Filozofická fakulta Univerzity Palackého v Olomouci
Katedra obecné lingvistiky

Identifikace autora ve forenzní lingvistice

magisterská diplomová práce

Autor: Bc. Anna Tichá
Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Olomouc
2024

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Identifikace autora ve forenzní lingvistice“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne 8. 5. 2024 Podpis

Poděkování: Chtěla bych poděkovat Mgr. Vladimíru Matlachovi, PhD., za vedení mé diplomové práce, cenné rady a odborný dohled.

Abstrakt

Název práce: Identifikace autora ve forenzní lingvistice
Autor práce: Bc. Anna Tichá
Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.
Počet stran a znaků: 100 s. (121 396 znaků)
Počet příloh: 0
Abstrakt (minimálně 900 znaků):

Tato diplomová práce se zaměřuje na identifikaci autorů v oblasti forenzní lingvistiky a skládá se ze dvou částí – teoretické a praktické. Teoretická část obsahuje úvod do forenzní lingvistiky, představuje klíčové koncepty a historický přehled této disciplíny. Následuje analýza různých typů textů v této oblasti, definice kritérií pro jejich analýzu a představení korpusu textů, představení lingvistického softwaru QUITA a vysvětlení konceptu logistické regrese.

Praktická část se věnuje analýze textů a metodám identifikace autorů, Nejprve proběhne kvantitativně-kvalitativní analýza s definovanými kritérii hodnocení anonymních textů. Následuje rozvoj dosavadních metod identifikace autorů s využitím vícerozměrných metod, konkrétně MDS. Dále budou použity Bag-of-Words modely pro analýzu textů. Kapitola dále obsahuje analýzu pomocí lingvistického indexu entropie a opětovné použití logistické regrese pro vyhodnocení úspěšnosti metod. V závěru navrhneme možnosti dalšího rozšíření výzkumu v této oblasti.

Klíčová slova: forenzní lingvistika, kvantitativní lingvistika, identifikace autora, kvantitativní a kvalitativní analýza, Bag-of-Words model, vícerozměrné škálování, QUITA, logistická regrese, indexy.

Abstract

Title: Author identification in forensic linguistics
Author: Bc. Anna Tichá
Supervisor: Mgr. Vladimír Matlach, Ph.D.
Number of pages and characters: 100 p. (121 396 characters)
Number of appendices: 0
Abstract (900 characters):

This master's thesis focuses on author identification in the field of forensic linguistics and consists of two parts – theoretical and practical. The theoretical part includes an introduction to forensic linguistics, presenting key concepts and a historical overview of this discipline. It is followed by an analysis of various types of texts in this area, definition of criteria for their analysis, and presentation of a text corpus, introduction of the linguistic software QUITA, and explanation of the concept of logistic regression.

The practical part deals with text analysis and methods of author identification. First, a quantitative-qualitative analysis with defined criteria for evaluating anonymous texts will be conducted. This is followed by the development of existing methods of author identification using multidimensional methods, specifically MDS. Bag-of-Words models will also be used for text analysis. The chapter further includes analysis using the linguistic entropy index and reapplication of logistic regression to evaluate the success of methods. In conclusion, we will propose possibilities for further research expansion in this area.

Keywords: forensic linguistics, quantitative linguistics, author identification, quantitative and qualitative analysis, Bag-of-Words model, multidimensional scaling, QUITA, logistic regression, indices.

Obsah

Úvod

Tématem této diplomové práce je identifikace autora ve forenzní lingvistice. Práce se skládá ze dvou částí – teoretické a praktické. První kapitola teoretické části se zaměřuje na úvod do oblasti forenzní lingvistiky. Představuje čtenářům klíčové koncepty a principy tohoto oboru. Následující kapitola přináší historický přehled forenzní lingvistiky, ve světě a následně se zaměřuje na její vývoj v českém prostředí, představuje významné osobnosti, práce a případy, ve kterých byla tato disciplína aplikována v praxi. Předposlední kapitola teoretické části analyzuje různé typy textů používané ve forenzní lingvistice a definuje kritéria, jež musí tyto texty splňovat, aby byly vhodné pro analýzu, a dále představí čtenáři korpus textů, se kterým budeme pracovat. V poslední kapitole je představen lingvistický software QUITA, jenž slouží k analýze textů, a zároveň je čtenářům vysvětlen koncept logistické regrese.

Praktická část se zabývá především analýzou textů a popisem metod, které v naší práci používáme k identifikaci autorů. Tato část je rozdělena do sedmi kapitol. Cílem praktické části bude vytvořit stručný a přehledný souhrn metod, které je možné použít k identifikaci autora. Následně je aplikujeme na námi zvolené anonymní texty. Jako první bude provedena kvantitativně-kvalitativní analýza, v níž si stanovíme tři kritéria hodnocení anonymních textů. Tato kritéria budou detailně popsána v kapitole 5. Následně provedeme vyhodnocení výskytu těchto kritérií v námi zkoumaných textech. Za účelem zlepšení orientace pro čtenáře bude vytvořeno a popsáno několik tabulek. Pro zobrazení výsledků využijeme vícerozměrných metod (MDS), jejich využití je však čistě ilustrativní. Budeme využívat lingvistického softwaru QUITA. Tato metoda bude detailně popsána v podkapitole 6.2. Analytická část práce bude koncipována tak, že provedeme analýzu a pomocí MDS vytvoříme náhledy na podobnost anonymních textů a pokusíme se identifikovat shluky jednotlivých textů a určit, zda je možné od sebe jednotlivé autory jednoduše oddělit a následně náš odhad statisticky ověříme pomocí logistické regrese. V kapitole šest budeme dále hodnotit, která kritéria se pro naši analýzu ukážou jako přínosná a která nám naopak

neposkytnou informace užitečné k rozlišení jednotlivých autorů. Další metodou, kterou v práci použijeme, bude Bag-of-Words model, pomocí níž provedeme analýzu pro stejný počet tokenů, 618, jednou však půjde o tokeny vybrané náhodně a jednou o tokeny, které v textu jdou po sobě. Na konci této kapitoly vyhodnotíme, zda byla tato metoda úspěšná v rozlišení jednotlivých autorů. Poslední metodou, kterou v práci použijeme, bude znovu Bag-of-Words model, tentokrát však budou zkoumané soubory obsahovat pouze hapax legomenon jednotlivých textů. Provedeme analýzu pro stejný počet tokenů, 618, jednou však půjde o tokeny vybrané náhodně a jednou o tokeny, které v textu jdou po sobě. Na konci této kapitoly vyhodnotíme, zda byla tato metoda úspěšná v rozlišení jednotlivých autorů. Jako poslední metodu použijeme analýzu pomocí indexu entropie. Znovu pak provedeme analýzu dat pomocí logistické regrese a stejně tak provedeme znovu i vyhodnocení úspěšnosti metody.

V závěru diplomové práce shrneme získané poznatky a zhodnotíme, která metoda určování autorství anonymních textů se ukázala být nejúspěšnější a navrhneme možnosti dalšího rozšíření výzkumu.

1. Forenzní lingvistika a určování autorství ve světě a u nás

Forenzní lingvistika je multidisciplinární obor, který se nachází na pomezí lingvistiky a forenzních věd (práva a kriminalistiky). Zahrnuje mnoho sub-disciplín teoretické a aplikované lingvistiky, jejichž poznatky a metodologie jsou použity při analýze jazykového materiálu souvisejícího s právním a forenzním kontextem.

V souvislosti s tím, jaký druh jazyka je analyzován a jaké metody zkoumání jsou zvoleny, můžeme hovořit o různých podoborech forenzní lingvistiky. Například forenzní fonetika se zabývá analýzou zvuků řeči, forenzní stylistika zkoumá styl psaní, forenzní sémantika se zaměřuje na význam slov a vět, forenzní pragmatika studuje, jak lidé používají jazyk v konkrétních situacích, a forenzní diskurzivní analýza se věnuje studiu celých textů nebo rozhovorů. Forenzní lingvistika také zahrnuje některé specializované obory, jako je grafologie (studium rukopisu) a analýza fyzických vlastností dokumentů. Existuje i vědecký obor zvaný

softwarová forenzika, který se zabývá analýzou způsobu, jakým programátoři píší svůj kód[1].

Historické kořeny forenzní lingvistiky jsou obtížně vystopovatelné. Již od pradávna se lidé snažili určit autorství uměleckých textů, zejména v případech podezření z plagiátorství. Mezi významné příklady patří pokusy o identifikaci autorů her Williama Shakespeara. Forenzní lingvistika se však začala vědecky ukotvovat až koncem 70. let 20. století v Německu a v anglosaských zemích. Odborný termín "forenzní lingvistika" pak byl poprvé použit v roce 1968 profesorem lingvistiky Janem Svartvikem[2] při analýze čtyř výpovědí Timothyho Evanse. Podezřelý Evans se ve svých výpovědích přiznal k vraždě své manželky a dcery, a byl za to následně popraven. Svartvikova analýza spolu s dalšími důkazy však prokázala, že Evans nemohl být autorem přiznání, která mu byla připisována. Součástí odborné práce profesora Svartvika byla kvalitativní a kvantitativní jazyková analýza zfalšovaného výslechu. Jeho zjištění pak představují jeden z průlomových momentů v historii forenzní lingvistiky.

Od 90. let 20. století se forenzní lingvistika stala celosvětově uznávanou metodou a začala se vyvíjet rychleji. V roce 1993 byla založena Mezinárodní asociace forenzní lingvistiky (IAFL) a také International Association for Forensic Phonetics (IAFP) zaměřující se na forenzní fonetiku. Tyto organizace pomáhají zlepšovat právní systémy skrze lepší porozumění mezi jazykem jako složitou a mnohovrstevnatou entitou a zákonem jehož cílem je jednoznačnost a přesnost. Kromě toho propagují používání jazyka jako důkazu v občanskoprávních sporech a šíří znalosti o jazykové analýze. Také shromažďují materiály, jako jsou přiznání a policejní výslechy, a zpřístupňují je v online korpusu k dalšímu vědeckému zkoumání.[3]

---

1 JURKA, Michal a FALTÝNEK, Dan. *Forenzní lingvistika*. Online. CzechEncy: Nový encyklopedický slovník češtiny. Brno, 2024. Dostupné z: https://www.czechency.org/slovnik/FORENZN%C3%8D%20LINGVISTIKA. [cit. 2024-05-09]. URL: https://www.czechency.org/slovnik/FORENZNÍ LINGVISTIKA (poslední přístup: 21. 4. 2024)

2 Svartvik, J. *The Evans Statements: A Case for Forensic Linguistics,* 1968.

3 *International AssociationofForensicLinguistics* [online]. [cit. 2024-04-21]. Dostupné z: https://www.iafl.com/about/

Mezi významné osobnosti zabývající se forenzní lingvistikou patří ve světovém měřítku Malcolm Coulthard, Jan Svartvik, John Olsson, Janet Cotterill a Hannes Kniffka. John Olsson založil v roce 1994 ve Velké Británii Institut forenzní lingvistiky[4] a v současnosti je možné získat magisterský titul v tomto oboru na třech evropských univerzitách – v Cardiffu, Birminghamu a Barceloně.

V české historii známe více než jedno století se vlekoucí spor u určování autorství dvou rukopisných skladeb Rukopisu Královédvorského a Zelenohorského jakoby náhodně objevených v letech 1816 a 1818. Tento případ zaujal nejen vědeckou komunitu, ale také politiky, včetně T. G. Masaryka, a českou veřejnost. Cílem bylo ověřit autentičnost těchto dokumentů a odhalit případné padělky. Vědecké zkoumání tohoto případu zahrnovalo rozsáhlou historickou analýzu, prováděnou například Gebauerem, lingvistické posouzení, identifikaci anachronismů, a dokonce i chemické analýzy inkoustu a papíru. Zájem o autentičnost těchto dokumentů nebyl omezen pouze na odborníky, ale rozšířil se i na širší veřejnost, která projevovala zájem z vlasteneckých důvodů.[5]

Co se týče forenzní lingvistiky jako takové, ta se u nás začala rozvíjet v 50. letech 20. století v souvislosti s písmoznalectvím. Jako samostatný expertizní obor s názvem jazyková expertiza psaných textů vznikl teprve roku 1967. Od 70. let jsou pak zkoumané komunikáty psané i mluvené. Jednou z předních odbornic tohoto oboru je bezesporu PhDr. Václava Musilová, soudní znalkyně specializující se na jazykovou expertizu a písmoznalectví. Její odbornost zahrnuje určování druhu, modelu a značky psacího stroje použitého k vyhotovení strojopisu, expertizu ručního písma, pravosti platidel a cenin a technickou expertizu písemností, tedy zjišťování pravosti nebo způsobu vyhotovení padělaných či pozměněných písemností.[6] Druhou významnou osobností je doc. PhDr. Alena Aigner, CSc., působící na Pedagogické fakultě Jihočeské univerzity v Českých Budějovicích. V osmdesátých letech minulého století se určováním autorství

[4] *FORENSIC LINGUISTICS INTELLIGENCE*. Online. Obituary: John Olsson. Dostupné z: https://www.thetext.co.uk/obituary. [cit. 2024-04-22].

[5] Česká společnost rukopisná. *www.rukopisy-rkz.cz* [online]. [cit. 2024-05-01]. Dostupné z: https://www.rukopisy-rkz.cz/rkz/histsporu/hist-sp.htm

[6] *Soudní znalci z oboru kriminalistiky*. Online. Praha, 2024. Dostupné z: https://www.grafickeexpertizy.com/. [cit. 2024-05-09].

literárních textů věnoval také doc. PhDr. Pavel Vašák, DrSc., literární teoretik a textolog. Další významnou osobností oboru je Mgr. et Mgr. Tatiana Tkačuková, Ph.D., která v současnosti působí na Masarykově univerzitě v Brně a zkoumá jazykový styk laiků a zástupců práva a dále mluvený soudní diskurz. Svůj článek o křížovém výslechu publikovala také v zahraničním sborníku editovaném Malcolmem Coulthardem.

Sumarizace

Tato kapitola měla za cíl seznámit čtenáře s pojmem forenzní lingvistika, včetně její definice a historického vývoje jak ve světě, tak i v českém prostředí. Představili jsme významné osobnosti této oblasti a jejich přínosy, a také jsme zmínili literaturu zabývající se touto disciplínou.

2. Kvantitativní lingvistika

Kvantitativní lingvistika je odvětví matematické lingvistiky, které aplikuje metody matematické statistiky, teorie pravděpodobnosti, teorie informace, teorie fraktálů, entropie a teorie chaosu pro zkoumání kvantitativních vlastností přirozeného jazyka Pro přehlednost a zřetelnost oblasti kvantitativní lingvistiky můžeme definovat oblast širší, tedy matematickou lingvistiku. Matematická lingvistika stojí na pomezí lingvistiky, matematiky a informatiky (včetně umělé inteligence) a je disciplínou zkoumající přirozený jazyk pomocí matematických metod. Tato disciplína se dělí na algebraickou a kvantitativní lingvistiku, o níž nám zde jde především. Algebraická lingvistika, někdy označovaná jako teorie matematických modelů jazyka, se často vyskytuje v generativní a transformační gramatice a za jejího zakladatele bývá označován Noam Chomsky. Jedna z částí algebraické lingvistiky se zabývá vytvářením řetězců symbolů, které odpovídají jazykovým řetězcům zastupujícím jazykové útvary. Kvantitativní lingvistika je naopak věda založená na empirickém testování hypotéz.

V rámci kvantitativní lingvistiky je jazyk chápán jako pravděpodobnostní systém, který je ovlivněn různými funkcemi a existuje v našem vědomí. Tento systém se neřídí absolutními zákony a každý jednotlivec ho používá trochu jinak. Jedním ze

zásadních pojmů je zde kvantum čili nutná složka jazykové skutečnosti. Záleží nejen na tom, zda je určitý prvek v jazyce přítomen, ale také na tom, jak je využit. Proto se kvantitativní analýza zaměřuje na zkoumání funkčního zatížení prvků v jazyce. Hlavním cílem kvantitativní lingvistiky je formulovat obecná pravidla jazyka, odvodit z těchto pravidel hypotézy a následně tyto hypotézy ověřovat pomocí empirických testů.

Kvantitativní lingvistika, jako disciplína s vlastním předmětem, metodami a úkoly, začala být zkoumána z kvantitativních hledisek ve významnější míře počátkem 20. století. Tento výzkum úzce souvisel s konstrukcí různých těsnopisných soustav a se snahou o pokrok v didaktické oblasti. Za hlavní milník v evoluci lingvistiky, zejména v kontextu s matematikou, můžeme na začátku 20. století považovat publikování Kurs obecné lingvistiky Ferdinanda de Saussura v roce 1916. Jeho teorie jazykového symbolu jako základního stavebního prvku jazyka přinesla vnímání jazyka jako systému, jehož komponenty jsou vzájemně propojeny specifickými vztahy. Tyto vztahy mají být zkoumány pomocí strukturní komparativní analýzy v jedné časové rovině. Jazyk jako systém symbolů sloužících k přenosu informací se začal stávat předmětem studia strukturalistů, označovaných dále jako Ženevská, Kodaňská a Pražská škola.

Filozofické základy české kvantitativní lingvistiky byly položeny již Vilémem Mathesiem, který ve své práci O potenciálnosti jevů jazykových (1911) představil strukturální pohled na jazyk. Kvantitativní lingvistika jako samostatná disciplína se pak začala rozvíjet v rámci funkčního strukturalismu především ve spojitosti s Pražskou školou (přičemž tento termín se v roce 1931 upevnil na mezinárodních odborných fórech jako označení pro Pražský lingvistický kroužek). Členové Kroužku (Bohuslav Havránek, Roman Jakobson, Jan Rypka, Bohumil Trnka aj.) také postupně formulovali program synchronní lingvistiky a rozpracovali de Saussurovy koncepty. Tyto koncepty zahrnovaly termíny langue a parole, signifié a signifiant a jejich arbitrárnost v opozici k ikoničnosti, stejně jako koncepty centra a periferie, synchronie a diachronie či syntagma a paradigma.

V průběhu 60.–80. let vznikl rozsáhlý soubor kvantitativních charakteristik současné spisovné češtiny v její variantě psané i mluvené, který se opíral o původní, speciálně pro daný účel sestavený počítačově zpracovaný textový korpus o rozsahu 540 000 výskytů slov. Tento korpus byl později modernizován a využívá se v novém formátu a anotačních schématech v rámci české korpusové lingvistiky pod názvem Český akademický korpus.

Za jednu z významných postav kvantitativní lingvistiky poloviny 20. století je v českém kontextu bezesporu považována Marie Těšitelová, která se poprvé setkala s kvantitativní lingvistikou ve Výzkumném ústavu pedagogickém. V té době provedla detailní analýzu knihy Karla Čapka „Život a dílo skladatele Foltýna“ a své závěry publikovala v roce 1948 v periodiku Naše řeč pod názvem Frekvence slov a tvarů ve spise „Život a dílo skladatele Foltýna“ od Karla Čapka. Když v roce 1956 přešla do Ústavu pro jazyk český, pracovala v různých odděleních a kvantitativní lingvistice se věnovala spíše jako svému koníčku. Až v roce 1965 se přesunula do oddělení matematické lingvistiky a o dva roky později se stala jeho vedoucí. Její práce je charakterizována širokým záběrem a kritickým přístupem, který prezentuje také ve své bibliografie Kvantitativní lingvistika, publikované v letech 1965 až 1972.

Od počátku 90. let 20. století můžeme sledovat odklon od ryze empirického zaměření kvantitativní lingvistiky směrem k teoretickému. Tuto tendenci můžeme pozorovat například u autorů publikujících v časopisu Journal of Quantitative Linguistics, který je oficiálním fórem Mezinárodní asociace pro kvantitativní lingvistiku.

Sumarizace

Tato kapitola měla za cíl seznámit čtenáře s pojmem kvantitativní lingvistika, včetně její definice a historického vývoje jak ve světě, tak i v českém prostředí. Představili jsme významné osobnosti této oblasti a jejich přínosy, a také jsme zmínili literaturu zabývající se touto disciplínou.

2.1. Korpus textů

Podle kriminalisty Jiřího Strause lze texty rozdělit do čtyř kategorií podle jejich délky. V jeho knize "Kriminalistická technika" z roku 2012 jsou tyto kategorie definovány jako texty velmi krátké (s méně než 170 slovy), texty krátké (s 170–380 slovy), texty dlouhé (s 380–750 slovy) a texty velmi dlouhé obsahující více než 750 slov.[7]

Pokud jde o analýzu textů, kde délka nepřesahuje 500 slov, je vhodné provést kvalitativní analýzu, protože kvantitativní (statistická) analýza by mohla poskytnout nepřesné výsledky. Naopak u delších textů je možné provést kvantitativní analýzu.

Bude provedena analýza dvaceti textů, které pocházejí od čtyř různých autorů, přičemž každý autor přispěl pěti texty. Každý text má délku přibližně 5 tisíc znaků. V odborných textech se délka slov vyjadřovaná v počtu znaků pohybuje od 5,2626 (v textech mluvených) až po 5,8593 (v textech psaných), s průměrnou hodnotou přesahující 5,5.[8] Údaje o průměrné délce internetového komentáře nejsou známé, ale pomocí lingvistického softwaru QUITA, který bude představen v následující kapitole, jsme schopni získat informace o průměrné délce tokenů v jednotlivých textech. Průměrná délka tokenu u autora A je přibližně 5,0355, autora B je přibližně 4,8144, u autora C je přibližně 4,7099 a u autora D je přibližně 4,3670. Každý text, který používáme k analýze má přibližně tisíc slov.

Tyto texty jsou složeny z komentářů různých délek, které autoři napsali pod články na webové stránce, která funguje na principu, že kdokoliv může napsat článek na jakékoliv téma, které ho zajímá, jakkoliv je toto téma kontroverzní, například *" Turecko, nástroj pro vyvolání III. světové?", " Tyto volby jsou poslední,kdy můžeme zvolit normálního člověka.Pak už budou jen pravdoláskaři?", " Černá barva kůže – indikátor inteligence?"* a *"A kdo se postará o ty,co pečují?."*

---

[7] STRAUS, Jiří. Kriminalistická technika. 3., rozš. vyd. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012. ISBN 978-80-7380-409-1.

[8] KRÁLÍK, Jan. Statistika českých grafémů s využitím moderní výpočetní techniky. Online. *Slovo a slovesnost*. 1983, roč. 44, č. 4, s. 295-304. Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=2913. [cit. 2024-05-09].

Cílem těchto článků je podnítit diskusi pod článkem, kde mohou všichni čtenáři svobodně vyjádřit své názory na dané téma. Každý z těchto textů představuje svébytný pohled a přístup k různorodým tématům, které zahrnují kontroverzní otázky a diskutovaná témata současné společnosti.

Každý z autorů, kteří přispěli těmito texty, má svůj vlastní styl psaní a způsob vyjadřování názorů. Tato rozmanitost přispívá k bohatosti analyzovaných textů. Tímto přístupem se snažíme přispět k hlubšímu poznání internetového prostředí jako prostoru veřejné diskuse a výměny názorů. Veškeré texty byly přijaty i s gramatickými, stylistickými a pravopisnými chybami, které autor udělal.

Zvolený korpus textů má několik výhod, z nichž jednou je skutečnost, že se jedná o reálné osoby a jejich autentické diskuse na internetu. Komentáře na internetu jsou většinou krátké, psané rychle, autoři často používají slang, emotikony, můžou obsahovat překlepy nebo chyby v interpunkci. Internetové komentáře bývají přímým dílem autora, protože na rozdíl od jiných literárních děl jsou needitované, neprošly korekturou provedenou další osobou.

Zmíněná analýza internetových komentářů může být zajímavou zejména pro Policii české republiky nebo kriminalisty, protože nenávistné projevy na internetu, tzv. hate speech, můžou být kvalifikovány jako přestupek, nebo i trestný čin. "Kdo veřejně podněcuje k nenávisti k některému národu, rase, etnické skupině, náboženství, třídě nebo jiné skupině osob nebo k omezování práv a svobod jejich příslušníků, bude potrestán odnětím svobody až na dvě léta."[9]

Nenávistné projevy nejsou fenoménem posledních let, existuje několik známých manifestů napsaných zločinci, které získaly širší pozornost veřejnosti a médií. Zde jsou některé příklady:

Manifest Theodora Kaczynskiho: Ted Kaczynski byl americký terorista, který byl znám jako "Unabomber". V letech 1978 až 1995 posílal bomby na různá místa v USA a způsobil několik smrtelných útoků. Před svými činy napsal rozsáhlý

[9] ČESKO, 2009. Zákon 40/2009 Sb. (trestní zákoník). Online. In: Sbírka zákonů. Částka 11. Dostupné také z: https://www.e-sbirka.cz/sb/2009/40?zalozka=text.

manifest nazvaný "Industrial Society and Its Future", ve kterém kritizoval moderní technologickou společnost a propagovat přírodní životní styl.[10]

Manifest Elliota Rodgera: Elliot Rodger byl pachatel masové střelby v Isla Vista v roce 2014, při které zabil šest lidí a zranil dalších 14 před tím, než spáchal sebevraždu. Před útokem napsal a zveřejnil video a manifest nazvaný "My Twisted World", ve kterém se zaměřoval na své pocity osamělosti, frustrace a nenávisti vůči ženám.[11]

Manifest Christopha Dornera: Christopher Dorner byl bývalý policista v Los Angeles, který se stal známým svou sérií vražd a útoků na policisty v roce 2013. Před svými činy zveřejnil na internetu dlouhý manifest, ve kterém popsal své pocity nespravedlnosti a obvinění vůči policii a systému.[12]

Manifest Andrese Breivika: Anders Behring Breivik byl pachatelem masového vraždění v Norsku v roce 2011, kdy zabil 77 lidí. Před útokem napsal 1500stránkový manifest s názvem "2083: A European Declaration of Independence", ve kterém vyjádřil své extrémní nacionalistické a protiislámské názory.[13]

Je patrné, že rozsah textů poskytovaných odsouzenými zločinci je často výrazně rozsáhlejší než krátké komentáře, které nalezneme na sociálních sítích či v diskusích pod články na internetu. Tento fakt představuje významnou výzvu v oblasti analýzy textů pro účely identifikace autorství, protože rozsáhlejší textové korpusy nabízejí bohatší a komplexnější soubor lingvistických a statistických informací, které je možné využít k identifikaci jednotlivých autorů. Naopak, krátké komentáře často poskytují pouze omezený kontext a mnohem menší množství dostupných dat, což může ztížit úspěšnou identifikaci autorství. Z tohoto důvodu je důležité přistupovat k analýze krátkých textů s vhodnými metodami

---

10 *Federal bureau of investigation: The Unabomber. Online. In: FBI Bureau of Investigation. Dostupné z: https://www.fbi.gov/history/famous-cases/unabomber. [cit. 2024-05-02].*

11 *Elliot Rodger: How misogynist killer became 'incel hero'*. Online. In: BBC News Services. 2018. Dostupné z: https://www.bbc.com/news/world-us-canada-43892189. [cit. 2024-05-02].

