Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra systémového inženýrství

Diplomová práce

Text mining pro analýzu publikací konference

Mgr. Martina Routner

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Mgr. Martina Chvílová

Systémové inženýrství a informatika
Informatika

Název práce

Text mining pro analýzu publikací konference

Název anglicky

Text mining for conference proceedings analysis

Cíle práce

Cílem práce je provést analýzu textových souborů konferenčních příspěvků. Výsledky analýz budou sloužit k identifikaci aktuálních a rozvíjejících se témat a predikci jejich vývoje v čase. Práce bude dále zaměřena na posouzení, zda existuje mezi publikovanými tématy geografická závislost a zda skladba autorů a autorských kolektivů vykazuje trendy a anomálie. Budou analyzovány dokumenty z posledních 5 ročníků konference Efficiency and Responsiblity in Education (ERIE) pomocí platformy Tovek Tools Analyst Pack.

Informace získané analýzami budou využitelné při sestavování příštích ročníků konference.

Metodika

Cíle práce bude dosaženo pomocí následujícího postupu:

1. Literární rešerše

- text mining
- metody text miningu
- obsahová analýza
- kontextová analýza

2. Praktická část

- charakteristika publikací
- obsahová analýz a kontextová analýza dokumentů
- geografická a časová analýza

3. Návrhy, doporučení, závěr

Doporučený rozsah práce

60 – 80

Klíčová slova

text mining, obsahová analýza, kontextová analýza, Tovek Tools, konferenční příspěvek

---

Doporučené zdroje informací

BERRY, Michael W. a Jacob KOGAN. Text Mining: Applications and Theory [online].2. Aufl. New York: Wiley, 2010. ISBN 0470749822;9780470749821.

DÖMEOVÁ, Ludmila; HOUŠKA, Milan; HOUŠKOVÁ BERÁNKOVÁ, Martina. Systems Approach to Knowledge Modelling. 1st ed. Prague: Graphical Studio, 2008. 282 s. ISBN 978-80-86703-30-5.

JO, Taeho. Text Mining: Concepts, Implementation, and Big Data Challenge [online].Cham: Springer International Publishing AG, 2018. ISBN 9783319918143;3319918141.

Předběžný termín obhajoby

2020/21 LS – PEF

Vedoucí práce

Ing. Martina Houšková Beránková, Ph.D.

Garantující pracoviště

Katedra systémového inženýrství

Konzultant

Ing. Kristýna Mudrychová

Elektronicky schváleno dne 29. 10. 2020

doc. Ing. Tomáš Šubrt, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 5. 11. 2020

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 23. 03. 2021

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Text mining pro analýzu publikací konference" jsem vypracoval(a) samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autorka uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušila autorská práva třetích osob.

V Praze dne 30.3.2021 _____

Poděkování

Ráda bych touto cestou poděkovala své vedoucí práce Ing. Martině Houškové Beránkové, Ph.D. za trpělivé vedení a vždy rychlou a velmi užitečnou pomoc. Dále bych velmi ráda poděkovala Ing. Kristýně Novákové Mudrychové, Ph.D. za velmi efektivní pomoc při práci se softwary a zasvěcení do vědeckého světa. Poděkování patří i mému manželovi Milanu Routnerovi za veškerou podporu v průběhu mého studia.

Text mining pro analýzu publikací konference

Abstrakt

Tato diplomová práce se věnuje text miningu - zpracování a získávání informací z nestrukturovaných textových dat, která ve světě dominují nad strukturovanými numerickými daty.

Diplomová práce je rozdělena na dvě hlavní části, na literární rešerši a výzkumnou část diplomové práce.

Literární rešerše se zabývá úvodem do problematiky znalostí a vzdělávání a teoretickým popisem procesu text miningu. Popis procesu text miningu obsahuje definici pojmů spojených s text miningem, popis přípravy textů pro jejich zpracování, výčet konkrétních analytických metod a popis následného vyhodnocení a vizualizace výsledků.

V praktické části byly metodami text miningu analyzovány příspěvky vědecké konference ERIE z let 2016 – 2020. V analýze textových příspěvků byl důraz kladen na odhalení nejvíce aktuálních a rozvíjejících se témat, predikci vývoje frekventovaných témat v čase, zda existuje geografická závislost mezi publikovanými tématy a zda autoři a autorské kolektivy vykazují anomálie a trendy. Analýzy byly provedeny v softwarech Statistica, Tovek Tools and MonkeyLearn.

Závěrem jsou shrnuty dosažené výsledky a představeny získané poznatky, které by mohly být využity při sestavování příštích ročníků konference.

Klíčová slova

text mining, obsahová analýza, kontextová analýza, Tovek Tools, konferenční příspěvek

Text mining for conference proceedings analysis

Abstract

This diploma thesis deals with text mining - processing and obtaining information from unstructured text data, which dominates the world over structured numerical data.

The diploma thesis is divided into two main parts, a literary review and a research part.

The literature review consists of introduction to the terms of knowledge and education and a theoretical description of the text mining process. The description of the text mining process contains the definition of terms associated with text mining, a description of the preparation of texts for the processing, a list of specific analytical methods and a description of the subsequent evaluation and visualization of results.

In the practical part, the proceedings of the ERIE scientific conference from 2016 - 2020 were analysed using text mining methods. Analysis such as development of frequent topics at a time, whether there is a geographical dependence between published topics and whether authors and groups of authors show anomalies and trends were performed. Analyses were run in Statistica software, Tovek Tools software and MonkeyLearn software.

The conclusion summaries the achieved results which could be used in compiling the next year proceedings of the conference.

Keywords

text mining, content analysis, context analysis, Tovek Tools, conference proceedings

Obsah

1. Úvod

V současném světě produkujeme a zároveň jsme obklopeni obrovským množstvím dat. Tato data mohou mít strukturovanou nebo nestrukturovanou podobu. Strukturovaná data, nejčastěji numerická, je možné bez dalších úprav použít k počítačovým analýzám.

Tato diplomová práce se věnuje zpracování a získávání informací z nestrukturovaných textových dat. Textová data ve světě dominují nad numerickými. Ze statistiky vyplývá, že téměř 80% stávajících textových dat je nestrukturovaných, což znamená, že nejsou uspořádána předem definovaným způsobem, nelze je prohledávat a je téměř nemožné z nich získat dále využitelné informace.

V obchodním kontextu jsou jako nestrukturovaná textová data chápány e-maily, příspěvky na sociálních sítích, chaty, zákaznické průzkumy atd. V této diplomové práci budou jako nestrukturovaná textová data uvažovány příspěvky vědecké konference o vzdělávání ERIE.

Text mining neboli počítačové získávání a zpracování komplexních dat z textových souborů se ukázal jako velmi spolehlivý a nákladově efektivní způsob, jak dosáhnout přesnosti, škálovatelnosti a rychlé doby odezvy nad informacemi obsaženými v nestrukturovaném textu v přirozeném jazyce. Zde jsou některé z jeho hlavních výhod podrobněji. Pomocí text miningu je možné analyzovat velké objemy dat obsažené v textech během několika sekund. Analýza textových dat v reálném čase přináší institucím možnost reagovat velmi rychle na potřeby zákazníků a uživatelů. Rychlá analýza také umožňuje prioritizovat úkony podle jejich aktuálních potřeb vyjádřených textem. Text mining oproti manuálním rešerším prováděným lidmi poskytuje konzistentní objektivní výsledky zatížené malým množstvím chyb. Pokud by se měla všechna tato data manuálně zpracovat, bylo by to extrémně nákladné, časově náročné a také velmi neefektivní a nepřesné, protože by docházelo k vysoké chybovosti.

Diplomová práce je rozdělena na dvě hlavní části, na literární rešerši a výzkumnou část diplomové práce.

Literární rešerše se zabývá úvodem do problematiky znalostí a vzdělávání a teoretickým popisem procesu text miningu. Popis procesu text miningu obsahuje definici pojmů spojených s text miningem, popis přípravy textů pro jejich zpracování, výčet konkrétních analytických metod a popis následného vyhodnocení a vizualizace výsledků.

V praktické části je za pomoci nabytých teoretických znalostí podrobně zpracována samotná analýza konferenčních příspěvků na základě předem definovaných postupů. V analýze textových příspěvků je důraz kladen především na odhalení nejvíce aktuálních a rozvíjejících se témat, predikci vývoje frekventovaných témat v čase, zda existuje geografická závislost mezi publikovanými tématy a zda autoři a autorské kolektivy vykazují anomálie a trendy.

2. Cíl práce a metodika

2.1 Cíl práce

Cílem práce je provést analýzu textových souborů konferenčních příspěvků a na základě výsledků těchto analýz posoudit:

- Aktuální a rozvíjející se témata a pokusit se o predikci jejich vývoje v čase.

  Témata budou hodnocena na základě frekvence výskytu klíčových slov v nadpisech a textech jednotlivých textových souborů. Na základě výsledků této analýzy budou stanovena nejvíce frekventovaná témata a témata s rostoucí frekvencí výskytu v čase.

- Zda existuje mezi publikovanými tématy geografická závislost

  Pomocí geografické analýzy bude provedeno posouzení, zda jsou témata rozprostřena geograficky homogenně či zda vyskytující se témata vykazují místní anomálie a v některých geografických regionech se vyskytuje vyšší četnost určitých témat než v jiných.

- Zda skladba autorů a autorských kolektivů vykazuje trendy a anomálie

  Cílem tohoto bodu je vyhodnotit, zda je skladba publikujících autorů napříč ročníky homogenní, či zda a jakým způsobem se v průběhu času proměňují.

- Jak mohou být informace získané analýzami využitelné při sestavování příštích ročníků konference.

2.2 Metodika

2.2.1 Proces analýzy textů

Příprava textů pro softwarovu analýzu

Texty konferenčních příspěvků jsou získány z archivu konference ERIE umístěné na webových stránkách konference ve formátu .pdf, každý článek v samostatném souboru. Pro analýzu softwary Statistica, Tovek tools a MonkeyLearn, které budou pro účely diplomové práci použity, bude potřeba nejdříve převést texty do formátu .txt. Texty v tomto formátu mají nejmenší velikost a jejich softwarové zpracování je nejméně časově náročné. Pro převod textů do formátu .txt bude použit open software Bulk PDF to text Extractor od společnosti Google.

Pro následné geografické analýzy a analýzy autorských kolektivů bude potřeba z článků manuálně odstranit hlavičku, seznam použitých zdrojů a dedikci. Tyto části článků obsahují geografické údaje a údaje o autorech, které by analýzu, jak byla navržena, zkreslovaly.

Dále bude potřeba vytvořit stop list, tj. seznam, který obsahuje nepodstatná slova a slova, která nenesou v textu význam.

Obsahová analýza pomocí softwaru Statistica

Texty budou softwarem Statistica analyzovány metodou frekvenční analýzy a metodou klasifikace textu. Na základě výsledků těchto analýz budou definovány tematické skupiny, do kterých budou jednotlivé příspěvky kategorizovány. Frekvenční analýza pracuje na principu nejčastěji se vyskytujících slov a výrazů. Metoda klasifikace zařazuje obsahově podobné články do stejných, předem zvolených kategorií. Pro zvýšení přesnosti analýzy bude nadefinován tezaurus a slovník frází. Tezaurus je slovník synonym a ekvivalentních výrazů. Slovník frází je textový dokument obsahující víceslovné pojmy. Výsledky softwaru Statistica jsou pouze jednoslovné.

Rozdělení příspěvků do tematických skupin bude využito kromě analýzy témat i pro další analýzy prováděné softwarem Tovek Tools tj. geografickou analýzu, analýzu autorských kolektivů a analýzu vztahů mezi tématy.

Obsahová analýza pomocí softwaru MonkeyLearn

V softwaru MonkeyLearn bude provedena frekvenční analýza dvou vybraných článků. Analýza bude vizualizována pomocí textového mraku. Tato analýza bude pouze ilustrační, protože výsledky nejsou přesně měřitelné a nelze s nimi dále exaktně pracovat, slouží pouze pro vizualizaci a jako prvotní podklad k dalším analýzám.

Obsahová a kontextová analýza pomocí softwaru Tovek Tools

Před zahájením analýzy prostřednictvím platformy Tovek Tools bude potřeba absolvovat školení o používání softwaru přímo v sídle společnosti Tovek a konzultovat jednotlivé kroky analýz, nastavení platformy a očekávané výsledky s interním expertem.

Geografická analýza bude provedena v nástroji *Query Editor* pomocí operátorů dotazovacího jazyka Tovek. Metodou extrakce textu budou vybrány geografické údaje a přiřazeny k jednotlivým tematickým skupinám pomocí nástroje *Inforating*.

2.2.2. Užitý software

Bulk pdf to text extractor je bezplatný nástroj, který lze použít k převodu dokumentů ve formátu .pdf na upravitelné textové soubory. Aplikace funguje pouze lokálně a není potřeba nahrávat data na žádný server. Texty extrahuje v reálném čase a umožňuje vkládat texty z lokální stanice nebo Google drive. Se softwarem se pracuje online.

Statistica 13 (Tibco) je statistický a analytický software. *Program umožňuje uživatelům vytvářet analytické pracovní postupy, které jsou shrnuty a prezentovány v různých formách uživatelům. Program také umožňuje interaktivně prozkoumat a vizualizovat, vytvářet a*

implementovat statistické a prediktivní modely, zároveň umožňuje tzv. data mining a jeho podskupinu text mining. (Mudrychová, 2020). Se softwarem se pracuje lokálně na vlastní stanici, licenční klíč je používán školní.

Tovek Tools *software se využívá ve všech studiích a analýzách, kde je potřeba analyzovat velké množství nesourodých dat. Program je využíván ve zpravodajství, v podpoře výzkumu a vývoje, při mapování konkurenčního prostředí, při vyhodnocování komunikace se zákazníky nebo při odhalování podvodů* (Mudrychová, 2020). Pro potřeby diplomové práce bude využit modul *Inforating*. Texty budou analyzovány pomocí platformy Tovek Tools Analyst Pack ve školní licenci v počítačových laboratořích Provozně ekonomické fakulty ČZU, ke kterým se autorka připojuje vzdáleně přes VPN.

MonkeyLearn je freeware platforma pro text mining a textovou analýzu. V diplomové práci byl použit online generátor tzv. „textových mraků“ anglicky “word cloud“. Klíčová slova z textu jsou touto vizualizační technikou podle důležitosti a frekvence uspořádána do grafů, které připomínají mraky.

2.2.3. Charakteristika textů

V rámci této diplomové práce jsou analyzovány příspěvky konference Efficiency and Responsiblity in Education (ERIE). Celkem je analyzováno 291 příspěvků z posledních 5 ročníků konference, tj. z let 2016 až 2020. Články jsou v anglickém jazyce. Příspěvky jsou ve formátu .pdf a každý příspěvek odpovídá jednomu souboru. Podrobnější charakteristika analyzovaných textů bude následovat v praktické části této práce v kapitole 4.1.

3. Literární rešerše

3.1. Data, informace, znalosti

Diplomová práce pracuje s pojmy data, informace a znalosti. Vysvětlení těchto termínů a vztahů mezi nimi je uvedeno v této kapitole.

Data jsou samostatné kousky informací obvykle formátovaná určitým způsobem. *Mohou existovat v řadě forem, například jako čísla nebo text na útržcích papíru, jako bity a byty uložené v elektronických paměťových nosičích a jako kusy informací v lidských myslích.* (Dömeová, Houška, Houšková Beránková, 2008). *Data bez dalšího popisu nebo kontextu nedávají smysl, ale jsou surovinou, z níž mohou vyvstávat informace* (Cooper, 2014).

