGPwrkroU"
"@Riverford @Greenpeace I hadn't, and now have. Good to acknowledge, but action is needed no? Will the beech nets replace the plastic bags? \nCarbon footprint is an important aspect of course, but must think about plastic pollution. \nWhat's holding you back?"
"It is so important for our world leaders to acknowledge the plastic pollution epidemic. One step closer to making sustainable living commonplace! https://t.co/SKoeUEvSav"

Tabulka č. 17: *Tweety* obsahující termíny „acknowledge“

První z *tweetů* z tabulky č. 17 zmiňuje Den Země a iniciativu obeznámení veřejnosti s problémem znečištění plasty, dále akcentuje potřebu změnit postoj obyvatel k plastům a jejich chování. Druhý jde dále za pouhou osvětu, a pokládá otázku, zdali síťky nahradí plastikové sáčky. A třetí *tweet* tabulky č. 17 tvrdí, že je důležité, aby o „epidemii znečištění plasty“ hovořili světoví lídři.

Tímto jsme rozkryli významy *tweetů* druhé části síťového grafu, pojďme se v něm proto nyní posunout dále směrem dolů a doprava.



Graf č. 8: Třetí část hlavní komponenty síťového grafu

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.
Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší.

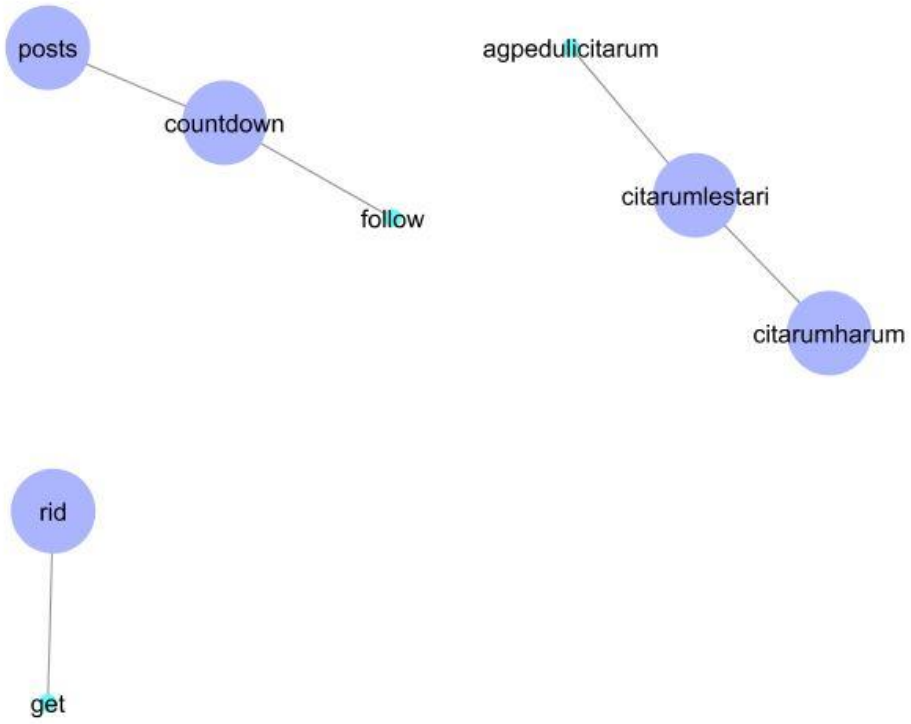
Třetí části síťového grafu dominuje vrchol náležející termínu „earthday“ a vidíme, že se k němu pojí hlavně termíny „endplasticpollution“, „lessplastic“, jejichž význam je nám již jasný. Společně pak odkazují k tomu, že tématem Dne Země 2018 je znečištění plasty. Termín „good“ se vyskytuje celkem ve 177 unikátních *tweetech*, tři z nich jsme náhodně vybrali a znázornili v tabulce č. 18. Můžeme z ní vyčíst, že každý z trojice *tweetů* hodnotí jako dobré či nedobré („not good“) zcela odlišné jevy. Pisatel prvního *tweetu* spatřuje v enzymu schopného pojídat plast potenciální hrozbu např. pro zateplení budov s tím, že by je enzym mohl poškodit. Druhý pisatel oproti tomu hodnotí pozitivně bambusové zubní kartáčky, kovové holící strojky či znovupoužitelné láhve. Konečně třetí *tweet* z tabulky č. 18 hodnotí negativně skutečnost, že na následky znečištění plasty zahynula velryba.

"This is NOT good news, should these bacteria escape \"into the wild\" & do damage to, for example insulation. Scientists Accidentally Create Super Enzyme That Can Devour Plastic Pollution https://t.co/zKGoUFQ3wI"
"Plastic toothbrushes, razors, and water bottles are the leading causes of plastic pollution! Bamboo toothbrushes, metal razors, and reusable water bottles and really good substitutes for these common items !! https://t.co/BPONt1j9CX*"*
"Plastic pollution kills whale. Not good at all... \n\n https://t.co/4SoXlI8i1U"

Tabulka č. 18: *Tweety* obsahující termín „good“

Jak vidíme, hodnotící výrazy užívají pisatelé *tweetů* obsažených v našem korpusu v rozličných kontextech. A kupříkladu přestože jsme se v předešlých vzorcích setkali pouze s pozitivními konotacemi enzymu schopného pojídat plast, první *tweet* z tabulky č. 18 nám ukázal, že se v korpusu vyskytují rovněž odlišné úhly pohledu na danou událost.

Nyní jsme rozkryli významy většiny hlavních vrcholů hlavní komponenty síťového grafu, proto se nyní přesuneme do jeho spodní části obsahující celkem devět samostatných komponent. Při jejich obsahové analýze budeme postupovat zleva doprava.



Graf č. 9: Tři malé komponenty síťového grafu

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.
Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší.

První z komponent obsahuje vrcholy s termíny „posts“, „countdown“ a „follow“. Abychom zjistil, k čemu se tato komponenta vztahuje, vyhledali jsme proto v korpusu všechny unikátní *tweety* obsahující zároveň termíny „posts“ a „countdown“. Jsou pouze dva a pro ilustraci je uvádíme v tabulce č. 19. Oba vybízejí čtenáře ke sledování odpočítávání času do začátku Dne Země 2018. První z nich krom toho vybízí čtenáře k připojení se k jisté úklidové události a podepsání petice. Druhý obdobně vybízí čtenáře k připojení se ke globální úklidové události a ke sledování příspěvků daného účtu za cílem osvěty o znečištění plasty, viz níže.

"Earth Day 2018 is just a week away! Follow our countdown posts on how you can help to end plastic pollution and join our clean up event: https://t.co/ChwthkaKZ6 \nSign the petition to end plastic pollution here: https://t.co/X7PwN8P9zu #sunmoney #EarthDay2018 https://t.co/PzASiKmHrz"
"Plastic Pollution Facts - Earth Day 2018\nFollow our countdown posts to learn more about what you can do to stop plastic pollution and join our worldwide clean up event: https://t.co/ChwthkaKZ6\nRead more about Earth Day here: https://t.co/XO0qIV9aSN #EarthDay2018 #sunmoney https://t.co/5cxzSjejPb"

Tabulka č. 19: *Tweety* obsahující termíny „posts" a „countdown"

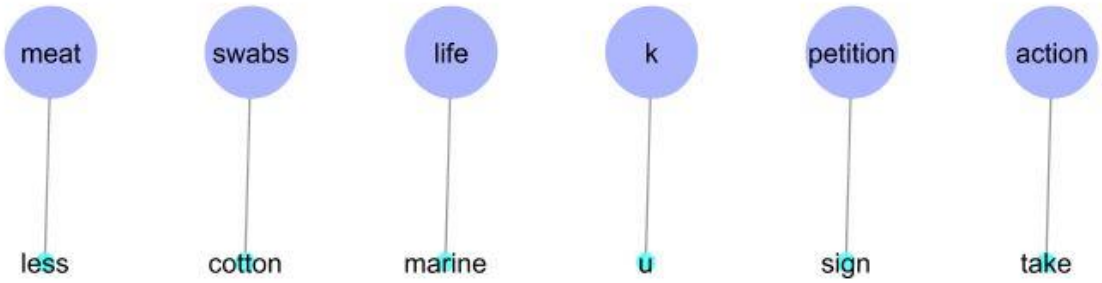
Druhou komponentu grafu č. 9 (vpravo nahoře) zde vzhledem k neznalosti termínů, které obsahuje, nejsme schopni vyjasnit. Komponenta třetí (vlevo dole) pak zastupuje *tweety* referující o potřebě zbavit se plastů, jak ilustrují tři náhodně z nich náhodně vybrané v tabulce č. 20. Přičemž druhý *tweet* této tabulky možnost zbavit se zcela plastů zpochybňuje a nabádá proto čtenáře k recyklování.

"I think this is more useful that turning out the lights for an hour. Let's get rid of #plasticwaste Plastic Pollution Calculator | #EarthDay Network https://t.co/uaLJ9Jof2l"
"This year's #EarthDay theme is: End Plastic Pollution.\nSince we cannot completely get rid of plastics in our homes, let us commit to reuse and recycle. \n\nFor our Mother Earth. https://t.co/bCyKRfgHwa"

"Great breakthrough! We still need to get rid of plastics. #ChadQuote… https://t.co/XC55uB1kDu"

Tabulka č. 20: *Tweety* obsahující termíny „get“ a „rid“

Nyní si projdeme popořadě jednotlivé komponenty grafu č. 10, přičemž budeme postupovat zleva doprava a vždy budeme ilustrovat jejich význam třemi náhodně vybranými *tweety*.



Graf č. 10: Šest malých komponent síťového grafu

Barvy vrcholů značí hodnotu jejich vstupního stupně – od světle přes tmavě modrou po červenou.
Barvy a šíře hran značí hodnotu jejich mezilehlosti – od užší a tmavší po širší a světlejší.

Termíny „meat“ a „less“ se vyskytují pouze ve třech unikátních *tweetech*, které uvádíme všechny v tabulce č. 21.

"Rich 'celebrities' trying to be all eco savy and celebrate earth day when they have 3 times the amount of cars necessary, wear fur, eat meat, contribute to pollution and plastic waste…etc etc etc https://t.co/4tu5Vv13Xc"
"Let's celebrate the #EarthDay together. This year's theme is 'End the plastic pollution', but we can do more. We can recycle more, plant a tree or eat less meat. We have to be the change. Let's take in our surroundings, appreciate them, and make an effort to be kind to our planet https://t.co/1C3cBStbcJ"
"@ImSophieDavies For goodness sake if you don't want to handle meat you have no business eating it. Man up and eat the corpse or go veggie and save the world form some plastic pollution and chickens from death. #SimpleDecision"

Tabulka č. 21: *Tweety* obsahující termíny „meat" a „less"

V tabulce č. 21 můžeme vidět, že první *tweet* se vyjadřuje ironicky o snaze bohatých celebrit slavit Den Země, protože mají údajně více aut, než potřebují, nosí kožešiny, jedí maso apod. Konzumace masa zde má tedy negativní konotaci. Druhý *tweet* zase vybízí čtenáře k tomu, aby více recyklovali, jedli méně masa a zasadili strom. Vše poté shrnuje poselstvím, že bychom si měli vážit svého okolí a starat se o ně. Konečně třetí *tweet* tabulky č. 21 obsahuje zmínku konkrétní osoby – Sophie Davis –, lze jej tedy považovat za veřejnou zprávu určenou právě jí. Pisatel v ní krom jiného vybízí, aby buď jedla „mrtvoly", nebo se stala vegetariánkou a přispěla tak k záchraně světa od znečištění plasty a zároveň aby ušetřila životy kuřat.