12 *KELLY, John. Christopher Dorner: What made a police officer kill? Online. BBC News. 2013. Dostupné z: https://www.bbc.com/news/magazine-21476904. [cit. 2024-05-09].*

13 *The Anti-Islamist: Anders Behring Breivik's Manifesto. Online. In: The International Centre for Counter-Terrorism. 2012. Dostupné z: https://www.icct.nl/publication/anti-islamist-anders-behring-breiviks-manifesto. [cit. 2024-05-02].*

a technikami, které dokážou efektivně zpracovat omezený obsah a získat relevantní informace pro identifikaci autorů.

Zaměřujeme se na otázku, zda je možné využít i tak krátké texty, jako jsou komentáře na internetu, k identifikaci autorství. Tato otázka má klíčový význam zejména v oblasti forenzní lingvistiky a strojového učení, kde se snažíme využít různé lingvistické a statistické metody k identifikaci jednotlivých autorů, například za účelem zlepšení a zjednodušení detekce podvodů. Analýza krátkých textů, jako jsou internetové komentáře, představuje výzvu z hlediska omezeného kontextu a dostupných informací. Nicméně, pokrok v oblasti analýzy jazyka a strojového učení dává naději na vytvoření postupů, které by mohly být schopné identifikovat charakteristické rysy jednotlivých autorů i v krátkých textech. Tímto způsobem lze potenciálně rozšířit aplikace forenzní lingvistiky na nové oblasti a přispět k větší efektivitě identifikace autorství textů v digitálním prostředí.

2.2. Použitý software

Lingvistický software se používá k různým účelům v oblasti lingvistiky a jazykovědné analýzy. Mezi nejčastější aplikace patří analýza textu, strojový překlad, syntaktická a sémantická analýza, extrakce informací a další. Mezi nejznámější lingvistické softwary patří například Natural Language Toolkit (NLTK), Stanford NLP, spaCy nebo GATE (General Architecture for Text Engineering).

Pro naši analýzu využijeme lingvistický software QUITA (Quantitative Index Text Analyser)[14]

2.3. Logistická regrese

Druhým nástrojem, který využijeme pro analýzu dat získaných pomocí QUITA, je program pro analýzu pomocí logistické regrese. Tento program je k dispozici

[14] *QUITA Online*. Online. Dostupné z: https://kol.ff.upol.cz/quita/. [cit. 2024-05-09].
[14] *Logistická Regrese*. Online. Dostupné z: http://kol-apps.ff.upol.cz/log-reg/. [cit. 2024-05-09].

na webových stránkách katedry obecné lingvistiky Univerzity Palackého v Olomouci.[15]

Hlavním cílem logistické regrese je predikce pravděpodobnosti, že určitá událost nastane, na základě hodnot nezávislých proměnných. Výstupem logistické regrese je logaritmus šance pravděpodobnosti vzniku události. Tento logaritmus se poté transformuje zpět na pravděpodobnostní škálu pomocí logistické funkce.

P hodnota nám říká, jestli máme dostatek důkazů, abychom mohli rozhodnout, jestli je výsledek statisticky signifikantní. Aby byl model vyhodnocen jako statisticky signifikantní, nesmí p-hodnota přesáhnout 0,05.

Práh alfa 0,05, zvolený bez korekcí, představuje jednu z běžných praxí v statistické analýze. Nicméně, při opakovaných testech se doporučuje zvážit použití Bonferroniho korekce, aby se minimalizovalo riziko falešně pozitivních výsledků. Tato korekce zamítá nulovou hypotézu, když je její p-hodnota nižší nebo rovna hodnotě α/m, kde α představuje stanovenou hladinu významnosti testu (obvykle 0,05 nebo 0,01), a m je počet současně provedených testů.[16] Práh je stanoven tak striktně, že můžeme ztratit možnost účinného fungování některých metod.

V případě, že použijeme korekci, musí být p hodnota menší než 0,05/48, kdy 48 je počet provedených pokusů. P hodnota tedy musí být nižší než 0,001,

Vědecký tým zabývající se testováním metod pro identifikaci autorství v beletrii má výhodu v tom, že disponuje rozsáhlejšími textovými korpusy. Tímto způsobem mají možnost provést důkladnější a komplexnější analýzy, které umožňují

---

16 *P-hodnota a její interpretace*. Online. Institut biostatistiky a analýz lékařské fakulty masarykovy univerzity. Portal.matematickabiologie.cz. 2024. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--uvod-do-testovani-hypotez--p-hodnota-a-jeji-interpretace. [cit. 2024-05-06].

přesněji posoudit účinnost zkoumaných metod. Tím, že se zaměřují na delší texty, mohou lépe zhodnotit, zda daná metoda skutečně funguje v praxi.

Naopak, v našem případě, kdy analyzujeme krátké texty, jako jsou komentáře na internetu, se setkáváme s trojí nejistotou ohledně neúčinnosti zkoumaných metod.

Existuje nejistota, zda metoda nefunguje z důvodu její neúčinnosti, zda neúčinnost souvisí s příliš krátkými texty, kde může být omezená dostupnost lingvistických znaků a statistických informací, nebo zda obě tyto faktory hrají roli současně, což způsobuje neúspěch metody ve spojení s krátkými texty.

Použití korekce, jako je například Bonferroniho korekce, může v takových případech vést k příliš striktním kritériím, která mohou omezit schopnost identifikovat účinné metody. To může mít za následek ztrátu potenciálně užitečných metod, které by mohly fungovat i v kontextu krátkých textů.

Sumarizace

Cílem této kapitoly bylo představit čtenářům různé typy textů, které jsou vhodné pro analýzu, a navrhnout, které jsou vhodné pro kvantitativní nebo kvalitativní analýzu. Dále představuje korpus textů, se kterým budeme pracovat a vysvětlit, jaké jsou výhody a nevýhody používání internetových komentářů k lingvistické analýze. Dále seznamuje čtenáře s pojmem lingvistický software a představuje nejznámější typy.

Poslední část této kapitoly představuje čtenářům metodu logistické regrese, vysvětluje její fungování a použití pro lingvistický výzkum.

P r a k t i c k á č á s t

3. Kvantitativní metody určování autorství

Metody identifikace autorství se často ověřují na literárních dílech, především na románech. Romány poskytují rozsáhlý textový korpus, který umožňuje hlubší analýzu stylu psaní a jazykových charakteristik. Důležitým faktorem při využití románů je to, že prochází profesionální editací před vydáním. Tento proces zahrnuje korekci a úpravu textu ze strany odborníka s cílem odstranit případné chyby, které by mohly ovlivnit konečnou podobu díla.

Při analýze krátkých textů, jako jsou internetové komentáře, je množství dostupných materiálů omezenější. Tím pádem není dostatek textu pro podrobnou stylistickou analýzu, jakou lze provést na románech či jiných delších textů v jiných funkčních stylech, například novinách. Nicméně, i na krátkých textech mohou být klíčové prvky, které napovídají o autorství. Mezi tyto prvky patří například charakteristické chyby, použití interpunkce, délka vět, frekvence emotikon, konzistence v používání velkých a malých písmen a další jazykové charakteristiky. Tyto znaky mohou být využity k identifikaci a analýze autorství, ačkoliv pravděpodobně s menší přesností a jistotou než u rozsáhlejších textových korpusů, jako jsou romány.

Cílem praktické části bude vyzkoušet, zda známé metody určování autorství dokáží rozlišit autorství i u kratších textů a stanovit jejich míru přesnosti.

4. Vstupní data pro analýzu

Jak jsme zmínili v předchozích kapitolách, pro náš výzkum využijeme dvacet textů čtyř autorů. Každému autorovi náleží 5 textů. Všechny texty jsou sestaveny z krátkých, přirozeně stylizovaných komentářů, které jsou pospojovány tak, aby vytvářely souvislý text. Aby bylo možné s daty dále pracovat, bylo potřeba je převést do formátu, který je vhodný pro využití softwaru QUITA. Textové soubory jsou formátovány v rozhraní .txt a využívají kódování do UTF-8. Toto kódování umožňuje reprezentovat širokou škálu znaků a symbolů z různých

jazyků. Použití kódování UTF-8 je vhodné pro svou univerzální kompatibilitu a schopnost zachytit různorodost jazykových znaků.

Tyto texty jsou dostatečně rozsáhlé a zaměřují se na podobná témata, jelikož je autoři napsali zhruba ve stejném časovém rámci pod stejné články na stejné webové stránce. Existuje možnost rozdělení textů na individuální komentáře a následné rozdělení do většího množství krátkých souborů. Nicméně délka těchto komentářů je různá, což může představovat problém pro další kvantitativní analýzy. S cílem minimalizovat tento vliv na výsledky byly komentáře rozděleny do souborů přibližně stejné délky, což vedlo k vytvoření pěti souborů pro každého autora. Tento přístup umožňuje aplikovat kvantitativní metody s větší přesností a konzistencí, což usnadňuje komparativní analýzy mezi různými autory a jejich texty.

Aby byla zachována anonymita autorů, rozhodli jsme se nepoužít přezdívky, které používají na webu svice.cz, ale použili jsme pro jednotlivé autory označení A, B, C, D. Texty tedy nesou označení A1-A5, B1-B5, C1-C5 a D1-D5.

Abychom mohli pokračovat v analýze, je třeba analyzované texty pomocí QUITA převést na tokeny.

4.1. Token

Token je nejmenší jednotka textu, často jedno slovo nebo jeho varianta zobrazená graficky. V oblasti korpusové lingvistiky může být jedno slovo rozděleno na více tokenů (například "mohu-li"), a interpunkce je často oddělena od sousedních slov pro snazší vyhledávání (například "řekl", "že"). Jednotlivé tokeny v korpusu jsou také označovány jako pozice. Velikost korpusu se obvykle udává v textových slovech. Proces rozdělení textu na tokeny se nazývá tokenizace a je výsledkem tohoto procesu. Tento proces provádí zařízení známé jako tokenizér.[17]

[17] CVRČEK, Václav. *Token*. Online. CzechEncy: Nový encyklopedický slovník češtiny. Brno, 2024. Dostupné z: https://www.czechency.org/slovnik/TOKEN. [cit. 2024-05-09].

5. Hodnotící kritéria

Jako hodnotící kritéria jsme si zvolili tři znaky jazykového vyjadřování, které bývají typické pro autory internetových komentářů. Tato kritéria jsou používání emotikonů, výskyt dvou a více teček za větou a používání CapsLocku nebo jiného zvýraznění slov. Tato tři kritéria patřila mezi třináct zkoumaných kritérií v naší bakalářské práci, která se zabývala hlavně kvalitativní analýzou anonymních textů.[18] Mezi další kritéria patřil například rasismus a projevy nenávisti vůči minoritním skupinám, konfliktnost, nespisovnost nebo množství gramatických chyb. V následující části si detailně představíme každé kritérium, které budeme zkoumat v této diplomové práci.

Emotikony:

Toto kritérium posuzuje míru používání emotikonů anonymními autory textů. Jde o symbolické vyjádření emocí, které jsou často přítomné v internetových diskusích a komentářích. Emotikony jako :-) nebo :-(jsou běžným prostředkem, jak sdělit své pocity či náladu v textu. V našem výzkumu nepřidělujeme specifický význam různým typům emotikonů; zahrnujeme všechny, bez ohledu na jejich výraz, včetně usmívajících se i mračících se emotikonů s nosem i bez něj. Nicméně při podrobnějším zkoumání jsme zjistili, že samotné používání emotikonů je pro jednotlivé autory velmi charakteristické. Například pokud někdo používá emotikony s nosem jako :-), pravděpodobně nebude používat emotikony bez nosu jako :)

Příklad od autora C1:

To je pravda, trendy jsou dnes jiné. Inovované školství, nové typy škol, kde se neznámkuje, děti nesedí v lavicích a členění času je taky volné. Vše musí děti bavit, říká se tomu kreativní výuka. Jenomže škola je práce a pro někoho i dřina. Vzpomínám na popis výuky na prestižní japonské obchodní škole. Je to současná škola. Student nedává pozor, učitel ho přetáhne rákoskou přes záda, žák se postaví a po-

[18] TICHÁ, Anna. *Identifikace autora ve forenzní lingvistice*. Online, bakalářská, vedoucí Mgr. Lukáš Zámečník, Ph.D. Olomouc: Univerzita Palackého Olomouc, 2019. Dostupné z: https://theses.cz/id/fumpwq/33270788. [cit. 2024-05-09].

děkuje. Jiný kraj, jiný mrav. Efektivnost takové školy je ovšem předem vyzkoušená..... Nějak si neumím představit, že kreativitou se dá tento model překonat :-)

Počet teček za větou:

Používání více teček za větou je součástí interpunkce, ale je to tak výrazný jev, že je vhodné ho zvýraznit samostatně. Tento jev je typický pro internetovou komunikaci a neformálních prostředích, jako jsou sociální sítě nebo chatovací aplikace, může být použití více teček za větou běžné. Mnoho lidí vnímá tento styl jako přirozený a nepříliš formální způsob vyjadřování. Časté používání více teček může naznačovat váhání, dramatický efekt nebo neúplně dokončenou myšlenku. Čtenář tak může být podnícen k doplnění či interpretaci nevyřčené části informace. Toto kritérium posuzuje, jak často anonymní autoři textů používají za větou více teček než jednu.

Příklad od autora A2:

Dokonale s Vámi souhlasím, pane Ivane. Dokonce i V.Č. si myslí, že dominujícím jazykem nebude čeština. Jenom bych si to přál.

A s tou morálkou to jste trefil bezvadně... Souhlas, souhlas, souhlas....

CapsLock nebo jiné zdůraznění slova:

V internetové komunikaci je běžné využívání funkce CapsLock k zvýraznění obsahu sdělení nebo jeho části, často je interpretováno jako ekvivalent hlasitého projevu. Lidé mohou chtít, aby jejich sdělení bylo více nápadné nebo výrazné, a proto používají velká písmena. Dalším způsobem, jak diskutéři zvýrazňují nebo zdůrazňují slova, je vkládání mezer mezi každý znak zdůrazněného slova.

Příklad od autora C4, kde můžeme sledovat oba způsoby zvýraznění slova. Tento způsob projevu je pro autora C typický:

Manželka rodinného přítele se n a d ý c h a l a při úklidu přípravku Domestos a měla z toho život ohrožující stav. Máme to všude kolem sebe. Pro mě to je pokyn ničeho se nebát. Vzorem je mi Vaše statečnost.Nepospíchejte na další operaci,ať si to mozek a rozmrznutý statečný oko navzájem mají čas zkonzultovat. Je to moc

podnětů a ZVRATŮ NAJEDNOU.Skoro to ani nemá lidskou dimenzi. No - mohlo by se to před sto lety přihodit? A to je vlastně před chvilkou..... Naviděnou!

Pro hodnocení anonymních textů jsme vytvořili Tabulku 1. V levém sloupci jsou seřazeny námi hodnocené texty a v horním řádku jsou vypsána jednotlivá kritéria, která u textů hodnotíme.Pro náš výzkum nejdříve využijeme plnou délku textů, které máme k dispozici. Díky tomu získáme lepší představu o jednotlivých autorech a jejich stylu. V Tabulce 1 můžeme vidět, že nejkratší text má 618 tokenů a nejdelší text má 1600 tokenů čili je více než dvakrát delší.

	Počet tokenů	Emotikony	Tečky za větou	Zvýraznění textu
A 1	825	5	21	2
A 2	752	1	20	1
A 3	885	2	39	1
A 4	618	3	31	2
A 5	1031	2	26	2
B1	618	1	0	0
B 2	708	0	2	0
B 3	620	0	0	0
B 4	788	0	1	0
B 5	622	0	1	0
C1	1443	10	21	0
C 2	1310	13	29	0
C 3	1539	7	26	0
C 4	1600	2	15	2
C 5	1448	8	17	2
D 1	1151	0	0	0
D 2	1368	0	1	0
D 3	1137	0	0	0
D 4	1226	0	2	0
D 5	1429	0	0	0

Tabulka 1: *Výskyt sledovaných jevů v textech autorů A, B, C a D a délka jednotlivých textů v tokenech.*

Z výsledků v Tabulce 1 je patrné, že je možné vizuálně odlišit čtyři autory. Znamená to, že se v používání jednotlivých znaků liší. I přesto, že jsou texty

různě dlouhé, můžeme sledovat, že délka textu nekoresponduje s počtem námi sledovaných jevů. Například autor A má texty téměř nejkratší ze všech, ale objevují se v nich všechny tři námi sledované jevy, a to navíc v nejvyšší míře ze všech. Autor B má také krátké texty, ale sledované jevy se v nich nevyskytují téměř vůbec. Autor C má texty dlouhé a objevují se v nich hojně všechny tři jevy, autor D má texty také dlouhé, a sledované jevy se v nich objevily pouze minimálně. Abychom zjistili, zda výsledky korespondují i pokud texty zkrátíme na stejnou délku, nejdříve vydělíme jednotlivé jevy počtem tokenů, které máme k dispozici. Výsledky zapíšeme do Tabulky 2.

	Počet tokenů	Emotikony/tokeny	Tečky za větou/tokeny	Zvýraznění textu/tokeny
A 1	825	0,0061	0,0254	0,0024
A 2	752	0,0014	0,0266	0,0014
A 3	885	0,0023	0,044	0,0011
A 4	618	0,0049	0,0631	0,0032
A 5	1031	0,0019	0,0252	0,0019
B1	618	0,0016	0	0
B 2	708	0	0,0028	0
B 3	620	0	0	0
B 4	788	0	0,0012	0
B 5	622	0	0,0016	0
C1	1443	0,0087	0,0145	0
C 2	1310	0,0099	0,0221	0
C 3	1539	0,0045	0,0169	0
C 4	1600	0,0013	0,0094	0,0013
C 5	1448	0,0055	0,0117	0,0013
D 1	1151	0	0	0
D 2	1368	0	0,0007	0
D 3	1137	0	0	0
D 4	1226	0	0,0016	0
D 5	1429	0	0	0

Tabulka 2: *Výskyt sledovaných jevů v textech autorů A, B, C a D přepočítáno na počet tokenů jednotlivých textů.*

Získané výsledky vynásobíme číslem 618, což je počet tokenů nejkratšího textu, který máme k dispozici. Tím získáme údaj, kolikrát se průměrně objeví námi sledovaný jev v textu dlouhém 618 tokenů.

	Emotikony	Tečky za větou	Zvýraznění textu
A 1	3,77	15,69	1,48
A 2	0,87	16,44	0,87
A 3	1,4	27,2	0,7
A 4	3	39	2
A 5	1,2	15,58	1,2
B1	1	0	0
B 2	0	1,75	0
B 3	0	0	0
B 4	0	0,79	0
B 5	0	0,99	0
C1	4,28	8,99	0
C 2	6,13	13,68	0
C 3	2,81	10,44	0
C 4	0,77	5,79	0,77
C 5	3,41	7,26	0,85
D 1	0	0	0
D 2	0	0,45	0
D 3	0	0	0
D 4	0	0,78	0
D 5	0	0	0

Tabulka 3: *Výskyt sledovaných jevů v textech autorů A, B, C a D přepočítáno na 618 tokenů pro každý text.*

Můžeme sledovat, že výsledky z Tabulky 3 korespondují s výsledky Tabulky 1. Stále je velice dobře možné vizuálně odlišit čtyři autory, ale je lépe pochopitelná pro čtenáře, protože díky přepočtu z desetinných čísel je lépe čitelná.

Autor C se ukázal být nejvíce emocionálním a expresivním, ve svých textech použil celkem nejvíce emotikonů. Tento častý výskyt emotikonů může naznačovat autorovu snahu vyjádřit své pocity a nálady v textu.

O něco méně emotikonů použil autor A, což svědčí o mírnější, ale stále výrazné tendenci k emocionálnímu vyjadřování. Oba autoři použili emotikony v každém ze svých textů.

Ukázka z textu C3:

Celé tělo je zázračné.Taky tím,že spolu jednotlivé orgány spolupracují. Céčko,rutin,pohanka,klid a nohy v teple!:-):-)

Naopak autor B použil emotikony pouze ve třech textech, což naznačuje jeho nečastější používání tradičnějšího a formálnějšího stylu bez tohoto typu symbolických výrazů.

Autor D je naopak charakterizován absencí emotikonů ve svých textech, což může být interpretováno jako jeho preference k jasnější a striktnější formě komunikace bez použití nonverbálních prvků.

Dalším z velice úspěšných kritérií hodnocení je použití více než jedné tečky za větou. V textech, které jsme zkoumali se objevily varianty dvou teček (..), trojice teček (…) nebo dokonce čtveřice teček (….) na konci věty.

Autoři A a C mají použití více teček za větou jako jeden z charakteristických rysů jejich stylu. U autora A je používání více teček velmi výrazné, použil tuto formu závěrečných teček ve všech svých textech několikrát. Autor C má také ve zvyku používat větší počet teček za větou v každém ze svých pěti textů a množstvím se blíží autorovi A. Autoři B a D rovněž použili více teček za větou, avšak v mnohem menší míře. Autor B tuto formu použil celkem ve třech textech z pěti, zatímco autor D ve dvou textech z pěti.

Jako příklad uvádím úryvek z textu A3, který měl nejvyšší počet výskytů teček za větou, můžeme sledovat, že je to opravdu charakteristický prvek pro autora A:

Koňové mají krásné oči....
mimochodem, taky pštrosové...
A prý se na pštrosech dá i jezdit... ???? .... neviděl jsem, slyšel jsem... obhajovat to nebudu...

V uvedeném úryvku vidíme, že i počet otazníků za větou se nabízí jako další možný charakteristický jev u autorů internetových komentářů.

Třetím z námi hodnocených kritérií je zvýraznění textu buď pomocí CapsLock, nebo pomocí vložení mezer mezi každý znak zvýrazňovaného slova.

Jako příklad uvádíme úryvek z textu C 5:

C5 Tento lektvar je z a r u č e n ě léčivý.:-)Přeju výdrž a věřím v ten nejlepší konec. Bezinky jsou ještě zelený,ale až dozrajou...

Z tabulky je patrné, že zvýraznění textu používají pouze dva autoři ze čtyř. Pro autora A je zvýrazňování slov velmi typické, použil ho ve všech pěti textech. Autor C použili zvýraznění ve dvou ze svých pěti textů, autoři B a D ani jednou. Pro případné rozšíření výzkumu by vhodné mít více autorů, protože není jasné, jestli to, že zvýraznění používají dva ze čtyř autorů, znamená, že je to mezi lidmi píšící internetové komentáře běžné nebo naopak unikátní.

5.1. Statistická analýza

V následující části se pokusíme výsledky statisticky zpracovat pomocí logistické regrese a srovnat. Budeme analyzovat data zaznamenaná v Tabulce 4. Tato data budou vložena do levého sloupce logistické regrese, určeného pro nezávislé proměnné, tj. prediktory. V našem případě budeme zkoumat naměřené hodnoty námi zvolených kritérií, tedy emotikon, teček za větou a zvýraznění textu. Tyto hodnoty budou graficky znázorněny na ose X. Do pravého sloupce logistické regrese vložíme závislé proměnné, které nabývají hodnot 0 nebo 1, odpovídající výskytu sledovaného jevu, v našem případě autora textu. Například autor A bude označován jako 0, autor B jako 1. Přidělení těchto hodnot je libovolné, ale pro náš případ vždy autor výše v abecedě dostane přidělenou nulu a autor níže v abecedě dostane přidělenou jedničku. Toto rozdělení je pouze pro lepší přehlednost a nemá vliv na výsledky.

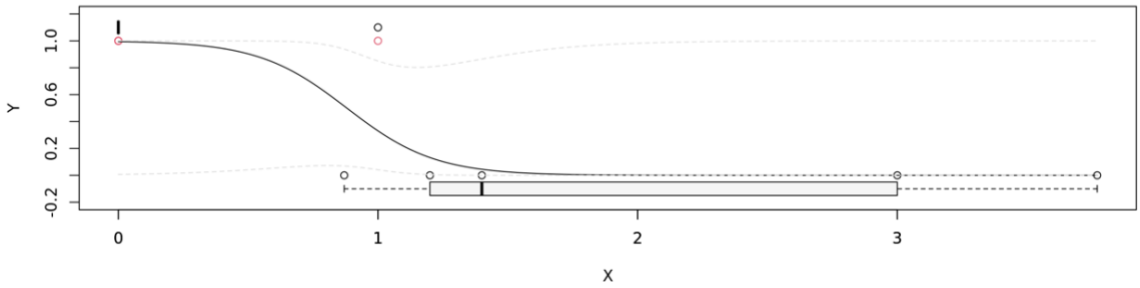
Dalším důležitým údajem pro naši analýzu je p-hodnota (p-value). V kontextu logistické regrese se p-hodnota používá k posouzení statistické významnosti

vztahu mezi nezávislými proměnnými (vstupními faktory) a závislou proměnnou (výstupem, kategorickou proměnnou). Abychom mohli model považovat za statisticky signifikantní, musí být p-hodnota nižší než 0,05.

Emotikony

V následující kapitole provedeme analýzu pomocí logistické regrese. Provedeme srovnání šesti párů textů tak, abychom srovnali každý text s každým. Výsledkem bude šest grafů.

Emotikony – srovnání autorů A a B



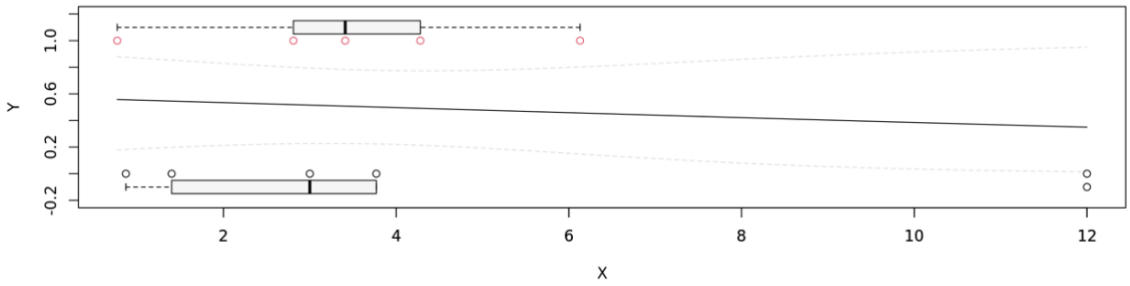
Obrázek 1: Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.001762

P-hodnota je u tohoto modelu 0.001762, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Emotikony – srovnání autorů A a C



Obrázek 2: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

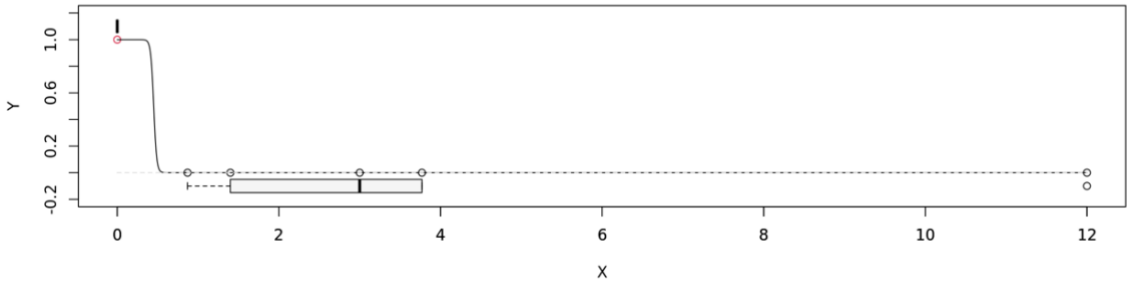
[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).

p-hodnota ≈ 0.711828

P-hodnota je u tohoto modelu 0.711828, což znamená, že tento model není funkční, a proto není statisticky významný.