Přidáním kontextu k čistým datům vznikne informace. Jedná se o data, která byla „očištěna" od chyb a dále zpracována způsobem, který usnadňuje měření, vizualizaci a analýzu pro konkrétní účel. I pouhým uspořádáním dat způsobem, který odhaluje vztahy mezi různými zdánlivě odlišnými a odpojenými datovými body, mohou vzniknout informace. Podobnou definici jako Ambuehl a Shengwu (2018) uvažuje i Han a Sangiorgi (2018), kteří informace považují za data, kterým právě uživatel přiřadil nějakou důležitost a význam při jejich interpretaci.

Pokud jsou informace nejen chápány jako popis shromážděných faktů, ale také jako prostředek k dosažení cílů, jsou klasifikovány jako znalosti. *Znalosti jsou informace, které byly uchovány s porozuměním o významu těchto informací. Znalosti zahrnují něco získaného zkušenostmi, studiem, sdružením, povědomím a / nebo porozuměním* (Kempe 2013).

3.2. Konference ERIE

Konference ERIE je mezinárodní konference o efektivitě a odpovědnosti ve vzdělávání pořádaná Katedrou systémového inženýrství Provozně ekonomické fakulty České zemědělské univerzity v Praze. Konference se koná každoročně začátkem června v Praze. Mezi témata konference se řadí například: teorie a metodologie pedagogiky a vzdělávání, informace a znalosti v celoživotním vzdělávání a odborné přípravě, ICT ve vzdělávání, aplikace, praxe a zkušenosti, odpovědnost ve vzdělávání, etické otázky (About the coference, 2020).

3.3. Text mining

Text mining je definován jako proces získávání implicitních znalostí z textových dat (Feldman a Sanger, 2007). Text mining se vyvinul jako jedno z odvětví data miningu, které

zpracovává texty v přirozeném nestrukturovaném jazyce a zároveň také vyhledává znalosti v nich obsažené a převádí do měřitelných dat.

V souvislosti se získáváním dat a informací z nestrukturovaného textu se používají dva termíny. Text mining a textová analýza. *Text mining identifikuje relevantní informace uvedené v textu a na jejich základě poskytuje kvalitativní výsledky a informace. Textová analýza se zaměřuje na hledání vzorců a trendů napříč velkou sadou dat a textových dokumentů a výsledkem je spíše kvantitativní analýza. Na základě textových analýz jsou data vizualizována do podoby grafů, tabulek* (Garreta, 2020). Strojové učení je nástroj, na základě kterého, je textová analýza schopná učit se z tréninkových dat a na základě získané předchozí zkušenosti predikovat výsledky nových informací. Ve většině případů je nutné kombinovat text mining, textovou analýzu i strojové učení k získání co možná nejkomplexnějších a nejpřesnějších výsledků.

V práci budou dále termíny text mining a textová analýza používány jako synonyma a nebudou všude striktně rozlišovány, a to zejména proto, že spolu v praxi úzce souvisejí a v textu je striktně nerozlišuje ani většina citované literatury.

3.3.1. Shromáždění textu pro analýzu

Prvním krokem k úspěšné textové analýze je získání a shromáždění potřebného vzorku textových dat. Čím větší vzorek dat a čím více si budou jednotlivé texty po strukturní stránce podobné, tím přesnější a podrobnější výsledky lze získat.

Podle původu se data dělí na interní a externí. Interní data jsou data získaná z databáze určité společnosti, jsou jejím majetkem a tato společnost k nim má výhradní přístup. Jedná se např. o texty emailů, průzkumy mínění nebo data v databázích. Externími daty se rozumí volně šiřitelná data jako např. informace ze sociálních médií a zpravodajských serverů. Konferenční příspěvky se pohybují na rozhraní obou zmiňovaných původů.

3.3.2. Formát textu

Vstupem pro textovou analýzu je jakýkoliv čistý text. *Čistý text lze definovat jako sekvenci alfanumerických, interpunkčních, oddělovacích grafických znaků a některých speciálních symbolů (např.%, &, *, apod.). Alfanumerické znaky mají fonetickou a lexikální hodnotu, jsou tedy přímými nositeli obsahu textu. Oddělovací a interpunkční znaky (např. mezera, tabulátor, tečka, čárka, pomlčka, závorky, lomítko atd.) se používají na členění textu. Čistý text lze získat z formátovaného elektronického dokumentu tak, že se z něj odstraní všechny typografické značky a netextové informace, jako jsou například označení velikosti a typu písma, obrázky a grafika, vzorce, tabulky, grafy, a podobně* (Paralič, 2010).

Texty mohou být uložené v různých datových formátech, nejběžnějšími jsou *„txt „doc“, „docx“, „ppt“, „xls“, „pdf“. Formát „pdf“ je nejsnáze akceptovatelný hlavně proto, že je nezávislý na systémových konfiguracích pro zobrazení dokumentu* (Taeho 2018). Po účely

diplomové práce byly texty převedeny do formátu .txt, nejsnáze s ním pracuje použitý software.

XML (Extensive Markup Language) lze považovat za další textový formát. *Text v XML formátu je opatřen značkami jasně určujícími strukturu textu a je ho tedy možné klasifikovat mezi semi-strukturovaná data* (Taeho 2018).

3.3.3. Příprava textu pro analýzu

Dalším krokem, který následuje po shromáždění textových dat, je jejich příprava pro samotnou analýzu. Metody text miningu zpracovávají velký objem dat a vyhledávají informace v textu, který je napsán v přirozeném jazyce. Z toho bohužel vyplývá nutnost předpřipravit texty před analýzou pomocí rozsáhlých úprav do co nejvíce vhodné podoby pro strojové zpracování.

Normalizace

Normalizace tokenů je proces sjednocení tokenů, aby se předešlo tomu, že v souboru slov se bude nacházet více identických a pouze jinak zapsaných slov. *Nejstandardnějším způsobem normalizace je implicitní vytvoření tříd ekvivalence, které jsou obvykle pojmenovány po jednom členovi sady* (Manning, Raghavan a Schütze, 2008). Příkladem mohou být významově identická, ale jinak zapsaná slova „Mr.“ a „mister“.

Tokenizace

Tokenizace je proces rozdělení celků na menší části (Baeza-Yates a Ribeiro-Neto, 1999). Tato technika úpravy textu rozděluje textové řetězce na jednotlivé morfémy, tedy nejmenší jednotky nesoucí význam. Ty mohou být reprezentovány jednotlivými slovy nebo i většími větnými celky. Části, které tokenizací vzniknou, se nazývají tokeny*. Kromě oddělení lexikálních jednotek textu tokenizace odstraňuje ze souboru dat interpunkci* (Indurkhya a Damerau, 2010). Jde v podstatě o formální oddělení významových celků, aby se daly dále počítačově využít jako smysluplné jednotky. Tokenizace je zvlášť užitečná při procesu indexování slov a počítání jejich výskytu. Výsledný text po úpravě tokenizací může vypadat například takto: („this“ „is“ „an“ „example“ „of“ „a“ „text“ „after“ „tokenization“ „there“ „is“ „no“ „punctuation“ „anymore“).

Segmentace

K tokenům je v rámci dalších úprav možno přiřadit jejich slovní druh. Zjednodušeně to znamená, že se u jednotlivých slov určí, zda se jedná o podstatné jméno, přídavné jméno, sloveso… atd. Tato úprava textu se nazývá segmentace, v angličtině parsing (Garreta, 2020).

Izolace kořene slova

Cílem dalšího kroku úprav textu před jeho samotnou analýzou je redukovat inflexní tvary a derivačně příbuzné tvary slova na běžný základní tvar. Pro úpravu slov na jejich základní

tvar se používají dva přístupy, stemming a lemmatizace. Na příkladu to znamená, že z podstatných jmen je získán první pád jednotného čísla, ze sloves infinitiv…atd. Dále jsou v rámci této úpravy ze slov odstraněny přepony a přípony.

příklady úpravy slov
am, are, is ⇒ be
car, cars, car's, cars' ⇒ car

příklad úpravy věty
the boy's cars are different colors ⇒ the boy car be differ color

Oba přístupy stemming a lemmatizace mají stejný cíl, ale trochu odlišný postup k jeho dosažení. Stemming využívá hrubý heuristický proces, který odřezává konce slov v naději, že většinou bude dosaženo správného výsledku (Manning, Raghavan a Schütze, 2008). Název této metody byl odvozen z anglického výrazu pro kořen slova – stemm. Výsledkem hrubého ořezání slov pomocí algoritmů stemmingu, nemusí vždy nutně být smysluplné slovo. Stemming je oproti lemmatizaci metoda rychlá a výpočetně relativně nenáročná, ovšem dochází při ní k větší chybovosti. *Při izolaci kořene slov může dojít k jedné z těchto dvou chyb a to je over-stemming a under-stemming. Over-stemming je jev kdy jsou dvě slova, která mají rozdílný kořen slova, izolována na stejný základní tvar. Jako příklad můžeme uvést slova "caring" a "cars". Tato slova mají různý význam a různé kořeny slova a měly by být izolovány na slova „care“ a „car“. Hrubý proces odstřižení konců slov, jaký využívá stemming, z obou slov vytvoří chybně „car“. Naopak under-stemming nastane, když ze dvou slov, která by měla být izolována na stejný kořen slova vzniknou dva rozdílné kořeny. Jako příklad uvádím slova cry" a "cried". Obě slova mají stejný kořen „cry“. Při under-stemmingu budou ale chybně z tohoto slova vytvořeny kořeny dva – „cry“ a „cri“* (Rusinková, 2015). Nejběžněji využívaným algoritmem pro úpravu anglických textů pomocí stemmingu je Porterův algoritmus. *Zjednodušeně popsáno se Porterův algoritmus skládá z pěti fází redukce přípon slov aplikovaných postupně za současné kontroly zvývajícího kořene slova. Předpony slov Porterův algoritmus ponechává nezměněné* (Manning, Raghavan a Schütze, 2008).

Cílem procesu lemmatizace je stejně jako u stemmingu odstranění flexe slov a izolace kořene slova. *Základním rozdílem mezi těmito dvěma metodami je, že lemmatizace získává kořen slova na základě dostupné slovní zásoby a morfologické analýzy slov, obvykle s cílem odstranit pouze inflexní konce a vrátit základní nebo slovníkovou formu slova, které se označuje jako lemma. Vložíme-li do procesu úpravy slovo „saw“, stemming by mohl vrátit jen „s“, zatímco lemmatizace by pravděpodobně vrátila buď „see“ nebo „saw“ v závislosti na tom, zda použité původní slovo bylo slovesem nebo podstatným jménem* (Manning, Raghavan a Schütze, 2008).

Odstranění stop slov

Slova v textu, která sama o sobě nenesou žádný význam a zaručují pouze syntaktickou správnost textu se nazývají stop slova (stop words) a v průběhu přípravy textu pro analýzu je cílem je z textu vyloučit. Tato slova bývají velice běžná, avšak pro význam věty nedůležitá. Nejčastěji bývají zastoupena slovními druhy jako jsou zájmena, předložky, spojky nebo částice. V následující tabulce 1 je seznam nejběžnějších stop slov vyskytujících se v anglických textech.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Tabulka 1: Seznam nejčastějších stop slov vyskytujících se v anglickém jazyce. Zdroj: Manning, Raghavan a Schütze, 2008; vlastní zpracování

Jako česká stop slova by se dala například označit slova jako „by“ „se“ „tedy“. Pro každý jazyk existuje sada stop slov, která je brána jako referenční pro jejich odstranění z textu. Seznam stop slov se označuje jako stop list. Odstraněním stop slov z textu se docílí snížení počtu tokenů, které je třeba analyzovat a zároveň se z textu získají pouze slova, která budou pro analýzu významná. Sníží se tím časová náročnost zpracování daného textu a zároveň i nároky na výpočetní výkon a paměť.

Některé speciální typy textů mohou být výskytem stop slov nepřiměřeně ovlivněny. Pro zajímavost některé názvy písní a dobře známé verše se skládají výhradně ze slov, která jsou běžně na seznamech stop slov (To be or not to be, Let It Be, I dont want to be, ...) (Manning, Raghavan a Schütze, 2008).

Klíčová slova

Klíčová slova jsou jedno nebo víceslovné výrazy, které extrahované z dokumentu představují ve zhuštěné formě jeho obsah. *Klíčová slova jsou široce používána k definování dotazů v systémech vyhledávání informací, protože jsou snadno definovatelná, revidovatelná, zapamatovatelná a sdílená* (Berry, 2010).

Přes svou užitečnost pro analýzu, indexování a vyhledávání nemá většina dokumentů přiřazená klíčová slova. Většina stávajících přístupů se zaměřuje na manuální přiřazování klíčových slov profesionálními kurátory, kteří mohou použít pevnou taxonomii nebo se opírají o úsudek autorů, aby poskytli reprezentativní seznam. Současný výzkum se proto zaměřil na metody automatického extrahování klíčových slov z dokumentů jako pomůcky buď k navrhování klíčových slov pro profesionální indexer, nebo ke generování souhrnných funkcí pro dokumenty, které by jinak byly nepřístupné (Berry, 2010).

3.3.4. Obsahová a kontextová analýza

V následující kapitole jsou popsány dva základní teoretické směry analýzy textů – obsahová a kontextová analýza. Vybrané metody text miningu popsané v kapitole 3.3.5. jsou pak většinou kombinací obou těchto směrů.

Obsahová analýza

V tomto obsahovém neboli statistickém přístupu k analýze textu jsou slova v textu kategorizována do několika předdefinovaných skupin podle toho, jaký nesou význam. *Zásadní pro obsahovou analýzu je identifikace klíčových slov, která nesou význam textu* (Wachsmuth, 2015). Do obsahové analýzy vstupuje samotný text i jeho metadata jako je například čas vytvoření textu, obsah dokumentu, údaje o autorovi, atd.

Kontextová analýza

Jako kontext je označováno významové spojení slov nebo částí textu. Kontextová analýza se zabývá hledáním souvislostí mezi slovy, větnými celky, větami, nebo částmi textu (Kubiš, 2019). Na rozdíl od obsahové analýzy se nezkoumá absolutní výskyt těchto částí analyzovaného textu, ale jejich vzájemná souvztažnost. Pokud by došlo ke změně uspořádání slov v konkrétním textu, došlo by i ke změně kontextu, přičemž z pohledu obsahové analýzy by text zůstal nezměněn.

Kontextový přístup je podstatně komplexnější a náročnější na výpočetní výkon než přístup obsahový, a tak k jeho většímu nasazení do rozsáhlých analýz dochází až v poslední době s jeho masivním rozvojem.

3.3.5. Vybrané metody text miningu

Následující kapitola se zabývá jednotlivými metodami aplikovatelnými při text miningu. První tři uvedené metody - frekvence slov, kolokace a concordance neboli shoda, jsou základní nebo se používají jako dílčí, následující čtyři metody – analýza sentimentu, klasifikace textu, shlukování a extrakce textu jsou pokročilejší metody.

Frekvence slov

Metoda frekvence slov identifikuje nejběžnějších termíny nebo pojmy v textu. Nalezení nejfrekventovanějších výrazů v nestrukturovaném textu se používá například u konverzací na sociálních sítích nebo při analýze zpětné vazby od zákazníků (Garreta, 2020).

Kolokace

Kolokace označuje slovní spojení dvou nebo více slov, které spolu gramaticky a sémanticky souvisejí. *Nejběžnějšími typy kolokací jsou dvouslovná slovní spojení - bigramy (v*

angličtině například sousloví „get started“, „save time“ or „decision making“) a trojslovná slovní spojení – trigramy (například „within walking distance“ nebo „keep in touch“) (Garreta, 2020). Mezi kolokace patří sousloví, termíny, frazémy nebo víceslovná místní označení.