Termíny vrcholů druhé a třetí malé komponenty jsme rozkryli již dříve: první dvojice se týká vatových tyčinek do uší a druhá podmořským životem, který je mimo jiné

ohrožen právě těmito hygienickými pomůckami, které jsou vyhazovány do moří a oceánů.

Komponenta o vrcholech s termíny „petition“ a „sign“ reprezentuje *tweety*, které čtenáře vyzývají k podpisu petice pro ukončení problému znečištění plasty; apelují na jedince, organizace i vlády, viz tabulka č. 22 níže. Vzhledem k druhému a třetímu *tweetu* v této tabulce se můžeme domnívat, že petice je směřována k zákazu plastových sáčků a láhví.

"Join us in celebrating #EarthDay2018. This year's theme is ending #plasticpollution. You can sign the petition to help end plastic pollution in our oceans here: https://t.co/hXifIdxcFm https://t.co/j2cma5Z94f"
"We can all make a difference. Individuals, organisations, governments - #PlasticPollution can be stopped. Cut down on plastic bag and bottle use, buy milk in bottles from a milk delivery service, sign the petition! #EarthDay2018 #SimpleChanges\nhttps://t.co/ytcRQP060Z https://t.co/FAOLoUO9eu"
"PLEASE retweet our petition to reduce plastic pollution and #BanTheBottle - sign it, share it and help support our hope for a better future https://t.co/ByXnAvnTgS"

Tabulka č. 22: *Tweety* obsahující termíny „petition“ a „sign“

Dále *tweety*, které jsou zde reprezentovány komponentou o vrcholech s termíny „take“ a „action“ vybízí čtenáře, aby rovněž jednal za účelem vyřešení problému znečištění plasty, viz tabulka č. 23.

"Sea mammals and birds are choking on plastic. Australia needs to take action & help end plastic pollution @TurnbullMalcolm #EndPlasticPollution #Oceanplastic #breakfreefromplastic https://t.co/tacR1q2H1k"
"Plastic pollution is poisoning our oceans and land, injuring marine life, and affecting our health! \n\nThis #EarthDay, take action to end this pollution.\n\n#EarthDay2018 #Earth #Marine #Animals #Water #Plastic https://t.co/xhIlZy5lE3"
"Stand up and take action against plastic pollution! https://t.co/IFBe1kZJXc via @greenpeaceusa"

Tabulka č. 23: *Tweety* obsahující termíny „take“ a „action“

Nyní, když jsme na náhodně vybraných příkladech *tweetů* popsali, k jakým obsahům se jednotlivé vrcholy síťového grafu vztahují a vyjasnili jsme si jejich význam, podíváme se na ně v kontextu celého síťového grafu. Díky tomu budeme moci nahlédnout, jak daleko od sebe jsou, a jak důležité jsou některé hrany, jež je spojují, z hlediska mezilehlosti.

Můžeme si povšimnout, že vrcholy reprezentující termíny, které se týkají vědeckého objevu enzymu údajně schopného pojídat plast – „enzyme“, „eat“, „eating“, „scientists“ –, jsou seskupeny v těsné blízkosti vrcholu „plastic“, zatímco vrchol reprezentující „earthday“ je od „plastic“ vzdálen podstatně více. Krom toho každá cesta od vrcholu reprezentujícího „earthday“ směřující k „plastic“ musí projít vrcholem reprezentujícím „pollution“. O vrcholech vztahujících se k vědeckému objevu to neplatí. Můžeme tedy usuzovat, že vědecký objev je zde primárně vztahován přímo k tématu

plastů, a ke znečištění plasty až sekundárně. Oproti tomu Den Země je zmiňován primárně spolu s tématem znečištění plasty.

Podívejme se nyní na cestu: „pollution“, „fight“, „help“, „enzyme“, „eating“, „plastic“. Hrany spojující tyto vrcholy jsou velmi široké a světlé, tj. mají vysoký stupeň mezilehlosti. To znamená, že tato cesta sehrává ústřední roli při propojení většiny termínů vztahujících se ke znečištění s nejfrekventovanějšími termíny vztahujícími se k enzymu a plastům obecně. Jinými slovy: hovoří-li se ve zprávách v rámci Dne Země o znečištění oceánů, krizi a o nejhorších světových problémech, jejich nejčastěji zmiňovaným řešením je enzym schopný pojídat plast.

Dále si můžeme všimnout, že do vrcholu „plastic“ směřují cesty zahrnující plastové výrobky jako např. „straws“ a „bottles“ či „waste“. Lze si proto dovodit následující závěr: enzym schopný pojídat plasty by mohl být nejspíše aplikován na rozklad odpadu, který tvoří plastové slámky či lahve. Krom toho vidíme, že k vrcholu „plastic“ vedou cesty od vrcholů „stop“, „reduce“ a „ban“, což odkazuje k omezení až úplnému zákazů plastů. V předchozí kapitole jsme si to konkretizovali na záměru zakázat vatové tyčinky a plastová míchadla nápojů ve Velké Británii.

Nyní již víme, jaká nejvýraznější témata se v rámci našeho korpusu vážnou k zastřešujícímu tématu znečištění plasty. Jsou jimi Den Země a vědecký objev enzymu údajně schopný pojídat plast. Víme také, jaké jsou nejfrekventovanější tokeny a bigramy, a které z nich se vážou spíše k prvnímu, či spíše k druhému z hlavních témat. Vyhledáním skupin zpráv obsahujících ať už některé tokeny nebo bigramy jsme dále byli schopni nahlédnout některá menší témata a udělat si konkrétní představu o mimojazykových jevech, k nimž referují.

Pro komplexnější náhled na sto nejfrekventovanějších bigramů korpusu a vazby mezi nimi jsme použili síťový graf. Tento graf nám pomohl odhalit, že termíny odkazující k plastovým výrobkům či vědeckému objevu se v rámci korpusu vážou přímo na termín „plastic“. Oproti tomu ty z okolí termínu „earthday“ se váží k plastům zprostředkovaně přes termín „pollution“. V posledním kroku naší analýzy se podíváme na termíny s nejvíce pozitivním a negativním sentimentem a na postoje, které jsou jimi vyjadřovány.

9.9 Detekce sentimentu a postojů

Nalézt v korpusu postoje pisatelů zpráv k entitám je nelehký úkol, i když víme, na co přesně se máme zaměřit. Přestože jsme už v rámci naší analýzy rozkryli jak hlavní témata, tak některá menší témata vztahující se ke znečištění plasty, rozhodli jsme se v posledním kroku analýzy při dolování postojů vycházet od sentimentu.

Pro tento účel jsme použili knihovnu syuzhet[131] od Matthewa Jockerse. Knihovna umožňuje krom jiného měření sentimentu tokenů, a to pomocí pěti rozdílných ručně anotovaných lexikonů, ve kterých bylo každému slovu přiřazeno jeho sentimentální skóre. Pro naše potřeby jsme zvolili lexikon AFINN[132] od Finna Årupa Nielsena. Tento lexikon zahrnuje celkem 2477 anglických slov a frází, kterým je v závislosti na intenzitě jejich sentimentu připsáno číslo od -5 do 5, tj.: -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5. Nielsen vytvořil AFINN speciálně pro potřeby analýzy sentimentu textů pocházejících z internetu, primárně z mikroblogovacích sítí jako je např. Twitter. Vzhledem k tomuto účelu do něj zahrnul jak internetový slang, tak některá „obscénní" slova, která mohou internetové texty obsahovat.[133] Právě z těchto důvodů jsme zvolili lexikon AFFIN pro naši analýzu.

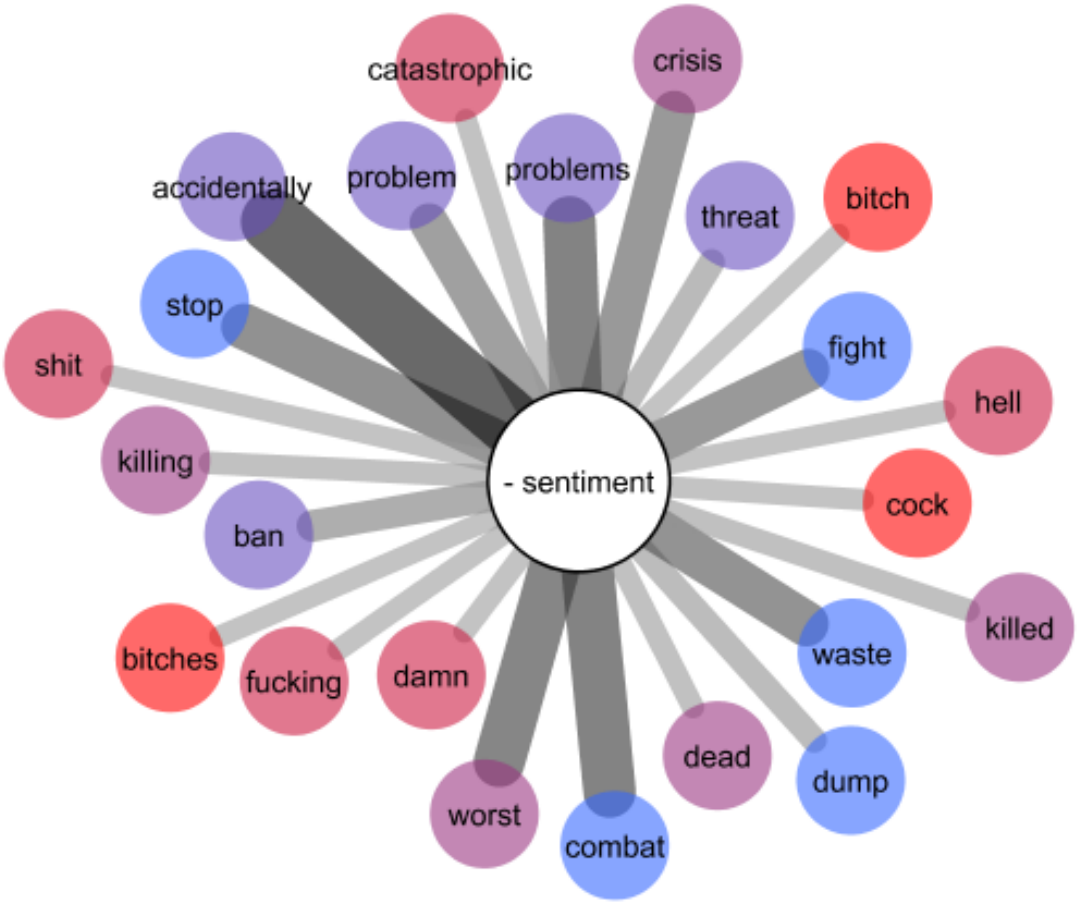
Pomocí funkce get_sentiment[134] jsme získali skóre sentimentu pro každý token z tokenizovného korpusu *tweetů*, který neobsahoval stop slova, nicméně zahrnoval „plastic" a „pollution". Dále jsme jednotlivé tokeny spárovali s jejich skóre a vydělili z výsledné tabulky menší celky obsahující vždy pouze tokeny se stejným skóre, tj. např. 1. Následně jsme z takovéto tabulky vytvořili tabulku frekvenční a vydělili z ní pět nejfrekventovanějších termínů. Tento postup jsme opakovali pro všechny skupiny termínů rozdělené podle intenzity jejich sentimentu vyjma sentimentu neutrálního, tj. majícím skóre 0.