Logistická regrese dokáže správně přiřadit '0' a '1' v 40 % případů.

Emotikony – srovnání autorů A a D



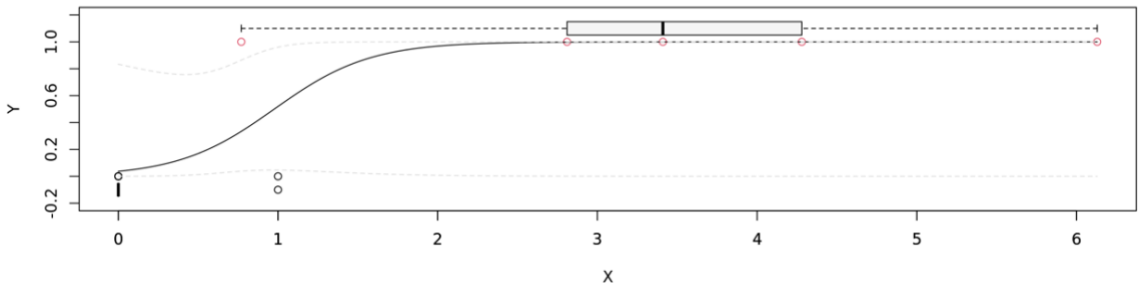
Obrázek 3: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Emotikony – srovnání autorů B a C



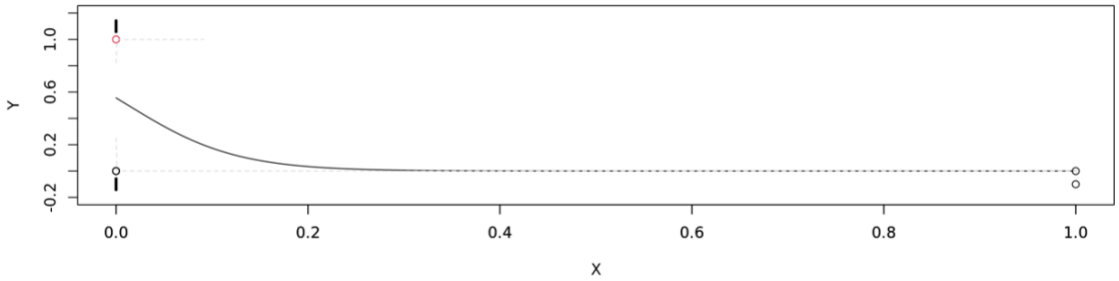
Obrázek 4: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.001655

P-hodnota je u tohoto modelu 0.001655, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Emotikony – srovnání autorů B a D



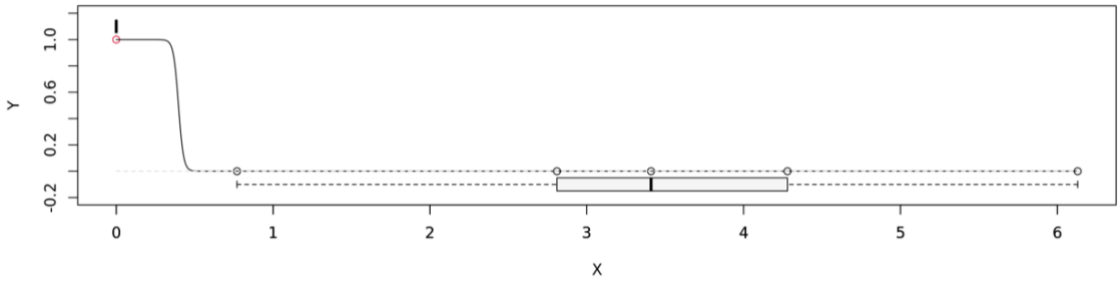
Obrázek 5: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).

p-hodnota ≈ 0.221036

P-hodnota je u tohoto modelu 0.221036, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.

Emotikony – srovnání autorů C a D



Obrázek 6: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých emotikon. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky velmi významný.

Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Hodnocení úspěšnosti metody

EMOTIKONY	A	B	C	D
A		ANO, 80 %	NE, 40 %	ANO, 100 %
B			ANO, 80 %	NE, 60 %
C				ANO, 100 %
D				

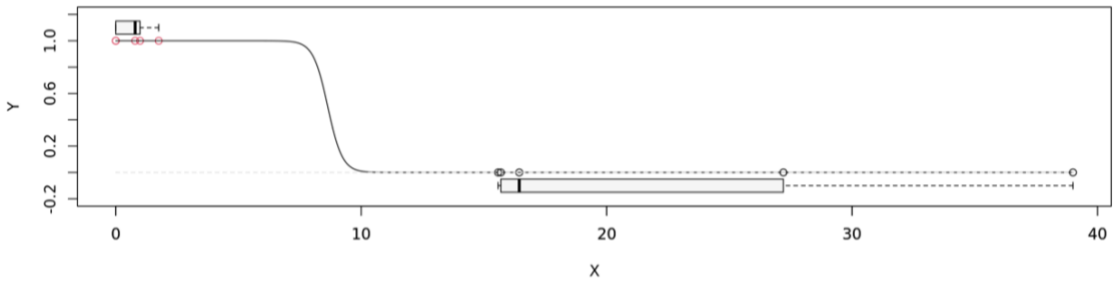
Tabulka 4: *Tabulka úspěšnosti jednotlivých pokusů při zkoumání použití emotikon v textu.*

Za účelem zlepšení přehlednosti pro čtenáře byla vytvořena Tabulka 4, která prezentuje, zda se podařilo vytvořit funkční, tj. statisticky významný model (v tabulce označeno jako ANO/NE) pomocí logistické regrese jeho úspěšnost v procentech. Z této tabulky vyplývá, že čtyři z šesti analyzovaných modelů byly funkční, tedy dosáhly statistické významnosti. Z toho vyplývá, že metoda byla úspěšná v přibližně 67 % pokusů. Dále v tabulce můžeme vidět, že průměrná úspěšnost při přiřazení je přibližně 77 %.

Tečky za větou

Znovu provedeme analýzu šesti párů textů pomocí logistické regrese, tentokrát budeme vkládat do levého sloupce hodnoty naměřené pro počet teček za větou. Výsledkem bude 6 grafů a tabulka úspěšnosti jednotlivých pokusů, kterou můžeme srovnat s úspěšností předchozí metody, množství použitých emotikon.

Tečky za větou – srovnání autorů A a B

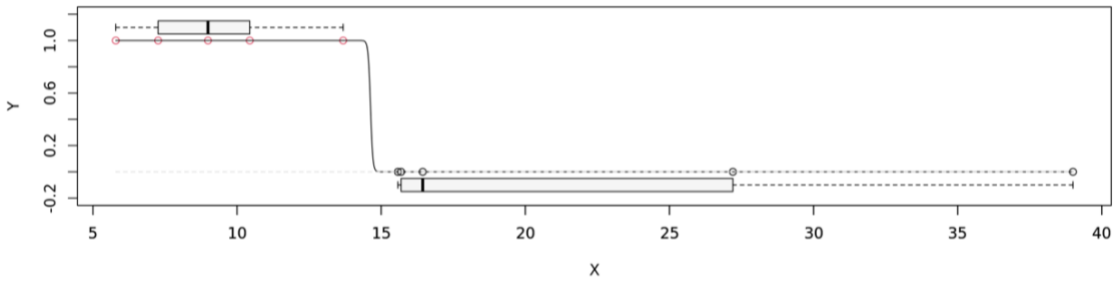


Obrázek 7: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Tečky za větou – srovnání autorů A a C



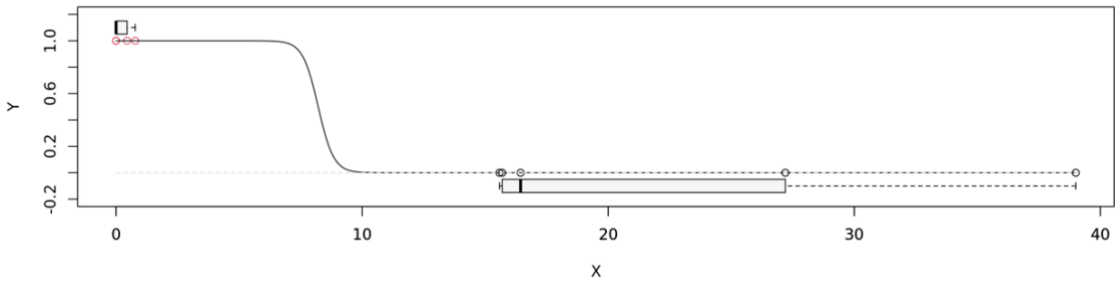
Obrázek 8: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Tečky za větou – srovnání autorů A a D



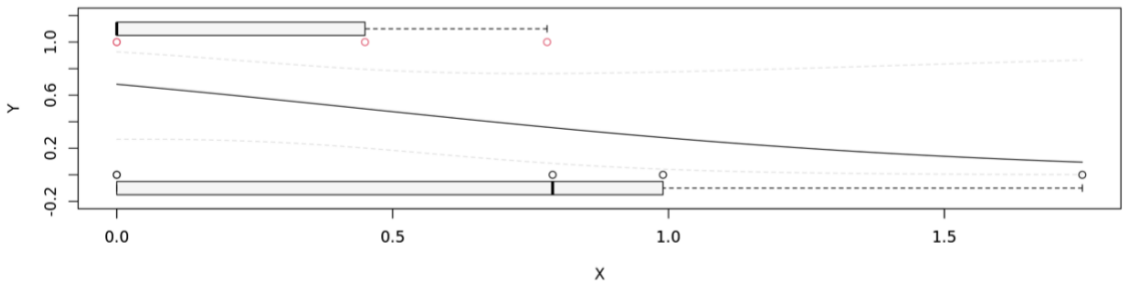
Obrázek 9: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Tečky za větou – srovnání autorů B a C

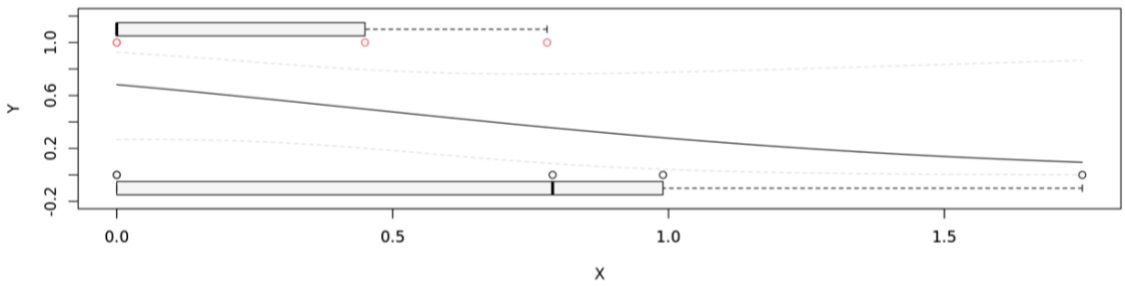


Obrázek 10: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.179661

P-hodnota je u tohoto modelu 0.179661, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.

Tečky za větou – srovnání autorů B a D

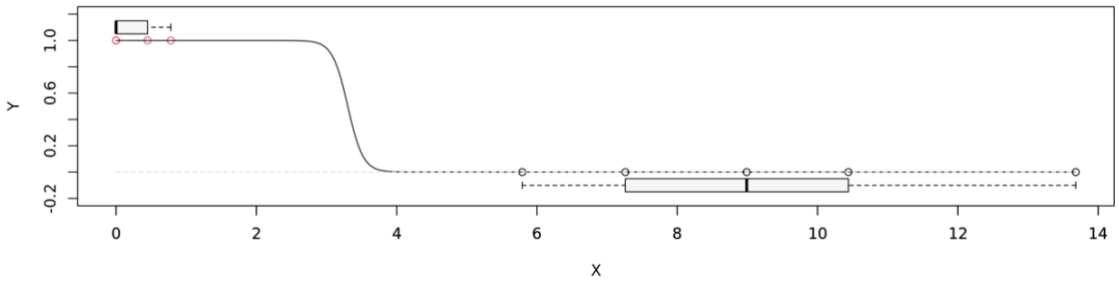


Obrázek 11: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.179661

P-hodnota je u tohoto modelu 0.179661, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.

Tečky za větou – srovnání autorů C a D



Obrázek 12: *Grafické znázornění výsledků logistické regrese při srovnání množství použitých teček za větou. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je velmi dobře funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Hodnocení úspěšnosti metody

POČET TEČEK	A	B	C	D
A		ANO, 100 %	ANO, 100 %	ANO, 100 %
B			NE, 60 %	NE, 60 %
C				ANO, 100 %
D				

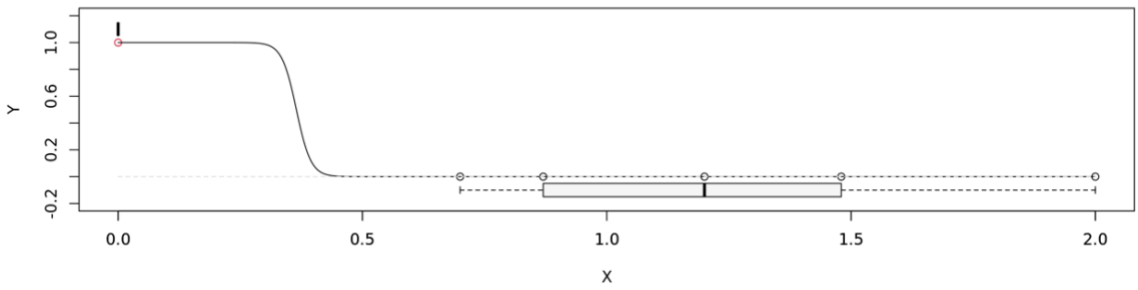
Tabulka 5: *Tabulka úspěšnosti jednotlivých pokusů při zkoumání použití více než jedné tečky za větou.*

Z tabulky úspěšnosti námi zkoumané metody je patrné, že bylo možné vytvořit pomocí logistické regrese funkční čili statisticky významný model ve čtyřech případech ze šesti. Metoda tedy byla úspěšná v přibližně 67 % pokusů. Její průměrná úspěšnost je přibližně 87 %, je tedy úspěšnější než předchozí pokus s emotikony.

Zvýraznění slov

Znovu provedeme analýzu šesti párů textů pomocí logistické regrese, tentokrát budeme vkládat do levého sloupce hodnoty naměřené pro zvýraznění slov. Výsledkem bude 6 grafů a tabulka úspěšnosti jednotlivých pokusů, kterou můžeme srovnat s úspěšností předchozích metod.

Zvýraznění slov – srovnání autorů A a B



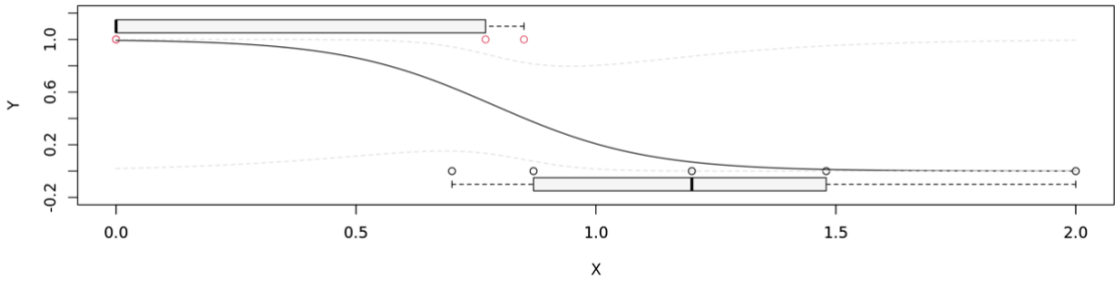
Obrázek 13: *Grafické znázornění výsledků logistické regrese při srovnání množství zvýrazněných slov. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Zvýraznění slov – srovnání autorů A a C



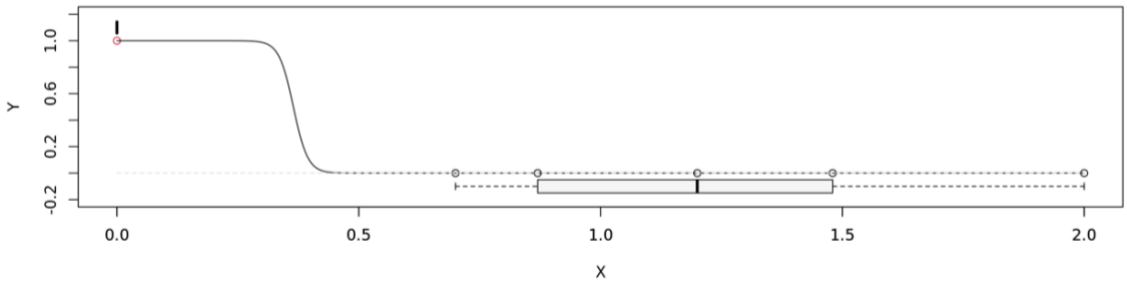
Obrázek 14: *Grafické znázornění výsledků logistické regrese při srovnání množství zvýrazněných slov. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.005812

P-hodnota je u tohoto modelu 0.005812, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Zvýraznění slov – srovnání autorů A a D



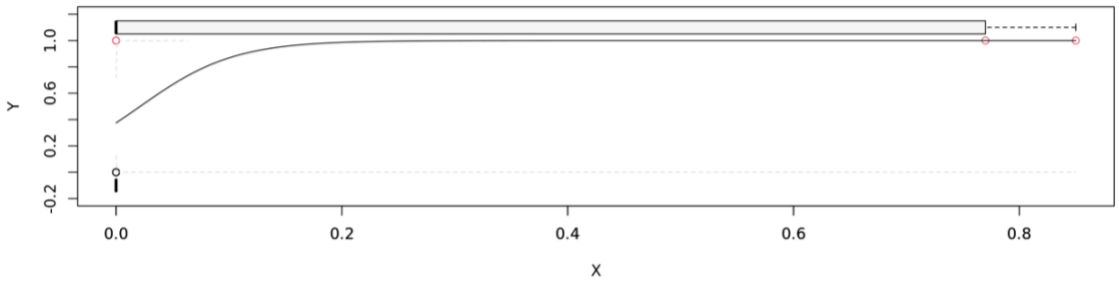
Obrázek 15: *Grafické znázornění výsledků logistické regrese při srovnání množství zvýrazněných slov. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

Zvýraznění slov – srovnání autorů B a C



Obrázek 16: *Grafické znázornění výsledků logistické regrese při srovnání množství zvýrazněných slov. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).

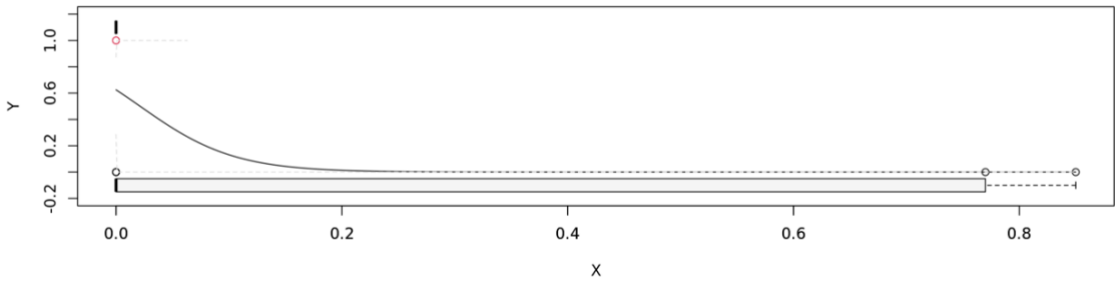
p-hodnota ≈ 0.070217

P-hodnota je u tohoto modelu 0.070217, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

Zvýraznění slov – srovnání autorů B a D

Ani jeden z autorů zvýraznění slov nepoužívá, proto nemáme data, která by mohla být vložena do levého sloupce aplikace logistické regrese, určeného pro nezávislé proměnné, tj. prediktory. Není tedy možné pomocí logistické regrese vytvořit statistický model.

Zvýraznění slov – srovnání autorů C a D



Obrázek 17: *Grafické znázornění výsledků logistické regrese při srovnání množství zvýrazněných slov. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.070217

P-hodnota je u tohoto modelu 0.070217, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

Hodnocení úspěšnosti metody

ZVÝRAZNĚNÍ	A	B	C	D
A		ANO, 100 %	ANO, 80 %	ANO, 100 %
B			NE, 70 %	nelze vytvořit
C				NE 70, %
D				

Tabulka 6: *Tabulka úspěšnosti jednotlivých pokusů při zkoumání použití zvýraznění textu autorem.*

Z tabulky úspěšnosti námi zkoumané metody je patrné, zda bylo možné vytvořit pomocí logistické regrese funkční čili statisticky významný model. V tabulce můžeme vidět, že tři z pěti vytvořených modelů byly funkční, proto je úspěšnost 80 %. Jeden model nebylo možné vytvořit, protože autoři B a D zvýraznění slov nepoužívali vůbec, proto nebyla k dispozici žádná použitelná data. Průměrná úspěšnost modelu je přibližně 84 %.

Sumarizace a diskuse

V této kapitole jsme se zaměřili na statistické zpracování výsledků pomocí logistické regrese a jejich srovnání. K analýze jsme využili data z Tabulky 4, která obsahuje informace o použití emotikon, teček za větou a zvýraznění textu v textech autorů A, B, C a D. Tyto údaje byly využity jako nezávislé proměnné v logistické regresi, zatímco závislou proměnnou byly hodnoty 0 nebo 1 přidělenou k odlišení dvou autorů od sebe.

Pro každý pár autorů jsme provedli analýzu zvlášť, což nám poskytlo grafické znázornění výsledků a hodnotu p-hodnoty, která určuje statistickou významnost vztahu mezi nezávislými a závislou proměnnou. Pokud je p-hodnota nižší než 0,05, považujeme model za statisticky významný.

Nejdříve jsme provedli analýzu pro použití emotikon, kde jsme získali statisticky významný model ve čtyřech případech ze šesti. Průměrná úspěšnost těchto modelů byla kolem 77 %.

Podobně jsme provedli analýzu s použitím teček za větou a zvýrazněním slov. V obou případech bylo možné vytvořit statisticky významné modely mezi

některými autory, a to ve čtyřech případech u teček za větou, kdy průměrná úspěšnost modelů byla kolem 87 % a ve třech případech u zvýraznění slov, kdy průměrná úspěšnost těchto modelů byla kolem 84 %.

Nejúspěšnější se tedy jeví metoda analýzy počtu teček za větou. Znamená to, že je to rys natolik charakteristický, že je možné ho použít k odlišení autorů.

6. Analýza pomocí kvantitativních metod

V této kapitole se zaměříme na metodu Bag-of-Words, vysvětlíme její fungování a provedeme analýzu anonymních textů za pomoci vícerozměrného škálování.

Interpretací jednotlivých grafů se pokusíme zjistit, jestli je možné autory od sebe jednoduše oddělit a určit, které texty jsou si nejvíce podobné a které jsou si nejméně podobné. Náš odhad následně statisticky ověříme pomocí logistické regrese.

Texty, které máme pro tento výzkum k dispozici, mají různou délku, což by ovlivnilo výsledky měření. Proto byly texty pomocí QUITA zkráceny na stejnou délku, podle nejkratšího textu, který jsme měli k dispozici, tedy na 618 tokenů.

V této i následujících kapitolách pracujeme nejprve s tokeny, které byly z původního textu vybrány náhodně, následně analýzu zopakujeme s tokeny, které jdou v textu po sobě.

Náhodný výběr tokenů má několik výhod. První z výhod je, že výsledky budou nezávislé na délce textu, všechny komentáře mají možnost se projevit. Původně komentáře měly časovou linku a konzistenci, protože odpovídaly samy na sebe. Náhodným výběrem tuto závislost na časové lince rozbijeme. Nevýhoda: výsledky budou ovlivněny náhodným výběrem tokenů.

6.1. Množina slov (Bag-of-Words model)

Při použití Bag-of-Words modelu není důležité pořadí slov v dokumentu, ale frekvence jejich výskytu.[19]

Slova nejsou jediným možným prvkem; lze rovněž využít n-tice slov (n-gramy), což jsou sekvence n po sobě jdoucích prvků.

U analýzy pomocí BOW zahrnujeme do určování autorství i jména. Někteří autoři mají oslovování ostatních účastníků diskuse jako výrazný charakteristický rys. Jména se potom v jejich textech opakují, autor je ve svém vyjadřování konzistentní. Nevýhoda pro takový typ analýzy je v tom, že už to může poukazovat na konkrétního autora.

6.2. Metoda vícerozměrného škálování

Díky použití vícerozměrného škálování (**M**ulti**D**imensional **S**caling, MDS) jsme schopni provádět objektivní srovnání textů a následně je vizualizovat. Data se na základě kvantifikovaných charakteristik shlukují do skupin, což nám umožňuje přiřadit je k sobě na základě jejich vlastností.[20] Výsledkem je tedy graf, jakási mapa, kde můžeme sledovat různé vlastnosti objektů, které se na ní nachází.

Body, které reprezentují nejpodobnější objekty, se nacházejí blízko sebe. S narůstající mírou nepodobnosti se tyto body nachází dále od sebe.

Při použití této metody je nezbytné definovat způsob, jakým se měří vzdálenost mezi texty. Běžně využíváme dvě běžné metody: euklidovskou vzdálenost a kosinovou podobnost. Hlavní rozdíl mezi těmito dvěma přístupy spočívá v jejich přístupu k rozdílům mezi jednotlivými kritérii. Euklidovská vzdálenost bere v úvahu celkové rozdíly mezi kritérii, zatímco kosinová nepodobnost ignoruje absolutní rozdíly a zaměřuje se na jejich poměry. Pro náš výzkum budeme používat kosinovou nepodobnost.