Identifikace kolokací a nakládání s nimi jako s jedním tokenem zlepšuje granularitu textu, umožňuje lepší pochopení jeho sémantické struktury a nakonec vede k přesnějším výsledkům dolování textu (Garreta, 2020).

Shoda

Tato metoda se zaměřuje na hledán homonym v textu. Jazyk může být nejednoznačný a stejné slovo lze použít v mnoha různých kontextech. Analýza shody slova může pomoci pochopit jeho přesný význam na základě kontextu (Garreta, 2020).

Zde je například několik vět extrahovaných ze sady recenzí, včetně slova „work“:

Preceding context	**Target**	**Following context**
It saves time and help teams	work	more efficiently.
Some advanced features only	work	in one language (English).
It enables us to	work	towards better conversion and retention.
We recommend this to several of the small businesses we	work	with, and they are all happy with the results.

Tabulka 2: Příklady různého významu slova „work“ v různých kontextech. Zdroj: Garreta, 2020

Analýza sentimentu

Analýza sentimentu nebude pro účely této diplomové práce stěžejní, nicméně pro zachování komplexity pohledu na přístupy k analýze textu je důležité ji zařadit.

Analýza sentimentu pracuje s emočně zabarvenými slovy v textu. Podstatou tohoto přístupu je emočně zabarvená slova vyhledat a podle jejich emočního obsahu roztřídit. Na základě výskytu těchto slov je pak celý text zařazen do tří skupin:

- pozitivní (positive),
- negativní (negative),
- neutrální (neutral).

Pokud lze z textu vyvodit sentiment, text je nazýván subjektivním, v opačném případě je nazýván objektivním. Metodu je možné použit samostatně nebo jakou součást pokročilejších přístupů.

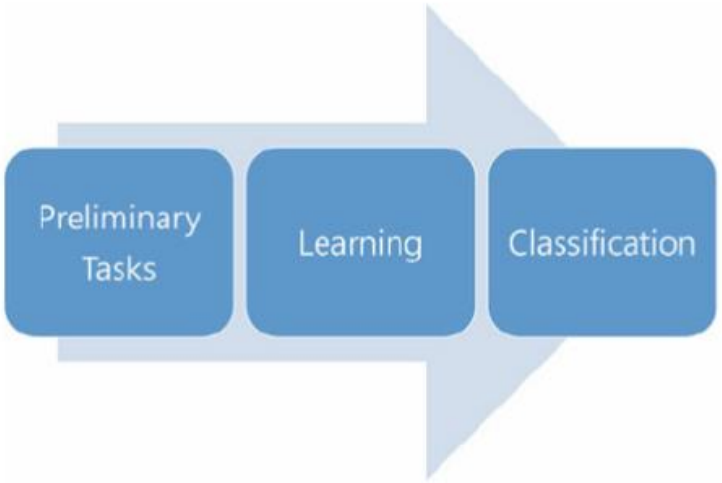
Klasifikace textu

Klasifikace textu je proces přiřazování kategorií nestrukturovaným částem textu na základě jejich obsahu. Tento proces mění text v nestrukturovaném přirozeném jazyce na dále strojově zpracovatelná a smysluplná data. Pokud by měla být metoda klasifikace textu zařazena do jednoho z výše uvedených směrů, odpovídala by obsahové analýze.

Díky automatizované klasifikaci textu je možné označit velkou sadu textových dat, dosáhnout výsledků ve velmi krátkém čase, aniž by bylo nutné do textu manuálně zasahovat. K nejčastějším úkolům klasifikace textu patří analýza témat, analýza sentimentů a detekce jazyka.

Klasifikace textu někdy označovaná i jako kategorizace se provádí v zásadě prostřednictvím tří kroků, které jsou znázorněny na obr. 1. Před vlastní kategorizací je nutné provést několik přípravných kroků. Jedním z nich je předem určit seznam nebo strom kategorií kterými se budou jednotlivé tokeny klasifikovat. Dalším přípravným krokem je připravit pro každou kategorii ukázkové tokeny, podle kterých se bude řídit algoritmus klasifikace. Dále je třeba rozhodnout o klasifikačním algoritmu a typu klasifikace, který bude použit. *Nejčastěji se používá hierarchický algoritmus a výlučná nebo překrývající se klasifikace* (Taeho, 2018).

Ve druhé fázi, v procesu učení, se z ukázkových textů, které jsou přiřazeny ke kategoriím v předchozím kroku, vytvoří vzorce jako jsou rovnice, symbolická pravidla nebo optimalizované parametry, na základě kterých se v dalším kroku bude text automaticky kategorizovat.



Obrázek 1: Fáze klasifikace textu. Zdroj: Taeho 2018

Výsledky procesu učení lze ověřit pomocí ověřovací sady příkladů. V závislosti na výsledcích z ověřovací sady se rozhodujeme, zda použijeme zvolený klasifikační algoritmus (Taeho, 2018). Pokud se ověřovací sadou podaří verifikovat dostatečnou přesnost algoritmu, nastává poslední fáze procesu a tou je samotná klasifikace.

Konceptů klasifikace je mnoho, první z nich je binární klasifikace, jejímž základem je roztřídit výrazy pouze do dvou kategorií například pozitivní a negativní nebo týkající se jednoho nebo druhého tématu. Vícečetné klasifikace dělí výrazy v textu do *m* skupin podle složitějších algoritmů. Většinou se jedná o nadefinované předlohy a posuzuje se, zda dané slovo té předloze odpovídá. Pokud odpovídá, zařadí se, pokud ne, je posuzováno vůči jiné skupině a proces se opakuje, dokud se nepovede výraz začlenit.

Klasifikace se dále dělí na měkkou a tvrdou v závislosti na tom, zda každá položka smí patřit do více než jedné kategorie, či nikoli. Dále se klasifikace dělí na hierarchickou a nehierarchickou podle toho, zda je povoleno vnořování kategorií do sebe či nikoliv. Dělení klasifikace na závislou a nezávislou se vyznačuje tím, zda je aktuální klasifikace ovlivněna výsledky klasifikace položek předchozích.

Shlukování

Shlukování textu (clustering) označuje proces segmentace textu do shluků na základě jejich obsahové nebo významové podobnosti. Pokud by měla být metoda shlukování textu zařazena do jednoho z výše uvedených směrů, odpovídala by kontextové analýze.

Shlukování textu může být prováděno na různých úrovních granularit, kde klastry mohou být dokumenty, odstavce, věty nebo výrazy. *Segmenty textu nejsou předem nijak klasifikovány a algoritmus učení probíhá na rozdíl od metody klasifikace bez verifikace ověřovací sadou* (Allahyari a kol, 2017).

Shlukování textu probíhá silně v závislosti na podobnosti textů, klíčový je tedy způsob definice metrik podobnosti. Prototypy klastrů jsou inicializovány náhodně nebo libovolně a jsou optimalizovány tak, aby maximalizovaly podobnosti položek v každém klastru a minimalizovaly podobnosti mezi klastry.

Existuje mnoho clusteringových algoritmů, které lze použít k práci s kontextem textových dat. Algoritmy mohu pracovat na principu binárního vektoru, tj. s ohledem na přítomnost nebo nepřítomnost slova v dokumentu nebo je možné použít rafinovanější reprezentace, které zahrnují metody vážení, jako je TF-IDF.

Obě metody klasifikaci a shlukování lze účinně v metodách kombinovat. Rozdíly mezi oběma metodami jsou patrné z tabulky 3. Shlukování oproti klasifikaci textu nepotřebuje manuální definici kategorií, shluky jsou vytvářeny automaticky podle významu obsaženého v textu. Výsledky strojového učení, které probíhá v metodě klasifikace, jsou verifikovány pomocí ověřovací sady, u shlukování žádná verifikace neprobíhá. Vstupem pro shlukovou analýzu je vždy celý textový soubor najednou, do klasifikace textu vstupují jednotlivé tokeny samostatně. Výsledkem shlukovací analýzy jsou strojově vytvořené podskupiny dat s podobným významem, výsledkem klasifikace textu jsou tokeny roztříděné do předem daných kategorií. Metoda klasifikace textu bývá označována jako statická, metoda shlukování jako dynamická.

	Clustering	**Classification**
Classification categories	Automatic definition	Manual definition
Machine learning type	Unsupervised learnng	Supervised learning
Input	Group of data items	Single data item
Output	Subgroups of similar data items	Category or some categories

Tabulka 3: Hlavní rozdíly mezi shlukovací analýzou a klasifikací textu (Taeho, 2018).

Extrakce textu

Extrakce textu je technika textové analýzy, která extrahuje konkrétní části dat z textu, jako jsou klíčová slova, názvy entit, adresy, e-maily atd. Pokud by měla být metoda extrakce textu zařazena do jednoho z výše uvedených směrů, odpovídala by obsahové analýze.

Pomocí extrakce textu lze získat z textu klíčové informace a zároveň se vyhnout veškerým problémům se spojeným s ručním tříděním dat.

Extrakce informací zahrnuje tři základní úkoly, extrakci klíčových slov, extrakci entit a extrakci funkcí.

Extrakce klíčových slov

Klíčová slova jsou nejdůležitější pojmy v textu a lze je použít k shrnutí jeho obsahu. Použití extraktoru klíčových slov umožňuje indexovat data, která mají být prohledávána, shrnout obsah textu nebo vytvořit tagy (značky).

Extrakce entit

Entita se rozumí posloupnost slov, která identifikují nějakou skutečnou entitu, jako jsou jména společností, organizací nebo osob z textu, např. „Google Inc“, „USA“, „Barack Obama“. Úkolem rozpoznávání entit je najít a klasifikovat entity ve volném textu do předem definovaných kategorií, jako je osoba, organizace, umístění atd. *Extrakci entit nelze zcela jednoduše provést porovnáním slov proti slovníku, a to z následujících dvou důvodů. Slovníky jsou obvykle neúplné a neobsahují všechny formy pojmenovaných entit. Jednoznačnost entity*

často závisí na kontextu, například „velké jablko“ může být ovoce, nebo jiné pojmenování města New Yorku (Allahyari a kol, 2017).

Extrakce funkcí

Extrakce funkcí pomáhá identifikovat specifické vlastnosti produktu nebo služby v sadě dat. Pokud se například analyzují popisy produktů, dají se snadno extrahovat funkce jako je barva, značka, model atd.

Ve stejné analýze může být často užitečné kombinovat extrakci textu s klasifikací textu.

3.3.6. Vizualizace

Vizualizace je v případě text miningu velmi silným nástrojem. *Text mining je schopen z velkého množství nestrukturovaných dat vytvořit podstatně menší a strukturovanou podmnožinu, tato podmnožina je často stále příliš velká na to, aby lidský analytik rozumně zpracoval, pochopil, detekoval trendy a vyvodil z nich závěry. Nástroje pro vizualizaci mají zásadní význam pro interpretaci informací obsažených v textech* (Berry, 2010).

Jednou z nejběžnějších forem vizualizace textových dat je textový mrak, anglicky word cloud, tag cloud, text cloud. Vstupními údaji pro textový mrak je četnost slov v daném textu. Jednotlivá slova analyzovaného textu jsou poté umístěna do prostoru v různých tvarech a četnost jejich výskytu a relevance je zohledněna ve velikosti písma. Čím větší písmo, tím vyšší počet výskytů slova v textu. Pro vizualizaci pomocí word cloudu je zásadní kvalitní extrakce klíčových slov z textu a odstranění stop slov. Četnost stop slov v textu je vyšší než četnost klíčových slov, a tak by při jejich nedostatečné izolaci došlo ke znehodnocení celého diagramu (Frydrychová 2020). Textové mraky poskytují rychlé a jednoduché vizuální informace, které mohou vést k podrobnějším analýzám. *Mezi jejich nevýhody patří, že neumožňují porovnávání jednotlivých textů mezi sebou, mají nepřesné kódování velikosti, a nezobrazují strukturu textu* (Heer 2020).

Další z jednodušších vizualizací je matice slov. *Znázorňuje počet výskytů klíčových slov v jednotlivých článcích* (Heer 2020). Tento způsob zobrazení textových dat je velmi přesný, ale u větších souborů dat může být matice obsáhlá a nepřehledná.

Typem vizualizace počtu slov v textu je i word count graf (tabulka). *Slova nejsou v takovém grafu umístěna do mraku ani jiného tvaru, ale jednoduše vedle sebe, jako v textu, přičemž může, ale nemusí být zachován parametr velikosti fontu textu jako indikátor počtu výskytů výrazu v analyzovaném dokumentu. Rozdílem oproti word cloudu je to, že počet výskytů je u každého slova zpravidla uveden, a tak není kladen takový důraz na velikost slova a je možné zanechat všechna slova ve stejné velikosti. Co je ale ve vizuálu podstatné je to, že slova jsou seřazena podle počtu jejich výskytů. Stále tedy platí důležitost vyřazení stop slov,*

jako ostatně u téměř všech typů vizualizací, aby v grafu nebyla na prvních příčkách zobrazena slova neposkytující hodnotnou informaci (Frydrychová 2020).

Dalším typem grafu, pomocí kterého je možné vizualizovat textová data je history flow graf. History flow je vizualizační nástroj pro časovou posloupnost snímků dokumentu v různých fázích jeho vytváření. Tento nástroj podporuje sledování příspěvků do článku různými uživateli a dokáže identifikovat, které části dokumentu zůstaly nezměněny během mnoha revizí úplných dokumentů.

4. Praktická část diplomové práce

4.1 Charakteristika textů

Pro účely diplomové práce bylo analyzováno celkem 291 vědeckých příspěvků z pěti ročníků konference Efficiency and Responsiblity in Education (ERIE), tj. z let 2016 až 2020. Počty příspěvků v jednotlivých ročnících jsou patrné z tabulky 4.

Ročník	Počet příspěvků
2016	85
2017	66
2018	53
2019	43
2020	44

Tabulka 4: Počty příspěvků v jednotlivých ročnících. Zdroj: vlastní zpracování

Články jsou v anglickém jazyce. Příspěvky jsou ve formátu .pdf a každý příspěvek odpovídá jednomu souboru. Články jsou uloženy pod jménem autora/autorů a číslem stránek, na kterých se v konferenčním sborníku jejich článek vyskytuje ve formátu 001-005_Příjmení,Příjmení.pdf

4.2. Frekvenční analýza a klasifikace textu

4.2.1. Příprava textů pro softwarovou analýzu

Před samotným softwarovým zpracováním bylo potřeba texty převést do formátu .txt. Texty v tomto formátu jsou pro počítačové zpracování softwary Statistica a Tovek tools nejvhodnější. Pro převod textů do formátu .txt byl použit open software Bulk PDF to text Extractor od společnosti Google.

Dále byla z článků manuálně odstraněna hlavička, dedikce a zdroje. Tyto části článků obsahovaly geografické údaje a údaje o autorech, které by analýzu, jak byla navržena, zkreslovaly.