Poté jsme spojili výsledné tabulky do dvou skupin podle orientace jejich sentimentu: na pozitivní, tj. od 1 do 5, a na negativní, tj. od -1 do -5. Tímto nám vznikly

[131] Jockers ML (2015). _Syuzhet: Extract Sentiment and Plot Arcs from Text_. Dostupné z https://github.com/mjockers/syuzhet, 26. 04. 2018.
[132] www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010, cit. 2. 05. 2018.
[133] Nielsen, F. A., 2011, cit. 02. 05. 2018.
[134] Jockers ML (2015). _Syuzhet: Extract Sentiment and Plot Arcs from Text_. Dostupné z https://github.com/mjockers/syuzhet, cit. 26. 04. 2018.

dvě tabulky. Každá z nich obsahuje 23 termínů, neboť jak termíny se sentimentem -5, tak 5 se v korpusu vyskytují pouze tři. Tyto tabulky, které jsme poté znázornili formou grafu v programu Cytoscape, si nyní ukážeme a podíváme se užitím funkce findTweetsContaingSpecificTerm, k jakým entitám se jednotlivé termíny vztahují. Začneme sentimentem s negativní orientací.

9.9.1 Termíny s negativním sentimentem



Graf č. 11: 23 nejfrekventovanějších termínů s negativním sentimentem

Barvy vrcholů značí intenzitu jejich sentimentu – od modré, přes fialovou po červenou.

Odstín a tloušťka hran značí frekvenci vrcholů, z nichž vychází – od světle tenkých po tmavě hrubé.

V grafu č. 11 vidíme, že nejvíce frekventované termíny s negativním sentimentem mají jeho nižší intenzitu: „stop“, „accidentally“, „problems“, „fight“, „waste“, „combat“ a „worst“. Se všemi z nich jsme se setkali již dříve a víme, že se vztahují přímo ke znečištění plasty jako jednomu z nejhorších světových problémů, který musíme přemoci. Termín „stop“ se vyskytuje např. v některý *tweetech* tabulek č. 16, 19. První z nich informuje o státech, které zastavily lov velryb z důvodu znečištění plasty; druhý vybízí čtenáře k poučení se o tom, čím mohou přispět k zastavení tohoto znečišťování. Pro úplnost jsme tyto dva *tweety* doplnili třetím náhodně vybraným, který obsahuje

termín „stop“ a zobrazili jej tabulkou č. 24. Vybízí čtenáře, aby přestali vyrábět a prodávat plasty na jedno použití, viz níže.

"Just deal with the problem at source - stop making and selling single use plastics https://t.co/Krrf5YUpAy"

Tabulka č. 24: *Tweet* obsahující termín „stop“

Termíny „problem“, „problems“, „combat“ a „worst“ odkazují ke *tweetům* referujícím o vědeckém objevu enzymu schopného údajně pojídat plast, který by mohl být použit pro boj s jedním z nejhorších světových problémů – znečištění plasty – viz tabulka č. 8. Termín „fight“ užívají místo „combat“ pisatelé v obdobně laděných *tweetech* – viz např. tabulky č. 2, 3 a 7. Můžeme tedy tvrdit, že přestože jsme výše zmíněným termínům přiřadili na základě lexikonu AFFIN negativní hodnoty sentimentu, vyskytují se ve *tweetech*, které jsou laděny spíše pozitivně, neboť předestírají řešení problému znečištění plasty. Dále termín „waste“ zahrnuje *tweet* informující o iniciativě v Den Země usilující rozšířit povědomí o plastovém odpadu – viz tabulka č. 4; *tweet* referující o velrybě zahynulé na následky pozření plastového smetí – viz tabulka č. 11; a *tweet* osočující celebrity z nadměrné produkce plastového odpadu – viz tabulka č. 21. Ve všech výše zmíněných *tweetech* „waste“ negativní konotace.

Termín „accidentally“ měl pro nás dříve pozitivní konotaci, neboť se vyskytuje převážně ve *tweetech* referujících o vědeckém objevu – viz tabulky č. 2, 8 a 18 –, který byl učiněn náhodou. Můžeme tedy tvrdit, že v závislosti na kontextu má tento termín v rámci našeho korpusu sentiment spíše pozitivní.

Termín „ban“ zahrnují např. některé *tweety* z tabulek č. 13 a 14, které pojednávají o zákazu plastových slámek, míchadel a vatových tyčinek. Některé z nich přikládají informaci o plánování podobného kroku ve Velké Británii. Stejně jako v předchozím případě můžeme říci, že v závislosti na kontextu, kterým je pozitivní snaha omezit

znečištění plasty mají *tweety* obsahující „ban“ v našem korpusu sentiment spíše pozitivní.

Termín „dump“ obsahuje 188 unikách *tweetů*, přičemž na vzorku tří náhodně vybraných z nich vidíme, že se skládka vyskytuje buď v kontextu informace o objevení brouků pojídajících plast na explicitně blíže neurčené japonské skládce, či jako označení Velké tichomořské odpadkové skvrny. V prvním případě má vzhledem ke kontextu skládka konotaci pozitivní – je možným příslibem řešení problému, v druhém pak negativní, protože označuje obrovské množství plastů plovoucích v Tichém oceánu, viz tabulka č. 25.

"Scientists accidentally create mutant enzyme that eats plastic bottles\n\nThe breakthrough, spurred by the discovery of plastic-eating bugs at a Japanese dump, could help solve the global plastic pollution crisis\n\nhttps://t.co/oiv1CiLA8D"
"Scientists accidentally create mutant enzyme that eats plastic bottles | The breakthrough, spurred by the discovery of plastic-eating bugs at a Japanese dump, could help solve the global plastic pollution crisis • r/Kossacks_for_Sanders https://t.co/mkMYi2DoJI"
"\xf0\u009f\u0093\u0091 Recap | The Pacific plastic dump now bigger than #France https://t.co/dzkrYJHFRH #afp #pollution"

Tabulka č. 25: *Tweety* obsahující termín „dump“

Termín „threat“ je obsažen celkem v 73 unikátních *tweetech*, z nichž jsme náhodně vybrali tři a zahrnuli je do tabulky č. 26. Vidíme v ní, že jako hrozba je ve všech

případech považováno znečištění plasty: ať už jako hrozba pro život na zemi i ve vodě, jako další největší hrozba pro zemi, či jako hrozba pro želvy, které plasty pojídají.

"#Insightwithaci Plastic pollution has assumed holocaust proportion and has become a threat to life both on land and water"
"Plastic pollution, next greatest threat to earth. #Plastic \nYou touch any item, traces of plastic or plastic parts in it. https://t.co/5kZwBr6OVD"
"#Plastic is such a huge threat to turtles in the wild - more than 50 per cent of turtles globally have eaten plastic debris or other rubbish. \xf0\u009f\u0090#PassOnPlastic https://t.co/gzdHhQ1zJL"

Tabulka č. 26: *Tweety* obsahující termín „threat“

Zaměříme-li se na termíny s nejvíce negativním sentimentem, všimneme si, že se jedná výhradně o vulgarismy: „bitches“, „bitch“ a „cock“. Abychom zjistili cíl jejich hodnocení, podívali jsme se na *tweety*, jež tyto termíny obsahují. Termín „bitches“[135] se vyskytuje pouze v jednom a termín „bitch“ ve dvou. Všechny tři *tweety* jsme pro ilustraci zobrazili tabulkou následující tabulkou č. 27.

"EVOLUTION, BITCHES.\nPREDICTIVE POWER FTW\nhttps://t.co/xVb7D2LQNg https://t.co/xVb7D2LQNg"

[135] Pro vyhledání *tweetu* obsahujícího termín „bitches“ je potřeba do příkazu funkce FindTweetsContainingSpecificTerm zadat „BITCHES“, neboť je funkce citlivá na malá a velká písmena

"#EARTHDAY... It gave birth to us served everything for us untill our death...but we people plundered its chastity raped it several time with our fuckin pollution and bitch plastics...Stop using plastic as much as you can https://t.co/OJfKj1V3RM"
"Make a comment about straws in a work meeting because I post facebook videos about straws/plastic/pollution...excuse me bitch, sorry not sorry I care about the planet and you're a selfish bitch. \xf0\u009f\u0092\u0081\xf0\u009f\u0098\u0087 #gtfo"

Tabulka č. 27: *Tweety* obsahující termíny „bitches" a „bitch"

V tabulce č. 27 vidíme, že vulgarismus „bitches" použil pisatel coby oslovení – z kontextu ale není jasné, zda čtenářů, nebo někoho jiného. Vulgarismu „bitch" pak použil jeden pisatel jako negativní hodnocení plastů obecně, druhý jako oslovení blíže neurčené osoby, pravděpodobně však oslovení čtenářů daného *tweetu*.

Konečně termín „cock" je v korpusu zastoupen ve formě „cock-up" ve dvou identických *tweetech*. Pisatelé jím hodnotí stav přírodního prostředí, který mají údajně na svědomí lidé, jako průser, viz tabulka č. 28.

"I knew it was bad but when you see photos like this one, you really begin to see the hard evidence of what a monumental cock-up we are making of the natural world! Plastic pollution should be an easy thing for our species to put right, shouldn't it? https://t.co/Syrk6lwmu9"

Tabulka č. 28: *Tweet* obsahující termín „cock-up"

Termín „hell“ se vyskytuje celkem v pěti unikátních *tweetech*, z nichž jsme náhodně vybrali následující tři, viz tabulka č. 29.

"Frame the plastic bottle or else it will smack your face. Every time you miss, the rate of #pollution increases. Say yes to #PlasticBan \nThis #EarthDay18, Let us all and sundry pledge that the earth is ours and we mustn't make it hell. Let's vow to be true not plastic. https://t.co/ajTx2wgO2M"
"Oh hell .. Indonesia and plastic pollution @abcnews now- in parts they have had to call in the army to clear rivers.. villages surrounded by plastic waste- this is what can happen if gov do not act in time. Get it done - for the country & the world #auspol"
"What the hell: Trump Eliminates Plastic Water Bottle Ban in National Parks, Removes White House Bikeshare Station https://t.co/rCORIKV3rU #Millennials #UniteBlue #Veterans #VetsResistSquadron #environment #pollution"

Tabulka č. 29: *Tweety* obsahující termín „hell“

V prvním *tweetu* z tabulky č. 29 vybízí pisatel čtenáře, aby nedopustili, aby se ze Země stalo peklo – hodnotí negativně možný budoucí stav planety Země. V druhém jím pisatel hodnotí situaci v Indonésii, jež byla údajně znečištění plasty natolik závažná, že vláda povolala armádu, aby vyčistila tamní řeky od plastu. Jako zdroj této informace jsou ve zprávě formou zmínky uvedeny ABC News. Třetí pisatel zase hodnotí jako peklo jednání Donada Trumpa, který údajně zrušil zákaz plastových lahví v amerických národních parcích.

Dále termín „shit“ figuruje celkem v devíti unikátních *tweetech*, z nichž jsme náhodě vybrali tři a vyobrazili je v rámci následujíc tabulky č. 30. Můžeme v ní vidět, že se výše zmíněný termín vztahuje převážně k lidské lhostejnosti.