---

[19] Manning, C. D.; Raghavan, P.; Schütze, H.: Introduction to informationretrieval. Cambridge University Press, 2008, ISBN 0521865719

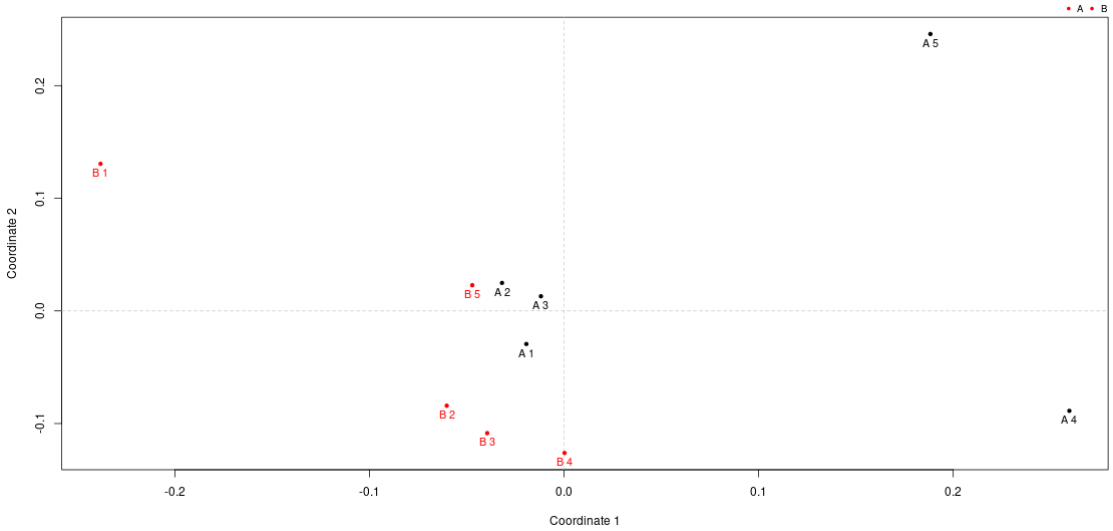
[20] FALTÝNEK, Dan, Dalibor PAVLAS, Ondřej VRABEL a Vladimír MATLACH. Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu. Olomouc: Univerzita Palackého, 2015

Náhled vytvořený pomocí vícerozměrného škálování je pouze ilustrativní, protože se jedná o dvojrozměrné zobrazení, ve kterém nejsou zahrnuty veškeré informace. U každého grafu tedy uvádíme, kolik procent variance vypočítaných vzdáleností graf zobrazuje.

Srovnání autorů

V této podkapitole se pokusíme mezi sebou porovnat dvojice autorů, pomocí multidimenzionálního škálování a následně odhad statisticky ověřit pomocí logistické regrese.

Srovnání autorů A a B – náhodně vybrané tokeny



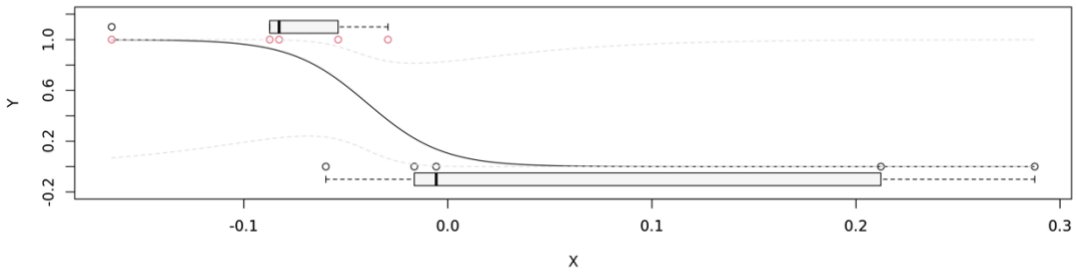
Obrázek 18: *Analýza deseti textů autorů A a Bo 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.3884712

V grafu můžeme vidět, že autory A a B od sebe nemůžeme jednoduše odlišit. Čtyři texty od autora B a tři texty od autora A se shlukly v těsné blízkosti ve středu grafu. Můžeme tedy předpokládat, že mají hodně společných vlastností. Texty B 1, A 5 a A 4 jsou rozptýlené po stranách grafu, z čehož vyplývá, že jsou od ostatních ve středu i od sebe navzájem odlišné. V případě této analýzy MDS

zrekonstruovalo přibližně 39% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



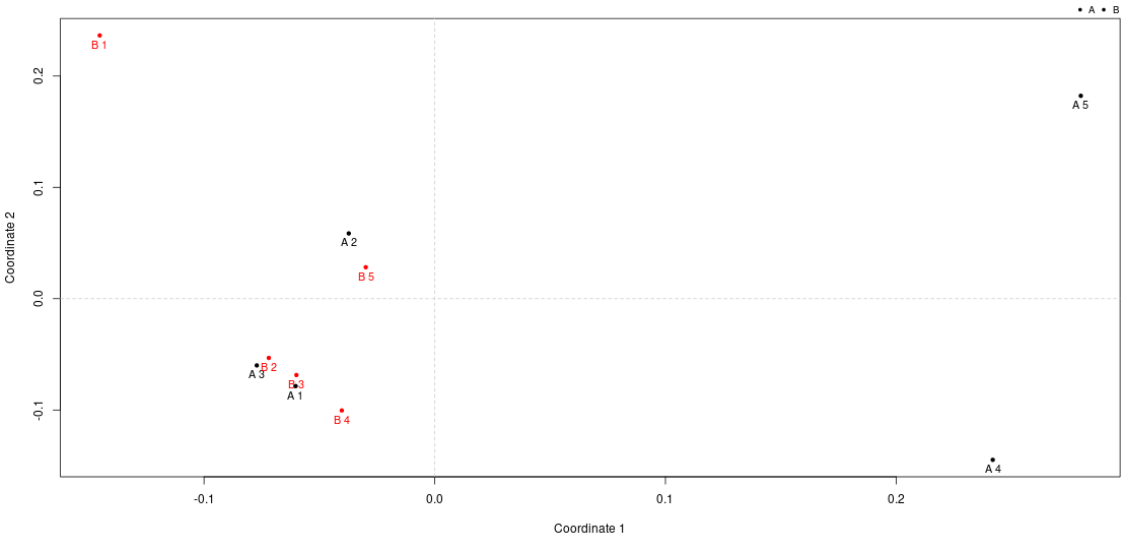
Obrázek 19: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 náhodně vybraných tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.007343

P-hodnota je u tohoto modelu 0.007343, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Srovnání autorů A a B – po sobě jdoucí tokeny



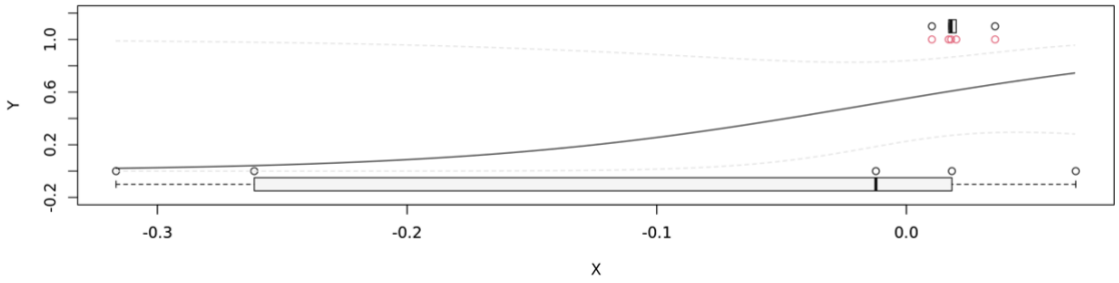
Obrázek 20: *Analýza deseti textů autorů A a B o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.4267059

V grafu můžeme vidět, že autory A a B od sebe nemůžeme jednoduše odlišit. Čtyři texty od autora B a tři texty od autora A se shlukly v těsné blízkosti ve středu grafu, některé z nich se téměř překrývají. Můžeme tedy předpokládat, že mají hodně společných vlastností. Texty B 1, A 5 a A 4 jsou rozptýlené po stranách grafu, z čehož vyplývá, že jsou od ostatních ve středu i od sebe navzájem odlišné.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 43% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



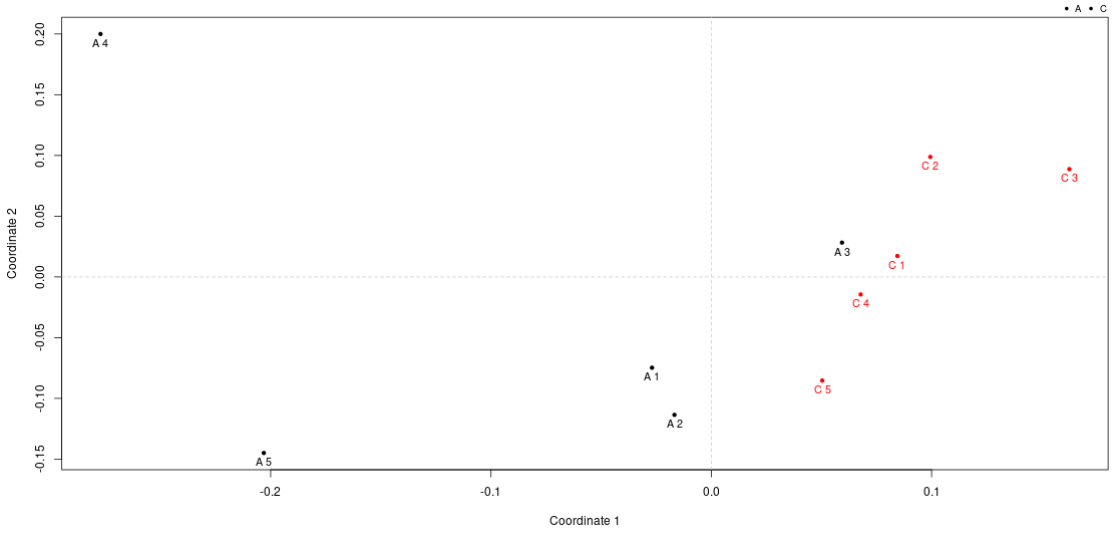
Obrázek 21: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.094549

Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 70 % případů, ale díky p-hodnotě 0.094549 víme, že tento model není úspěšnější než náhoda.

Srovnání autorů A a C – náhodně vybrané tokeny



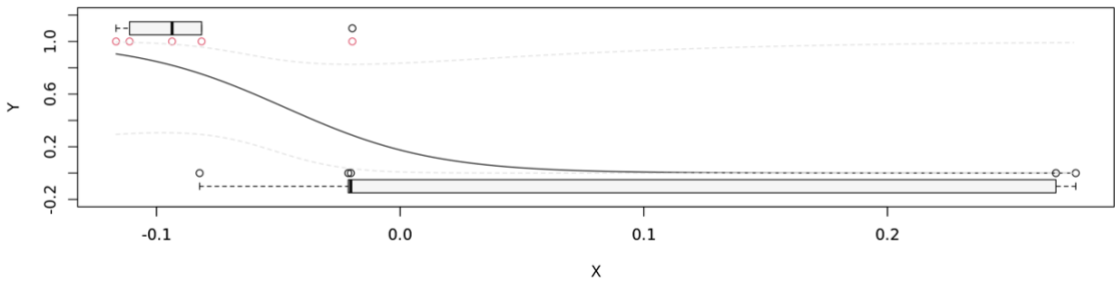
Obrázek 22: *Analýza deseti textů autorů A a C o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.4949457

Z grafu vyplývá, že texty autora A se shlukly spíše v levé části grafu a autora C se shlukly v pravé části grafu, ale nejsou tolik odlišní, aby vytvořili výrazně oddělené shluky. Text A 3 je velmi blízko autorovi C, proto můžeme předpokládat, že je mu nějakým způsobem podobný. Text A 5 se nejvíce liší od autora C a zároveň je i nejvíce vzdálený svému vlastnímu autorovi.

V případě této analýzy MDS zrekonstruovalo přibližně 49% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



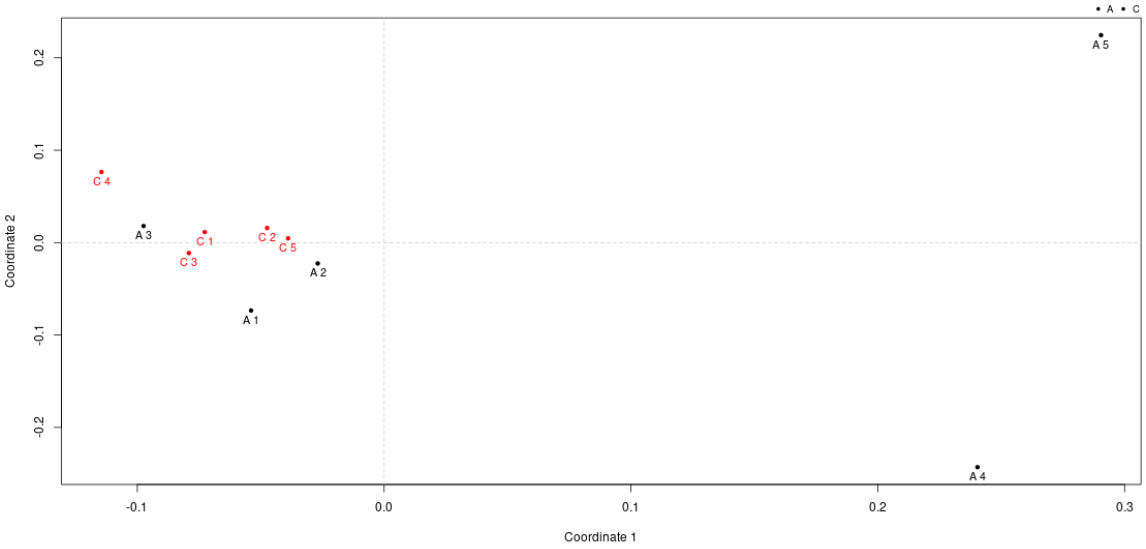
Obrázek 23: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 náhodně vybraných tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.016639

P-hodnota je u tohoto modelu 0.016639, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Srovnání autorů A a C – po sobě jdoucí tokeny



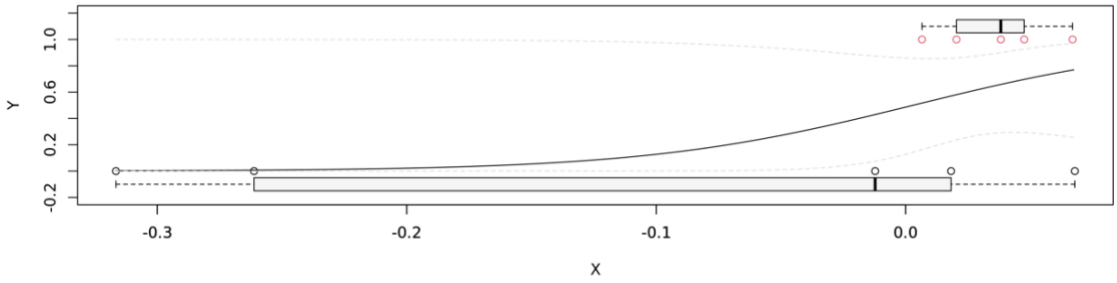
Obrázek 24: *Analýza deseti textů autorů A a C o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.4834524

Z grafu vyplývá, že tři texty autora A a všech pět textů autora C se shlukly v levé části grafu, ale nejsou tolik odlišní, aby vytvořili výrazně oddělené shluky. Texty A 5 a A 4 se odtrhly od ostatních a jsou umístěny v nejvzdálenějších rozích v pravé části grafu, můžeme tedy předpokládat, že nejvíce liší od autora C a zároveň jsou i nejvíce vzdálený svému vlastnímu autorovi a sobě navzájem.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 48% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



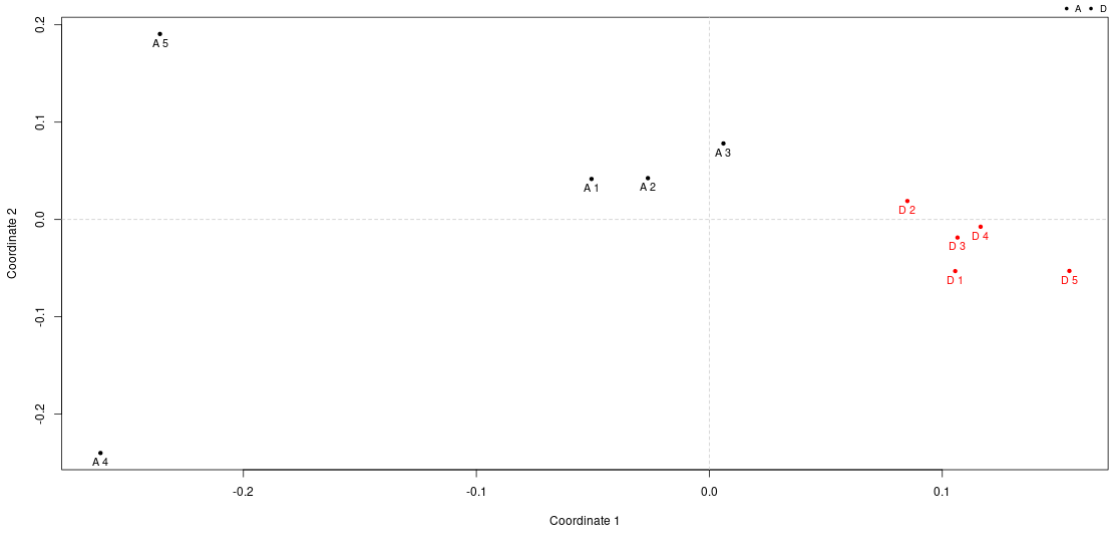
Obrázek 25: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj (nemá na urč ení odpovědí významný vliv).
p-hodnota ≈ 0.058094

Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 80 % případů, ale díky p-hodnotě 0.058094 víme, že tento model není úspěšnější než náhoda.

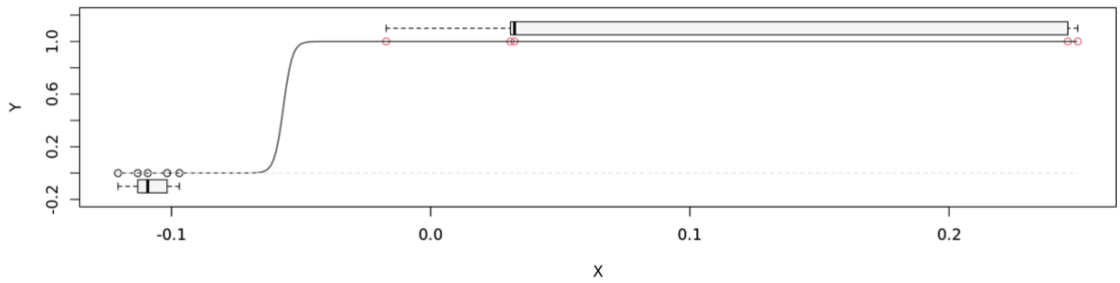
Srovnání autorů A a D – náhodně vybrané tokeny



Obrázek 26: *Analýza deseti textů autorů A a D o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

V grafu můžeme pozorovat, že je možné od sebe autory A a D velmi dobře lineárně oddělit. Autor D má kompaktní shluk textů v pravé části grafu. Texty jsou shluklé blízko k sobě, z toho můžeme usuzovat, že jsou si vzájemně podobnější, než texty autora A, které jsou rozptýleny po pravé části grafu. Texty A 1, A 2 a A 3 jsou k sobě blízko, texty A 4 a A 5 se od ostatních vzdálily, je tedy pravděpodobné, že jsou velmi odlišné od textů autora D, ale i od textů autora A.

V případě této analýzy MDS zrekonstruovalo přibližně 48% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



Obrázek 27: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 náhodně vybraných tokenů*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

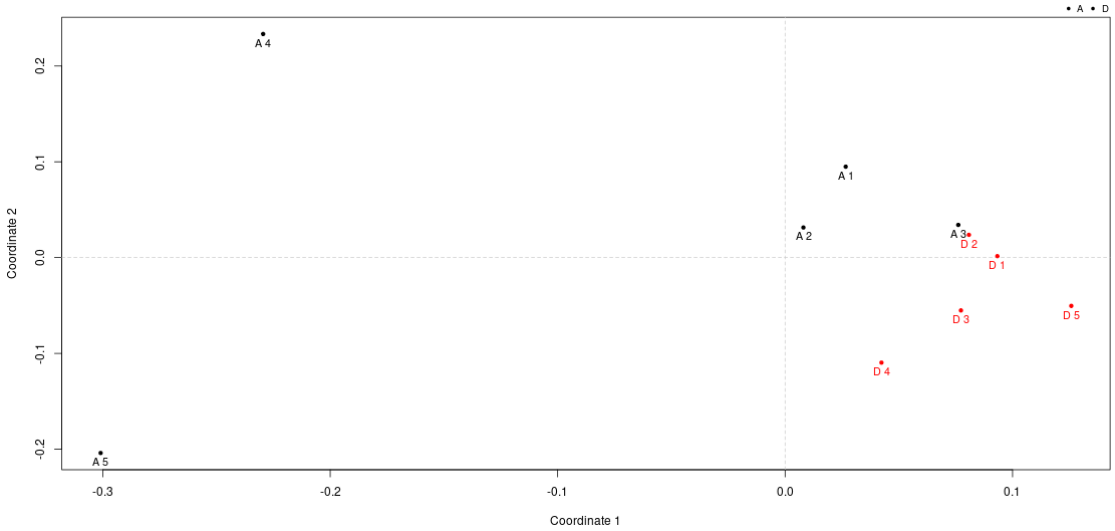
p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferroniho korekci, protože p-hodnota vychází nižší než 0,001.

$GOF

[1] 0.4805854

Srovnání autorů A a D – po sobě jdoucí tokeny



Obrázek 28: *Analýza deseti textů autorů A a D o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

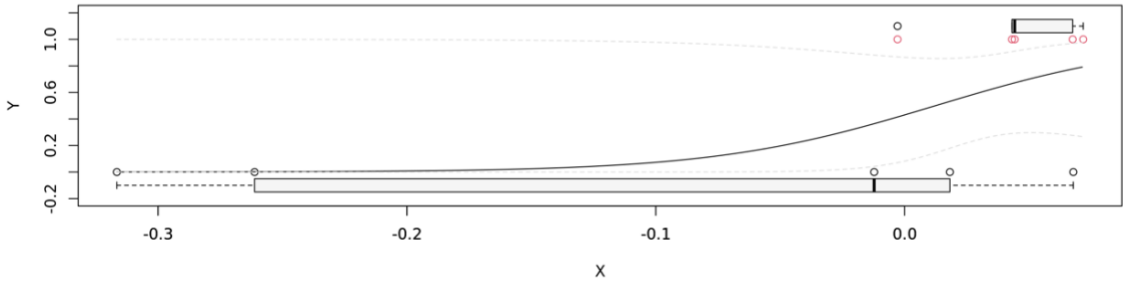
$GOF

[1] 0.4491856

V grafu můžeme pozorovat, že není možné od sebe autory A a D velmi dobře lineárně oddělit. Autor D má kompaktní shluk textů v pravé části grafu. Texty jsou shluklé blízko k sobě, z toho můžeme usuzovat, že jsou si vzájemně podobnější, než texty autora A, které jsou rozptýleny po pravé části grafu. Texty

A 1, A 2 a A 3 jsou k sobě blízko, texty A 4 a A 5 se od ostatních vzdálily a zůstaly v levé části grafu, je tedy pravděpodobné, že jsou velmi odlišné od textů autora D, ale i od textů autora A a od sebe navzájem.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 45% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



Obrázek 29: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

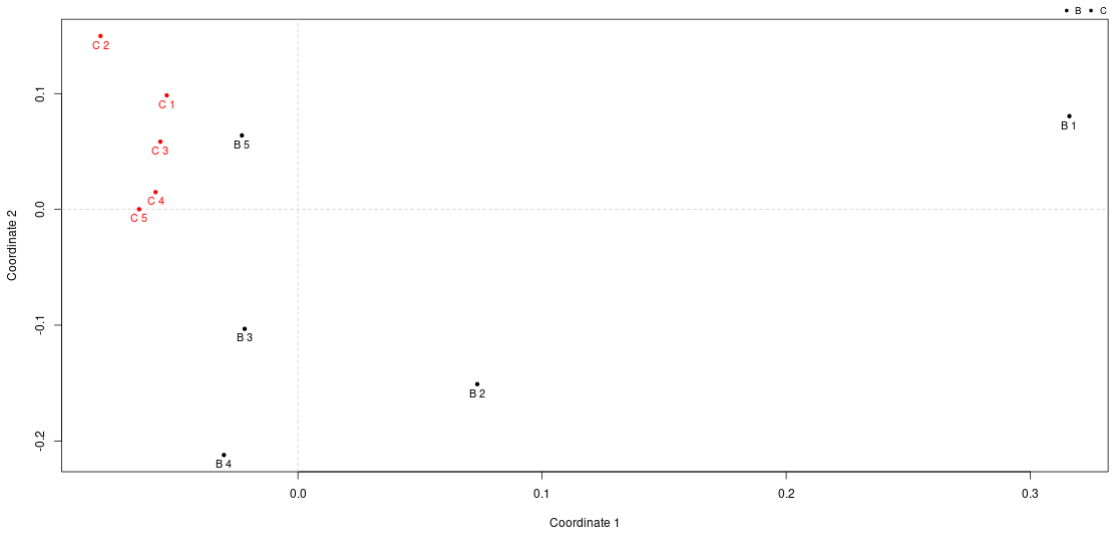
[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.04315

Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

P-hodnota je u tohoto modelu 0.04315 což znamená, že tento model je funkční. Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

Srovnání autorů B a C – náhodně vybrané tokeny



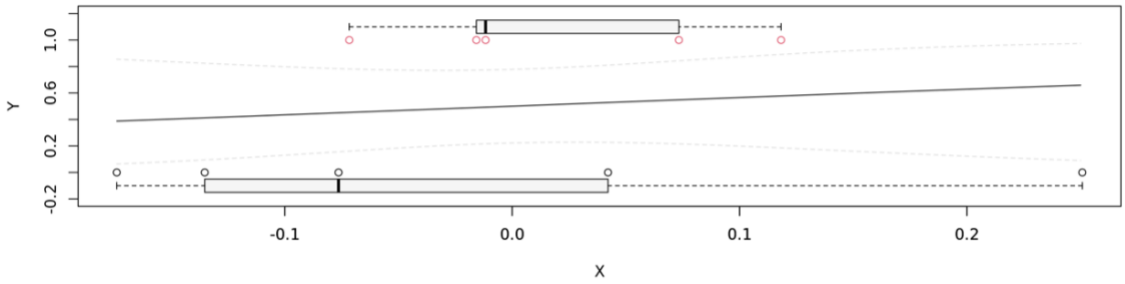
Obrázek 30: *Analýza deseti textů autorů B a C o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.5091738

Na tomto grafu můžeme sledovat, že texty autora C se shlukly v levém horním rohu a tvoří skupinu, která je možná oddělit od textů autora B, nicméně text B5 je velmi blízko skupině C. Můžeme tedy předpokládat, že je textům autora C nejpodobnější. Text B 1 se od ostatních oddělil a leží na opačné straně grafu, můžeme tedy předpokládat, že je od ostatních textů nejvíce odlišný.

V případě této analýzy MDS zrekonstruovalo přibližně 50% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.