4.2.2. Vytvoření stop listu

Dalším krokem přípravy před samotnou analýzou bylo odstranění slov, která nenesou v textech žádný význam. K odstranění nepodstatných slov bylo nutné vytvořit stop list. Odstraněním nepotřebných slov z textů se zmenší vzorek slov pro samotnou analýzu a tím se jednak usnadní následné kroky a také zpřesní celá analýza. Stop listy jsou volně k vyhledání na internetu nebo je již obsahuje používaný software. Software Statistica obsahuje anglický slovník, který je ale pro účely této práce málo obsáhlý. K odstranění stop slov byl jako základ použit stop list stažený online ze serveru GitHub od autora Diaz, 2016. Stop list byl na základě

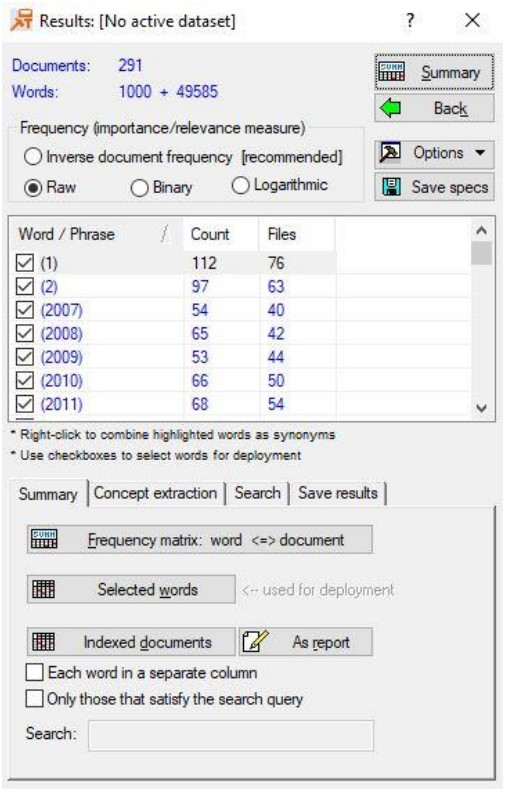
výsledků první frekvenční analýzy softwarem Statistica doplněn o další slova, která byla vyhodnocena pro řešené téma jako nadbytečná. Do toho spadaly zejména nadpisy částí textu (např. abstract, keywords apod.), dále spojky, předložky atd. Náhled stop listu je v tabulce 5. Z textu jsou také odstraněna čísla stránek a další netextové informace.

able	act	affects	ain't	already	amount	anything	approximately
ableabout	actually	after	aint	also	an	anyway	aq
about	ad	af	al	although	and	anyways	afterwards
above	added	ag	all	always	announce	anywhere	are
abroad	adj	again	allow	am	another	ao	area
abst	adopted	against	allows	amid	any	apart	areas
accordance	ae	ago	almost	amidst	anybody	apparently	aren
according	af	ah	alone	among	anyhow	appear	aren't
accordingly	affected	ahead	along	amongst	anymore	appreciate	arent
across	affecting	ai	alongside	amoungst	anyone	appropriate	arise

Tabulka 5: Stop list. Zdroj: Diaz 2016; vlastní zpacování

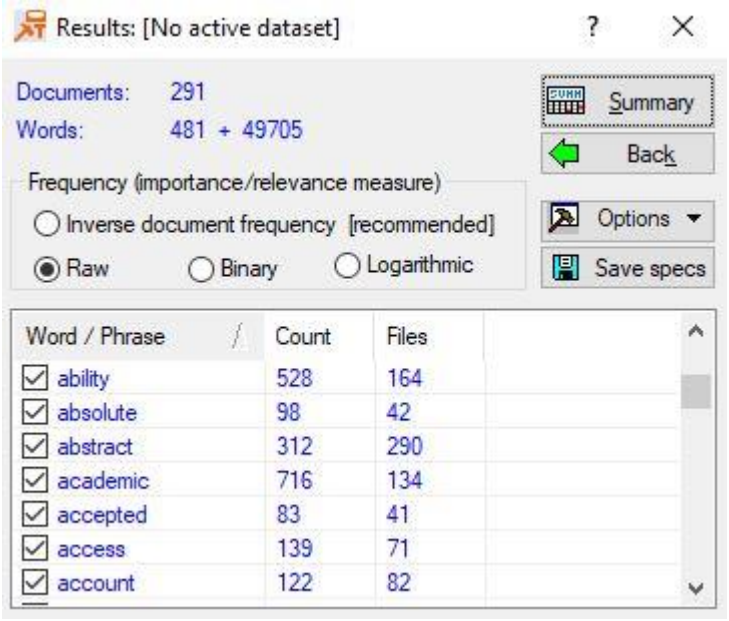
4.2.3. Frekvenční analýza

Kompletní soubor všech očištěných textů byl nahrán do modulu Text Miner softwaru Statistica. Jako seznam stop slov byl použit připravený stop list. Výsledky prvotní frekvenční analýzy jsou viditelné z obrázku 2.



Obrázek 2: Náhled na výsledky prvotní frekvenční analýzy softwarem Statistica. Zdroj: vlastní zpracování

Celkem bylo v prvním kroku analýzy zpracováno 291 textů, program identifikoval 1000 slov a 49585 jejich synonym. Po seskupení synonym nástrojem *combine words a* odstranění irelevantních slov a zbývajících číslic, zůstalo pro další analýzu 481 slov a 49705 synonym, jak je patrné z obrázku 3.



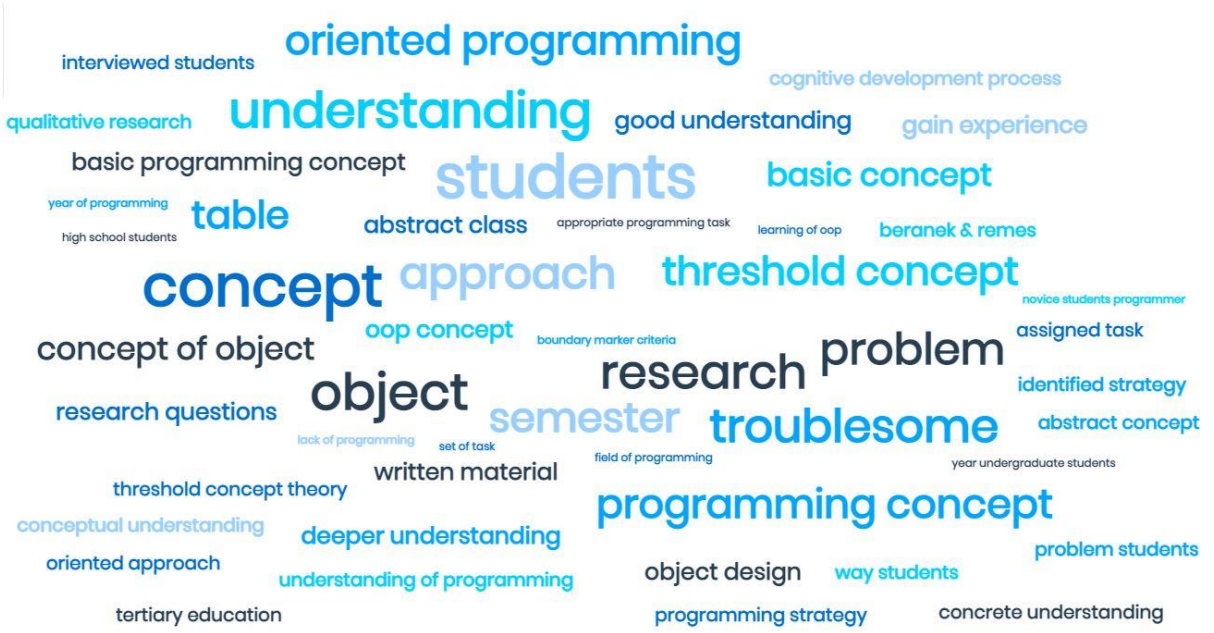
Obrázek 3: Náhled na výsledky analýzy po odstraněn synonym a irelevantních slov a znaků. Zdroj: vlastní zpracování.

S využitím modulu *Selected words* byla zobrazena jednotlivá slova a četnost jejich výskytů v textech. Tato metoda vizualizace se nazývá word count. V tabulce 6 je zobrazeno 25 nejčastěji se vyskytujících slov. V Příloze 1 je celý seznam klíčových slov.

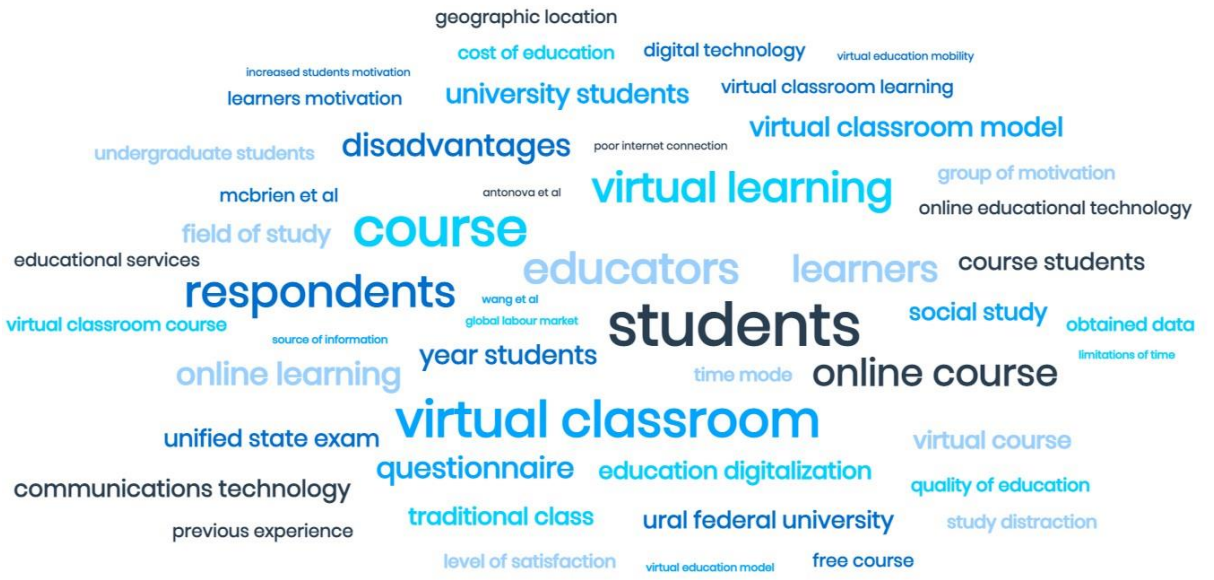
Pořadí	Slovo	Počet výskytů	Počet dokumentů
1	study	12333	286
2	teacher	5244	233
3	education	4816	272
4	question	2506	244
5	learn	2316	221
6	school	2141	196
7	university	2135	234
8	level	2106	270
9	analyse	2056	276
10	develop	1956	246
11	knowledge	1856	231
12	response	1790	199
13	data	1786	262
14	method	1724	291
15	process	1551	252
16	table	1545	234
17	mathematical	1413	106
18	active	1369	219
19	evaluation	1316	203
20	subject	1272	191
21	statistic	1264	193
22	based	1240	268
23	focus	1169	256
24	solution	1167	162
25	czech	1120	188

Tabulka 6: Výsledky metody Word count ze sw Statistica. Zdroj: vlastní zpracování.

Frekvenci slov a jejich relevanci v textu lze dále vizualizovat technikou word cloud. Zde v diplomové práci je tato technika uvedena, ale není s ní dále pracováno, protože její výsledky nejsou přesně měřitelné a nelze s nimi dále exaktně pracovat, slouží pouze pro vizualizaci a jako prvotní podklad k dalším analýzám. Pro analýzu byl použit SW MonkeyLearn, který pracuje s víceslovnými výrazy. Na obrázku 4 a 5 je word cloud z náhodně zvoleného článku 242-248_Remeš, Beránek.txt a 005-011_Antonova, Abramova, Popova.txt. Oba články jsou z roku 2020.



Obrázek 4: Word cloud vytvořený softwarem MonkeyLearn z textu: Remeš, Beránek. Zdroj: vlastní zpracování.



Obrázek 5: Word cloud vytvořený softwarem MonkeyLearn z textu: Antonova, Abramva, Popova. Zdroj: vlastní zpracování.

4.2.4. Klasifikace textů

Výběr kategorií

Prvním krokem metody klasifikace textů je výběr kategorií, do kterých jsou texty členěny. Na základě analýzy klíčových slov a tematických celků definovaných přímo konferencí ERIE bylo navrženo 7 tematických kategorií. Pro další analýzy bylo nutné kategorie jednoslovně označit a tato vybraná pojmenování kategorií zpětně doplnit do článků. Označení kategorií se tak v dalším kroku naindexovalo a bylo možné s klasifikovanými články dále pracovat. Pro označení kategorií bylo nutné vybrat taková slova, která se v anglickém jazyce běžně vyskytují. Pokud by byla zvolena neexistující slova nebo zkratky, program by je porovnal se slovníkem a mohl by je z další analýzy vyloučit. Dále bylo nutné zabezpečit, aby byla jako označení vybrána taková slova, která se jinde ve článcích nevyskytují, aby analýza nebyla kontaminována. Jako názvy kategorií byla zvolena řecká písmena, která byla vybrána tak, aby se v textech článků nenacházela. Pro další analýzy byly označení kategorií názvy řeckých písmen do textů zpětně doplněny.

Názvy kategorií a jejich označení zobrazuje tabulka 7.

Označení	**Kategorie**
omega	Teorie a metodologie pedagogiky a vzdělávání
beta	Teorie a metodologie vědy
gamma	Lidské zdroje a pracovní uplatnění
delta	Řízení znalostí a znalostní inženýrství
kappa	Systémové inženýrství a informační inženýrství, software
sigma	Kvantitativní metody pro vzdělávání a / nebo vědu
epsilon	Chování a motivace studentů

Tabulka 7: Názvy kategorií a jejich označení. Zdroj: vlastní zracování

Popis kategorií

- Teorie a metodologie pedagogiky a vzdělávání. Tato kategorie sdružuje všechny články týkající se obecně pedagogiky a vzdělávání.
- Teorie a metodologie vědy.
- Lidské zdroje a pracovní uplatnění. Články z této kategorie pojednávají o školních výsledcích studentů a jejich souvislosti s pracovním uplatněním, situaci na trhu práce, zaměstnání při studiu a jiných tématech souvisejících s vzděláním jako prerekvizitou pro pracovní uplatnění.
- Řízení znalostí a znalostní inženýrství. Tato kategorie zahrnuje články zabývající se problematikou vytěžování informací z dat, statistickými metodami analýzy a zpracování dat a ostatními tématy z oboru řízení znalostí a znalostního inženýrství.

- Systémové inženýrství a informační inženýrství, software. Do této kategorie spadají články zabývající se návrhem a realizací informačních systémů a články zabývající se využitím určitých systémových řešení a softwarů.
- Kvantitativní metody pro vzdělávání a / nebo vědu. Tato kategorie sdružuje články, které se věnují výzkumu postavenému na kvantitativních metodách jako je například sběr dat formou dotazníků, testů a sebehodnocení.
- Chování a motivace studentů.

Ke klasifikaci textů neboli rozdělení textů do kategorií byl využit modul *Indexed documents.* Výsledkem analýzy tímto modulem je tabulka, která zobrazuje klíčová slova v jednotlivých textech. Náhled této tabulky zobrazuje obrázek 6.