"@johnpodesta Every day you dump your ritually sacrificed children on your pig farm. The plastic pollution is not the real problem. Corrupt politicians, pedofile rings, agenda 21 etc.. don't act like you give a shit about the environment. You only care about your adrenachrome fix"
"@LeMattSaphirium @OrwellNGoode holy shit. “reducing pollution is anti-murican it’s muh god given right to have plastic straws instead of paper”\nfuck off lmao"
"@richiegill70 @ErikSolheim @BBCEarth I know that is what you WANT to be true, so you can blame the western world and force us to pay. But still your narrative IS a lie: Plastic waste is a global problem caused by African/Asian cultures that don‘t give a shit about plastic pollution. Perod. \nhttps://t.co/TBxpMRviGk"

Tabulka č. 30: *Tweety* obsahující termín „shit“

První *tweet* tabulky č. 30 obsahuje zmínku uživatele johnpodesta, amerického politického konzultanta, který byl mezi lety 2014–2015 poradcem amerického prezidenta. Pisatel v něm mimo jiné obviňuje Podestu z rituálního obětování dětí a z užívání omamné látky adrenachrom. Do kontrastu pak staví konzultantův údajný nezájem o životní prostředí. Druhý *tweet* je uvozen frází „holy shit“ vyjadřující údiv a dále obsahuje rovnou dvě zmínky uživatelů a nám neznámou citaci výroku o tom, že omezování znečištění je anti-americké a právo užívat plastové slámky je nám dáno od boha. Lze si odvodit, že pisatel zde zaujímá vůči citovanému textu spíše negativní

než pozitivní postoj a je v údivu. Konečně třetí *tweet* tabulky č. 30 vyjadřuje termínem „shit“ lhostejnost Afrických a Asijských kultur, které dle jeho autora zaujímají k problému znečištění plasty a jsou proto jeho původci. Jak vidíme, náhodně vybraný vzorek *tweetů* s termínem „shit“ má výrazné politické a kulturní negativní konotace.

Termínem „catastrophic“ vyskytujícím se celkem ve čtyřech unikátních *tweetech*, z nichž náhodně vybrali tři, zase pisatelé hodnotí proporce blížící se situace znečištění plasty ve světových oceánech, případně dopad plastového smetí na oceány, viz tabulka č. 31.

"8 million metric tons of plastic escapes into the world's oceans each year. The situation is approaching catastrophic proportions. Read how governments, companies, and other organizations are hoping to stop the flow of plastic pollution:\nhttps://t.co/ZCU9DByqYz https://t.co/okdWGjavl5"
"Hope important and unique work by @heidiacampora on catastrophic impact of plastic pollution in our oceans continues to be supported by @MfrcGmit https://t.co/rnLnSstOWT"
"8 million metric tons of plastic escapes into the world's oceans each year. The situation is approaching catastrophic proportions. Read how governments, companies, and other organizations are hoping to stop the flow of plastic pollution:\nhttps://t.co/1L8lMouFfE https://t.co/ajXZSz8OJu"

Tabulka č. 31: *Tweety* obsahující termín „catastrophic“

Dále termín „fucking“ je užit pouze ve třech unikátních *tweetech*: u prvního z nich nejsme schopni rozklíčovat jeho význam, v druhé jej pisatel užívá jako intenzifikační částici v souvislosti s potřebou zmírnit znečištění plasty v oceánech. Třetí pisatel zase používá tento termín k vyjádření negativního názoru na množství pozlátka ve světě, viz tabulka č. 32.

"Ffs\nA) get a fucking life\nB) there's this thing called soap\nC) yes let's deal with the amount of plastic pollution in the world by USING MORE PLASTIC https://t.co/w5Zr7PMGMB"
"We ain't even clean out all the plastic can the ocean get a fucking break from y'alls pollution https://t.co/oEwOoz7OY1"
"THERE'S TOO MUCH FUCKING GLITTER IN THE WORLD!!! \xf0\u009f\u008c\u008d | MNN - Mother Nature Network https://t.co/cDzwrySZqi"

Tabulka č. 32: *Tweety* obsahující termín „fucking“

Termín „damn“ se byl použit celkem ve čtyřech unikátních *tweetech*, které ilustrujeme následují trojicí z nich náhodně vybraných a zanesených do tabulky č. 33.

"Breathing, eating and destroying the planet. Thanks #Plastic! and the Chemical Industries who don't give a damn. Stop using #plastics as much as u can - for your health and ur kids! https://t.co/rxrg5yerSC"
"I totally forgot that I have to interview someone to get another stakeholder perspective on plastic pollution… damn. It never stops."
"There's hope now…because without plastic bags, shopping in Austin is damn inconvenient and worse, autobot pollution machine Zane Barfwich will single-handedly cover the entire West Coast with his non-biodegradable litter as he plots and schemes how… https://t.co/nhCwUNyYps"

Tabulka č. 33: *Tweety* obsahující termín „damn“

Z prvního *tweetu* tabulky č. 33 lze vyčíst, že jeho pisatel zastává negativní postoj vůči chemickému průmyslu, který dle jeho slov neřeší dopady svého působení na planetu. Druhý *tweet* je příliš abstraktní na to, abychom z něj dokázali vyčíst něco konkrétního. Třetí pak působí ironicky, nicméně vzhledem k neznalosti kontextu osob a jevů, o nichž referuje, z něj bohužel rovněž nedokážeme odvodit více.

Jako na poslední termíny s negativním sentimentem nyní nahlédneme na: „killing“, „killed“ a „dead“. Nejprve se zaměříme na vzorek tří náhodně vybraných *tweetů* obsahujících „killing“, viz tabulka č. 34.

"Plastic pollution is killing our wildlife. The Nuneaton and Warwickshire Wildlife Sanctuary @WarksWildlife save hundreds of animals a year who are caught up in plastic. For more on our #PlasticWaste series click here: https://t.co/wHMBx9RWlx https://t.co/QZqrCHLBeq"
"Plastic pollution is slowly killing our planet. It's poisoning marine life and destroying landscapes. Together we can make a difference. Help spread awareness on the responsible use and disposal of plastic. Save Mother Earth. #EarthDay https://t.co/ZexllOb3iL"
"Plastic pollution is killing our marine wildlife. Here are a few ways you can help https://t.co/XVTilKGQ1n https://t.co/WWuvoH7OOu"

Tabulka č. 34: *Tweety* obsahující termín „killing"

V tabulce č. 34 vidíme, že termín „killing" se vztahuje a) k divoké přírodě, b) k planetě Zemi, c) k mořským zvířatům. Ve všech *tweetech* má přitom negativní konotaci, neboť jej pisatelé užívají k vyjádření skutečnosti, že znečištění plasty zabíjí jak divoce žijící živočichy na zemi a ve vodě, tak Zemi obecně.

Termín „killed" jsme mohli zahlédnout již dříve, a sice v tabulkách č. 10 a 11. V obou případech jej obsahují *tweety* – celkem tři – referující o mořských živočiších a ptácích zahynulých na následky znečištění plasty.

Na závěr této kapitoly rozkryjeme význam *tweetů* s termínem „dead", kterých je v našem korpusu celkem 47 unikátních, tři z nich náhodně vybrané ukazuje tabulka č. 35. Všechny hovoří o velrybách, jež zahynuly kvůli znečištění plasty.

"I added a video to a @YouTube playlist https://t.co/dmGogtyU8P Plastic pollution killed sperm whale found dead on Spanish beach"
"A sperm whale found dead in #Spain had 30 kilos of #plastic\nin its stomach https://t.co/yAh6UZzh7X via @zmescience #pollution #saynotoplastic #environment"
"Whales are Dying from Plastic Pollution – Urge your State to End Single-Use Garbage, Now! The Great Pacific Garbage Patch is now twice as large as Texas. A juvenile sperm whale was just found dead with 64 pounds of plastic trash in its stomach. #care2 https://t.co/h4dcsdMjlC"

Tabulka č. 35: *Tweety* obsahující termín „dead“

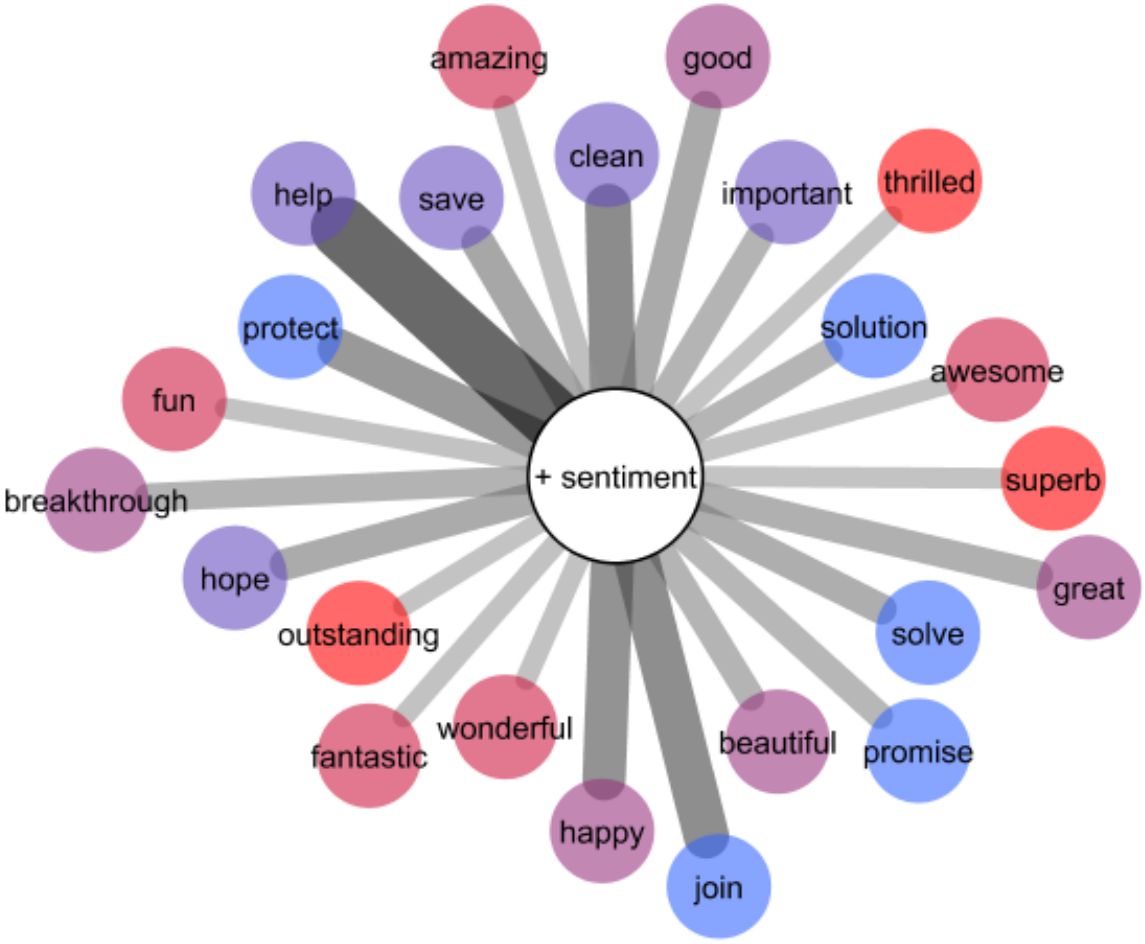
Tímto jsme rozklíčovali k čemu se vztahují *tweety* obsahující nejfrekventovanější termíny s negativním sentimentem. Část z těchto *tweetů* pojednává vědeckém objevu enzymu schopného pojídat plast, a tudíž termíny, jež se v nich vyskytují a jsou anotované dle lexikonu AFFIN jako negativní, mají vzhledem ke kontextu konotace spíše pozitivní. Obdobně můžeme nahlížet na *tweety* referující o objevu brouků pojídajících plast.

Další část termínů z grafu č. 11 figurovala v *tweetech*, které hodnotily negativně problém znečištění plasty a jeho fatální důsledky na mořskou, ale též zemskou faunu, životní prostředí či dokonce celou planetu Zemi. Konkrétně byla pisateli jako negativní hodnocena např. situace v Indonésii. Oproti tomu jiné *tweety* obsahující negativní termíny hodnotily pozitivně iniciativy bojující proti znečištění plasty, případně k těmto iniciativám čtenáře vybízely.

Část *tweetů* s negativními termíny se vyjadřovala negativně k některým americkým politikům (Donald Trump, John Podesta) a jejich postojům či lhostejnosti vůči problému znečištění plasty. Dále vulgarismy figurují v našem korpusu pouze v malém množství a pisatelé je užívají buď jako oslovení, intenzifikační částice, nebo k vyjádření silně negativního názoru.