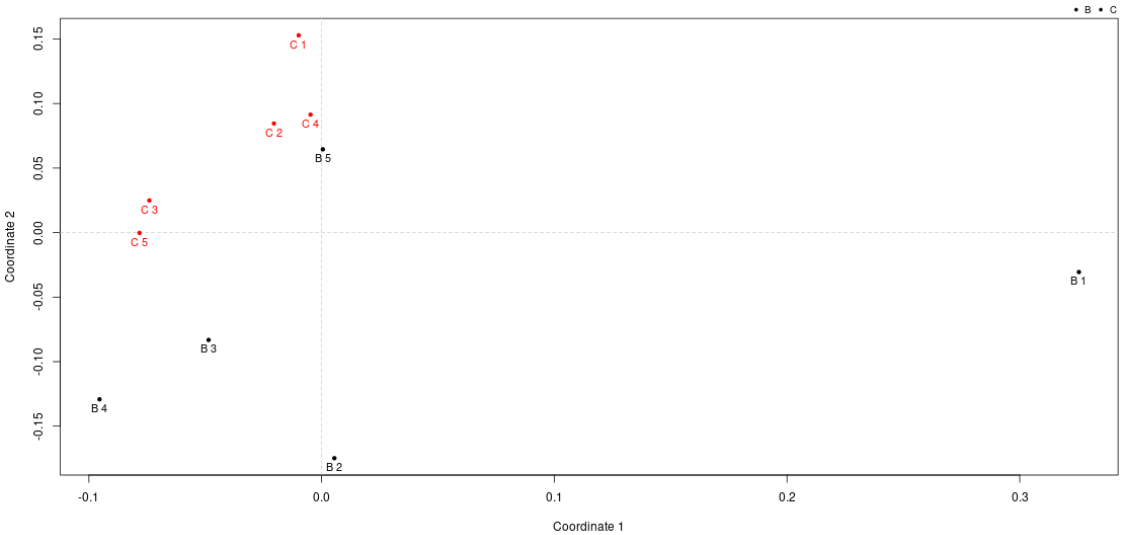


Obrázek 31: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 náhodně vybraných tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.623764

P-hodnota je u tohoto modelu 0.623764, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' pouze v 50 % případů.

Srovnání autorů B a C – po sobě jdoucí tokeny



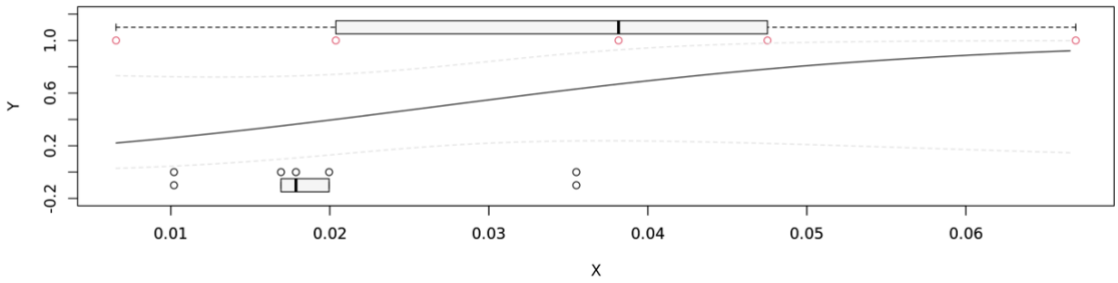
Obrázek 32: *Analýza deseti textů autorů B a C o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.465145

Na tomto grafu můžeme sledovat, že texty autora C se shlukly v levém horním rohu a tvoří skupinu, ale není možné je jednoduše oddělit od textů autora B. Text B5 je velmi blízko skupině C. Můžeme tedy předpokládat, že je textům autora C nejpodobnější. Text B 1 se od ostatních oddělil a leží na opačné straně grafu, můžeme tedy předpokládat, že je od ostatních textů nejvíce odlišný.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 47% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



Obrázek 33: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

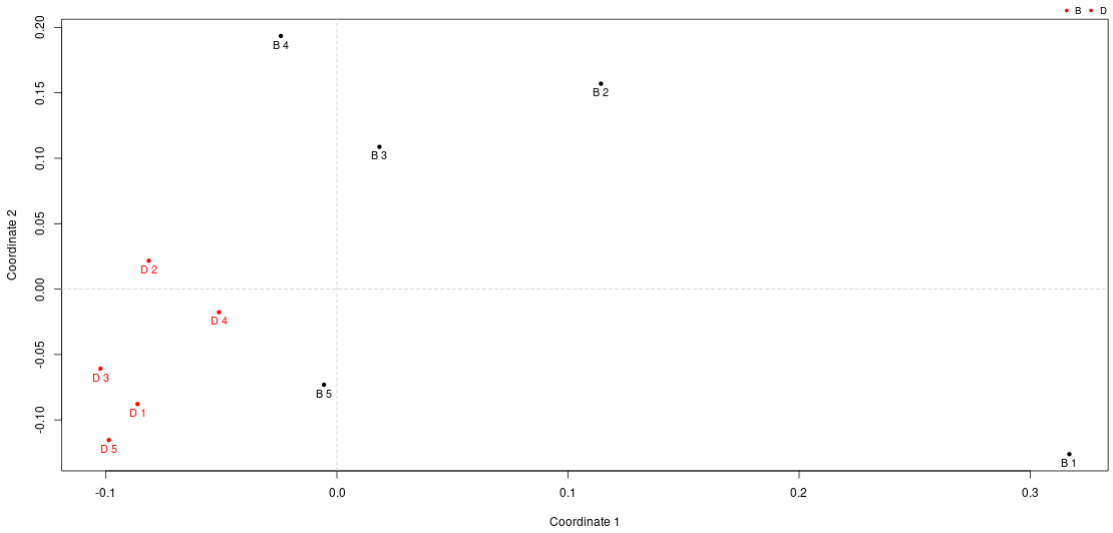
[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).

p-hodnota ≈ 0.139333

Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 70 % případů, ale díky p-hodnotě 0.139333 víme, že tento model není úspěšnější než náhoda.

Srovnání autorů B a D – náhodně vybrané tokeny

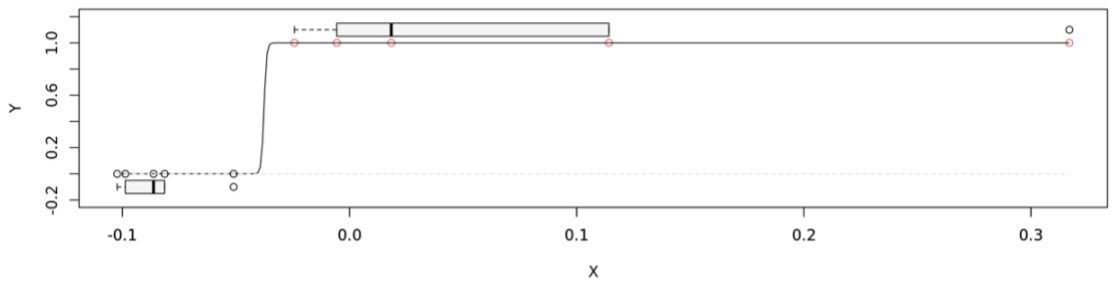


Obrázek 34: *Analýza deseti textů autorů B a D o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.4945283

Texty autorů B a D od sebe lze lineárně oddělit a je možné sledovat, že texty autora D se shlukly v levém dolním rohu a jsou blízko u sebe. Text B 1 je umístěn v pravém dolním rohu nejdále od ostatních, můžeme tedy předpokládat, že se nevíce liší od textů autora D, ale zároveň je i velmi odlišný od textů autora B. V případě této analýzy MDS zrekonstruovalo přibližně 49% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



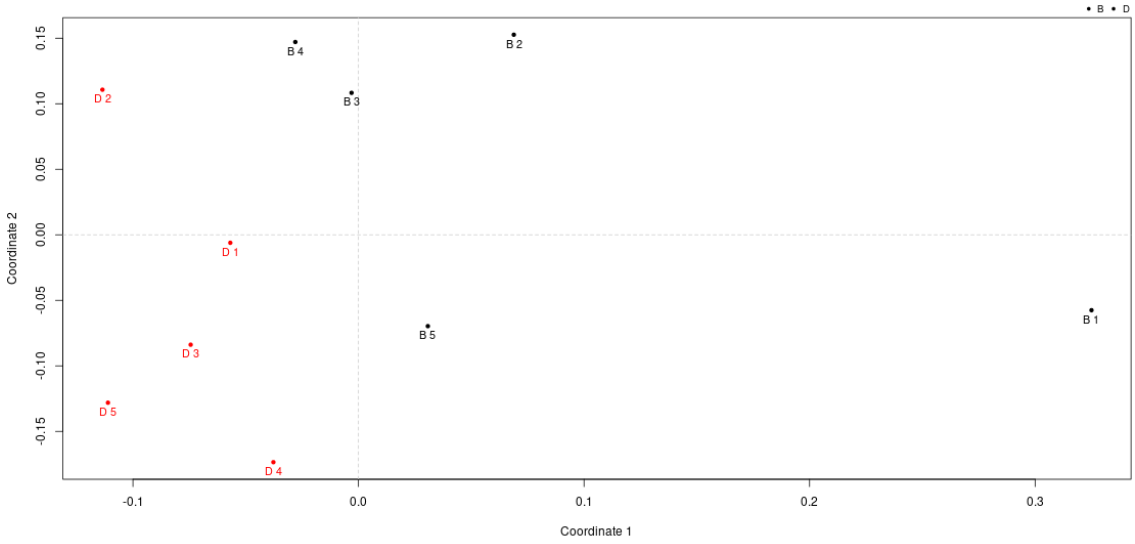
Obrázek 35: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 náhodně vybraných tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů B a D – po sobě jdoucí tokeny



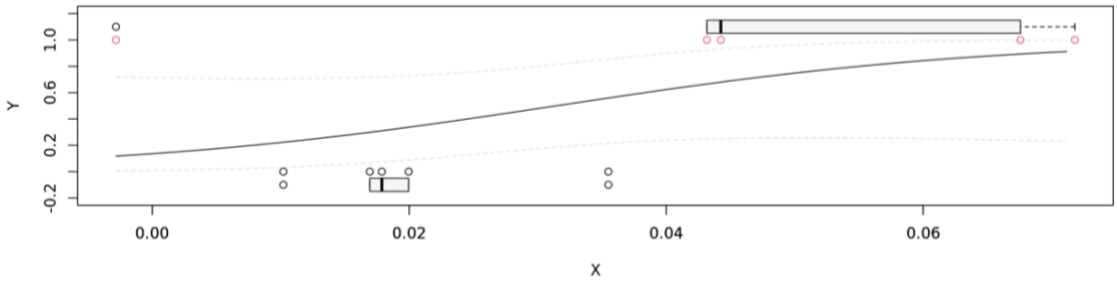
Obrázek 36: *Analýza deseti zkoumaných textů o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

[1] 0.4948347

Texty autorů B a D od sebe lze lineárně oddělit a je možné sledovat, že texty autora D se shlukly v levé části grafu. Tři texty od autora B jsou umístěny také v levé části grafu a jsou autorovi D blízko, můžeme tedy odhadovat, že jsou si některé texty obou autorů podobné. Text B 1 je umístěn v pravém dolním rohu nejdále od ostatních, můžeme tedy předpokládat, že se nevíce liší od textů autora D, ale zároveň je i velmi odlišný od textů autora B.

V případě této analýzy MDS zrekonstruovalo přibližně 49% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.

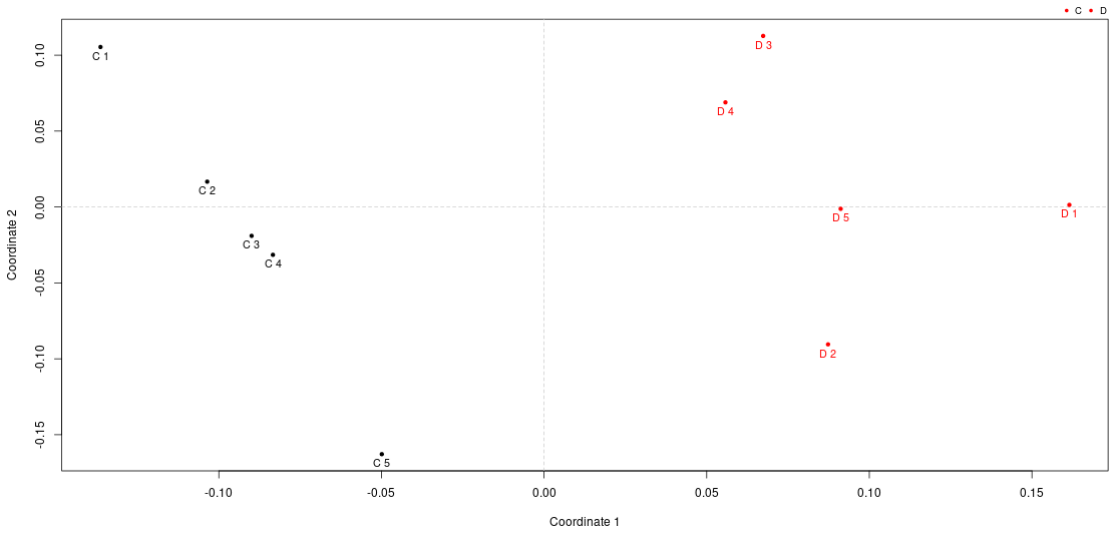


Obrázek 37: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.074828

Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Srovnání autorů C a D – náhodně vybrané tokeny



Obrázek 38: *Analýza deseti textů autorů C a D o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

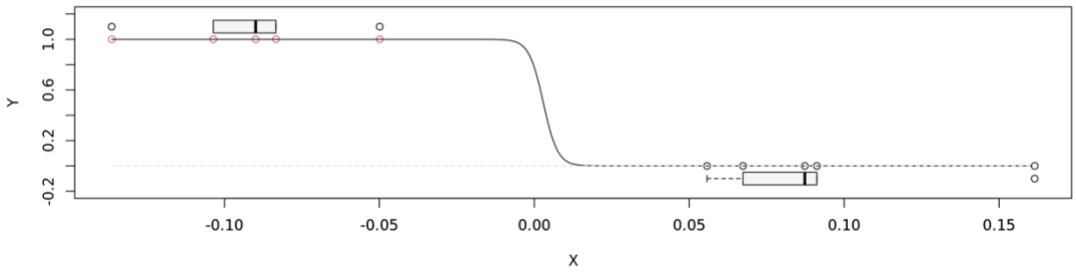
$GOF

[1] 0.4696219

V případě této analýzy MDS zrekonstruovalo přibližně 46% variance vypočítaných vzdáleností.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů autora C a v levé části také shluk pěti textů od autora D. Texty je možné lineárně oddělit. Znamená to, že se od sebe autoři výrazně liší.

V případě této analýzy MDS zrekonstruovalo přibližně 46% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



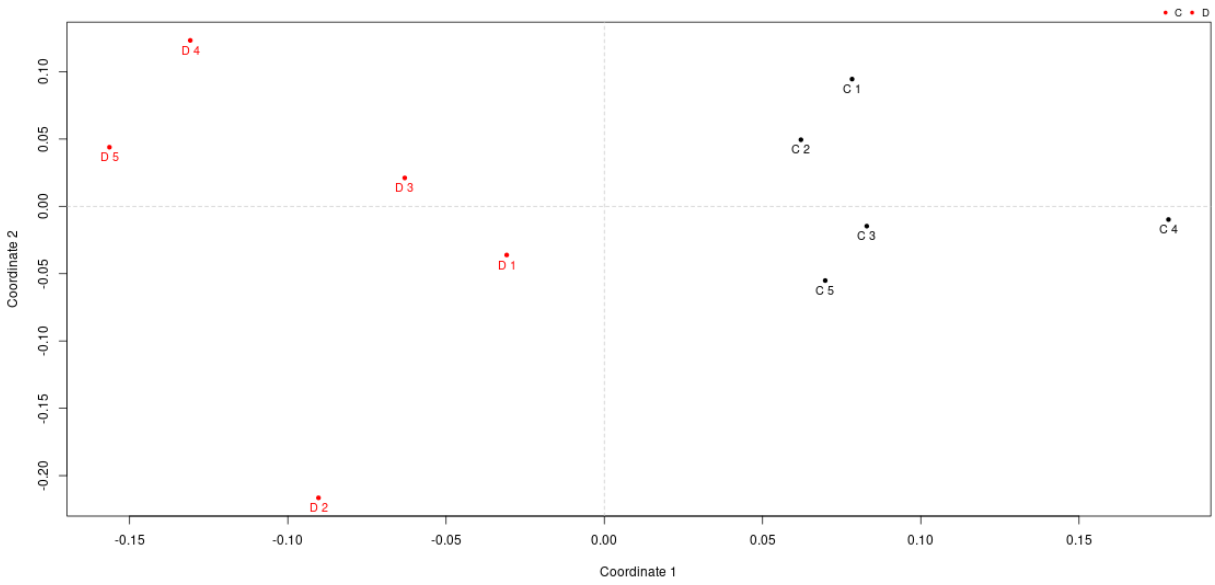
Obrázek 39: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 po náhodně vybraných tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů C a D – po sobě jdoucí tokeny

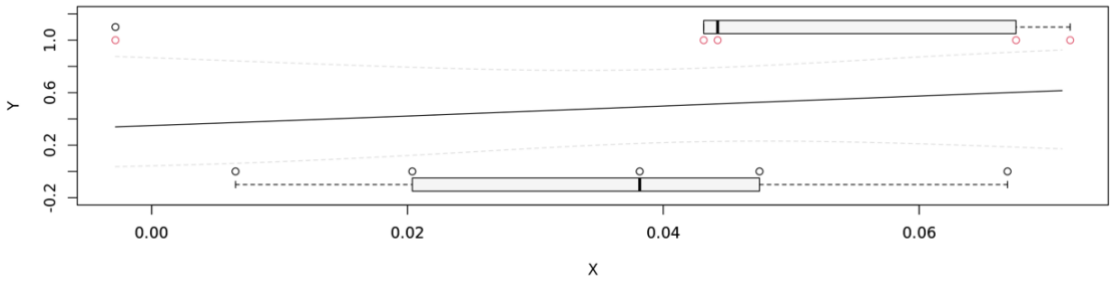


Obrázek 40: *Analýza deseti textů autorů C a D o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.4406701

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 44% variance vypočítaných vzdáleností. Následně provedeme statistické ověření pomocí logistické regrese.



Obrázek 41: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 po sobě jdoucích tokenů.*

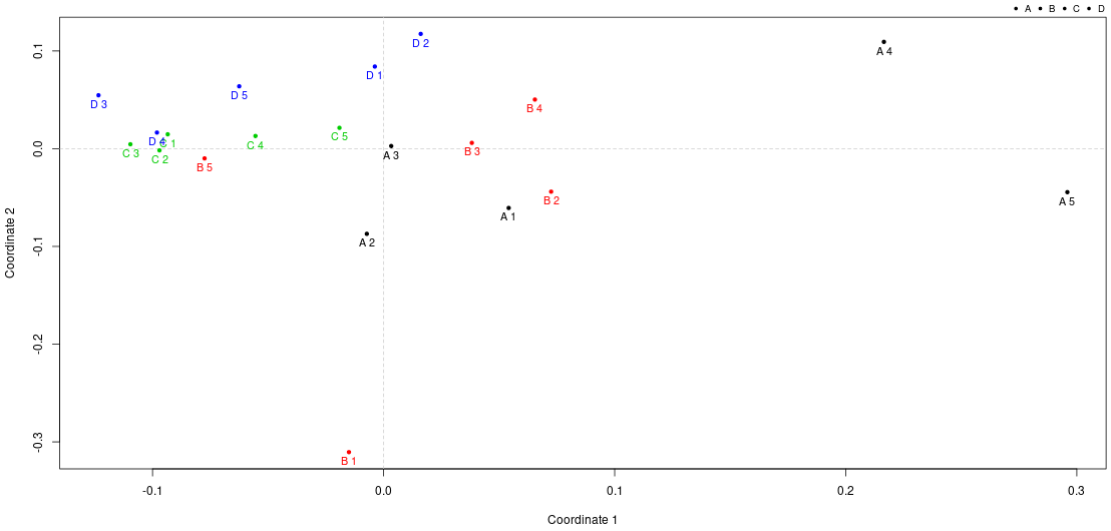
[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).

p-hodnota ≈ 0.561374

Logistická regrese dokáže správně přiřadit '0' a '1' v 70 % případů.

S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 70 % případů, ale díky p-hodnotě 0.561374 víme, že tento model není úspěšnější než náhoda.

Srovnání všech autorů – náhodně vybrané tokeny



Obrázek 42: *Analýza dvaceti textů všech zkoumaných autorů o 618 náhodně vybraných tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

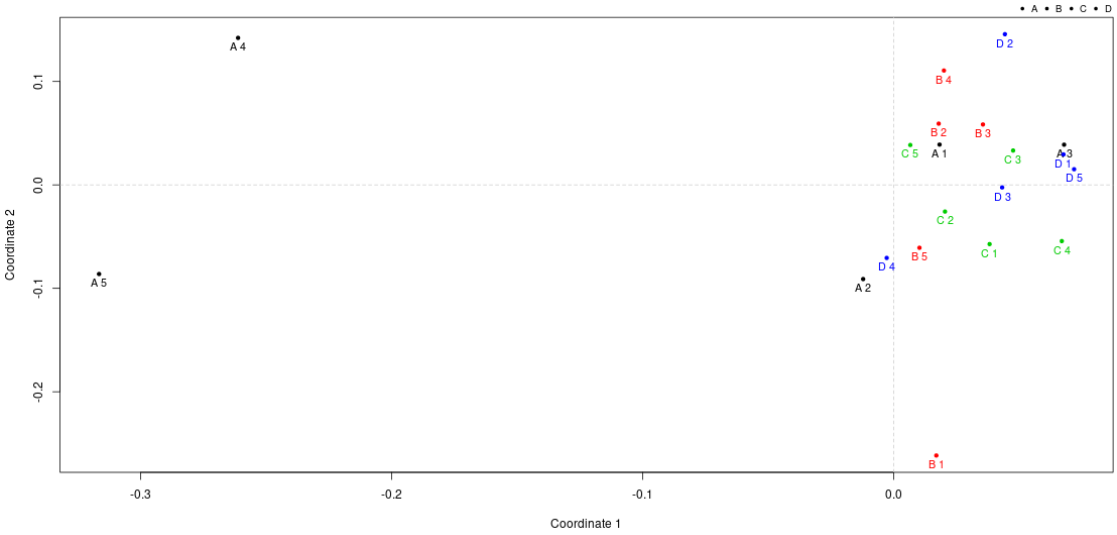
$GOF

[1] 0.2993429

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 29% variance vypočítaných vzdáleností.

Jednotlivé shluky textů od sebe nelze snadno oddělit. Texty A4 a A5 byly vyhodnocen jako nejvíce odlišné, a proto leží v pravém horním rohu nejvíce vzdáleny od všech ostatních textů.

Srovnání všech autorů – po sobě jdoucí tokeny



Obrázek 43: *Analýza dvaceti textů všech zkoumaných autorů o 618 po sobě jdoucích tokenech. Analýza byla provedena pomocí modelu Bag-of-Words a výsledky zobrazeny pomocí metody vícerozměrného škálování (MDS).*

$GOF

[1] 0.2827677

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. V případě této analýzy MDS zrekonstruovalo přibližně 28% variance vypočítaných vzdáleností. V optimálním scénáři by se očekávalo, že se v grafu zobrazujícím shluky textů odlišených podle autorů zobrazí čtyři homogenní shluky, což by naznačovalo jasnou separaci mezi texty jednotlivých autorů. Nicméně v tomto konkrétním případě není možné snadno oddělit shluky textů od sebe.

Texty označené jako A 4 a A 5 vykazují největší míru odlišnosti, a proto se nacházejí v pravém horním rohu grafu, což znamená, že jsou nejvíce vzdáleny od ostatních textů v souboru.

Je zajímavé, že texty A 4 a A 5 se oddělily od ostatních při každé analýze, pokusíme se tedy identifikovat, v čem by se mohly lišit od ostatních, co by mohlo ovlivnit analýzu BOW natolik, že jsou v grafu vždy zobrazeny mimo hlavní shluk svého autora.

Sumarizace a diskuse

V kapitole 5 jsme si všímali tří kritérií, která mohou být natolik charakteristickým rysem, že by mohla být použita k odlišen jednotlivých autorů od sebe. V Tabulce 3 jsme mohli sledovat, že autor A má u všech tří kritérií nejvyšší hodnoty. U teček za větou je výskyt dokonce mnohonásobně vyšší než u ostatních autorů. Pro Bag of Words model je relevantní počet teček za větou a emotikony, zvýraznění slov Caps Lockem se v modelu BOW nepromítne, nicméně autor A je v používání zvýraznění slov velmi kreativní, zvýrazňuje nejen CapsLockem, ale i spojováním slov pomlčkami.

Příklad z textu A 5:

Ale tady? Ono by to mohlo fungovat jako MEGA-GIGA-SUPER-HUPER-VELE splachovadlo, podobné, jako máme na Dlouhých stráních. Akorát s tím, že tu vodu by Židi nemuseli do nádržky čerpat. Přiteče jim tam sama.

Druhým typem zvýraznění je vložením mezery mezi každý znak slova.

Příklad z textu C4:

Ten čas,kterej nám zbývá,má mnoho kvalit.Já chtěla napsat z b e j v á - a ono mě to opravilo na zbývá,che che. I ten čas- chronos,ten,co nám namlouvá,že čas je měřitelný,ten,který požírá svoje děti,ta obluda,na kterou se po čase každej rozumnej člověk vykašle a zaměří se na k v a l i t n í čas,kterej se neměří,ale ž i j e. Zázračné vidění se uskutečňuje srdcem a pak začne automaticky fungovat každý oko!Dejme očím laskavý čas.

Bag of Words v tomto případě vyhodnotí každý znak oddělený mezerou jako samostatné slovo. Toto však pravděpodobně není důvodem, proč texty A4 a A5 tolik vyčnívají mezi ostatními, protože autor A tento způsob zvýraznění nepoužívá v žádném ze svých textů.

Je tedy otázkou, jestli emotikony a vyšší počet teček můžou ovlivnit BOW natolik, že by se při náhodném výběru tokenů pokaždé texty A 4 a A 5 oddělily od všech ostatních analyzovaných textů.

Při snaze o zjištění, proč texty A 4 a A 5 vybočují, jsme se podívali do souborů s Bag of Words, které vytvořila QUITA. Oba texty vykazovaly znaky, které bychom očekávali od všech textů, tedy např. vysoká frekvence předložek, spojek a zájmen.

Text A 4 obsahoval 14x a, 11x to, 9x v. Text A 5 obsahoval 12x a, 11x v, 6x to.

V ostatních textech autora A je frekvence velice podobná, toto tedy nemůžeme považovat za důvod, proč texty A 4 a A 5 vyčnívají. Při dalším pohledu do Bag of Words vidíme, že se pravděpodobně jednalo o komentáře pod články s lodní, námořní nebo válečnou tématikou, protože se v obou textech objevovala častěji slova z tohoto okruhu.