	File summary (Spreadsheet1)		
	Document length	Number of words	Indexed text
007-012_Antonova, Merenkov, Popova_146-618-1-RV.txt	21256	57	assessment students abstract paper learning innovative pedagogical technol
013-020_Beranová, Navrátilová, Šíma_159-606-1-RV.txt	23471	60	evaluation knowledge finance students selected czech universities abstract
021-026_Berka, Vrabec, Marek_147-585-2-RV.txt	16058	61	bachelor students university prague abstract completing study aâ students
027-033_Boguslavskaya_170-616-1-RV.txt	22352	56	benefits project abstract paper approach development university author appli
034-042_Cocca, Cocca, Dimas Castro, Espino_176-574-1-RV.txt	21541	56	program professional abstract professional contribute increasing quality proc
043-049_Fajčíková, Fejfarová, Hlavsa, Fejfar_209-654-1-RV.txt	20083	59	success rate degree programme matter abstract article relationship success
050-056_Flégl, Jiménez-Bandala, Andrade Rosas_215-641-1-RV.txt	19858	59	measure period introductory analysis abstract main tasks education instituti
057-067_Gunina, Komárková, Marková, Tripes_166-628-1-RV.txt	27158	58	project management view abstract study analyses attitude project managem
068-075_Höfrová, Moore de Peralta_188-564-1-RV.txt	27165	57	faculty satisfaction gender differences public university abstract time limited
076-082_Homan, Hanzal_172-572-1-RV.txt	20648	53	services secondary schools selected region abstract secondary schools ser
083-089_Horáková, Houšková Beránková, Mudrychová, Houška_217-648-1-RV.txt	22062	57	influence selected factors learning outcomes knowledge abstract paper aims
090-096_Hrubý, Staňková.txt	19985	61	future czech abstract paper aâ study data monitoring study university studer
097-104_Chýlová, Michálek, Krejčová, Natovová_161-597-1-RV.txt	23546	61	students â€" academic success abstract aâ change university student acad
105-112_Jančaříková, Novotná_133-570-2-RV.txt	24856	57	questions â€" questions abstract questions introduced (2010) tools creation
113-119_Jiménez-Bandala, Flegl, Luis_211-555-1-RV.txt	17429	59	education positive impact developing countries abstract developing countries
120-128_Jordánová, Koldová, Petrášková, Rosa_179-620-1-RV.txt	25426	65	factors influencing financial abstract main goal paper aâ research, examined
129-135_Krejčová, Chýlová, Michálek, Vydrová_143-550-1-RV.txt	22209	59	impact education abstract family aâ primary psychological development thec
136-145_Kuchařová, Pfeiferová, Prášilová, Šimůnková_180-580-1-RV.txt	29208	56	academic students accounting subjects abstract scale measure academic s
146-153_Kukalová, Ječmínek, Moravec, Bína Filipová_155-662-1-RV.txt	23741	63	finance education prague abstract paper focuses exams success rate stude
154-160_Kvasničová Stanislavská, Kvasnička, Paníková_199-643-1-RV.txt	13690	58	social universities czech comparison situation 2011 2018 abstract concept s
161-167_Lánský, Mildeová_152-655-1-RV.txt	23047	58	practice subject application software university finance administration abstra

Obrázek 6: klíčová slova v jednotlivých textech. Zdroj: vlastní zpracování.

Podle souboru klíčových slov v jednotlivých textech byly články do kategorií rozřazeny. V některých textech nebylo zařazení do kategorií pouze na základě klíčových slov jednoznačné. Pro větší jistotu v kategorizaci byla analýza doplněna o jednoznačné víceslovné fráze, které byly do softwaru doplněny.

Seznam frází je uveden v tabulce 8.

methodology of pedagogy	Theory of science	Knowledge unit	whiteboards
methodology of education	methodology of science	knowledge text	online tool
Theory of pedagogy	human resources	knowledge transfer	computer labs
theory of education	human relations	Knowledge sharing	quantitative methods
Performance prediction	human relations management	Tacit knowledge	self-reflection
learning-teaching process	employers	Explicit knowledge	university assessment
education effectiveness	First job	systems engineering	statistical methods
oral assessment	earnings	information engineering	survey
written assesment	organizational structure	applications in education	student reflection
Explicit knowledge	Career decision	applications in science	teacher's abilities
tertiary education	labour market	Computer-aided assessment	Assessment tool
educational process	job market	LMS Moodle	soft skills
teaching methods	knowledge management	E-learning	higher education
learning competences	knowledge engineering	mobile devices	

Tabulka 8: Seznam frází použitých k doplnění klasifikace textů na základě klíčových slov. Zdroj: vlastní zpracování

Byla zvolena měkká klasifikace, tj. že jeden text mohl spadat do více kategorií.

Ověření správnosti klasifikace

Druhým krokem metody klasifikace textů je ověření, zda klasifikační algoritmus, který byl zvolen v prvním kroku analýzy, rozřazuje články do kategorií správně. Tento ověřovací krok je velmi důležitý, neboť s kategoriemi pracují dále všechny následující analýzy prováděné v rámci této diplomové práce. Ze sady klasifikovaných článků bylo náhodně vybráno 10. Kategorie, do kterých byly články zařazeny na základě seznamu klíčových slov a frází byly porovnány se zařazením na základě zobrazení celého textu článku. Výsledky prvních tří testovaných článků jsou zobrazeny níže.

Text - 380-386 Šálková, Navrátilová 2018

- Klíčová slova definovaná programem Statistica:

„experience success graduates job market abstract specific experience studies increasing university experience essential future success job market. paper partial survey conducted students degre"

Zařazení do kategorií na základě klíčových slov:
- Gamma: Lidské zdroje a pracovní uplatnění

- Výběr z celého textu článku:

WORK EXPERIENCE AS A PREREQUISITE FOR THE SUCCESS OF GRADUATES IN THE JOB MARKET

Daniela Šálková, Miroslava Navrátilová

Abstract
The acquisition of specific work experience during studies is proving to be of increasing importance for university graduates. Work experience is currently an essential prerequisite for future success in the job market. Paper presents the partial results of a survey conducted among students in the bachelors' degree program at the Economics and Management Faculty of the CULS over a five-year period from 2013 to 2017. The goal of the paper is to define the degree to which students were engaged in employment during their studies at university, their reasons for entering employment and the areas where the students most frequently worked. Most students are aware of the significance of acquiring work experience and skills in parallel with theoretical preparation. The students were most frequently employed in the form of part-time jobs which represent a flexible form of employment and at the same time can be well combined with university studies.
Keywords
Employment, graduate, job market, studies, university, work experience
Introduction
The issue of the employment of graduates is proving to be highly significant worldwide. As Dimian (2011) points out, the unemployment of young people is negatively related to per capita GDP, but positively correlated in the case of a delayed unemployment rate….

Zařazení do kategorií na základě posouzení celého textu článku:

- Gamma: Lidské zdroje a pracovní uplatnění

Text – 061-068 Gadušová, Ďurková 2018

- Klíčová slova definovaná programem Statistica:

"responsible teacher assessment abstract paper tools project key professional assessing identify psychological factors learning. based qualitative method analysis interpretation collected data. Data"

Zařazení do kategorií na základě klíčových slov:

- Omega: Teorie a metodologie pedagogiky a vzdělávání
- Sigma: Kvantitativní metody pro vzdělávání a / nebo vědu

- Výběr z celého textu článku:

RESPONSIBLE SELF-ASSESSMENT AS PART OF TEACHER ASSESSMENT

Zdenka Gadušová, Simona Ďurková

Abstract
The paper presents the results of piloting the tools (developed within the project APVV-14-0446 for 10 key teachers' professional competences) for assessing the teachers' competence Can identify the psychological factors of pupil learning. It is based on qualitative research using the method of analysis and interpretation of the collected data. Though the piloting is still in progress,

the data analysis presented in the paper and its results have shown that neither the management nor the staff of Slovak schools is keen on trying new methods of teacher assessment. They do not like filling in forms which are longer than one page and hate questions which ask them to use theoretical knowledge to solve practical problems. This creates a lot of dilemmas for the research team of the project as any of the closed questions in the sheets would need an open sub-question asking for giving of evidence.

Keywords

Assessment, assessment tool, assessor, professional competence, self-assessment, teacher

Introduction

Evaluation of teacher competences has been one of the most frequently discussed topics in recent years. There are many methods of teacher evaluation (Magová et al., 2016).

Zařazení do kategorií na základě posouzení celého textu článku

- Omega: Teorie a metodologie pedagogiky a vzdělávání
- Sigma: Kvantitativní metody pro vzdělávání a / nebo vědu

Text – 175-182 Martiník 2019

- Klíčová slova definovaná programem Statistica:

"technologies applied abstract plays crucial role implementation learning process students special students difficult participate learning process. students support technologies participate process"

Zařazení do kategorií na základě klíčových slov:

- Kappa: Systémové inženýrství a informační inženýrství, software
- Omega: Teorie a metodologie pedagogiky a vzdělávání

- Výběr z celého textu článku:

APPLE MOBILE TECHNOLOGIES APPLIED TO SHARING AND RECORDING OF REMOTE LECTURES

Ivo Martiník

Abstract

Mobile equipment plays crucial role in the implementation of learning process for the students with special needs, especially for students with locomotive, visual and aural disability who may find it difficult to personally participate in the learning process. The realization of the teacher-students interaction with the support of mobile technologies and its recording that is available on-line or on-demand is also necessary if the educator cannot participate in the lecture process directly in the classroom but interacts with the students from a remote site. Apple mobile technologies used within the services of the classroom of the Apple Authorised Training Centres for Education (AATCe) worldwide program have been successfully deployed in fulfilling these objectives at the Faculty of Economics VŠB-Technical University of Ostrava. The process Petri nets theory was applied at the design and implementation of information barrier-free approach that is determined for the fulfillment of the above goals.

Keywords

FaceTime, MERLINGO, Mobile technologies, process Petri nets, remote group communication

Introduction

Mobile technologies of all kind have been increasingly used in the teaching process at all types of the schools in the Czech Republic.

Zařazení do kategorií na základě posouzení celého textu článku:

- Kappa: Systémové inženýrství a informační inženýrství, software
- Omega: Teorie a metodologie pedagogiky a vzdělávání

4.2.5. Zhodnocení výsledků

Analýzou článků, které byly upraveny vložením názvu kategorie, byly zjištěny počty výskytů jednotlivých kategorií. Výsledky jsou uvedeny v tabulce 9.

Kategorie						
	2016	**2017**	**2018**	**2019**	**2020**	**Celkem**
Teorie a metodologie pedagogiky a vzdělávání	54	44	30	29	30	187
Teorie a metodologie vědy	13	6	4	8	15	46
Lidské zdroje a pracovní uplatnění	17	4	7	7	7	42
Řízení znalostí a znalostní inženýrství	9	3	5	2	4	23
Systémové inženýrství a informační inženýrství, software,	13	10	3	5	11	42
Kvantitativní metody pro vzdělávání a/nebo vědu	14	15	8	7	10	54
Chování a motivace studentů	12	14	16	13	17	72

Tabulka 9: Počty výskytů jednotlivých kategorií v ročnících. Zdroj: vlastní zpracování

Následující tabulka 10 zobrazuje relativní četnosti výskytů jednotlivých kategorií v ročnících. Absolutní výsledky byly přepočítány, aby nedošlo ke zkreslení analýzy rozdílným počtem článků v ročnících.

Kategorie					
	2016	**2017**	**2018**	**2019**	**2020**
Teorie a metodologie pedagogiky a vzdělávání	64	67	57	67	68
Teorie a metodologie vědy	15	9	8	19	34
Lidské zdroje a pracovní uplatnění	20	6	13	16	16
Řízení znalostí a znalostní inženýrství	11	5	9	5	9
Systémové inženýrství a informační inženýrství, software	15	15	6	12	25
Kvantitativní metody pro vzdělávání a/nebo vědu	16	23	15	16	23
Chování a motivace studentů	14	21	30	30	39

Tabulka 10: Relativní počty výskytů jednotlivých kategorií v ročnících. Zdroj: vlastní zpracování

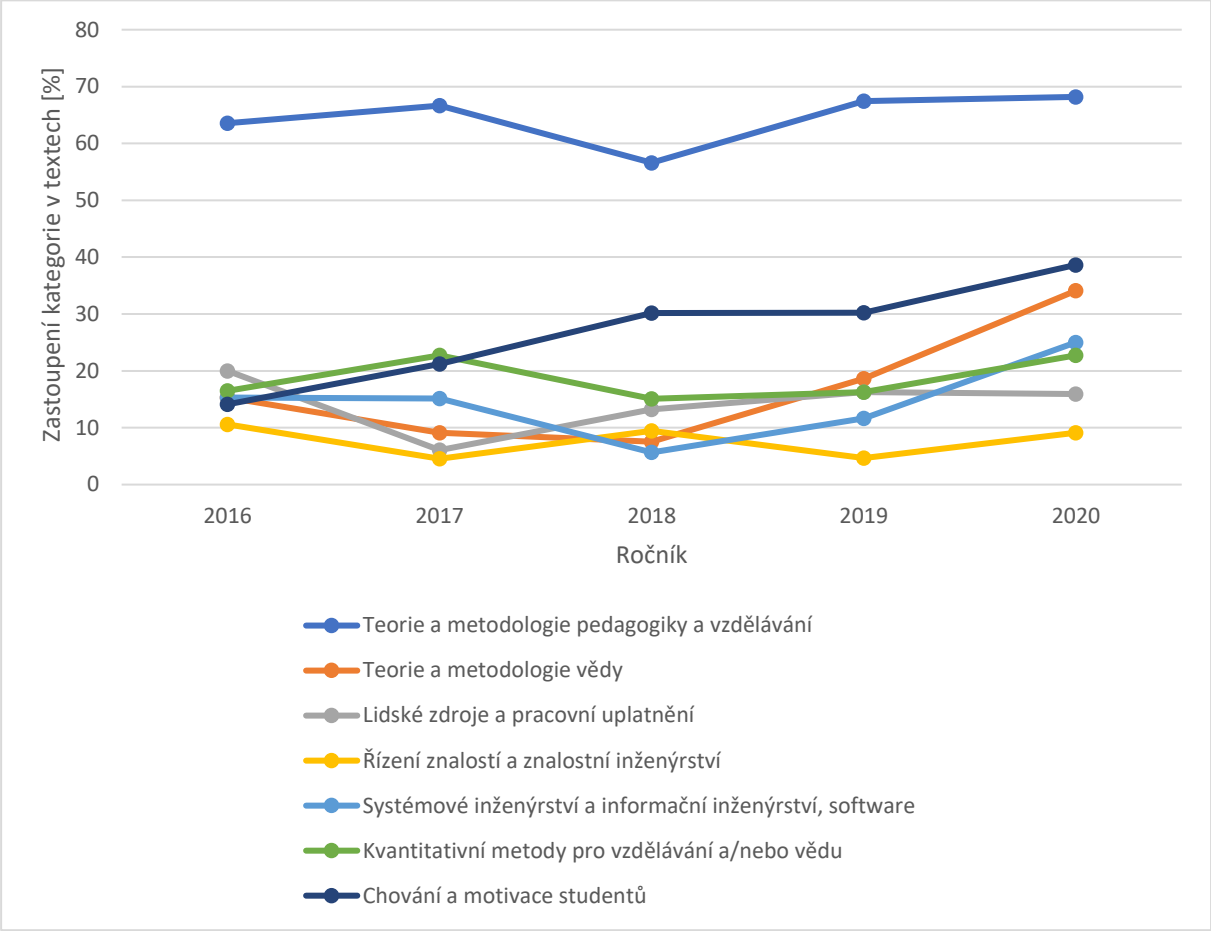
Tematický překryv

Články byly do kategorií rozřazovány pomocí měkké klasifikace, tj., že jeden článek mohl být klasifikován více než jednou kategorií, pokud to bylo třeba. V následující tabulce 11 je v řádku tematický překryv uvedeno, do kolika kategorií byl jeden článek z daného ročníku průměrně zařazen.

Ročník	2016	2017	2018	2019	2020
Počet článků v ročníku	85	66	53	43	44
Počet kategorií	132	96	73	71	94
Tematický překryv	1,55	1,45	1,38	1,65	2,14

Tabulka 11: Tematický přeryv. Zdroj: vlastní zpracování.

Graf 1 je grafickým znázorněním tabulky 10. Zobrazuje relativní četnost kategorií v jednotlivých ročnících.



Graf 1: Relativní četnost kategorií v jednotlivých ročnících

Výstupem z frekvenční analýzy provedené softwarem Statistica bylo 7 nejčastěji se v příspěvcích opakujících témat, které vytvořily kategorie, do kterých se jednotlivé příspěvky roztřídily. Na základě četností témat v jednotlivých ročnících byl vytvořen graf nejčastěji se vyskytujících témat a jejich rozložení v čase.