Na závěr naší analýzy obrátíme pozornost na světlejší stránku námi zkoumaného problému: podíváme se na nejfrekventovanější termíny s pozitivním sentimentem a na *tweety*, v nichž figurují.

9.9.2 Termíny s pozitivním sentimentem



Graf č. 12: 23 nejfrekventovanějších termínů s pozitivním sentimentem

Barvy vrcholů značí intenzitu jejich sentimentu – od modré, přes fialovou po červenou.

Odstín a tloušťka hran značí frekvenci vrcholů, z nichž vychází – od světle tenkých po tmavě hrubé.

Nejprve se zaměříme na termíny s nejméně pozitivním sentimentem: „protect“, „solution“, „solve“, „promise“, „join“. První z nich figuruje ve *tweetech*, s nimiž jsme se již setkali v tabulce č. 16, a které vybízejí čtenáře k ochraně planety a životního prostředí. Řešení se pak ve *tweetech* obsahujících „solution“ a „solve“, jako první nahlédneme pro lepší představu tři z nich náhodně vybrané, jež zahrnují „solution“, viz tabulka č. 36.

"To stop plastic pollution, let's all adopt a small solution to help solve a big problem. #plasticfree #CHOGM2018 https://t.co/9L2rNb5lbn"
"#amazing solution to #pollution! Scientists accidentally create mutant enzyme that eats plastic bottles https://t.co/pKGIvXubMV"
"#Bioplastics, which are made of things like algae and shrimp shells, could be a solution to the plastic pollution problem, say @tonyrwalker1 & @scharleb of @Dalnews\nhttps://t.co/wuc7davx78"

Tabulka č. 36: *Tweety* obsahující termín „solution“

V tabulce č. 36 vidíme, že řešením je zde ve všech případech míněno rozdílné řešení problému znečištění plasty: a) na úrovni jednotlivců – hashtag prvního *tweetu* odkazuje k úplnému zřeknutí se plastů, b) pomocí enzymu schopného pojídat plast, c) užíváním bioplastů. Všechny *tweety* z tabulky č. 36 jsou laděny pozitivně, stejně jako ty, jež obsahují „solve“. Setkali jsme se s nimi již dříve v tabulkách č. 16, 25 a 36. Tyto *tweety* buďto vybízejí čtenáře k řešení znečištění plasty, nebo informují o objevu enzymu schopného pojídat plast. Za naději v boji se znečištění plasty pak na základě námi vybraného náhodného vzorku, můžeme soudit, považují výhradně již tolikrát zmiňovaný enzym schopný pojídat plast, viz tabulka č. 37.

"Pac Man To The Rescue? Plastic-eating enzyme holds promise in fighting pollution, scientists say https://t.co/cjSXLQHs1Z via @Reuters \n\n@mrjamiereeves @HoneymoonGondol @MsRonnyB @EHFoundation247 @BeingFarhad @KAdnoub @Planetary_Sec @stewartboyle4 @arikring @HansLak @Sustainia https://t.co/CMuhkboaFk"
"Plastic-eating enzyme could hold promise in the fight against pollution via @BiFab"
"Plastic-eating enzyme holds promise in fighting pollution - scientists https://t.co/p7WgYt2LsW"

Tabulka č. 37: *Tweety* obsahující termín „promise“

Tweety obsahující termín „join“ jsme viděli již v tabulkách č. 19 a 22, přičemž vybízely čtenáře buď ke spoluúčasti na hromadném úklidu, nebo k připojení se k oslavám Dne Země 2018. Naděje („hope“) figuruje ve *tweetech* tabulek č. 2, 3, 22, 31 a 33. Některé z těchto *tweetů* zahrnují informaci o naději na ukončení znečištění plasty, již vědci vkládají do enzymu schopného pojídat plast. V jiných zas pisatelé doufají v lepší budoucnost, nebo doufají v důležitost a jedinečnost explicitně blíže neurčené práce při řešení důsledků znečištění plasty v oceánech. Jeden *tweet* obsahující termín „hope“ je však také laděn ironicky, když referuje o životě plastikových sáčků v Austinu a jisté osobě upřednostňující pravděpodobně opačný přístup.

Termín „help“ nalezneme v řadě *tweetů*, s nimiž jsme se rovněž setkali již dříve – viz tabulky 2, 3, 7, 13, 19, 22, 23, 25, 34 a 36 –, a které: a) informují o enzymu schopném pojídat plast, který by mohl pomoci ukončit znečištění plasty, b) tvrdí, že pochopení toho, jak bakterie pomáhá rozkládat plasty, by mohlo napomoci ukončení problému

znečištění plasty, c) vybízejí k záchraně ohrožených živočišných druhů, d) vybízejí k udržování čistých lokálních vodních toků a okolí konkrétních měst, e) vyzývají ke vzdělávání se o problému znečištění plasty, f) vybízí k podpisu petice za účelem pomoci ukončení znečištění plasty, g) obecně k ukončení znečištění plasty, h) pomoci se šířením povědomí o znečištění plasty.

Termín „save“ zpravidla figuruje ve *tweetech*, jež vybízejí k záchraně ohrožených živočišných druhů, nebo Země, viz tabulky č. 7, 12; případně informují o organizaci zachraňující zvířata chycená v plastu, viz tab. č. 34. Další *tweet* z téže tabulky pak vybízí k pomoci se šířením povědomí o znečištění plasty. Viděli jsme již ale též takové, jež vedle sebe staví záchranu světa před znečištěním plasty a záchranu kuřat před smrtí a obojí přitom spojují s vegetariánstvím, viz tab. č. 21.

S termínem „clean“ jsme se již setkali v tabulkách č. 13, 19 a 32. Obsahují jej *tweet* vybízející čtenáře k udržení čistých vod v okolí Muskoky, dva *tweety* vybízející k připojení se k úklidu („clean up event“) a konečně poslední vyjadřující se negativně ke skutečnosti, že doposud nebyly vyčištěny oceány.

Jako důležité („important“) považuje jeden pisatel „zbavení se“ amerického právníka a republikána Scotta Pruitta z Agentury pro ochranu životního prostředí (EPA) v USA, detailnější však ve svém *tweetu* není, viz tabulka č. 15. Jiný pisatel zase poukazuje na důležitost uhlíkové stopy a zároveň též znečištění plasty, viz tab. č. 17. V téže tabulce se nachází také *tweet* zdůrazňující toho, aby světoví lídři obeznámili veřejnost s „epidemií znečištění plasty“. Jeden *tweet* pak poukazuje na důležitou a jedinečnou práci s důsledky znečištění plasty, kterou se věnuje jistá organizace, viz tab. č. 31.

Na základě vzorku tří náhodně vybraných *tweetů* obsahujících termín „breakthrough“ můžeme dále soudit, že za průlom považují pisatelé především objev enzymu schopného pojídat plast, viz tabulka č. 38.

"Scientists have accidentally created an enzyme that has the ability to eat through PET plastic used for soft drink bottles. The breakthrough brings hope upon solving the current global plastic pollution crisis. \n\nhttps://t.co/oCNAdoWbeF"
"A mutant enzyme that eats plastic bottles in a few days – far faster than the centuries it takes in the oceans. A breakthrough that could help solve the global plastic pollution crisis. https://t.co/dI81m4A6qI via @guardianeco"
"Scientists have created a mutant enzyme that breaks down plastic by accident. The breakthrough could help solve the global plastic pollution crisis by enabling for the first time the full #recycling of bottles. https://t.co/lPje7HIz6g"

Tabulka č. 38: *Tweety* obsahující termín „breakthrough“

Na termín „good“ jsme se již zaměřili v tabulce č. 18, proto si zde příslušné *tweety* nyní pouze připomeneme. Pisatel prvního z nich spatřuje v enzymu schopného pojídat plast potenciální hrozbu např. pro zateplení budov s tím, že by je enzym mohl poškodit. Druhý pisatel oproti tomu hodnotí pozitivně bambusové zubní kartáčky, kovové holící strojky či znovupoužitelné láhve. Konečně třetí *tweet* z tabulky č. 18 hodnotí negativně skutečnost, že na následky znečištění plasty zahynula velryba.

Termín „great“ se vyskytuje v různých kontextech. V tabulce č. 39, viz níže, můžeme vidět, že jím jeden pisatel hodnotí plakát, který je výstupem jeho školního projektu, a nese název „End of Plastic Pollution“. Jiný pisatel zase spatřuje výborné řešení problému znečištění plasty v enzymu schopném pojídat plast. Jako veliký krok správným směrem v rámci boje se znečištěním plasty pak hodnotí další pisatel zákaz užívání plastových mikročástic, který vešel v platnost ve Velké Británii.

"I just finished my sch project for #EarthDay . The poster looks great and the title for the poster is ' End of Plastic Pollution ' . Take care this world for our life and the flora fauna ❤'
"Recycling hope for plastic-hungry enzyme could be a great solution for our plastic pollution.."
"Britains microbe ban has gone into affect!! This is a great step in the right direction, and now they want to take more steps to fight back against plastic pollution!! Help End #Plastic #Pollution #WednesdayWisdom #ocean https://t.co/soFdSQTP3x"

Tabulka č. 39: *Tweety* obsahující termín „great“

Jako krásný („beautiful“) je pak popisován buď nejmenovaný film věnující se problematice znečištění plasty, viz tabulka č. 6, nebo naše planeta, viz tab. č. 15. Termín „happy“ jsme viděli již v tabulkách č. 7, 15 a 17, přičemž ve všech případech byl součástí přání šťastného Dne Země.

Oproti tomu termín „fun“ se vyskytuje v různých *tweetech*, viz tabulka č. 40 níže. Jeden z nich vybízí čtenáře k připojení se k chatu na Twitteru na téma lokálních potravin a boji s plýtváním, který označuje za zábavu. Další zas považuje za zábavu blíže explicitně nespecifikovaný projekt, který údajně převádí plasty na olej. Poslední *tweet* z tab. č. 40 tvrdí, že někdo – z kontextu lze vyvodit, že pravděpodobně asijské národy – zostuzují Američany pro zábavu, moc a profit. Podrobnější informace, které by nám tento názor pomohly osvětlit, však pisatel explicitně neuvádí.

"Come and join us for a Tweetchat on eating local and fighting waste. We'll talk all things tips and personal experiences, what you love to eat locally, how you're tackling food waste, plastic pollution and striving for sustainability. Use the hashtag #EatsTheLot and join the fun! https://t.co/NyezmdMDPG"
"Wow this project looks fun...converting plastic back into oil \xf0\u009f\u0098\u0083 #EarthDay #plastic #pollution #climatechange #change #diy https://t.co/QPACDIA3sm"
"Asian countries and messy fishermen are destroying the world's oceans \nhttps://t.co/SBmcEaILo5\nBut shaming normal Americans for fun, power, and profit is the goal here, as with all lefty movements."

Tabulka č. 40: *Tweety* obsahující termín „fun"

Za úžasnou („amazing") považují autoři *tweetů* planetu Zemi, viz tab. č. 7 a objev enzymu schopného pojídat plast, viz tab. č. 36. Protože se však jedná pouze o dva *tweety*, doplnili jsme je třetím náhodně vybraným, viz tab. č. 41. Jeho pisatel považuje za úžasnou píseň o znečištění plasty, kterou nazpívaly děti.