Text A 4 obsahoval 8x loď, 7x gb (GB, zkratka pro Velkou Británii), 5x válku, 5x britannica, 4x ponorka.

Text A 5 neobsahoval žádné plnovýznamové slovo s tak vysokou frekvencí, nicméně většina byla velice úzce zaměřena na leteckou, lodní, nebo válečnou tematiku (motorizované, vzducholoď, maršál, hindemburgu, luftwaffe). Při pohledu do textů A 1, A 2 a A 3 už takovou frekvenci slov zaměřených na dopravu nebo válku nevidíme, v textech se objevují i další témata. Je tedy možné, že za výrazným oddělením textů A 4 a A 5 je jejich úzké tematické zaměření.

Hodnocení úspěšnosti metody

BOW náhodně	A	B	C	D
A		ANO, 80 %	ANO, 80 %	ANO, 80 %
B			NE, 50 %	ANO, 100 %
C				ANO, 100 %
D				

Tabulka 7: *Tabulka úspěšnosti jednotlivých pokusů u metody Bag of Words pro 618 náhodně vybraných tokenů.*

Při použití této metody jsme byli schopni od sebe úspěšně odlišit pět případů ze šesti, tedy byla úspěšná ve více než 83 % případů. Průměrná úspěšnost byla přibližně 82 %.

BOW po sobě jdoucí tokeny	A	B	C	D
A		NE, 70 %	NE, 80 %	ANO, 70 %
B			NE, 70 %	NE, 80 %
C				NE, 70 %
D				

Tabulka 8: *Tabulka úspěšnosti jednotlivých pokusů u metody Bag of Words pro 618 po sobě jdoucích tokenů.*

Tato metoda se ukázala jako nejméně úspěšná ze všech námi použitých metod, logistická regrese úspěšně odlišila pouze jeden případ ze šesti, tedy pouze necelých 17 %. Průměrná úspěšnost byla kolem 73 %, což je nejnižší ze všech testovaných.

Jedním z důvodů, proč byla metoda Bag of Word tak neúspěšná u pokusů, kdy jsme použili po sobě jdoucí tokeny, může být to, že jednotlivé textové soubory se skládají z různě dlouhých komentářů na různá témata. V případě, že by soubor obsahoval pouze jeden komentář na jedno téma, autor pravděpodobně nevyužije veškerou bohatost slovníku, kterou má k dispozici. Při pohledu do jednotlivých souborů autorů vidíme, že autoři C a D píší dlouhé a souvislé komentáře, autoři A a B píší spíše více kratších komentářů. Očekávali bychom tedy, že autoři A a B se budou jevit jako že mají bohatší slovní zásobu. Bohatostí slovníku se budeme detailněji zabývat v následující kapitole.

7. Entropie

Shannonova entropie je definována jako stupeň nejistoty v systému. Když se použije k analýze distribuce slov v textu, entropie ukazuje na míru diverzity – čím vyšší je hodnota entropie, tím větší je rozmanitost slovníku, což naznačuje, že bohatství slovníku je vyšší, protože jednotlivá slova obsahují více informací.[21] Nižší entropie je spojena s vyšší redundancí textu, což znamená častější opakování slov.

V následující tabulce udáváme naměřené hodnoty entropie u jednotlivých textů.

Text	ENTROPIE
A 1.txt	8.234119
A 2.txt	8.152646
A 3.txt	8.317986
A 4.txt	8.298599
A 5.txt	8.500156
B 1.txt	8.166745
B 2.txt	8.325651
B 3.txt	8.230491
B 4.txt	8.252603
B 5.txt	7.900882
C 1.txt	8.325525
C 2.txt	8.164557
C 3.txt	8.23544
C 4.txt	8.147802
C 5.txt	8.330193
D 1.txt	8.171092
D 2.txt	8.116657
D 3.txt	8.014911
D 4.txt	8.077282
D 5.txt	8.079145

Tabulka 9: *Naměřené hodnoty entropie u dvaceti zkoumaných textů pomocí lingvistického softwaru QUITA.*

[21] ČECH, Radek, Ioan-Iovitz POPESCU a Gabriel ALTMANN. Metody kvantitativní analýzy (nejen) básnických textů. Olomouc: Univerzita Palackého v Olomouci, 2014. Qfwfq. ISBN 978-80-244-4044–6.

V tabulce můžeme vidět, že nejnižší hodnotu entropie má text B 5 a to 7.900882. Znamená to, že ze všech zkoumaných textů má nejméně bohatý slovník. Při pohledu do textu zjistíme, že autor B napsal devět komentářů, ve kterých se opakovala dvě témata. Prvním tématem bylo domácí zvířectvo, koně, ovce a oslové a jejich chov. **Ukázka z textu B 5:**

***Osla** NE!*

Tvrdohlavej, nezvladatelnej, kope a kouše. Nikdy nevíš co ho napadne! Pravdou ale je, že se nažere i bodláků.

*Myslím, že právě pro tyto vlastnosti se kříží na **mezky** a **muly**. Dosáhnout na nenáročnost osla a ovladatelnost **koně**.*

Druhým opakujícím se tématem byl pravděpodobně oční zákrok, který autor B podstoupil v nemocnici.

***Oko** už začíná něco vidět, ale zatím nevím ještě co. Tak snad. Ráno jedu do ÚVN na kontrolu, snad bych se už mohl dozvědět jestli se **sítnice** chytla a přilnula. To rozhodne jestli ještě nějakou dobu vydrží nebo definitivní šlus. Po čtyřech **operacích** s pátou ani nepočítám, ale kdyby byla z jejich strany vůle, tak bych ji určitě neodmít.*

Z toho můžeme usuzovat, že autor B nevyužil svoji slovní zásobu v celém rozsahu, protože psal jen o omezeném počtu témat. Všimli jsme si také, že autor B na konci každého svého komentáře píše jako jediný podpis, Iv. Tento jev se pravděpodobně neodrazí na metodě Bag of Words, ale může být užitečný pro potenciální kvalitativní analýzu.

Kolik ovcí se vejde na hektar, aby byly sobstačné. Máte nástupce kterého to bude bavit tak jako Vás? Ať se dílo podaří. Iv.

Nejvyšší míru entropie má text A 5, a to 8.500156. Při pohledu do textu zjistíme, že autor napsal 12 komentářů, které se drží jednoho tématu, letectví za druhé světové války. Z toho bychom mohli usuzovat, že entropie bude nízká, protože autor bude používat stále stejná slova dokola. Při pohledu do textu zjistíme, že

jedním z důvodů, proč je entropie naopak vysoká může být, že autor sice používá stále stejná slova, ale dělá v nich chyby. Jako příklad uvádíme tři ukázky z textu A 5, kde autor používá slovo Zeppelin.

*Tentokrát vedle, vedle, vedle...Graf **Zeppelin** L127 byla naopak veleúspěšná vzducholoď.*

*Vy jste si to, pane Ivane, popletl asi se vzducholodí Hindenburg, která tak efektně a příhodně ztroskotala před americkou filmovou kamerou! Graf **Zepellin** měl číslo L127, kdežto nešťastný Hindenburg byl čísla L129.*

***Zepelíny** jsou zvány vzducholodě s pevnou kostrou a pláštěm. Čím více má gramatických chyb, tím víc má slov. Entropie může ouviset nejen s tématičností, ale i překlepy*

Dalším důvodem, proč by mohla být entropie u textu A 5 tak vysoká, je poměrně vysoké množství překlepů a chyb v komentářích autora A:

***H.Göring**, když naznal, že k lepšímu poválečnému postavení jeho **vojemsky** poražené vlasti jednáním nepřispěje, ukončil svůj život **vlatní** rukou také...*

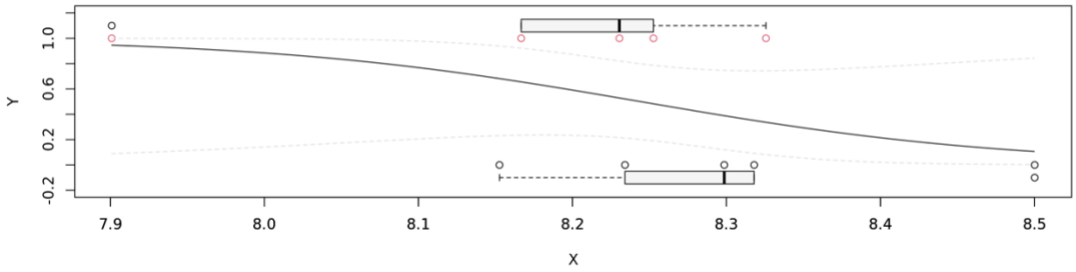
Jako poslední důvod vysoké entropie bychom mohli považovat časté používání odborných a německých výrazů, které se pojí s tématem druhé světové války:

*Jako příklad uvedu třeba nešťastné rozhodnutí, které vedlo ku zdržení nasazení **Me 262 -Schwalbe.** Vývoj nesčetného množství jistě pokrokových a moderních strojů, který však svými náklady podvazoval produkci běžných a docela dobrých letadel frontových. Kromě polského tažení bojovala **Luftwaffe** většinou proti početní přesile.*

Jak můžeme vidět v příkladech, které jsme uvedli v této kapitole, míru entropie ovlivňuje množství faktorů, jde nejen o bohatství slovníku zkoumaného autora, ale také o množství překlepů, používání cizích slov a odborných výrazů, gramatické chyby a další. Vzhledem k množství informací, které zjistíme u entropie očekáváme, že tato metoda bude úspěšná.

V následující části provedeme srovnání jednotlivých dvojic autorů pomocí logistické regrese a statisticky ověříme předpoklad, že rozlišení dvou autorů na základě entropie bude patřit mezi úspěšné metody.

Entropie – srovnání autorů A a B



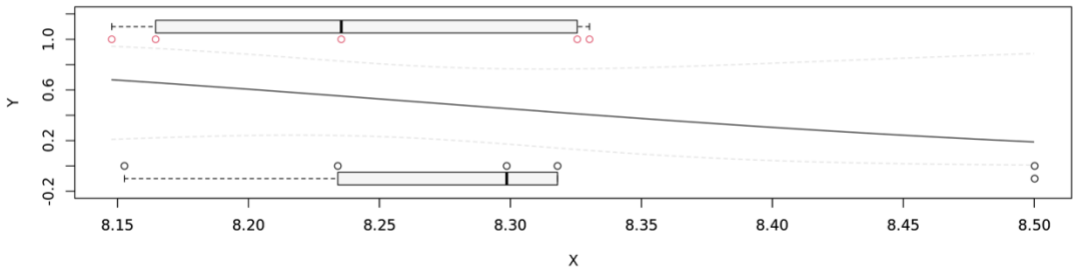
Obrázek 44: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.141486

Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.

Na Obrázku 44 můžeme vidět grafické znázornění výsledků analýzy indexu entropie pomocí logistické regrese, která byla provedena na datech z textů autorů A a B. Model nám umožňuje správně rozlišit autory v 60 % případů. Úspěšnost tohoto modelu není úspěšnější než čistě náhodné rozhodování, jak naznačuje p-hodnota 0.141486.

Entropie – srovnání autorů A a C



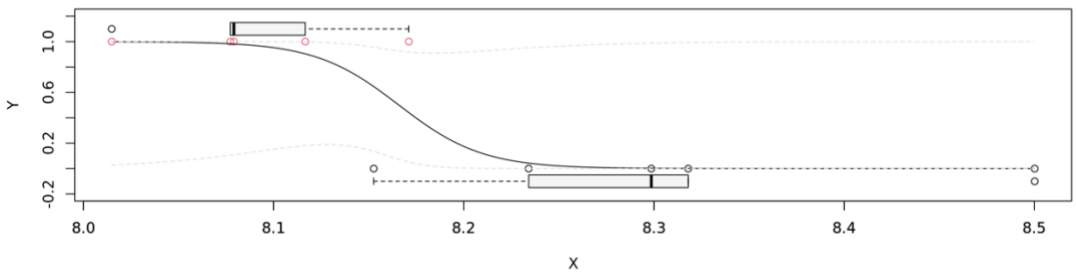
Obrázek 45: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.343659

Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.

Na Obrázku 45 můžeme vidět grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů A a C. S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit pouze v 60 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0.343659, což znamená, že tento model není úspěšnější než náhoda.

Entropie – srovnání autorů A a D



Obrázek 46: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

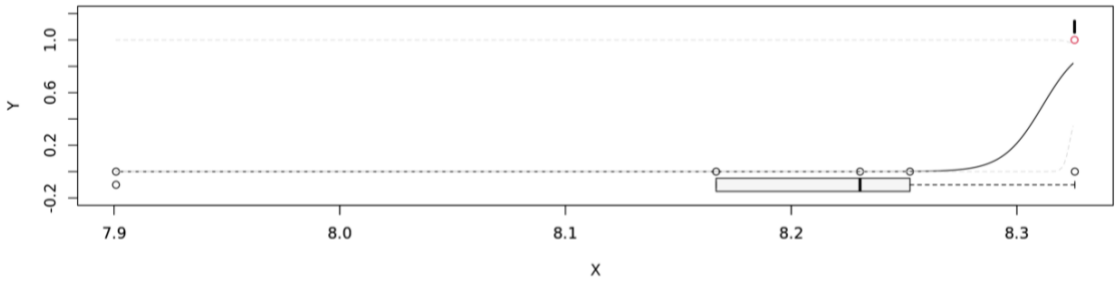
[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.001769

Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Na Obrázku 46 můžeme vidět grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů, které jsme v QUITA vybrali náhodně. S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 80 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0.001769, což znamená, že tento model je úspěšnější než náhoda.

Entropie – srovnání autorů B a C



Obrázek 47: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

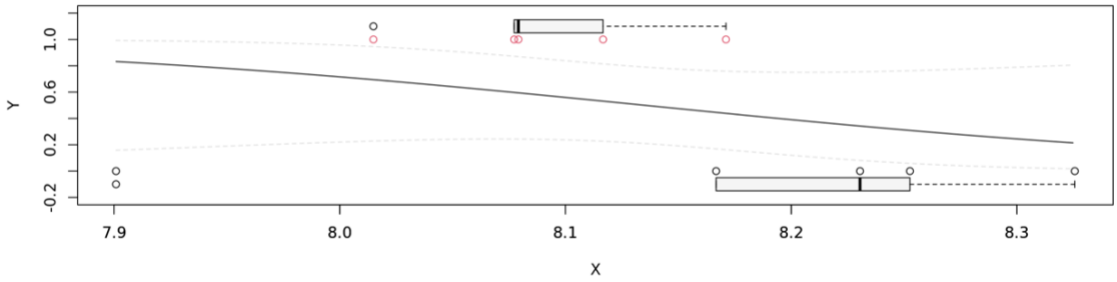
[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.003691

Logistická regrese dokáže správně přiřadit '0' a '1' v 90 % případů.

Na Obrázku 47 můžeme vidět grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů B a C. S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit v 90 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0.003691což znamená, že je tento model velmi úspěšný.

Entropie – srovnání autorů B a D



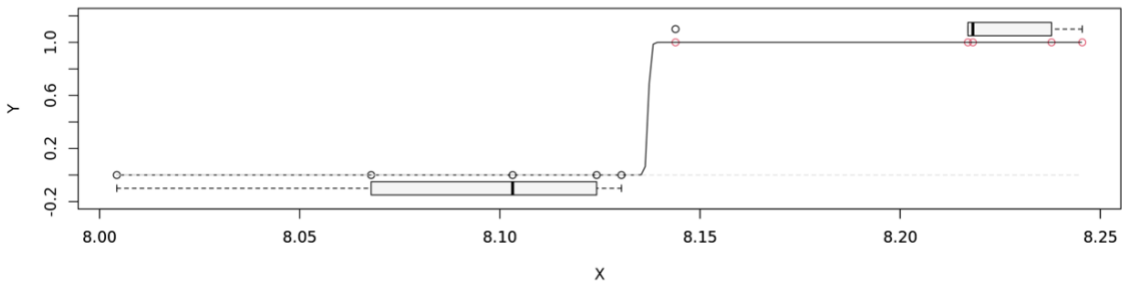
Obrázek 48: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 tokenů.*

[!] Model NENÍ s dodaným regresorem významně lepší než bez něj
(nemá na určení odpovědí významný vliv).
p-hodnota ≈ 0.246794

Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Na Obrázku 48 můžeme vidět grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů B a D. P-hodnota je u tohoto modelu 0.246794, což znamená, že tento model není funkční, a proto není statisticky významný. Logistická regrese sice dokáže správně přiřadit '0' a '1' ve 80 % případů, ale nemáme dost důkazů na to, abychom mohli tvrdit, že jde o úspěšný model, a ne o náhodu.

Entropie – srovnání autorů C a D



Obrázek 49: *Grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů.

Na Obrázku 49 můžeme vidět grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny v textech autorů C a D. S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit ve 100 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0.000197, což znamená, že je tento model velmi úspěšný, protože správně přiřadil všechny autory.

Hodnocení úspěšnosti metody

ENTROPIE	A	B	C	D
A		NE, 60 %	NE, 60 %	ANO, 80 %
B			ANO, 90 %	NE, 80 %
C				ANO, 100 %
D				

Tabulka 10: *Tabulka úspěšnosti jednotlivých pokusů u metody logistické regrese indexu entropie.*

Při použití této metody pro náhodně vybrané tokeny jsme byli schopni od sebe úspěšně odlišit 3 případy ze 6, tedy byla úspěšná v 50 % případů. Průměrná úspěšnost byla kolem 78 %.

Sumarizace a diskuse

V sedmé kapitole jsme definovali pojem entropie a pomocí QUITA jsme vytvořili tabulku míry entropie jednotlivých textů. Pokusili jsme se odhadnout, co je důvodem pro nejnižší entropii u textu B 5 a nejvyšší entropii u textu A 5. Následně jsme provedli srovnání šesti dvojic textů pomocí logistické regrese. I

přes to, že míru entropie ovlivňuje množství faktorů, jako například bohatství slovníku autora, používání cizích slov, množství překlepů a gramatických chyb, se naše očekávání, že entropie dokáže úspěšně odlišit jednotlivé autory, nepotvrdilo.

Tento neúspěch může mít více důvodů. Je možné, že někteří autoři mohou mít podobný styl psaní, který znesnadňuje odlišení jejich textů. Mohou mít podobnou slovní zásobu, větnou strukturu nebo preferované fráze, což snižuje rozmanitost a tím i entropii jejich textů. Výsledky mohou být ovlivněny i tím, že nebyly použity dostatečně reprezentativní vzorky textů od každého autora. Pokud je vzorek textů malý nebo nedostatečně různorodý, může to vést k nedostatečné přesnosti v odhadu entropie a tím i v identifikaci autorů.

Entropie v našem případě dokázala odlišit pouze polovinu autorů. Vzhledem k této nedostatečné spolehlivosti není tato metoda vhodná pro aplikaci ve forenzní praxi, kde je vyžadována vysoká míra spolehlivosti a přesnosti při identifikaci autorů na základě psaného materiálu.

8. Hapax legomenon

Slovo „hapax legomenon“ pochází z transliterace řeckého výrazu ἅπαξ λεγόμενον, což doslovně znamená „řečené pouze jednou“. Výskyt hapax legomenon ve starší literatuře je obtížně zjistitelný, zejména kvůli omezené dostupnosti dochovaných textů. Existuje tedy pravděpodobnost, že některá slova považovaná za hapaxy mohou být nesprávně identifikována kvůli nedostatečné evidenci. Znalost hapax legomenon by mohla být využita při identifikaci autorství literárních děl, protože výskyt hapaxů bývá často spojen s konkrétním dílem nebo částí autorské tvorby. Například dramatické dílo Williama Shakespeara obsahuje podobné procentuální zastoupení hapax legomenon v celé své rozmanitosti, což může sloužit jako jedno z identifikačních kritérií při analýze autorství.[22]

Pomocí QUITA jsme znovu vytvořili Bag of Words, ale ponechali jsme pouze hapaxy. V prvním případě jsou hapaxy vybrány náhodně, což znamená, že Bag of Words obsahuje pouze jedinečná slova z celého textu bez ohledu na jejich pořadí nebo výskyt v rámci textu. Tato metoda nám umožňuje získat přehled o celkovém bohatství slovní zásoby textu a identifikovat jedinečné výrazy, které nejsou běžně používány.

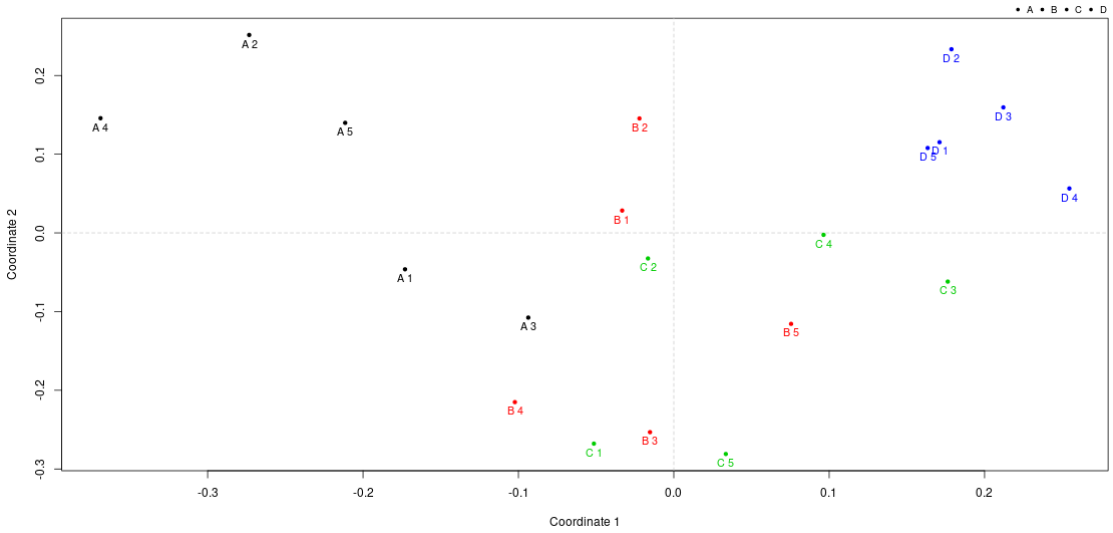
Ve druhém případě jsou vybrány hapaxy jdoucí po sobě, což znamená, že Bag of Words obsahuje pouze jedinečná slova, která jdou za sebou bez opakování. Tento přístup nám umožňuje analyzovat text z hlediska posloupnosti jedinečných výrazů a identifikovat sekvence slov, které mohou nést zvláštní význam nebo tematickou souvislost.

V podkapitole 8.1. se pokusíme mezi sebou vizuálně porovnat dvojice autorů pomocí multidimenzionálního škálování a následně odhad statisticky ověřit pomocí logistické regrese.

[22] VARJASSYOVÁ, Ivana. *Hapax legomenon vs. nonce word*. Online. In: Encyklopedie lingvistiky. 2014. Dostupné z: https://encyklopedieoltk.upol.cz/encyklopedie/index.php5/Hapax_legomenon_vs.html. [cit. 2024-05-02].

8.1. Hapaxy náhodně vybrané

V ideálním případě by graf obsahoval čtyři shluky textů jednotlivých autorů, které jdou od sebe snadno odlišit. To se v případě naší analýzy nestalo, ale i přesto můžeme v grafu vidět některé shluky, které odpovídají jednotlivým autorům.



Obrázek 50: *Analýza hapax legomenon dvaceti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF 0.128573

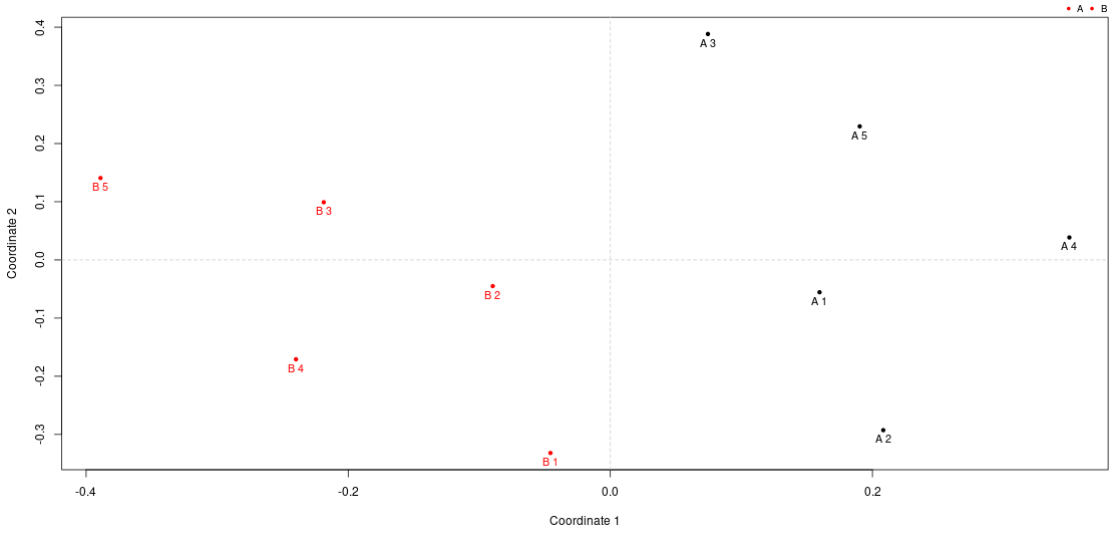
V případě této analýzy MDS zrekonstruovalo přibližně 12% variance vypočítaných vzdáleností, což znamená, že některé informace o podobnosti či vzdálenosti zadaných textů nejsou vidět.

Nejúspěšnější bylo odlišení autora D, jeho texty se shlukly v pravém horním rohu a tvoří samostatnou skupinu. Znamená to, že je autor ve svém vyjadřování konzistentní a veškeré jeho texty jsou si velmi podobné.

Texty autora A jsou všechny v levé části grafu, ale text a A3 už je velmi blízko autorům B a C a nejdou od nich dobře oddělit. Stále je ale autor A dobře odlišitelný od zbytku, takže je podobně jako autor D konzistentní a jeho texty jsou si podobné. Autoři B a C jsou společně ve shluku ve středu obrázku a není

možné je od sebe lineárně oddělit. Znamená to, že pravděpodobně mají výrazné společné rysy.

Srovnání autorů A a B – náhodně vybrané tokeny

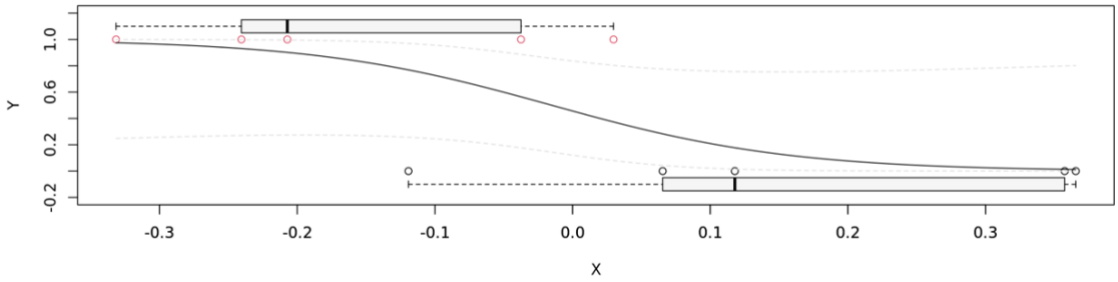


Obrázek 51: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF 0.2430741

Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora A a v levé části také shluk pěti textů od autora B. Autory je od sebe možné velice snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 24% variance vypočítaných vzdáleností.