Z výsledků vyplývá, že celkově relativně nejvíce je zastoupena kategorie Teorie a metodologie pedagogiky a vzdělávání, následuje kategorie Chování a motivace studentů, Kvantitativní metody pro vzdělávání a / nebo vědu, Teorie a metodologie vědy, Systémové inženýrství a informační inženýrství, software, Lidské zdroje a pracovní uplatnění a nejméně zastoupenou je kategorie Řízení znalostí a znalostní inženýrství.

Konstantní trend výskytu počtu článků vykazuje kategorie Teorie a metodologie pedagogiky a vzdělávání a Řízení znalostí a znalostní inženýrství. Jasně rostoucí trend výskytu vykazuje kategorie Teorie a metodologie vědy a Chování a motivace studentů.

Konstantní trend u kategorie Teorie a metodologie pedagogiky a vzdělávání lze vysvětlit tím, že se jedná o nejobecnější téma ze všech vybraných. Naopak kategorie Řízení

znalostí a znalostní inženýrství je velmi úzce specializované téma, věnuje se mu jen malé procento autorů, kteří se tématu věnují delší dobu.

Rostoucí trend u kategorie Teorie a metodologie vědy může značit průběžné rozšiřování tematického záběru konference i na jiné vědní disciplíny než na vědu o vzdělávání. Rostoucí trend u kategorie Chování a motivace studentů lze vysvětlit obecným trendem, kdy se pozornost zaměřuje spíše na lidi, jejich motivaci a životní komfort než na technologie.

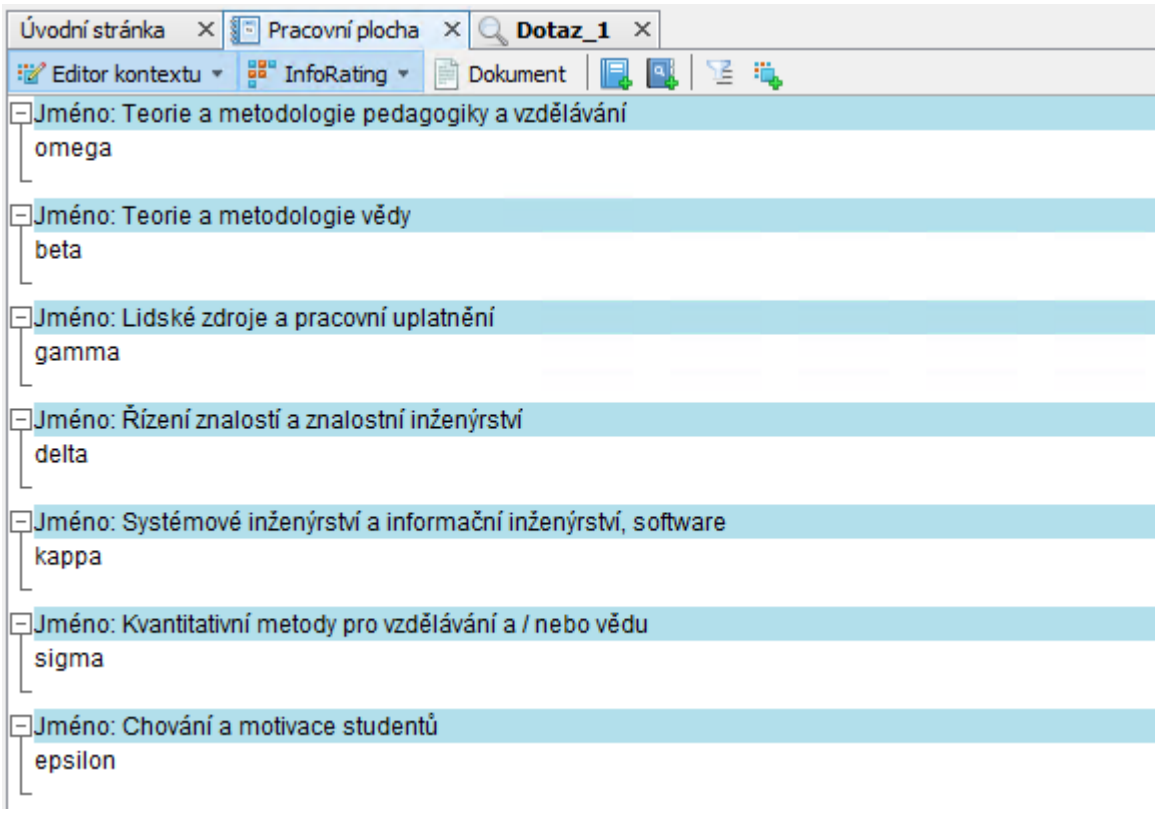
Relativně rostoucí křivky u všech kategorií lze vysvětlit tak, že články v ročníku 2020 spadaly častěji do více kategorií najednou než články z ročníků dřívějších. To lze přikládat širšímu tematickému záběru většiny článků a z toho plynoucího širšího tematického záběru klíčových slov. Na základě těchto tematicky diverzifikovaných klíčových slov bylo obtížnější klasifikovat článek pouze jednou kategorií.

Z nejobsáhlejší kategorie Teorie a metodologie pedagogiky a vzdělávání by se pro další analýzy tematicky daly vyčlenit dílčí kategorie Zkoušky a výsledky, Gamifikace jako učební metoda, Sociální media a Etika a morálka. Doplněním dílčích podkategorií do nejobsáhlejší kategorie by se klasifikace stala hierarchickou.

4.3. Kontextová analýza

4.3.1. Analýza vztahů mezi tématy

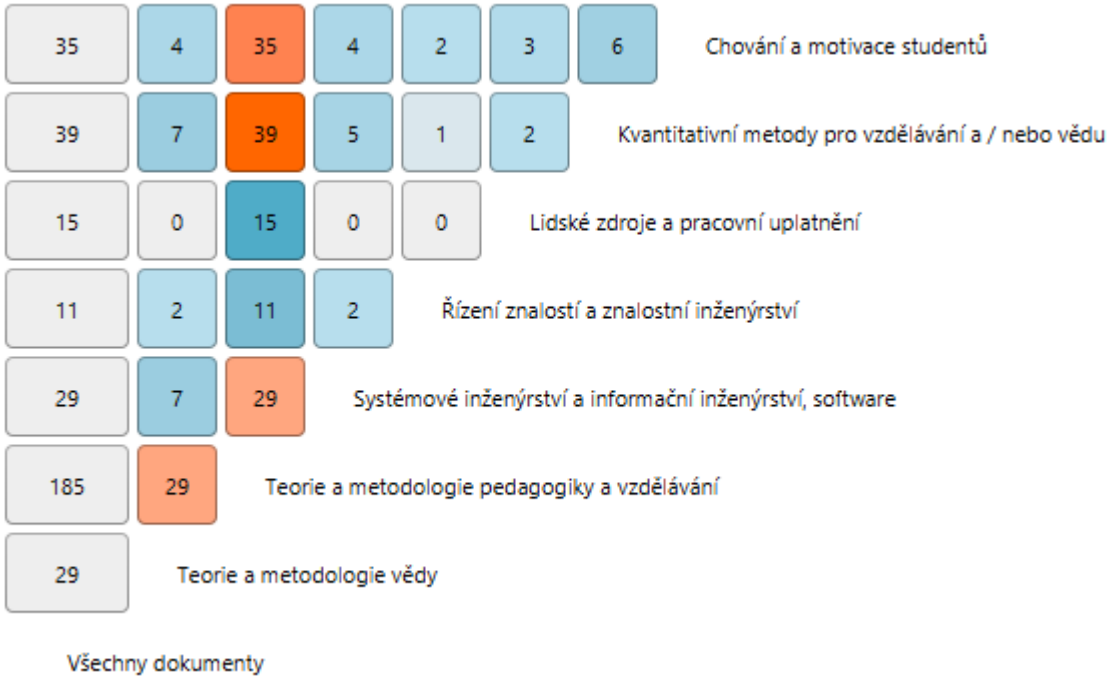
Kompletní soubor všech očištěných textů, do kterých bylo doplněno označení kategorie řeckým písmenem, byl nahrán do softwaru Tovek Tools. V modulu *InfoRating* v sekci *Editor kontextu* byly nadefinovány kontextové dotazy v jazyce *Tovek query language.* Pro každou ze zvolených kategorií byl nadefinován jeden kontextový dotaz. Jako texty dotazu byla použita řecká písmena přidělená kategoriím. Nadefinované dotazy jsou zobrazeny na obrázku 7.



Obrázek 7: Kontextové dotazy pro jednotlivé kategorie v Nástroji InfoRating. Zdoj: vlastní zpracování.

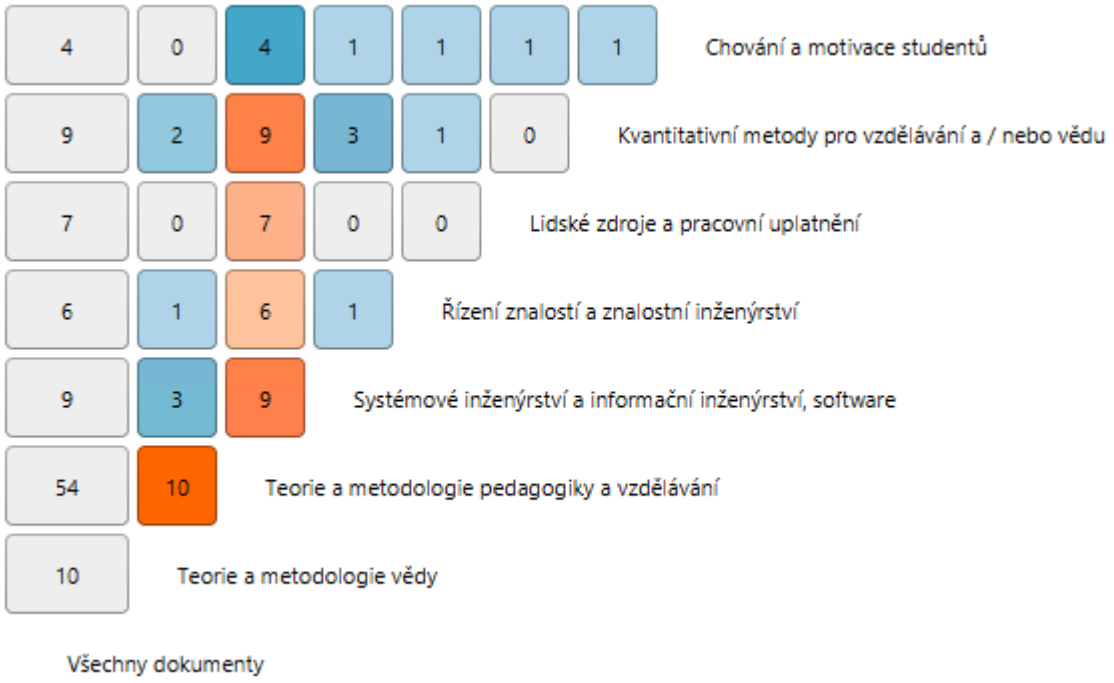
Kontextové dotazy vyhledají kategorie článků. Výsledkem kontextové analýzy je matice, která v každém poli zobrazuje, kolik článků z jedné kategorie bylo začleněno do kategorie jiné. Ze dvojic v této matici lze odvozovat vztahy mezi tématy.

Matice 1 zobrazuje vztahy mezi tématy napříč všemi zkoumanými ročníky.



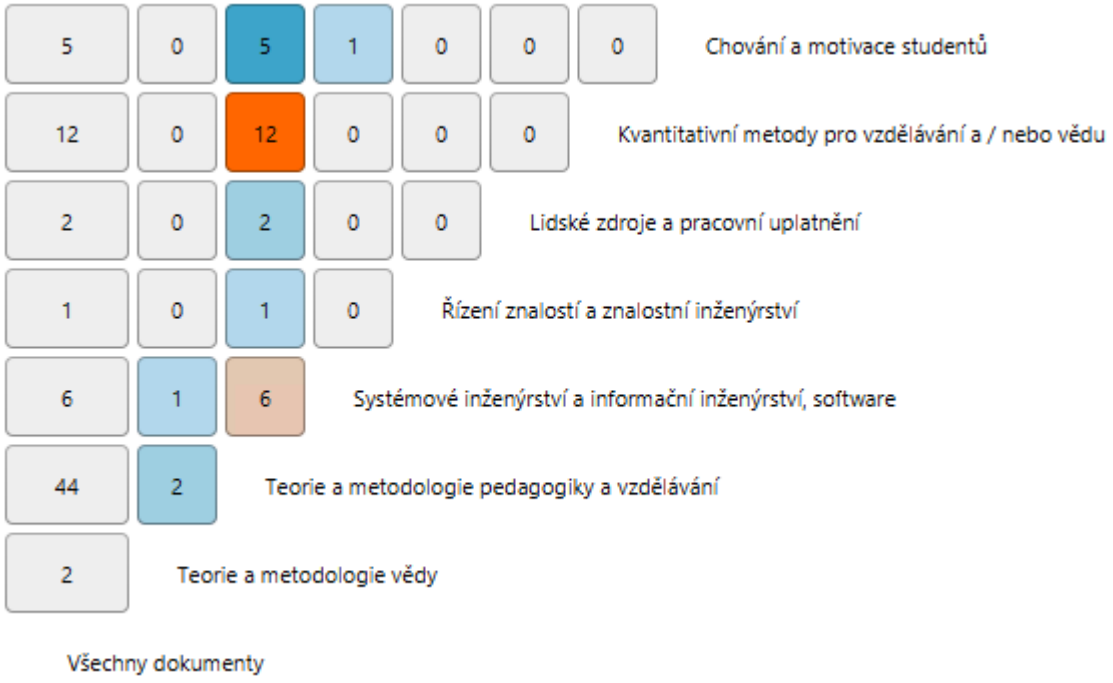
Matice 1: Vztahy mezi tématy ve všech zkoumaných ročnících. Zdroj: vlastní zpracování

Matice 2 zobrazuje vztahy mezi tématy v ročníku 2016.



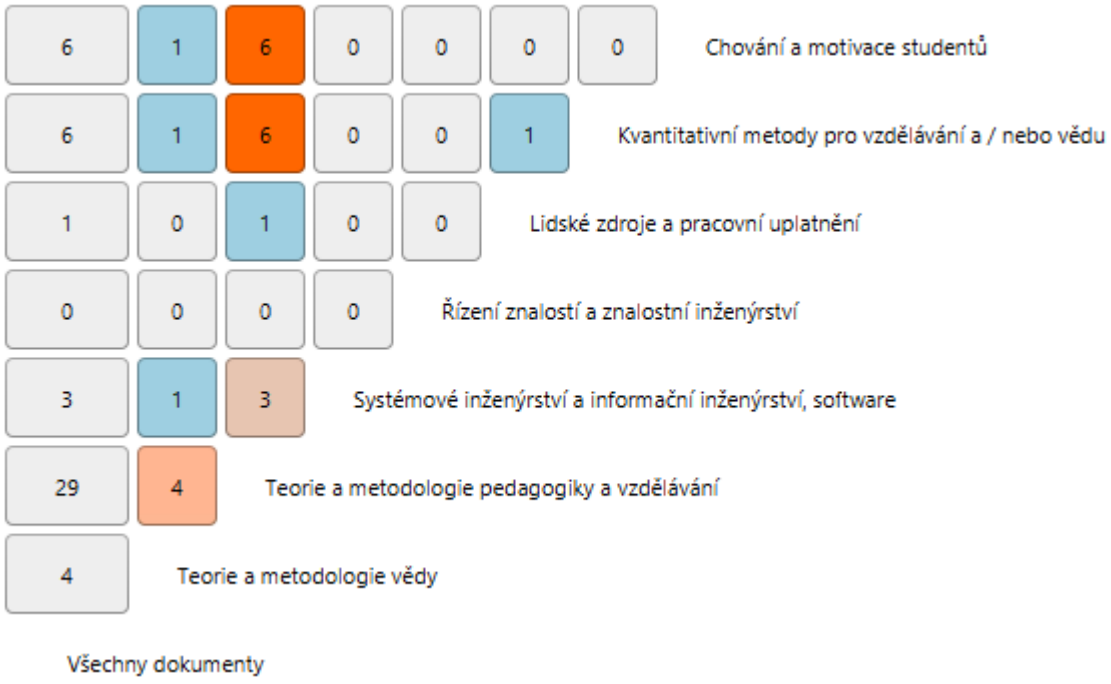
Matice 2: Vztahy mezi tématy v ročníku 2016. Zdroj: vlastní zpracování

Matice 3 zobrazuje vztahy mezi tématy v ročníku 2017.