"Absolutely amazing song by kids on plastic pollution! Listen to it for them. They are taking control of their future home. Our Earth. Love it. https://t.co/crcSATIZz7"

Tabulka č. 41: *Tweet* obsahující termín „amazing"

Význam *tweetů* obsahujících termín „awesome“ nám pomůže osvětlit tabulka č. 42. Jeden pisatel považuje za ohromný nápad vydělávat peníze hudbou o moři, přičemž tyto peníze by mohly být použity k zachování oceánů. Další zase podobně hodnotí skutečnost, že se plastové slámky objevily ve zprávách. Pisatel třetího *tweetu* pak považuje za ohromnou blíže neurčitou věc, pokud by byla prokázána její efektivita.

"This is an awesome idea. Make money with sea music and raise the money to perserve the oceans \xf0\u009f\u008c\u008a Love it! #sea #music #ocean #polution\nhttps://t.co/jIGj7qXXnu"
"Straws in the news again, Daniela! This is awesome. https://t.co/ndiaWmuyb1"
"This could be really awesome if proven effective. But how about we stop using single-use plastic and letting it pollute our oceans and nature in the first place! https://t.co/9mPkipScTB"

Tabulka č. 42: *Tweety* obsahující termín „awesome“

Nyní se podíváme na vzorek tří náhodně vybraných *tweetů* obsahujících termín „wonderful“, viz tabulka č. 43. Jeden pisatel považuje za báječné buďto pláže, což považují za důvod, proč bychom je měli zachovat v jejich současném stavu. Druhý *tweet* referuje o báječném týmu, který bojuje s hrozbou znečištění plasty. Třetí pak hodnotí jako báječnou možnost, že by eznym schopný pojídat plast přeměnil plasty na potravu pro mořské živočichy.

"Our beaches are wonderful - which is why we need to keep them that way https://t.co/oVilPsQQoG #plasticpollution #earthday"
"Nothing as challenging as Plastic Pollution today. A wonderful team is out there to beat the menace. Getting all round support is very motivating. Keep it up Chinu and the Team ! https://t.co/UnEW802sxq"
"Scientists hope new enzyme will 'eat' plastic pollution https://t.co/enNq9jGSrQ Would this be wonderful if it turned plastic into something edible for marine life?"

Tabulka č. 43: *Tweety* obsahující termín „wonderful“

Konečně jako fantastické („fantastic“) je ve *tweetech* zmíněno sbírání odpadu, avšak s dodatkem, že pokud je to vše, co dokážeme v rámci boje se znečištěním plasty, nikdy s ním neskončíme; pisatel tedy toto řešení pravděpodobně považuje za nedostačující. Jiný *tweet* dále hodnotí jako fantastický obecně úbytek plastu. Jako ne fantastické pak považuje plasty pisatel prvního *tweetu* z tabulky č. 44.

"\"Plastic is no longer fantastic, in fact it is out of control.\"\n\nRead more on #META \n@ZeroWasteFR @ECNnetwork @KeepBritainTidy @RethinkPlastic @PBarczak\nhttps://t.co/BDNJzoLZBD"
"Less plastic? That's fantastic! https://t.co/Dmxe7QdABh"

"Picking up litter is fantastic, but if picking up is all we do that is all we're ever going to do! #cleanseas #plastic #pollution #solution #ourocean https://t.co/DeRjV9rwsZ"

Tabulka č. 44: *Tweety* obsahující termín „fantastic"

Jak vidíme, kladně jsou pisateli *tweetů* hodnoceny především snahy bojovat proti znečištění plasty, šíření osvěty a novinky o vědeckých objevech, které by mohly znečištění plasty pomoci vyřešit. Pozitivně hovoří ale rovněž o planetě Zemi a životním prostředí a vybízejí k jejich ochraně spolu s ochranou živočišných druhů. Nyní se podíváme na entity a postoje, které se vážou k termínům s nejvíce pozitivním sentimentem: „outstanding", „thrilled" a „superb".

Termín „outstanding" obsahuje pouze jeden unikátní *tweet*, který zve čitatele na výjimečnou blíže explicitně nespecifikovanou událost, jejímž tématem je řešení krize znečištění plasty, viz tabulka č. 45.

"Don't forget about this event scheduled for Sunday, April 22.\nIt will be an outstanding day at @buffaloscience as we all declare \"I am the solution to plastic pollution.\"\nDetails can be found here:\nhttps://t.co/guce6iWxPz …\n#ErieBYObag\n#Buffalo\n#EarthDay2018\n#PlasticPollution"

Tabulka č. 45: *Tweet* obsahující termín „outstanding"

Podobně nadšení vyjadřuje jeden pisatel v souvislosti s vývojem Dne Země. Dozvídáme se ale také, že britský princ Harry a jeho snoubenka Meghan Markle se údajně rozhodli využít své svatby a podpořit a zviditelnit v jejím rámci práci na záchraně britského pobřeží. Další pisatel je údajně nadšen („thrilled") v souvislosti

s vyhlášením dubna jako měsíce pro zvýšení povědomí o znečištění plasty a polystyrenem v americkém státu Colorado, které údajně učinil tamní guvernér John Hickenlooper, viz tabulka č. 46.

"I remember the first #EarthDay. We talked about population, pollution, and recycling. We had teach-ins and parades. I am thrilled to see 'the beat go on'. This year's theme, #EndPlasticPollution is spot on. Plastic was a problem even before… https://t.co/FJspUV45Se"
"We are thrilled that Prince Harry & Ms. Meghan Markle care about protecting beaches and tackling plastic pollution, and have chosen to use their special day to help promote and support our work to safeguard the UK coastline! Thank you! #RoyalWedding #PlasticFreeCoastlines https://t.co/LSU069j41B"
"We are thrilled to announce that Governor John Hickenlooper @GovofCO Declares April as \"Plastic and Styrofoam Pollution Awareness Month\" in Colorado. Our Thanks to @GovofCO @RepPerlmutter This is a big step to #beatplasticpollution @UNEnvironment @algore #microplastics @wef https://t.co/QgAokcioEJ"

Tabulka č. 46: *Tweety* obsahující termín „thrilled“

Poslední z výše zmíněných termínů („superb“) pak obsahuje pouze jeden unikátní *tweet*, jenž krom něj zahrnuje jen přiložený hypertextovým odkaz, takže z něj nejsme schopni vyčíst, čeho přesně se týká, aniž bychom tento odkaz rozklikli, viz tabulka č. 47.

"Superb \xf0\u009f\u0091\u008d\n\nhttps://t.co/DIPtU3RqPY"

Tabulka č. 47: *Tweet* obsahující termín „superb“

Následování hypertextových odkazů ale ani nebylo cílem naší analýzy. Bylo jím nahlédnutí diskuze o znečištění plasty na mikroblogovací síti Twitter. Ještě, než provedeme shrnutí celé analýzy, pojďme si zrekapitulovat o čem pojednávají *tweety* obsahující nejfrekventovanější termíny s pozitivním sentimentem.

Jako pozitivní je pisateli *tweetů* hodnocena ochrana životního prostředí, potažmo celé planety Země a živočišných druhů zde žijících. Dále *tweety* obsahující termíny s pozitivním sentimentem vybízejí čtenáře k řešení problému znečištění plasty, přičemž jako možné způsoby tohoto řešení nabízejí pisatelé např. užívaní bioplastů, či úplné zřeknutí se plastů. Za dobré alternativy k plastovým výrobkům považují pisatelé např. bambusové zubní kartáčky, kovové holící strojky či znovupoužitelné láhve. Dále jsou v *tweetech* čtenáři vybízeni k udržování čistých vodních toků a okolí konkrétních měst, k vzdělávání se o znečištění plasty, a k šíření povědomí o tomto problému – dozvídáme se např., že duben byl vyhlášen měsícem pro zvýšení povědomí o znečištění plasty a polystyrenem v americkém Coloradu. Za skvělý krok v boji se znečištění plasty je považován zákaz užívání plastových částic ve Velké Británii. Avšak nejčastěji zmiňovaným – možným – řešením problému je enzym schopný pojídat plast, jehož objev je považován za průlom. Jako pozitivní jsou rovněž nahlíženy nejrůznější události počínaje Dnem Země, přes online chat o lokálních potravinách, až po hromadný úklid. Stejně hodnotí pisatelé rovněž umělecké počiny tematizující znečištění plasty, případně moře jako jsou např. film či hudba. V některých *tweetech* měly ale pozitivní termíny v závislosti na kontextu také negativní konotace, viz např. ne fantastické plasty, případně působily celé *tweety* ironicky. Jiné *tweety* obsahovaly zase příliš málo informací na to, abychom byli schopni rozklíčovat jejich význam.

9.10 Shrnutí analýzy

Na začátku datové analýzy diskuze o znečištění plasty na mikroblogovací síti Twitter jsme představili programovací jazyk R, v němž byla analýza převážně provedena. Dále jsme představili knihovnu jazyka R – rtweet – určenou pro dolování dat z mikroblogovací sítě Twitter a rozhraní pro programování aplikací Twitteru, k němuž se knihovna rtweet připojuje za účelem získání dat. Poté jsme se věnovali metodice sběru dat. Dataset analyzovaný v této práci tvoří celkem 64 299 anglicky psaných *tweetů* obsahujících výraz „plastic pollution", jež byly publikovány v časovém rozmezí od 13. 04. 2018 19:01:20 do 23. 04. 2018 18:24:05 letního českého času. Následně jsme z datasetu vytvořili užitím jak vlastních funkcí, tak pomocí funkcí knihoven jazyka R korpus *tweetů*.

Po vytvoření korpusu *tweetů* jsme přistoupili k detekci tematických celků pomocí tří různých metod: a) Latentní Dirichletovy alokace (LDA), b) frekvenčních tabulek tokenů a bigramů, c) síťového grafu. Kvůli komputačním nárokům algoritmu LDA a časové náročnosti jeho výpočtů jsme se jej rozhodli použít pouze pro odhalení dvou nejvýraznějších témat našeho korpusu. Těmi se ukázaly být Den Země a vědecký objev enzymu schopného pojídat plast. Díky tomu, že každý *tweet* obsahuje údaj o tom, kdy byl publikován, mohli jsme zastoupení dvou nejvýraznějších témat korpusu sledovat i v průběhu času. Největší koncentrace nejvýraznějšího tématu – Dne Země – nastala 22. dubna, kdežto téma vědeckého objevu enzymu schopného pojídat plast kumuluje nejvíce dubna 17. a následně jeho koncentrace stoupá o pět dní později spolu s tématem Dne Země.

Když jsme pomocí LDA detekovali dvě nejvýraznější témata korpusu, zaměřili jsme se na rozložení tematických celků na makroúrovni, přičemž jsme pro ilustraci uvedli deset nejfrekventovanějších hashtagů. Zjistili jsme, že jak mesovrstvě – celý korpus včetně hashtagů –, tak makrovrstvě – pouze hashtagy – dominuje téma Dne Země. Po nabytí těchto poznatků jsme se rozhodli rozkrýt rovněž další tematické celky korpusu, a sice pomocí frekvenčních tabulek tokenů a bigramů.

Nejprve jsme pro ilustraci nahlédli padesát nejfrekventovanějších tokenů, které jsme doplnili třiceti nejfrekventovanějšími bigramy. Nejfrekventovanějšími tokeny

jsou: „earthday“, „earth“, „day“, „enzyme“, „end“ a „scientists“, což odpovídá rozložení dvou hlavních témat v korpusu. Zjistili jsme dále, že znečištění plasty je pisateli *tweetů* považováno za jeden z největších světových problému a za globální krizi. Řada *tweetů* vybízí čtenáře v reakci na problém znečištění plasty ke zřeknutí se plastových slámek, lahví na vodu a k pokusu o omezení užívání jednorázových plastů. Plastové slámky a míchadla jsou často zmiňovány rovněž v souvislosti s jejich zákazem, ať už obecně na úrovni měst či národů nebo konkrétně – v případě Velké Británie. Někteří pisatelé pak apelují na lokálně situované čtenáře, aby udržovali čistotu vodních toků a ve svém okolí – např. Petrohrad či Muskoka – a okolí samotného.