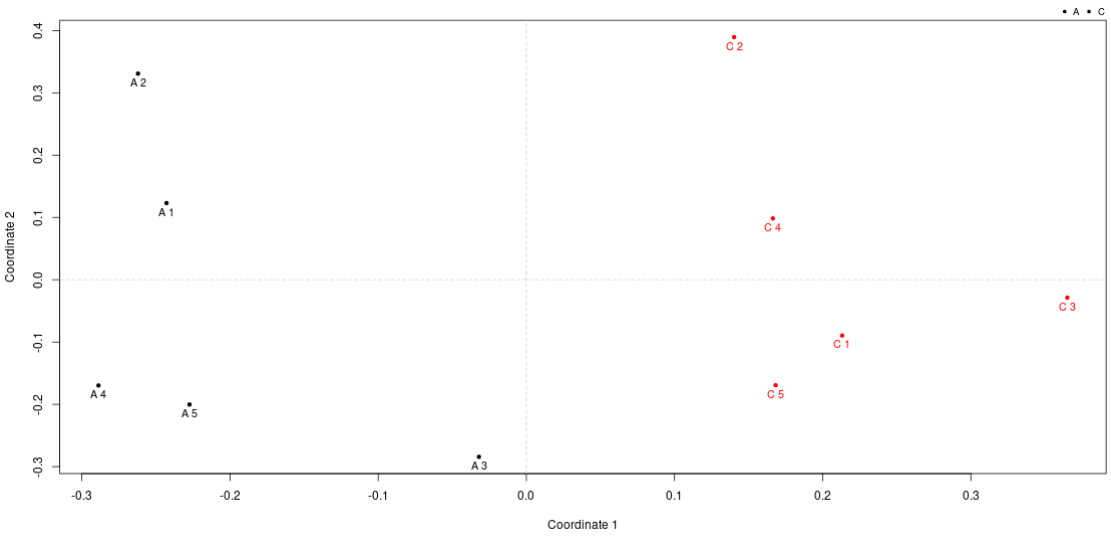
Obrázek 52: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.012

P-hodnota je u tohoto modelu 0.012, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 80 % případů.

Srovnání autorů A a C – náhodně vybrané tokeny

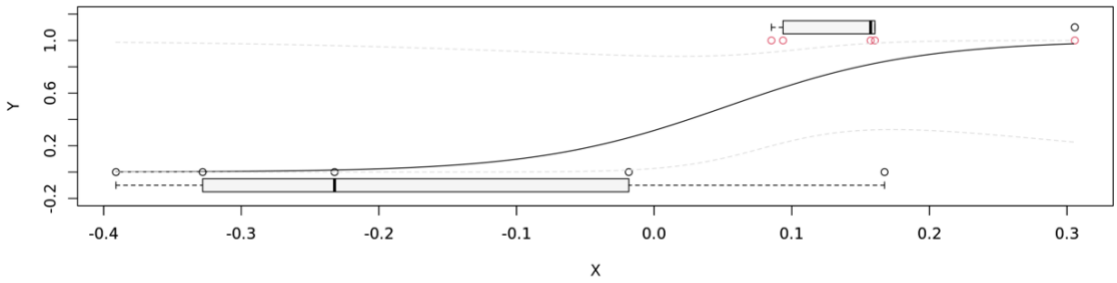


Obrázek 53: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF [1] 0.2518509

Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora C a v levé části také shluk pěti textů od autora A. Autory je od sebe možné velice snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



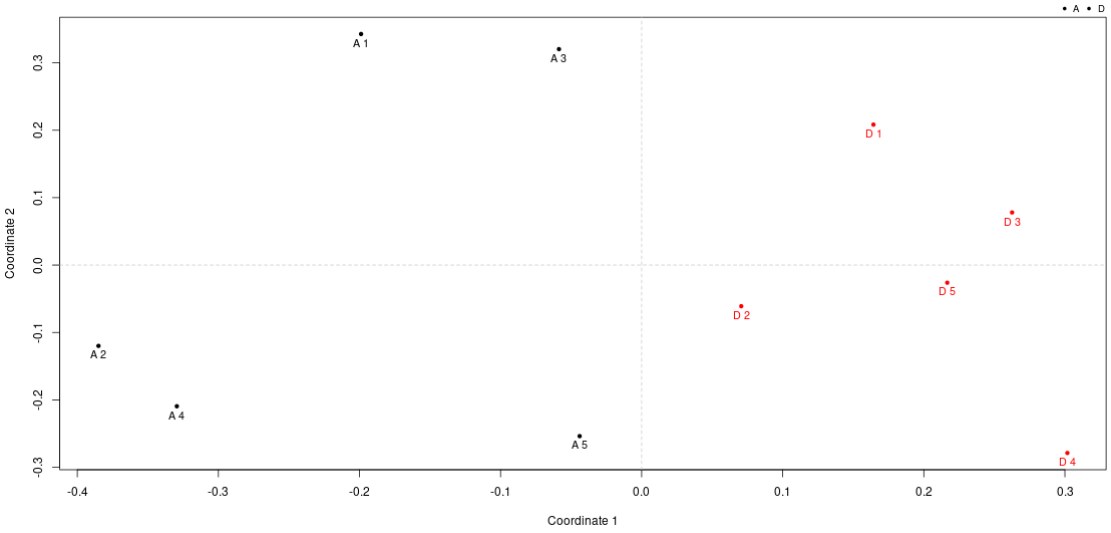
Obrázek 54: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.008821

P-hodnota je u tohoto modelu 0.008821, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 90 % případů.

Srovnání autorů A a D – náhodně vybrané tokeny



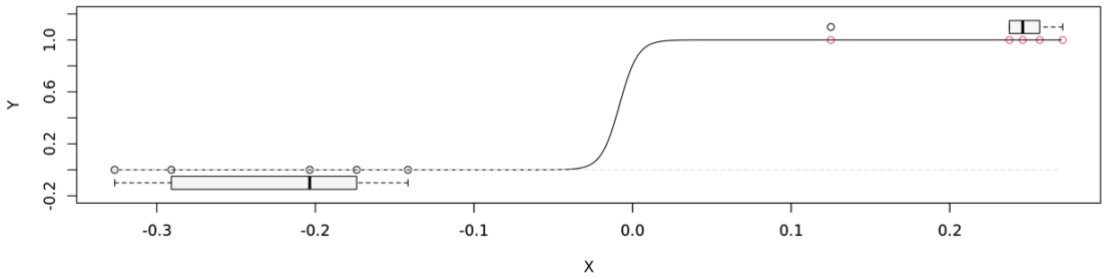
Obrázek 55: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS)*

$GOF

[1] 0.2552338

V grafu vidíme dva výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora D a v levé části také shluk pěti textů od autora A. Autory je od sebe možné velice snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



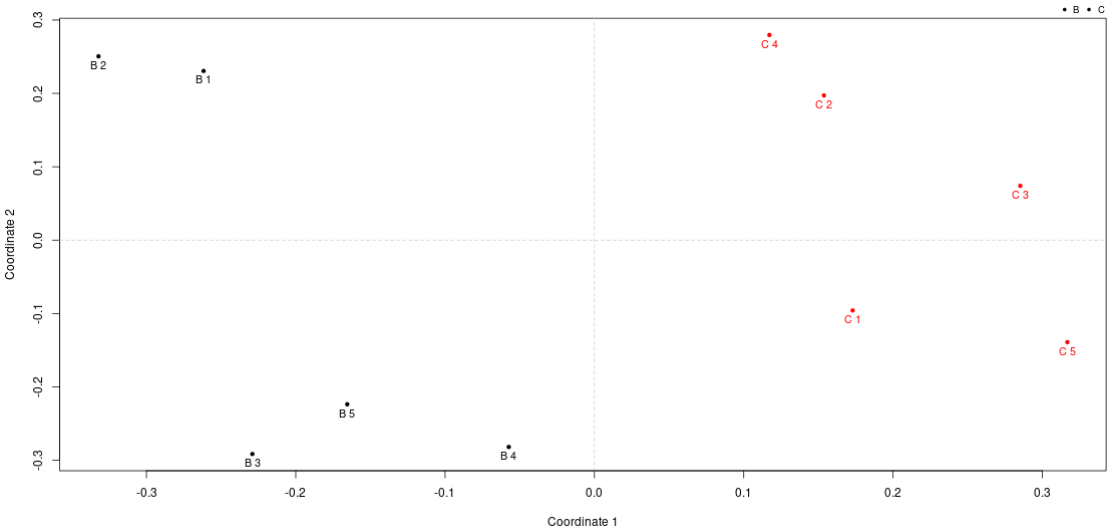
Obrázek 56: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů B a C – náhodně vybrané tokeny



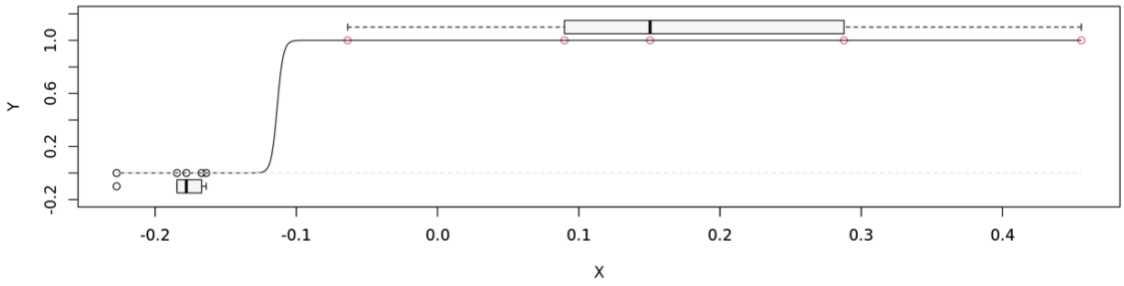
Obrázek 57: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

[1] 0.2523262

Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora C a v levé části také shluk pěti textů od autora B. Autory je od sebe možné velice snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



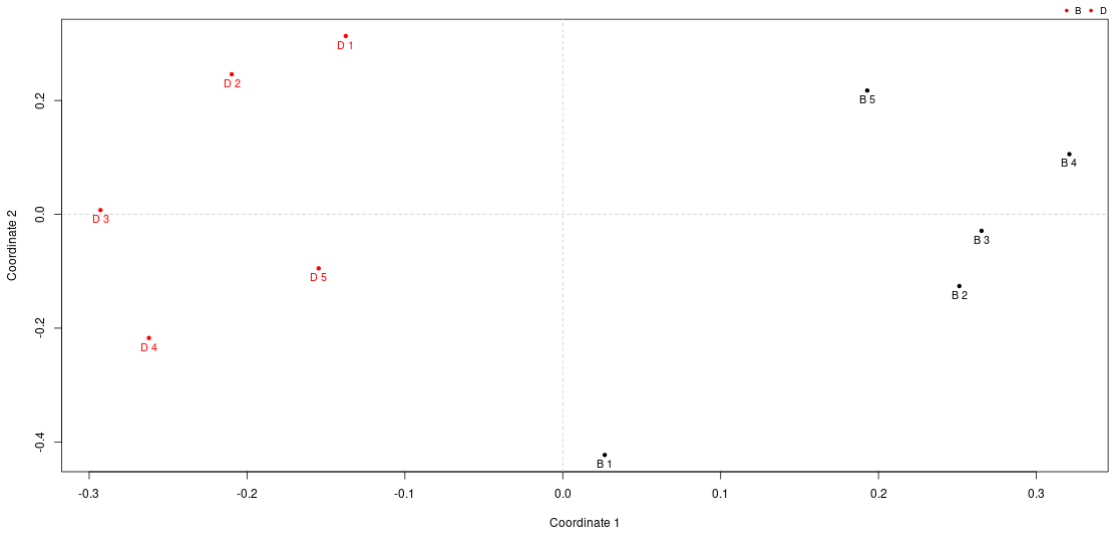
Obrázek 58: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů B a D – náhodně vybrané tokeny



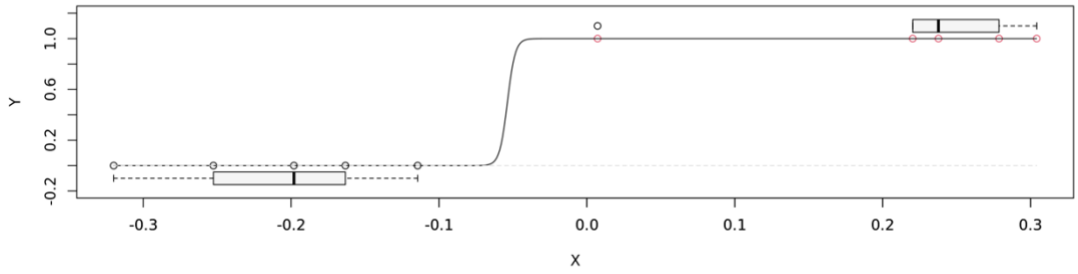
Obrázek 59: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

[1] 0.2503406

Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora B a v levé části také shluk pěti textů od autora D. Autory je od sebe možné snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



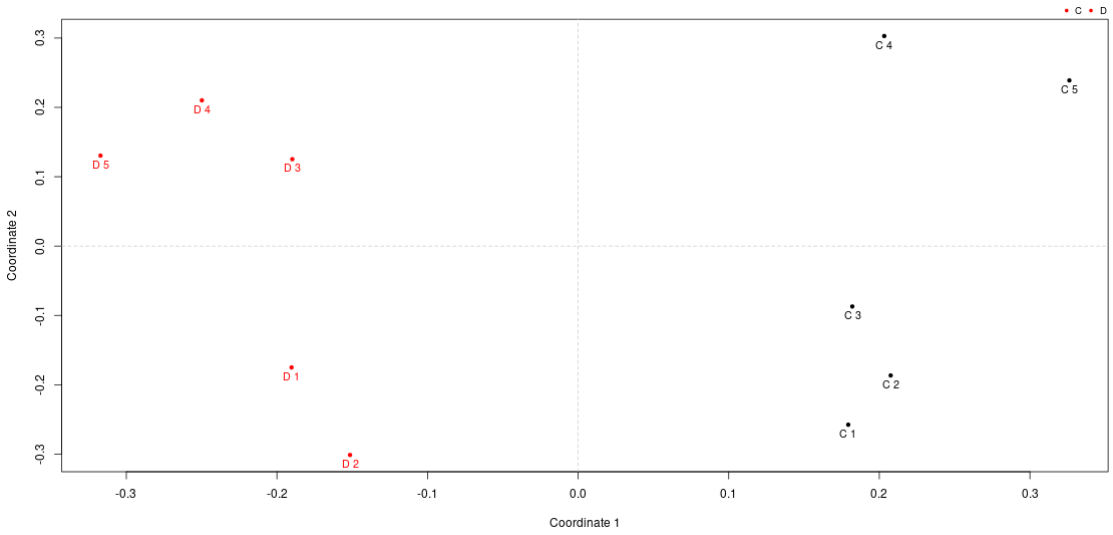
Obrázek 60: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů C a D – náhodně vybrané tokeny



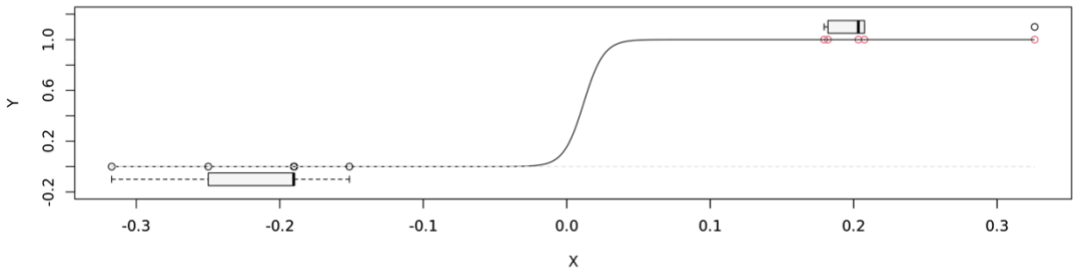
Obrázek 61: *Analýza hapax legomenon deseti zkoumaných textů o velikosti 618 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

[1] 0.2523933

Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk pěti textů od autora D a v levé části také shluk pěti textů od autora C. Jednotlivé autory je od sebe možné velice snadno lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



Obrázek 62: *Grafické znázornění výsledků logistické regrese při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Hodnocení úspěšnosti metody

HAPAXY náhodně	A	B	C	D
A		ANO, 80 %	ANO, 90 %	ANO, 100 %
B			ANO, 100 %	ANO, 100 %
C				ANO, 100 %
D				

Tabulka 11: *Tabulka úspěšnosti jednotlivých pokusů u metody Bag of Words pro náhodně vybrané hapax legomenon. Texty mají stejnou délku, 618 tokenů.*

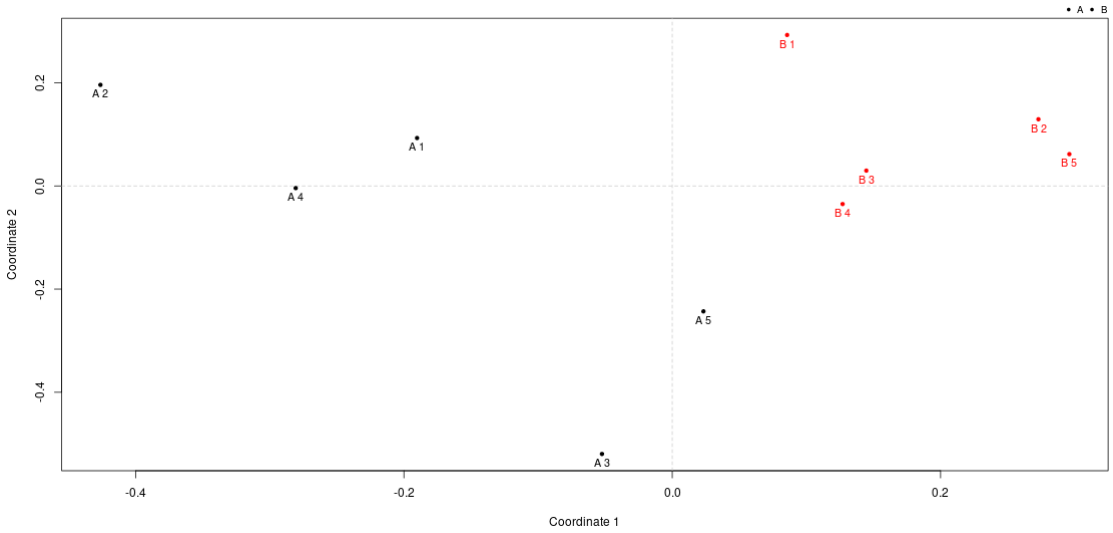
Při použití logistické regrese pro náhodně vybrané hapax legomenon jsme zjistili, že je tato metoda velmi úspěšná, protože dokázala odlišit autory v 6

případech ze 6 čili ve 100 % případů. Průměrná úspěšnost pokusů byla 95 %. Tato metoda je tedy z námi testovaných nejúspěšnější.

8.2. Hapaxy po sobě jdoucí

V této podkapitole se pokusíme mezi sebou porovnat dvojice autorů pomocí multidimenzionálního škálování a následně odhad statisticky ověřit pomocí logistické regrese.

Srovnání autorů A a B – po sobě jdoucí tokeny



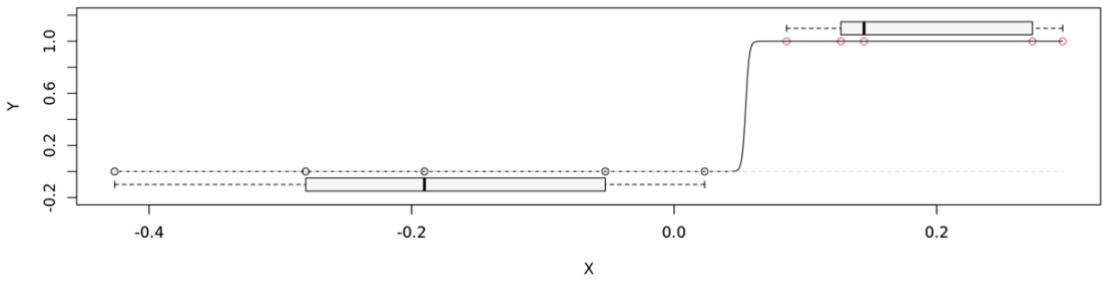
Obrázek 63: *Analýza hapax legomenon deseti zkoumaných textů autorů A a B o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

0.2460501

Texty byly rozděleny na dvě poloviny, v pravé části obrázku se nachází shluk pěti textů od autora B a v levé části také shluk pěti textů od autora A. Autory je od sebe možné poměrně snadno lineárně oddělit. Texty autora B jsou si navzájem blíž, než texty autora A, můžeme tedy odhadovat, že jsou si navzájem podobnější. Můžeme odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



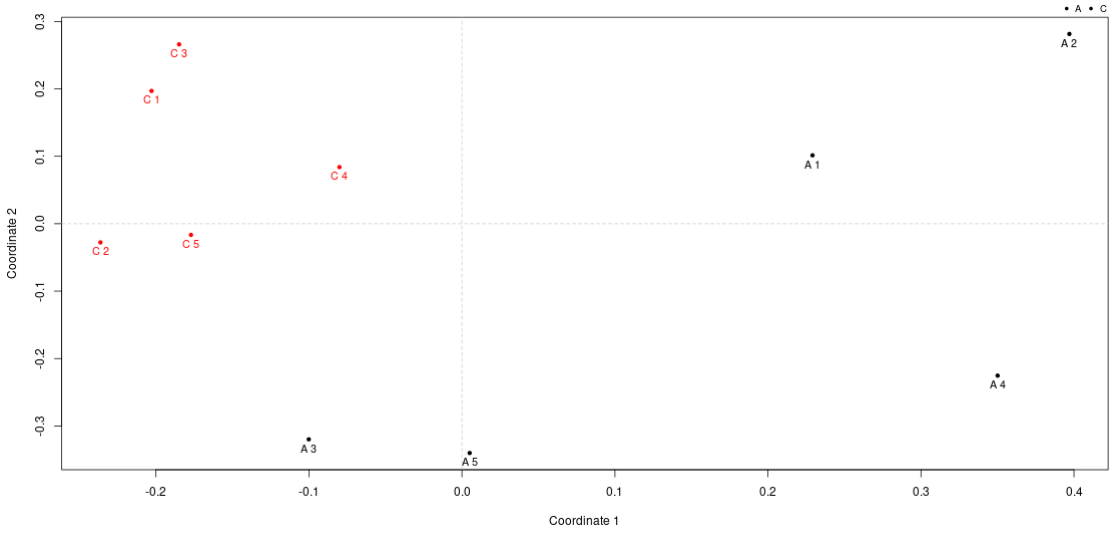
Obrázek 64: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a B, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů A a C – po sobě jdoucí tokeny



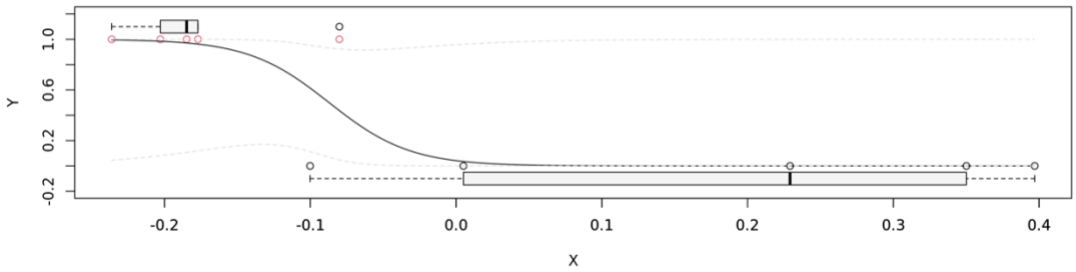
Obrázek 65: *Analýza hapax legomenon deseti zkoumaných textů autorů A a C o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

0.2473364

Můžeme si všimnout jednoho výrazně odděleného shluků textů v levém horním rohu grafu, kde se nachází všechny texty od autora C. Texty jsou si blízko, můžeme tedy předpokládat, že jsou si navzájem podobné. Texty autora A jsou rozmístěny po obou polovinách grafu, nepříliš blízko sobě navzájem ani autorovi C. Stále je však možné jednotlivé autory od sebe odlišit. Můžeme předpokládat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



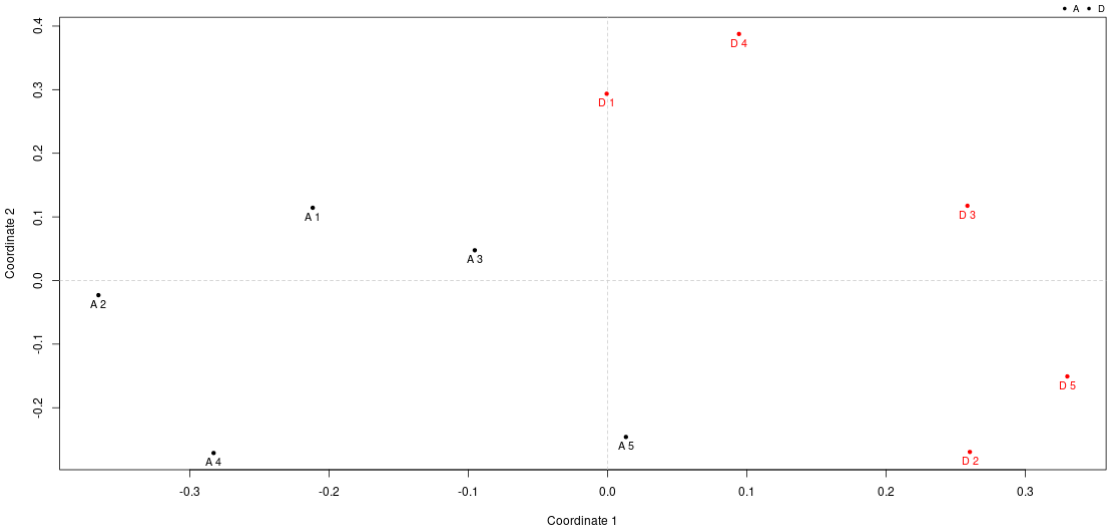
Obrázek 66: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.001523

P-hodnota je u tohoto modelu 0.001523, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Srovnání autorů A a D – po sobě jdoucí tokeny