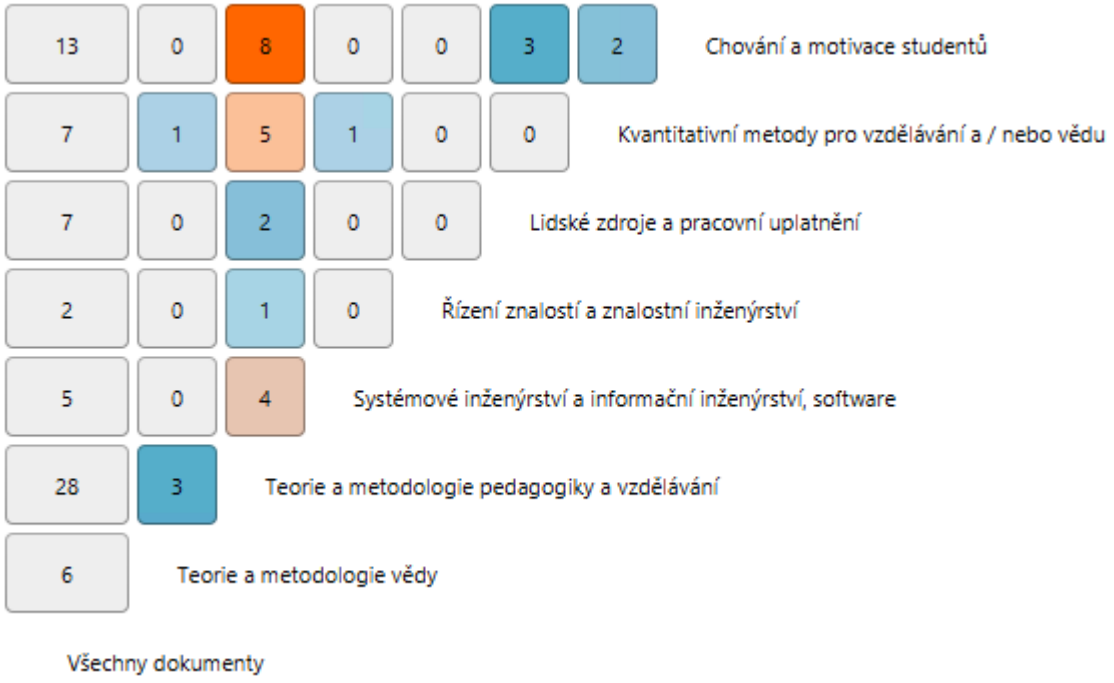
Matice 3 Vztahy mezi tématy v ročníku 2017. Zdroj: vlastní zpracování

Matice 4 zobrazuje vztahy mezi tématy v ročníku 2018.



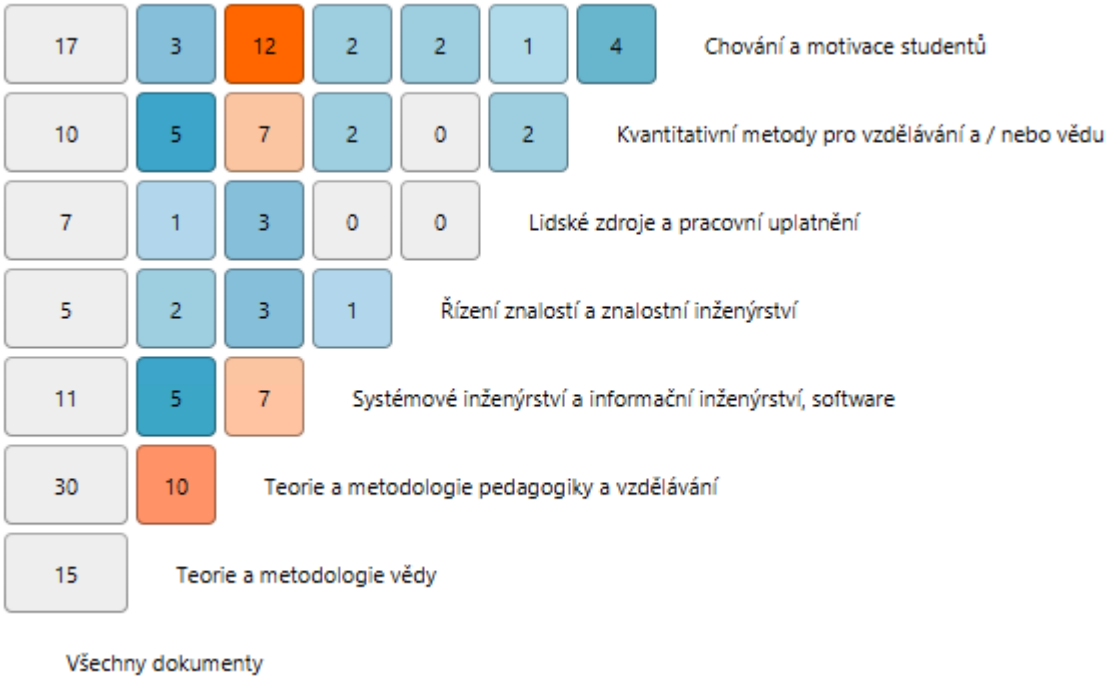
Matice 4 Vztahy mezi tématy v ročníku 2018. Zdroj: vlastní zpracování

Matice 5 zobrazuje vztahy mezi tématy v ročníku 2019.



Matice 5: Vztahy mezi tématy v ročníku 2019. Zdroj: vlastní zpracování

Matice 6 zobrazuje vztahy mezi tématy v ročníku 2020.



Matice 6: Vztahy mezi tématy v ročníku 2020. Zdroj: vlastní zpracování

Výsledky analýzy vztahů mezi tématy

Kategorií, která se nejčastěji vyskytuje spolu s ostatními kategoriemi, je Teorie a metodologie pedagogiky a vzdělávání. Naopak nejjednoznačnější kategorií je kategorie Lidské zdroje a pracovní uplatnění, vyskytuje se 15krát spolu s kategorií Teorie a metodologie pedagogiky a vzdělávání, spolu s ostatními kategoriemi se nevyskytuje.

Nejčastější dvojice kategorií, do které jsou články zařazeny, je Teorie a metodologie pedagogiky a vzdělávání a Kvantitativní metody pro vzdělávání a / nebo vědu, druhou nejčastější dvojicí je Teorie a metodologie pedagogiky a vzdělávání a Chování a motivace studentů, dále pak Teorie a metodologie pedagogiky a vzdělávání a Teorie a metodologie vědy spolu s dvojicí Teorie a metodologie pedagogiky a vzdělávání a Systémové inženýrství a informační inženýrství, software.

Skutečnost, že kategorie Teorie a metodologie vědy je nejčastější kategorií, která se vyskytuje spolu s ostatními kategoriemi, je jednak její relativně největší zastoupení v příspěvcích ale také její relativně nejširší tematický záběr.

4.3.2. Geografická analýza

Kompletní soubor všech očištěných textů, do kterých bylo doplněno označení kategorie řeckým písmenem, byl nahrán do softwaru Tovek Tools do modulu *Editor dotazů*. V Editoru dotazů byly pomocí nástroje *Extrakce entit* extrahovány názvy zemí příkazem *Entita Stát*, jak je vidět na obrázku 8.

Skóre	Titulek	Datum	Entita stát
100		2021-02-27	România
100		2021-02-27	France, România
100		2021-02-27	Czech Republic, Slovakia
100		2021-02-27	România
100		2021-02-27	Serbia
100		2021-02-27	Korea (South)
100		2021-02-27	Czech Republic
100		2021-02-27	European Union, Serbia
100		2021-02-27	Czech Republic
100		2021-02-27	Czech Republic, European Union, France, Germany, Italy, Netherlands, Norway, United States
100		2021-02-27	Czech Republic, European Union
100		2021-02-27	Turkey
100		2021-02-27	Czech Republic, Korea (South)
100		2021-02-25	
100		2021-02-27	Czech Republic
100		2021-02-27	Canada, Czech Republic
100		2021-02-27	Czech Republic, Poland, Slovakia, Spain
100		2021-02-27	
100		2021-02-27	Czech Republic, European Union, România, Spain
100		2021-02-27	Czech Republic
100		2021-02-25	
100		2021-02-25	Czech Republic, Netherlands
100		2021-02-25	Czech Republic, România, United Republic of Tanzania
100		2021-02-27	Czech Republic, European Union, Korea (South)
100		2021-02-27	Czech Republic
100		2021-02-27	Czech Republic
100		2021-02-25	Austria, Czech Republic, Germany
100		2021-02-27	
100		2021-02-27	Russia
100		2021-02-27	
100		2021-02-27	
100		2021-02-27	Czech Republic, European Union, United States
100		2021-02-27	Czech Republic
100		2021-02-25	Slovakia
100		2021-02-27	Czech Republic, Finland, Singapore
100		2021-02-27	Czech Republic, Spain

Obrázek 8: Entity extrahované nástrojem Extrakce entit softwaru Tovek Tools. Zdroj: vlastní zpracování.

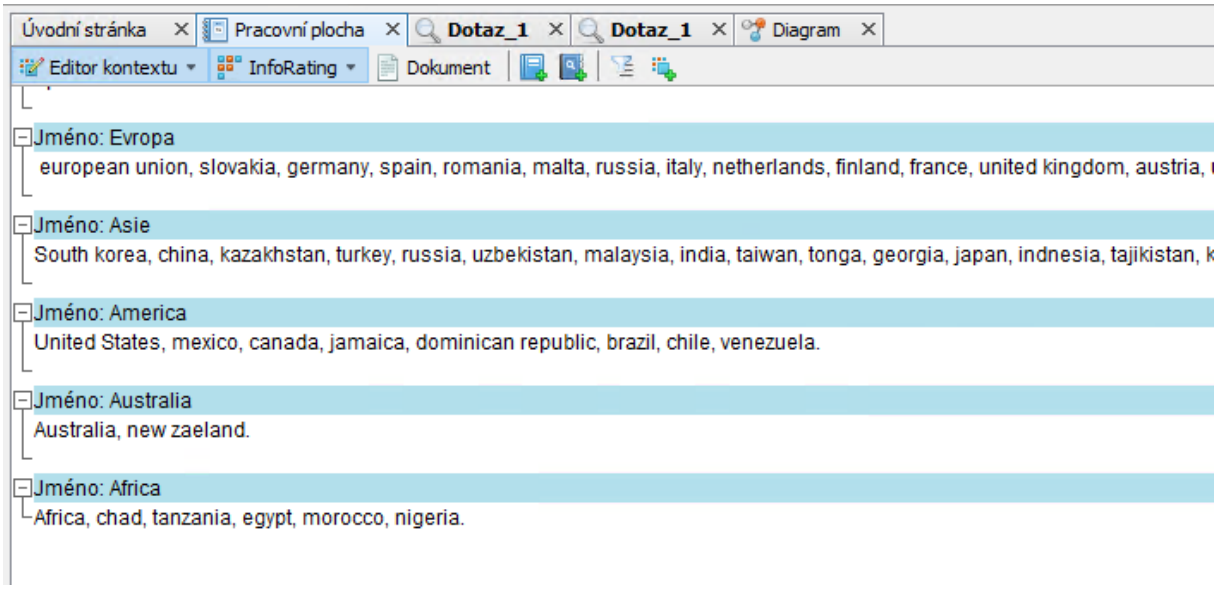
Výsledky ze softwaru Tovek Tools byly exportovány do formátu .xls. Pomocí platformy MS Excel byly výsledky seřazeny podle četnosti výskytu v článcích do tabulky 12. Bylo nalezeno celkem 83 zemí v 221 článcích. Mezi země program počítá i entitu „Evropská Unie“. Pokud byla v jednom článku nějaká země zmíněna vícekrát, tento modul ji započítal pouze jednou.

Země	Počet článků	Země	Počet článků	Země	Počet článků	Země	Počet článků
czech republic	142	portugal	6	croatia	3	jordan	1
european union	38	poland	6	latvia	3	jamaica	1
south korea	36	norway	5	georgia	3	united arab emirates	1
slovakia	32	greece	5	africa	3	guinea	1
germany	24	uzbekistan	5	japan	3	dominican	1
united states	15	switzerland	5	cyprus	3	morocco	1
spain	14	hungary	5	serbia	3	brazil	1
românia	14	malaysia	4	new zealand	2	pakistan	1
russia	13	malta	4	indonesia	2	moldova	1
china	12	ireland	4	tajikistan	2	bangladesh	1
italy	12	belgium	4	belarus	2	armenia	1
australia	11	sweden	4	kyrgyzstan	2	albania	1
netherlands	11	canada	4	iceland	2	angola	1
finland	10	india	4	singapore	2	azerbaijan	1
france	10	estonia	4	vietnam	2	turkmenistan	1
united kingdom	9	lithuania	4	brunei	2	chile	1
mexico	9	taiwan	4	denmark	2	luxembourg	1
kazakhstan	9	bulgaria	4	syria	2	macedonia	1
austria	8	egypt	4	thailand	1	nigeria	1
ukraine	7	tonga	4	chad	1	venezuela	1
turkey	7	slovenia	3	tanzania	1		

Tabulka 12: Četnosti výskytu názvů zemí v článcích. Zdroj: vlastní zpracování.