Odhalili jsme též *tweety* referující o fatálních následcích znečištění plasty pro oceánskou faunu, jež tvrdí, že ročně zemře v jeho důsledku na jeden milion mořských ptáků a 100 000 mořských savců. Některé *tweety* obsahují informace o velrybách a o jejich smrti způsobené plasty, které měly ve svých útrobách. Na tyto jevy následně dle *tweetů* reagují umělci sochami vyrobenými z plastového smetí.

Pouze na základě padesáti nejfrekventovanějších tokenů a třiceti nejfrekventovanějších bigramů jsme byli schopni rozkrýt tři různé typy reakcí – ať už na úrovni jedinců, komunit, či státních celků – na problém znečištění plasty. Ten dostal zároveň jasnější obrysy, když jsme zjistili, jakým způsobem ovlivňuje život mořských živočichů. V další části naší analýzy jsme se rozhodli nahlédnout vztahy jednotlivých bigramů ještě detailněji za užití síťového grafu.

Znázorněním sto nejfrekventovanějších bigramů jsme zhotovili orientovaný nesouvislý síťový graf o deseti komponentách. Vrcholem s největším vstupním stupněm celého grafu je „plastic“, následují jej: „earthday“, „earth“, „day“ „end“, „enzyme“, „scientists“ a „pollution“. Tento fakt odpovídá rozložení hlavních témat v korpusu – bigram „plastic pollution“ se nutně vyskytuje ve všech *tweetech*.

Díky síťovému grafu zhotovenému v programu Cytoscape jsme odhalili *tweety* vybízející čtenáře, aby chránil vodní plochy a toky, využíval efektivně energii, jedl méně masa a jedl lokální a organické produkty, chránil včely a ukončil znečištění plasty. Interpretovat si to můžeme tak, že čím efektivnější bude výroba energie, tím čistší bude vodní ekosystém na naší planetě. Narazili jsme rovněž na informaci, že státy Island,

Norsko, Japonsko a Dánsko spolu s Faerskými ostrovy zastavily lov velryb kvůli nebezpečí vzniklému údajně z důvodů znečištění plasty a klimatické změny. Na rozdíl od změny klimatu však, tvrdí pisatelé, nemá problém znečištění plasty své kritiky.

Další skupinu *tweetů* tvoří ty, jež se týkají Dne Země. Některé z nich vybízejí čtenáře ke sledování odpočítávání času do jeho začátku, či na ně apelují, aby se připojili se k jisté úklidové události a podepsání petice. Někteří pisatelé se ale uchylují v souvislosti se Dnem Země též k ironii. Hovoří ironicky o snaze bohatých celebrit slavit Den Země, protože mají údajně více aut, než potřebují, nosí kožešiny, jedí maso apod.

Největší výhoda síťového grafu nicméně nespočívá v možnosti odhalit tematické celky, nýbrž v rozkrytí vztahů mezi jednotlivými termíny – reprezentovanými vrcholy grafu – a v možnosti nahlédnout vztahy mezi celými tematickými celky. Zjistili jsme, že vědecký objev je zde primárně vztahován přímo k tématu plastů, a ke znečištění plasty až sekundárně. Oproti tomu Den Země je zmiňován primárně spolu s tématem znečištění plasty. Hovoří-li se v *tweetech* v rámci Dne Země o znečištění oceánů, krizi a o nejhorších světových problémech, jejich nejčastěji zmiňovaným řešením je enzym schopný pojídat plast. Dále jsme byli schopni vyčíst ze síťového grafu, že by enzym mohl být nejspíše aplikován na rozklad odpadu, který tvoří plastové slámky či lahve.

Jak vidíme, pomocí třech rozličných metod pro detekci tematických celků vyskytujících se v našem korpusu jsme došli k týmž výsledkům. Jak užitím Latentní Dirichletovy alokace (LDA), tak frekvenčních tabulek tokenů a bigramů a konečně též užitím síťového grafu jsme detekovali dva nejvýraznější tematické celky korpusu: *tweety* týkající se Dne Země a *tweety* referující o vědeckém objevu enzymu schopného pojídat plast. Vzhledem k tomu, že jsme algoritmus LDA kvůli jeho komputačním nárokům a časové náročnosti jeho výpočtů užili pouze k objevení dvou výše zmíněných témat, nemůžeme jej porovnávat s frekvenčními tabulkami a síťovým grafem co se dalších tematických celků týče.

Vzhledem k výsledkům analýzy v podobě detekovaných tematických celků *tweetů* však můžeme tvrdit, že frekvenční tabulky tokenů a bigramů přinesly téměř totožné výsledky jako síťový graf. Zároveň mají oproti LDA jak frekvenční tabulky,

tak síťový graf zřetelnou výhodu v malých komputačních nárocích a malé časové náročnosti výpočtů. Nevýhodou pak je, že výsledky obou těchto metod jsou na rozdíl od LDA hůře interpretovatelné – jedná se hlavně o určení hranice mezi jednotlivými tematickými celky, tzn. např. jaké termíny reprezentují tematický celek č. 1 a jaké již spadají do tematického celku č. 2 apod. Uživatelsky přívětivější je v tomto ohledu síťový graf, z něhož se dá v závislosti na jeho struktuře a způsobu, jakým ohodnotíme jeho vrcholy či hrany vyvodit řada informací, které z výstupu algoritmu LDA a frekvenčních tabulek nevyčteme. Můžeme tak získat např. přehled o tom, jak jsou spolu jednotlivé termíny – znázorněny vrcholy grafu – navzájem propojeny nebo zdali některé z nich sehrávají roli prostředníku mezi jinými termíny nebo tematickými celky.

Když jsme pomocí výše popsaných metod odhalili v korpusu *tweetů* nám dříve neznámé tematické celky, rozhodli jsme se podívat na nejfrekventovanější termíny jak s negativním, tak s pozitivním sentimentem a na *tweety*, v nichž se vyskytují. Skóre sentimentu jsme jednotlivým termínům připsali na základě lexikonu AFFIN, který vznikl speciálně pro potřeby analýzy sentimentu textů pocházejících z internetu, primárně z mikroblogovacích sítí jako je např. Twitter. Poté, co jsme znázornili grafy 23 nejfrekventovanějších termínů s negativním a 23 nejfrekventovanějších termínů s pozitivním sentimentem, podívali jsme se do *tweetů* na to, v jakých kontextech se vyskytují a jaké postoje jimi pisatelé vyjadřují.

Nejprve jsme se zaměřili na termíny s pozitivním sentimentem. Zjistili jsme, že část *tweetů*, které je obsahují, pojednává o vědeckém objevu enzymu schopného pojídat plast, a tudíž termíny, jež se v nich vyskytují a jsou anotované dle lexikonu AFFIN jako negativní, mají vzhledem ke kontextu konotace spíše pozitivní. Obdobně můžeme nahlížet na *tweety* referující o objevu brouků pojídajících plast.

Další část termínů s negativním sentimentem figurovala v *tweetech*, které hodnotí negativně problém znečištění plasty a jeho fatální důsledky na mořskou, ale též zemskou faunu, životní prostředí či dokonce celou planetu Zemi. Konkrétně je pisateli jako negativní hodnocena např. situace v Indonésii. Oproti tomu jiné *tweety* obsahující negativní termíny hodnotí pozitivně iniciativy bojující proti znečištění plasty, případně k těmto iniciativám čtenáře vybízejí.

Část *tweetů* s negativními termíny se vyjadřovala negativně k některým americkým politikům (Donald Trump, John Podesta) a jejich postojům či lhostejnosti vůči problému znečištění plasty. Termíny s nejvíce negativním sentimentem – vulgarismy – figurují v našem korpusu pouze v malém množství a pisatelé je užívají buď jako oslovení, intenzifikační částice, nebo k vyjádření silně negativního názoru.

Jako pozitivní je pak pisateli *tweetů* hodnocena ochrana životního prostředí, potažmo celé planety Země a živočišných druhů zde žijících. Dále *tweety* obsahující termíny s pozitivním sentimentem vybízejí čtenáře k řešení problému znečištění plasty, přičemž jako možné způsoby tohoto řešení nabízejí pisatelé např. užívaní bioplastů, či úplné zřeknutí se plastů. Za dobré alternativy k plastovým výrobkům považují pisatelé např. bambusové zubní kartáčky, kovové holící strojky či znovupoužitelné láhve.

Dále jsou v *tweetech*, jež obsahují termíny s pozitivním sentimentem, čtenáři vybízeni k udržování čistých vodních toků a okolí konkrétních měst, k vzdělávání se o znečištění plasty, a k šíření povědomí o tomto problému – dozvídáme se např., že duben byl vyhlášen měsícem pro zvýšení povědomí o znečištění plasty a polystyrenem v americkém Coloradu. Za skvělý krok v boji se znečištění plasty je považován zákaz užívání plastových částic ve Velké Británii. Avšak nejčastěji zmiňovaným – možným – řešením problému je enzym schopný pojídat plast, jehož objev je považován za průlom. Jako pozitivní jsou rovněž nahlíženy nejrůznější události počínaje Dnem Země, přes online chat o lokálních potravinách, až po hromadný úklid. Stejně hodnotí pisatelé rovněž umělecké počiny tematizující znečištění plasty, případně moře jako jsou např. film či hudba. V některých *tweetech* měly ale pozitivní termíny v závislosti na kontextu také negativní konotace, viz např. ne fantastické plasty, případně působily celé *tweety* ironicky. Jiné *tweety* obsahovaly zase příliš málo informací na to, abychom byli schopni rozklíčovat jejich význam.

Jak vidíme, téměř bez jakýchkoli apriorních znalostí o obsahu korpusu *tweetů* jsme byli schopni užitím metod dolování textových dat detekovat nejvýraznější tematické celky. Pomocí prostého náhodného výběru jsme zároveň vždy nahlédli vzorek *tweetů* referujících o daném tématu a zjistili tak, o čem konkrétně v rámci tohoto tématu pojednávají. Krom toho jsme v korpusu detekovali nejfrekventovanější termíny s negativním a pozitivním sentimentem. Opět pomocí prostého náhodného výběru

vzorku *tweetů* obsahujících tyto termíny jsme rozkryli, v jakých kontextech je pisatelé užívají, tzn. jaké postoje jimi vyjadřují.

Závěr

Cílem této práce bylo představit metodu dolování textových dat vycházející z tradice vědní disciplíny Digital Humanities a uplatnit ji při analýze textových dat pocházejících z mikroblogovací sítě Twitter. V teoretické části jsme se věnovali nejprve historii a vývoji Digital Humanities spolu s dolováním textových dat, nastínili jsme rovněž jejich současnou podobu. Dále jsme představili teorii sítí spolu se základními pojmy teorie grafů, které jsme využili v praktické části. Vzhledem k tomu, že jsme se v praktické části věnovali analýze dat pocházejících z mikroblogovací sítě Twitter, představili jsme v teoretické části také dolování sociálních médií, jejich historii a kategorie a Twitter spolu s jeho komunikačními vrstvami. Dále jsme představili dolování skrytých témat v textech, konkrétně metodu Lanentní Dirichletovy alokace, kterou jsme použili v praktické části. Teoretickou část jsme zakončili představením postojové analýzy a krátkým pojednáním o enviromentálním problému znečištění plasty, jehož se v praktické části analyzovaná textová data z Twitteru týkala.