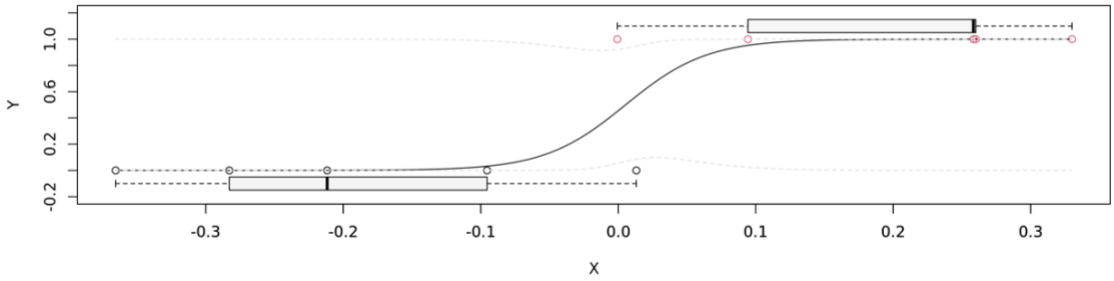
Obrázek 67: *Analýza hapax legomenon deseti zkoumaných textů autorů A a D o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

0.2506974

Texty byly rozděleny přibližně na dvě poloviny, v pravé části obrázku se nachází všech pět textů od autora D a v levé části také shluk pěti textů od autora A. Autory je stále od sebe možné lineárně oddělit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



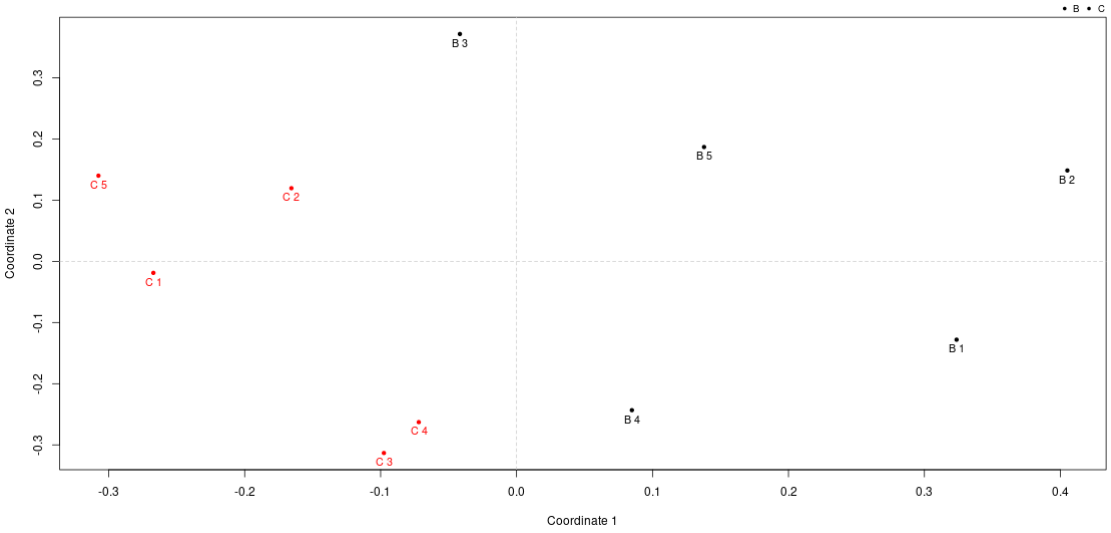
Obrázek 68: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů A a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.001239

P-hodnota je u tohoto modelu 0.001239, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Srovnání autorů B a C – po sobě jdoucí tokeny



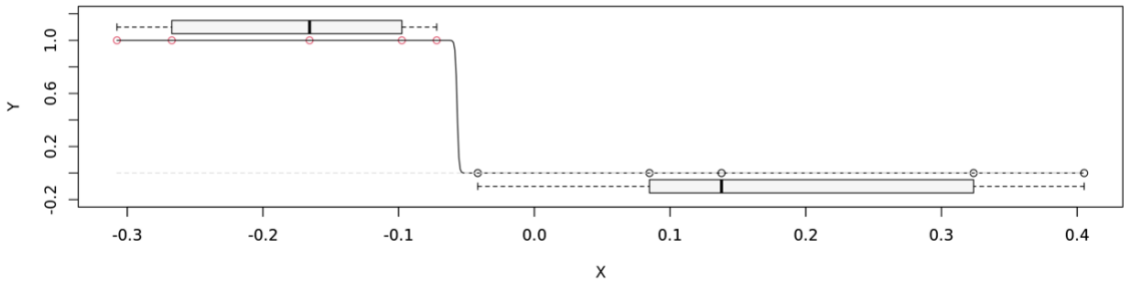
Obrázek 69: *Analýza hapax legomenon deseti zkoumaných textů autorů B a C o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

0.2474962

Texty byly rozděleny na dvě poloviny, v pravé části obrázku se nachází shluk pěti textů od autora B a v levé části také shluk pěti textů od autora C. Autory je od sebe možné lineárně oddělit. Texty autora C jsou si navzájem blíž než texty autora B, můžeme tedy odhadovat, že jsou si navzájem podobnější. Můžeme předpokládat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



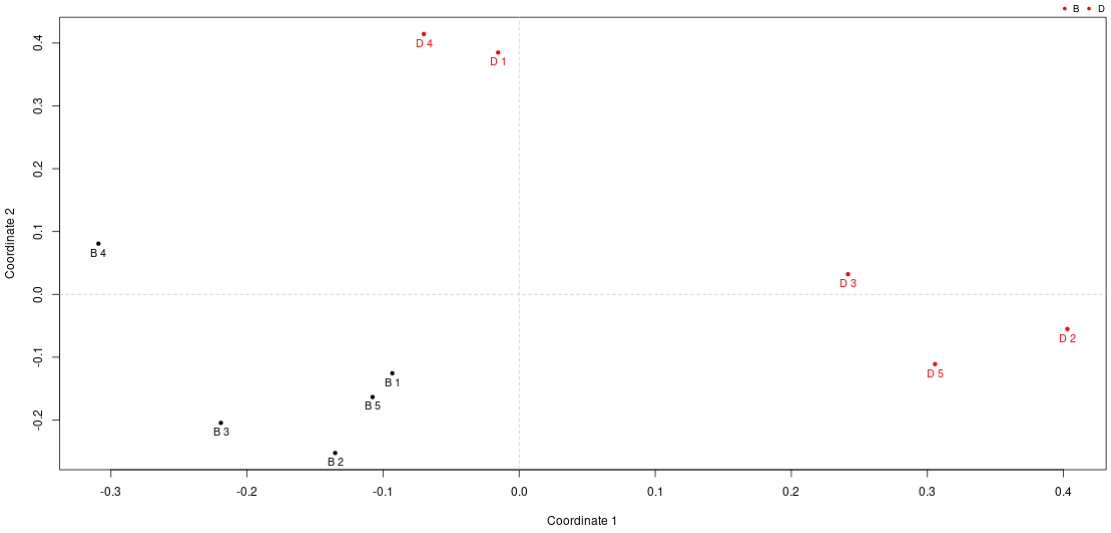
Obrázek 70: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a C, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů B a D – po sobě jdoucí tokeny



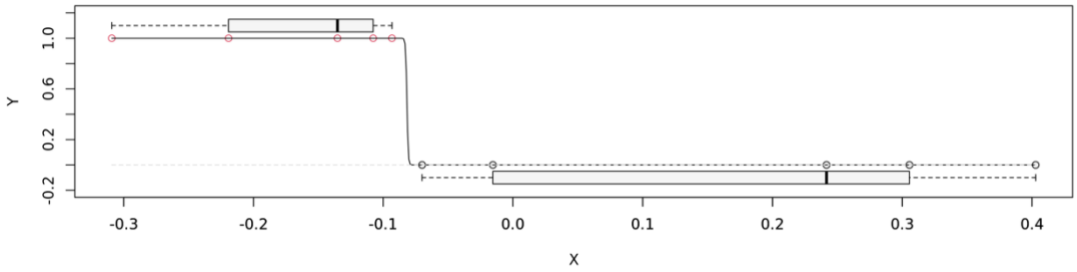
Obrázek 71: *Analýza hapax legomenon deseti zkoumaných textů autorů B a D o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF

0.2479899

Můžeme si všimnout jednoho výrazně odděleného shluků textů v levém dolním rohu grafu, kde se nachází všechny texty od autora B. Texty jsou si blízko, můžeme tedy předpokládat, že jsou si navzájem podobné. Texty autora D jsou rozmístěny po obou polovinách grafu, nepříliš blízko sobě navzájem ani autorovi B. Stále je však možné jednotlivé autory od sebe odlišit. Můžeme tedy odhadovat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 25% variance vypočítaných vzdáleností.



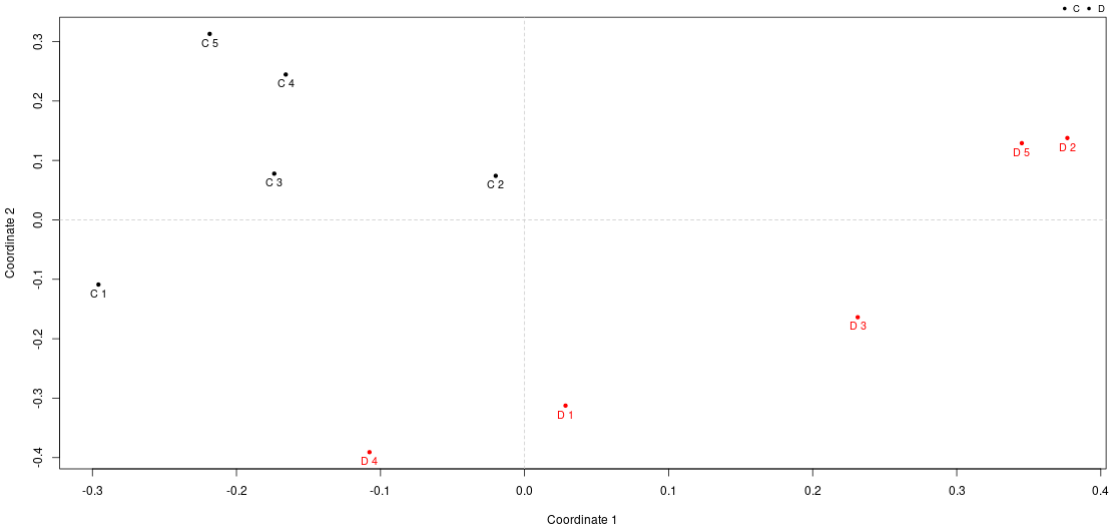
Obrázek 72: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů B a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.

p-hodnota ≈ 0.000197

P-hodnota je u tohoto modelu 0.000197, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' ve 100 % případů, a navíc je úspěšná i vůči striktní Bonferoniho korekci, protože p-hodnota vychází nižší než 0,001.

Srovnání autorů C a D – po sobě jdoucí tokeny

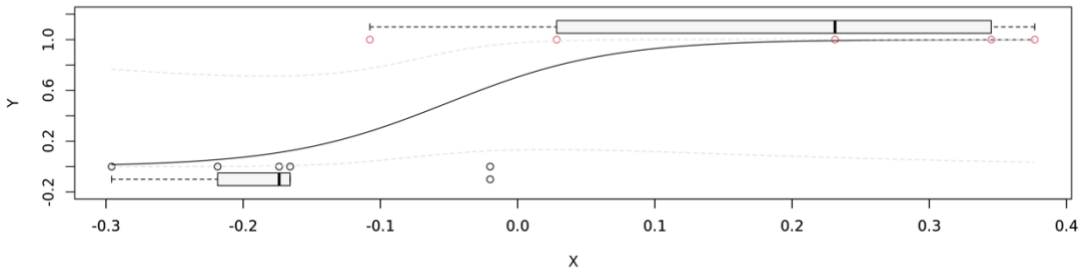


Obrázek 73: *Analýza hapax legomenon deseti zkoumaných textů autorů C a D o velikosti 618 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*

$GOF
0.2551415

Můžeme si všimnout jednoho výrazně odděleného shluků textů v levém horním rohu grafu, kde se nachází všechny texty od autora C. Texty jsou si blízko, můžeme tedy předpokládat, že jsou si navzájem podobné. Texty autora D jsou rozmístěny po obou polovinách grafu, nepříliš blízko sobě navzájem ani autorovi C. Stále je však možné jednotlivé autory od sebe odlišit. Můžeme předpokládat, že se pomocí logistické regrese podaří vytvořit funkční model.

V případě této analýzy MDS zrekonstruovalo přibližně 26% variance vypočítaných vzdáleností.



Obrázek 74: *Grafické znázornění výsledků logistické regrese pro po sobě jdoucí hapax legomenon při použití metody Bag of Words. Hodnoty byly naměřeny v textech autorů C a D, použité texty mají stejnou délku, 618 tokenů.*

[✓] Model je s dodaným regresorem významně lepší než bez něj.
p-hodnota ≈ 0.004225

P-hodnota je u tohoto modelu 0.004225, což znamená, že tento model je funkční, a proto je statisticky významný. Logistická regrese dokáže správně přiřadit '0' a '1' v 80 % případů.

Hodnocení úspěšnosti metody

HAPAXY jdoucí po sobě	A	B	C	D
		ANO, 100 %	ANO, 80 %	ANO, 80 %
B			ANO, 100 %	ANO, 100 %
C				ANO, 80 %
D				

Tabulka 12: *Tabulka úspěšnosti jednotlivých pokusů u metody Bag of Words pro po sobě jdoucí hapax legomenon. Texty mají stejnou délku, 618 tokenů.*

Při použití logistické regrese pro po sobě jdoucí hapax legomenon jsme zjistili, že je tato metoda velmi úspěšná, protože dokázala odlišit autory v 6 případech ze 6. Průměrná úspěšnost je 90 %. Stejně jako při použití náhodně vybraných hapax legomenon, logistická regrese dokázala autory odlišit ve 100 % případů, nicméně při použití náhodně vybraných tokenů měla úspěšnost 95 %.

Sumarizace a diskuse

V deváté kapitole jsme čtenáři představili pojem hapax legomenon čili slova, která se vyskytují pouze jednou v rámci zkoumaného textového korpusu. Provedli jsme zobrazení podobnosti dvojic textů pomocí MDS a následně analýzu pomocí logistické regrese.

Identifikace hapax legomenon může být užitečná při analýze autorství literárních děl, protože jedinečná slova často ukazují specifické autorské rysy, což má potenciální význam pro aplikaci v oblasti forenzní lingvistiky. Výsledky analýzy provedené v deváté kapitole naznačují, že tato metoda je velmi úspěšná i u mnohem kratších textů, než jsou knihy, články nebo další literární díla, protože metoda dokázala spolehlivě odlišit autory ve 100 % případů.

Přestože analýza hapax legomenon může být užitečným nástrojem pro rozlišení autorů v internetových komentářích, je důležité si uvědomit, že tato metoda má své limity. Například pokud autoři používají podobnou slovní zásobu nebo se zabývají podobnými tématy, může být obtížné získat dostatečné množství materiálu, protože slovo použité v textu vícekrát, než jednou přestává být hapaxem a tedy materiálem vhodným k analýze.

Porovnání všech výsledků praktické části

Pro lepší přehlednost při hodnocení a porovnávání v následující části budou veškeré výsledky získané v praktické části systematicky uspořádány podle úspěšnosti použitých metod, v dalším sloupci pak můžeme vidět úspěšnost metod vyjádřených procenty a úspěšnost jednotlivých pokusů vyjádřených procenty.

Zkoumaná metoda	**Signifikance modelu metod párů autorů (6/6 párů autorů = 100 %)**	**Průměrná úspěšnost modelu**
Hapaxy náhodně vybrané tokeny	100 %	95 %
Hapaxy po sobě jdoucí tokeny	100 %	90 %
BOW náhodně vybrané tokeny	83 %	82 %
Zvýraznění slov	80 %	84 %
Emotikony	67 %	77 %
Počet teček za větou	67 %	87 %
Entropie	50 %	78 %
BOW po sobě jdoucí tokeny	17 %	73 %

Tabulka 13: *Tabulka úspěšnosti všech testovaných metod.*

V tabulce úspěšnosti vidíme, že šest metod mělo vyšší úspěšnost než 50 % čili jsou úspěšnější než náhoda. Metody používající Bag of Words mohou být ovlivněny nejen slovníkem autora, ale i tématem, o kterém autor píše. Nemůžeme vědět s jistotou, zda je metoda úspěšná díky autorskému stylu, nebo volbou tématu. Metody, které reflektují autorův styl jsou zvýraznění slov, emotikony a počet teček za větou. Ty sice vykazují nižší úspěšnost, ale máme větší jistotu, že došlo k rozlišení autorů na základě jejich charakteristických rysů. Proto je tato metoda vhodná pro identifikaci autora ve forenzní lingvistice.

9. Závěr

Cílem praktické části bylo představit existující metody identifikace autorů a navrhnout způsoby, jak efektivně rozšířit stávající metodiku pomocí vícerozměrných přístupů. Provedli jsme analýzu, která kombinovala kvantitativní a kvalitativní přístupy, abychom stanovili tři kritéria hodnocení dvaceti textů čtyř autorů (označených A, B, C a D). Na základě této analýzy jsme formulovali hypotézy o tom, jaká kritéria by mohla být využita k odlišení mezi těmito autory. Následně jsme se pokusili pomocí kvantitativních metod tyto hypotézy ověřit nebo vyvrátit.

V šesté kapitole jsme popsali metodu Bag of Words a metodu vícerozměrného škálování. Interpretací jednotlivých grafů vícerozměrného škálování se jsme se pokusili odhadnout, jak dobře je možné jednotlivé autory oddělit od sebe. Ukázalo se, že některé texty výrazně vybočují od ostatních, podívali jsme se proto do jejich Bag of Words a pokusili se přijít na to, v čem jsou odlišné od ostatních. Podle našeho očekávání se potvrdilo, že autoři nejvíce používají předložky, spojky, zájmena, a další funkční slova. Při pohledu do textů jsme zjistili, že autoři si často odpovídají navzájem, odkazují se na sebe nebo na svoje komentáře. Toto zjištění ukazuje výrazný rozdíl mezi spontánně napsanými komentáři na internetu a dalšími texty, jako jsou například novinové články, knihy, hesla v encyklopediích a další, které prošly korekturou a editací. Je tedy důležité porovnávat mezi sebou pouze texty stejného typu. Tomuto tématu se velmi podrobně věnují Venglařová a Matlach v článku Beyond content: discriminatory power of function words in text type classification (2024). [23]

[23] VENGLAŘOVÁ, Klára a MATLACH, Vladimír. Beyond content: discriminatory power of function words in text type classification. Online. Digital Scholarship in the Humanities. 2024. Dostupné z: https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqae013/7634746. [cit. 2024-05-09].

Následně jsme v této kapitole provedli analýzu pomocí logistické regrese. Analýza pomocí Bag of Words se ukázala jako úspěšná, podařilo se nám správně odlišit autory v 5 případech ze 6.

Cílem sedmé kapitoly bylo popsat metodu hodnocení podle indexu entropie. Byla provedena analýza textů. Analýza pomocí entropie ukázala menší úspěšnost než analýza pomocí modelu Bag-of-Words, jelikož úspěšně rozdělila autory pouze ve třech z šesti případů. Důvodem, proč tato metoda nedosáhla očekávaných výsledků, může být skutečnost, že i když existuje rozdíl v entropii mezi různými autory, tento rozdíl není dostatečně výrazný nebo stabilní, aby umožnil spolehlivé odlišení. Ukázalo se, že hodnoty entropie mezi zkoumanými autory se překrývají, což komplikuje spolehlivou identifikaci autorů. Někteří autoři mohou mít proměnlivý styl psaní, což vede k většímu překryvu hodnot entropie mezi nimi.

Každý autor má tendenci používat jedinečná slova a fráze, které se mohou objevit pouze v jeho textech. Proto jsme v osmé kapitole seznámili čtenáře s hapax legomenon. Identifikace hapax legomenon může být užitečná při analýze autorství literárních děl, protože jedinečná slova často ukazují specifické autorské rysy. Cílem této kapitoly bylo zjistit, zda lze tuto metodu použít i na tak krátké texty. Znovu jsme provedli analýzu pomocí Bag-of-Words modelu, ale v tomto případě za použití pouze hapaxů získaných pomocí lingvistického softwaru QUITA. Při pohledu do Bag of Words jednotlivých autorů jsme zjistili, že část slov, které spadají pod hapaxy nejsou hapaxy, protože jde o jedinečný styl autora, ale protože se jedná o překlepy nebo chybně napsaná slova. Příkladem můžou být například *"oslovojí", "babtistů" , "vynázce"* od autora A. Dalším důvodem, proč byla některá slova označena za hapaxy byla chybějící interpunkce jako například u slov *"komentare", "ostatni"* nebo *"zidovska"*. Jak už bylo zmíněno v předchozí kapitole, literární díla často prochází editací a korekturou, které mají za cíl zdokonalit jejich kvalitu a konzistenci. Hapax legomena získaná například z románů, jsou v tomto kontextu pravděpodobně spíše výsledkem autorova stylu než pouhé náhody, jako bývají například překlepy nebo chyby v interpunkci.

Analýza hapax legomenon se ukázal jako nejvíce úspěšná, pomocí logistické regrese se podařilo odlišit autory v 6 případech ze 6 při použití náhodně vybraných i po sobě jdoucích tokenů. Ukázalo se tedy, že odlišení autorů pomocí hapaxů může fungovat i na kratších textech, jako jsou internetové komentáře. Nicméně je důležité zvážit, zda tento úspěch nebyl ovlivněn specifickým výběrem textů a autorů. Bylo by vhodné provést analýzu s jiným souborem internetových komentářů, aby se ověřila konzistence výsledků.

Vytyčeného cíle praktické části, kterým bylo seznámení čtenáře se způsoby využití vícerozměrných metod, použití logistické regrese a jejich aplikace na dané případy, byl splněn.

Výzkum by bylo možné rozšířit několika způsoby. Pokud bychom se rozhodli použít diskuse pod více různými články, byly by korpusy textů tematicky rozmanitější a získali bychom více informací o bohatosti slovníku jednotlivých autorů. Dále by bylo možné provést pokusy s více autory a tím bychom získali více párů pro testování. Jako poslední možnost navrhujeme zkusit stejné experimenty s ještě kratšími texty než 618 tokenů, a sledovat, zda se úspěšnost mění. Samozřejmě se nabízí možnost provádět pokusy i s delšími texty, nicméně výzkumů s identifikací autora dlouhých textů, například románů, už proběhlo větší množství.

Literatura a zdroje

CVRČEK, Václav. *Token*. Online. CzechEncy: Nový encyklopedický slovník češtiny. Brno, 2024. Dostupné z: https://www.czechency.org/slovnik/TOKEN. [cit. 2024-05-09].

ČECH, Radek; POPESCU, Ioan-Ioviț a ALTMANN, Gabriel. *Metody kvantitativní analýzy (nejen) básnických textů*. Qfwfq. Olomouc: Univerzita Palackého v Olomouci, 2014. ISBN 978-80-244-4044–6.

ČESKO, 2009. Zákon 40/2009 Sb. (trestní zákoník). Online. In: Sbírka zákonů. Částka 11. Dostupné také z: https://www.e-sbirka.cz/sb/2009/40?zalozka=text.

Elliot Rodger: How misogynist killer became 'incel hero'. Online. In: BBC News Services. 2018. Dostupné z: https://www.bbc.com/news/world-us-canada-43892189. [cit. 2024-05-02].

FALTÝNEK, Dan, Dalibor PAVLAS, Ondřej VRABEL a Vladimír MATLACH. Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu. Olomouc: Univerzita Palackého, 2015

Federal bureau of investigation: The Unabomber. Online. In: FBI Bureau of Investigation. Dostupné z: https://www.fbi.gov/history/famous-cases/unabomber. [cit. 2024-05-02].

Forensic linguistics intelligence. Online. Obituary: John Olsson. Dostupné z: https://www.thetext.co.uk/obituary. [cit. 2024-04-22].

Historie rukopisného sporu. Online. ČESKÁ SPOLEČNOST RUKOPISNÁ. Rukopisy Královédvorský a Zelenohroský. Dostupné z: https://www.rukopisy-rkz.cz/rkz/histsporu/hist-sp.htm. [cit. 2024-05-09].

International AssociationofForensicLinguistics [online]. [cit. 2024-04-21]. Dostupné z: https://www.iafl.com/about/

JURKA, Michal a FALTÝNEK, Dan. *Forenzní lingvistika*. Online. CzechEncy: Nový encyklopedický slovník češtiny. Brno, 2024. Dostupné z: https://www.czechency.org/slovnik/FORENZN%C3%8D%20LINGVISTIKA. [cit. 2024-05-09].

KELLY, John. *Christopher Dorner: What made a police officer kill?* Online. BBC News. 2013. Dostupné z: https://www.bbc.com/news/magazine-21476904. [cit. 2024-05-09].

KRÁLÍK, Jan. Statistika českých grafémů s využitím moderní výpočetní techniky. Online. *Slovo a slovesnost*. 1983, roč. 44, č. 4, s. 295-304. Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=2913. [cit. 2024-05-09].

Logistická Regrese. Online. Dostupné z: http://kol-apps.ff.upol.cz/log-reg/. [cit. 2024-05-09].

MANNING, C. D.; RAGHAVEN, P.; SCHÜTZE, H.: Introduction to informationretrieval. Cambridge University Press, 2008, ISBN 0521865719

P-hodnota a její interpretace. Online. Institut biostatistiky a analýz lékařské fakulty masarykovy univerzity. Portal.matematickabiologie.cz. 2024. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--uvod-do-testovani-hypotez--p-hodnota-a-jeji-interpretace. [cit. 2024-05-06].

QUITA Online. Online. Dostupné z: https://kol.ff.upol.cz/quita/. [cit. 2024-05-09].

Soudní znalci z oboru kriminalistiky. Online. Praha, 2024. Dostupné z: https://www.grafickeexpertizy.com/. [cit. 2024-05-09].

STRAUS, Jiří. *Kriminalistická technika*. 3., rozš. vyd. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012. ISBN 978-80-7380-409-1.

SVARTVIK, J. *The Evans Statements: A Case for Forensic Linguistics,* 1968.

The Anti-Islamist: Anders Behring Breivik's Manifesto. Online. In: The International Centre for Counter-Terrorism. 2012. Dostupné z: https://www.icct.nl/publication/anti-islamist-anders-behring-breiviks-manifesto. [cit. 2024-05-02].

VARJASSYOVÁ, Ivana. *Hapax legomenon vs. nonce word*. Online. In: Encyklopedie lingvistiky. 2014. Dostupné z: https://encyklopedieoltk.upol.cz/encyklopedie/index.php5/Hapax_legomenon_vs.html. [cit. 2024-05-02].

VENGLAŘOVÁ, Klára a MATLACH, Vladimír. Beyond content: discriminatory power of function words in text type classification. Online. *Digital Scholarship in the Humanities*. 2024. Dostupné z: https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqae013/7634746. [cit. 2024-05-09].