V praktické části jsme užitím metod dolování textových dat a užitím programovacího jazyka R nejprve vytvořili z datasetu *tweetů* korpus *tweetů*, v němž jsme následně detekovali pomocí Latentní Dirichletovy alokace dvě nejvýraznější témata. Každé z těchto témat jsme opatřili kvalitativním rozborem. Poté jsme detekovali další témata pomocí frekvenčních tabulek tokenů a bigramů a pomocí síťového grafu zhotoveného v programu Cytoscape. Síťový graf nám pomohl poodhalit rovněž vztahy mezi jednotlivými detekovanými tématy. Rovněž veškerá tato témata jsme opatřili kvalitativním rozborem. V následujícím kroku jsme užitím lexikonu AFFIN detekovali nejfrekventovanější termíny s nejvíce negativním a nejvíce pozitivním sentimentem. Následně jsme nahlédli, v jakých kontextech, tzn. *tweetech* se vyskytují a jaká témata reprezentují. Taktéž každé z těchto témat jsme opatřili kvalitativním rozborem. Výsledky analýzy jsme průběžně vizualizovali formou grafů. K interpretaci výsledků, tj. kvalitativnímu rozboru jsme použili prostý náhodný výběr.

Zvolené metody – Latentní Dirichletova alokace, frekvenční tabulky tokenů a bigramů, síťový graf – se ukázaly být pro dolování textových dat efektivními. Vytyčený

cíl práce – nahlédnout způsob, jakým lidé hovoří o enviromentálním problému znečištění plasty na mikroblogovací síti Twitter, témata, o nichž hovoří, a co hodnotí pozitivně a co naopak negativně – byl splněn.

Budoucí výzkum v této oblasti by se mohl zaměřit např. na automatické třídění pozitivně laděných *tweetů* od negativně laděných pomocí hlubokých neuronových sítí, či na predikci vzniku iniciativ bojujících proti znečištění plasty na základě diskuze o tomto problému v rámci sociálních médií a médií obecně.

Seznam použité literatury

Knižní a časopisecké zdroje

Aggarwal, C. C., ed. (2011). *Social Network Data Analytics*. New York: Springer.

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In D. M. Berry, ed., *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan.

Blei, D. M., & Lafferty, J., D. (2009). Topic Models. In Srivastava, A., Sahami, M., eds., *Text Mining: Classification, Clustering, and Applications.* Boca Raton: CRC Press.

Blei, D. M., Ng, A., Y., & Jordan, M., I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Brandes, U., & Erlebach, T. (2005). *Network analysis: Methodological foundations*. Berlin: Springer.

Bruns, A., & Moe, H. (2013). Structural Layers of Communication on Twitter. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C., eds., *Twitter and Society*. New York: Peter Lang.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities*. Cambridge, Massachusetts: The MIT Press.

Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, 14(2), 83–90.

Cotton, R. (2013). *Learning R: A Step-by-Step Function Guide to Data Analysis*. Sebastopol: O'Reilly Media.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, a member of the Perseus Books Group.

Estrada, E., & Knight, P. (2015). *A First Course in Network Theory*. Oxford: Oxford University Press.

Gardinier, E., & Musto, R. G. (2015). *The Digital Humanities: A Primer for Students and Scholars*. Cambridge: Cambridge University Press.

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1–30.

Hayles, N. K. (2012). How We Think: Transforming Power and Digital Technologies. In D. M. Berry, ed., *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan.

Kaplan, A., M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2), 105–113.

Kaplan, A., M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68.

Kwartler, R. (2017). *Text Mining in Practice with R*. Hoboken: Willey–Blackwell.

Lewis, T., G. (2008). *Network Science: Theory and Applications.* Hoboken: Willey–Blackwell.

Lin, Y. (2012). Transdisciplinarity and Digital Humanities: Lessons Learned from Developing Text-Mining Tools for Textual Analysis. In D. M. Berry, ed., *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge: Cambridge University Press.

Manning, Ch., D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: The MIT Press.

Manovich, L. (2001). *The Language of New Media*. Cambridge, Massachusetts: The MIT Press.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* New York: Elsevier.

Rogers, R. (2013). Foreword: Debanalising Twitter: The Transformation of an Object of Study. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C., eds., *Twitter and Society*. New York: Peter Lang.

Sanger, J., & Feldman, R. (2002). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press.

Schreibman, S., Siemens, R., & Unsworth, J. (2016). *A New Companion to Digital Humanities*. Hoboken: Willey–Blackwell.

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. Sebastopol: O'Reilly Media.

Small, H. (2012). *The Value of the Humanities*. Oxford: Oxford University Press.

Štícha, F. et al. (2003). *Akademická gramatika spisovné češtiny*. Praha: Academia.

Veselovská, K. (2017). *Sentiment Analysis in Czech.* Praha: ÚFAL.

Zafarani, R., Abbasi, M., A., & Liu, H. (2014). *Social Media Mining: An Introduction.* Cambridge: Cambridge University Press.

Zhai, Ch., & Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* New York: ACM Books.

Elektronické zdroje

Barabási, A-L. (2016). *Network Science*. Dostupné z http://networksciencebook.com/

Barnes, D. K. A., Galgani, F., Thompson, R. C., & Barlaz, M. (2009). Accumulation and fragmentation of plastic debris in global environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1526), 1985–1998. Dostupné z http://doi.org/10.1098/rstb.2008.0205

Beaumont, P. (2011, 25. února). *The truth about Twitter, Facebook and the uprisings in the Arab world.* Dostupné z https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya

Brandom, R. (2014, 12. února). *This algorithm can predict a revolution*. Dostupné z https://www.theverge.com/2014/2/12/5404750/can-a-database-predict-a-revolution

Digital humanities. (n.d.). Dostupné z https://en.wikipedia.org/wiki/Digital_humanities

Geyer, R., Jambeck, J. R., & Law, K., L. (2017). Production, use, and fate of all plastics ever made. *Science Advances*, 3(7).
Dostupné z http://advances.sciencemag.org/content/3/7/e1700782.full

Hearst, M. (2003). *What Is Text Mining?*
Dostupné z http://people.ischool.berkeley.edu/~hearst/text-mining.html

Horton, H. (2017, 3. února). *Whale found dying off coast of Norway with 30 plastic bags in its stomach*. Dostupné z https://www.telegraph.co.uk/news/2017/02/03/whale-found-dying-coast-norway-30-plastic-bags-stomach/

Chen, E. (2012). *Soda vs. pop with Twitter*.
Dostupné z http://blog.echen.me/2012/07/06/soda-vs-pop-with-twitter/

Kelly, R. (2009). *Twitter study*. San Antonio: Pear Analytics. Dostupné z http://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf

Kosuth M., & Mason S. A., Wattenberg E. V. (2018). Anthropogenic contamination of tap water, beer, and sea salt. *PLOS ONE*, 13(4): e0194970.
Dostupné z https://doi.org/10.1371/journal.pone.0194970

Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. [web]
Dostupné z http://www.korpus.cz

Lebreton, L. et al. (2018). Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. *Nature, Scientific Reports*, 8.
Dostupné z https://www.nature.com/articles/s41598-018-22939-w

Liebezeit, G., & Liebezeit, E. (2013). Non-pollen particulates in honey and sugar, *Food Additives & Contaminants: Part A*, 30(12), 2136–2140. Dostupné z https://www.tandfonline.com/doi/abs/10.1080/19440049.2013.843025?journalCode=tfac20

List of Websites and Apps Blocked in China [Updated April 2018]. (2018, 26. dubna). [web] Dostupné z http://startuplivingchina.com/list-websites-apps-blocked-china/

Lu, L., & Zhang, M. (2013). Edge Betweenness Centrality. In Dubitzky W., Wolkenhauer O., Cho KH., Yokota H., eds., *Encyclopedia of Systems Biology*. New York: Springer. Dostupné z https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_874

Mason, A. S., et al. (2018). *Synthetic Polymer Contamination in Bottled Water*. Dostupné z https://orbmedia.org/sites/default/files/FinalBottledWaterReport.pdf

MuleSoft Videos. (2016, 19. června). *What is an API?* [video]. Dostupné z https://www.youtube.com/watch?v=s7wmiS2mSXY

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Workshop Proceedings*: 93–98. Dostupné z https://arxiv.org/abs/1103.2903

Quoc V. Le, & M. Schuster, *A Neural Network for Machine Translation, at Production Scale*. (2016, 27. září). Dostupné z https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

Seven charts that explain the plastic pollution problem. (2017, 10. prosince). Dostupné z http://www.bbc.com/news/science-environment-42264788

Schneider, Ch. (2017, 16. května). *The biggest data challenges that you might not even know you have*. [web] Dostupné z https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know

Taiwan to ban disposable plastic items by 2030. (2018, 22. února) Dostupné z https://phys.org/news/2018-02-taiwan-disposable-plastic-items.html#jCp

Top 20 Facebook Statistics - Updated April 2018. (2018, 30. dubna). [web] Dostupné z https://zephoria.com/top-15-valuable-facebook-statistics/

Twitter Usage Statistics. (n.d.).
Dostupné z http://www.internetlivestats.com/twitter-statistics/

Whale found dead with 29kg of plastic in stomach. (2018, 12. dubna).
Dostupné z http://www.newshub.co.nz/home/world/2018/04/whale-found-dead-with-29kg-of-plastic-in-stomach.html

Zalasiewicz, J. et al. (2016). The geological cycle of plastics and their use as a stratigraphic indicator of the Anthropocene. *Anthropocene*, 13, 4–17.
Dostupné z http://www.sciencedirect.com/science/article/pii/S2213305416300029

Seznam obrázků

Seznam grafů

Seznam tabulek

Obsah přiloženého CD

- DP_Jaworek_Robert.pdf
    - Diplomová práce ve formátu PDF
- DP_Jaworek_Robert_code.R
    - Kód použitý k analýze ve formátu R
- Tweets.Rda
    - Tweety analyzované v této práci ve formátu RDA
- StopWords_pp_included.csv
    - Stop slova zahrnující termíny „plastic“ a „pollution“ použitá při analýze ve formátu CSV
- StopWords_without_pp.csv
    - Stop slova nezahrnující termíny „plastic“ a „pollution“ použitá při analýze ve formátu CSV
- 100bigrams_pp_included.xlsx
    - 100 nejfrekventovanějších bigramů včetně termínů „plastic“ a „pollution“ ve formátu XLSX
- Negative_sentiment_terms.xlsx
    - Nejvíce frekventované termíny s negativním sentimentem ve formátu XLSX
- Positive_sentiment_terms.xlsx
    - Nejvíce frekventované termíny s pozitivním sentimentem ve formátu XLSX

Příloha: Stop slova zahrnující rovněž „plastic“ a „pollution“

a	about	above	after	again
against	all	am	an	and
any	are	aren't	as	at
be	because	been	before	being
below	between	both	but	by
cannot	can't	co	could	couldn't
did	didn't	do	does	doesn't
doing	don't	down	during	each
few	for	from	further	had
hadn't	has	hasn't	have	haven't
having	he	he'd	he'll	her
here	here's	hers	herself	he's
him	himself	his	how	how's
i	i'd	if	i'll	i'm
in	into	is	isn't	it
its	it's	itself	i've	let's
me	more	most	mustn't	my
myself	no	nor	not	of
off	on	once	only	or
other	ought	our	ours	ourselves
out	over	own	plastic	pollution
rt	s	same	shan't	she
she'd	she'll	she's	should	shouldn't
so	some	such	t	than
that	that's	the	their	theirs
them	themselves	then	there	there's
these	they	they'd	they'll	they're
they've	this	those	through	to

too	under	until	up	very
was	wasn't	we	we'd	we'll
were	we're	weren't	we've	what
what's	when	when's	where	where's
which	while	who	whom	who's
why	why's	with	won't	would
wouldn't	you	you'd	you'll	your
you're	yours	yourself	yourselves	you've