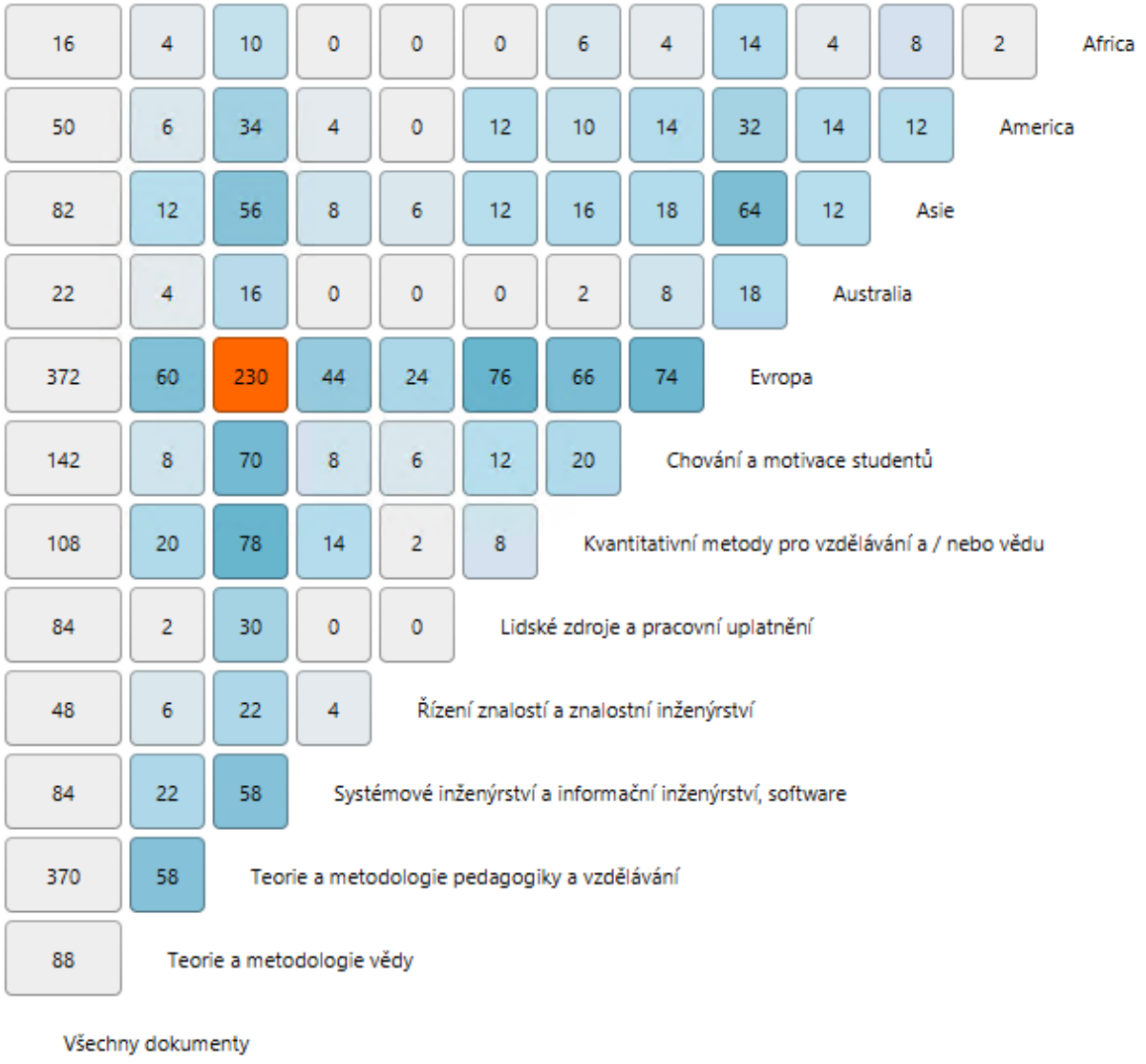
Kontextová matice pro jednotlivé nalezené státy by byla příliš obsáhlá a nepřehledná. Pro větší přehlednost byly státy rozděleny do skupin podle geografické příslušnosti k jednotlivým kontinentům. Pokud nebylo zcela jasné, k jakému kontinentu by se měl stát počítat, například z důvodu, že zeměpisně leží na více kontinentech, rozhodovala poloha jeho hlavního města. Severní a Jižní Amerika jsou počítány jako jeden kontinent. V modulu *InfoRating* v sekci *Editor kontextu* byly nadefinovány kontextové dotazy. Pro každý z kontinentů byl nadefinován jeden kontextový dotaz. Jako texty dotazu byly použity názvy států vygenerované v předchozím kroku.

Náhled na nadefinované dotazy je vidět na obrázku 9.



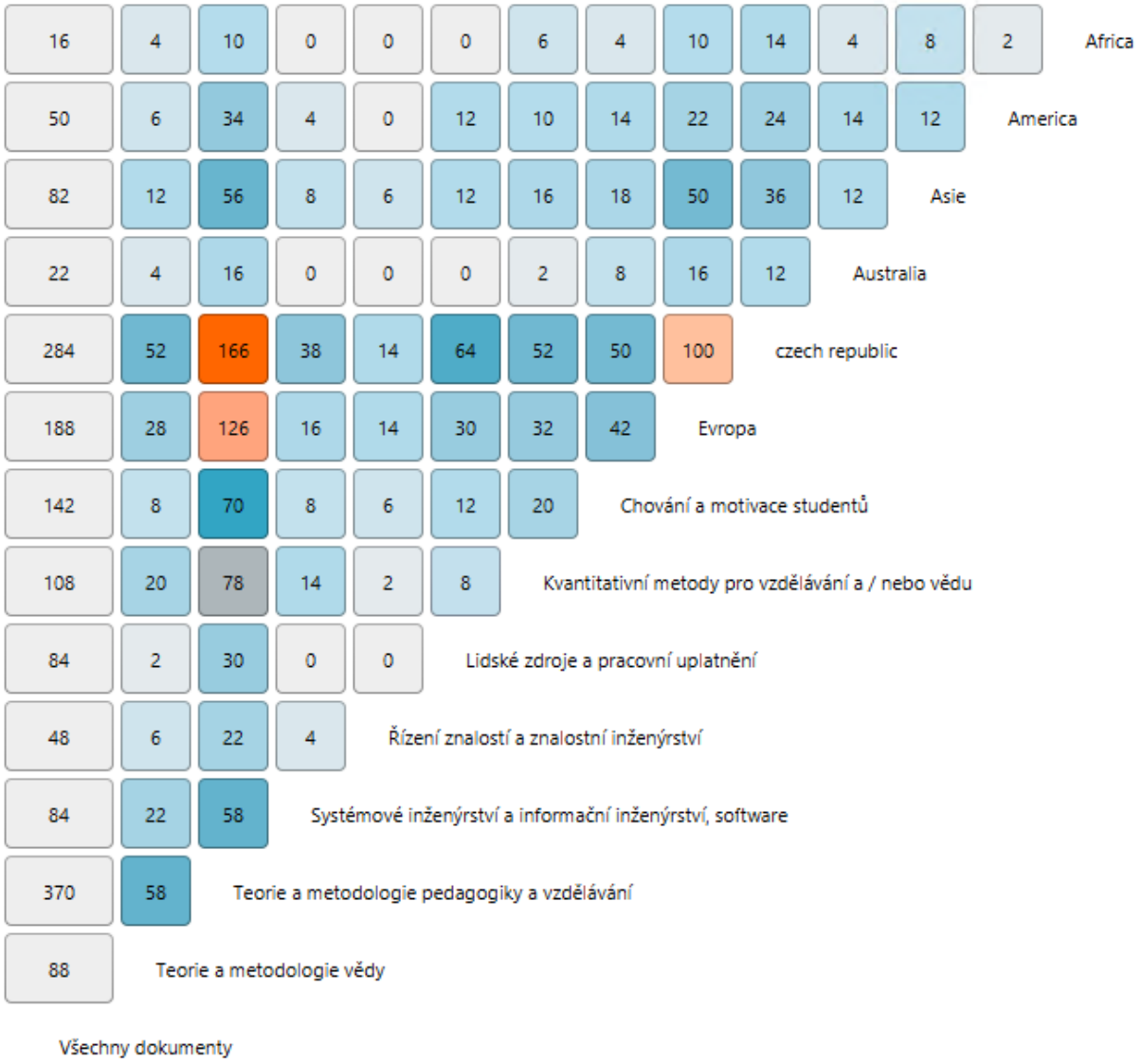
Obrázek 9: Kontextové dotazy pro jednotlivé kontinenty v nástroji InfoRating. Zdroj: vlastní zpracování.

Celkem bylo analyzováno 291 článků a 221 výskytů entity stát. Souhrnné výsledky ze všech pěti ročníků jsou patrné z matice 7. Nejvíce společných výskytů bylo nalezeno u kontinentu Evropa a kategorie Teorie a metodologie pedagogiky a vzdělávání.



Matice 7: Vztahy mezi kontinenty a kategoriemi. Zdroj: vlastní zpracování.

Pro větší přesnost byla entita Česká republika vyčleněna z entity Evropa. Výsledky jsou patrné z matice 8.



Matice 8: Vztahy mezi kontinenty a kategoriemi po vyčlenění entity Česká republika z entity Evropa. Zdroj: vlastní zpracování.

Města v ČR

Entita Česká republika byla v rámci analýzy dále granularizována na podle měst. V Editoru dotazů byly pomocí nástroje *Extrakce entit* extrahovány názvy měst příkazem *Entita město*.

Skóre	Titulek	Datum	Entita stát	Entita město
100		2021-02-27	Czech Republic	Poole, Praha
100		2021-02-27		Pereira, Ústí nad Labem
100		2021-02-27	Slovakia	Paterson, Xinzhou
100		2021-02-27		Pardubice, Praha
100		2021-02-25		Ostrava
100		2021-02-27		Ostrava
100		2021-02-27	Czech Republic	Ostrava
100		2021-02-27	China	Ostrava
100		2021-02-27	Czech Republic, Dominican Republic, E...	Ostrava
100		2021-02-27		Oral, Yekaterinburg
100		2021-02-27	Czech Republic, Norway	Oral, Praha
100		2021-02-27		Omsk, Rio Grande, The Valley
100		2021-02-25	Czech Republic, European Union, Ger...	Olomouc, Ostrava, Třinec, Šumperk
100		2021-02-27	Czech Republic, European Union, Net...	Okazaki, Praha, Saint George s, Warren
100		2021-02-27	Bulgaria, Finland, Hungary, Italy	Nehe
100		2021-02-27	Germany	Murcia, Praha, Reading
100		2021-03-05	Czech Republic	Mosul, Praha
100		2021-02-27	Russia, Tonga	Moscow, Samsun, Yekaterinburg
100		2021-02-27		Montréal, Praha
100		2021-02-27	Czech Republic, Korea (South), Ukraine	Montgomery, České Budějovice
100		2021-02-27	Australia, Brazil, Germany, New Zeala...	Mobile, Sousse
100		2021-02-27	Czech Republic	Mobile, Ostrava
100		2021-02-27	Czech Republic	Mobile
100		2021-02-25	Czech Republic, Netherlands	Milan, Praha, York
100		2021-02-27		Milan
100		2021-03-01		Milan
100		2021-02-27		Milan
100		2021-02-27		Milan
100		2021-02-27	China, Ukraine, United States	Mary, Xinxiang
100		2021-02-27		Male, Virginia
100		2021-02-27	Greece, România	Male, Thessaloníki, Timişoara
100		2021-02-27	Czech Republic	Male, Praha, Zhengzhou
100		2021-02-27	Czech Republic, Slovakia	Male, Praha, Stanley
100		2021-02-27		Male, Praha
100		2021-02-27	Australia, Brunei	Male, Praha
100		2021-02-27		Male, Praha

Výsledky ze softwaru Tovek Tools byly stejně jako v předchozím kroku exportovány do formátu .xls. Pomocí platformy MS Excel byly výsledky očištěny od chybných nálezů, a seřazeny podle četnosti výskytu v článcích. Celkově bylo nalezeno 175 měst v 235 článcích, kompletní výsledky analýzy měst jsou v Příloze 2 této DP. Pro další analýzu byla z kompletních výsledků vybrána pouze česká města.

V článcích bylo nalezeno celkem 28 českých měst, která byla podle četnosti výskytu seřazena do tabulky 13.

Město	Počet článků	Město	Počet článků	Město	Počet článků	Město	Počet článků
praha	102	hradec králové	5	tábor	2	klatovy	1
ústí nad labem	13	olomouc	3	jihlava	2	cheb	1
Ostrava	11	karlovy vary	3	žatec	1	třinec	1
české budějovice	11	liberec	3	karviná	1	libeň	1
bratislava	9	zlín	3	mladá boleslav	1	opava	1
brno	8	plzeň	2	jičín	1	děčín	1
pardubice	6	šumperk	2	litoměřice	1	teplice	1

Tabulka 13: Četnosti výskytu názvů českých měst v článcích. Zdroj: vlastní zpracování.

Výsledky geografické analýzy

Extrakcí entity stát bylo nalezeno celkem 83 zemí vyskytujících se v 221 článcích. Nejčastěji se v článcích vyskytovala entita Česká republika, druhou nejčastější byla Evropská unie, třetí v pořadí byla Jižní Korea.

Ze sedmi známých kontinentů byly v článcích nalezeny geografické údaje týkající se šesti z nich. Ze 44 evropských států byly v článcích nalezeny údaje týkající se 40 z nich.

Nejčastěji se v článcích vyskytoval kontinent Evropa společně s kategorií Teorie a metodologie pedagogiky a vzdělávání. Jakmile byla z entity Evropa vyčleněna Česká republika do samostatné entity, zaznamenala nejvyšší výskyt kategorie Teorie a metodologie pedagogiky a vzdělávání a Česká republika. Převažující výskyt České republiky v entitě stát lze vysvětlit tím, že články jsou psané převážně českými autory na území České republiky a samotná konference se koná na území České republiky.

Tematickými kategoriemi silně závislými na geografii jsou Teorie a metodologie pedagogiky a vzdělávání, Lidské zdroje a pracovní uplatnění, Kvantitativní metody pro vzdělávání a / nebo vědu a Chování a motivace studentů.

Nejméně geograficky závislými tématy jsou témata Řízení znalostí a znalostní inženýrství a Systémové inženýrství a informační inženýrství, software.

4.3.3. Analýza autorů a autorských kolektivů

Cílem této kapitoly bylo vyhodnotit, zda je skladba publikujících autorů napříč ročníky homogenní, či zda a jakým způsobem se v průběhu času proměňuje.

Pomocí softwaru Statistica byla z názvů souborů extrahována jména autorů. V platformě MS Excel byly výsledky očištěny a pomocí podmíněného formátování provedeny následující analýzy.

Analýza počtu autorů a článků

Celkem bylo v 291 článcích identifikováno 757 autorů. Někteří autoři publikovali opakovaně. Po odstranění duplicitních jmen bylo identifikováno dohromady 403 jedinečných jmen autorů. Tabulka 14 znázorňuje kvantitativní přehled, kolik autorů publikovalo jaké množství článků.

Počet článků celkem v ročnících od jednoho autora	**Počet autorů**
9	1
8	1
7	5
6	4
5	11
4	28
3	41
2	81
1	231

Tabulka 14: Přehled počtu článků publikovaných autory. Zdroj: vlastní zpracování.

Počet autorů v autorských kolektivech

Z předchozí analýzy autorů a článků vyplývá, že na jednom článku se průměrně podílelo 2,6 autora. Nejmenší analyzovaný autorský kolektiv má jednoho člena, největší má pět členů. Tabulka 15 znázorňuje průměrný počet autorů v autorském kolektivu v jednotlivých ročnících.

Rok	**Počet autorů**	**Průměrný počet autorů na jeden příspěvek**
2016	205	2,4
2017	176	2,7
2018	137	2,6
2019	130	3,0
2020	109	2,5

Tabulka 15: Průměrný počet autorů v autorských kolektivech v jednotlivých ročnících. Zdroj: vlastní zpracování.

Přesně se opakující autorské kolektivy

V této podkapitole bylo vyhodnoceno, zda existují autorské kolektivy, které v nezměněné podobě publikovaly ve více ročnících. V této kapitole nebylo rozlišováno, kolik

autorů tvoří autorský kolektiv a jako kolektiv se v tomto případě počítají i jednotlivci. V analýze bylo uvažováno pouze se jmény autorů, nikoliv jejich pořadí v publikujícím kolektivu.

V tabulce 16 jsou zaznamenány autorské kolektivy a počty ročníků, ve kterých v nezměněné podobě publikovaly.

Počet ročníků			
2	**3**	**4**	**5**
Kazda, Petraskova, Rosa	Fejfarova, Fejfar		Beranek, Remes
Samkova, Ticha	Vladisavljevic, Solesa, Stanisavljevic		Martinik
Stefl	Samková		Sigmund
Cihlář, Eisenmann, Hejnová, Přibyl			
Höfrová			
Husák, Hudečková			
Klůfa			
Moraová, Novotná			
Otavová, Sýkorová			
Beranová, Navrátilová, Šíma			
Hrubý, Staňková			

Tabulka 16: Autorské kolektivy a počty ročníků, ve kterých v nezměněné podobě publikovaly. Zdroj: vlastní zpracování.

Opakované autorství v rámci ročníku a mezi ročníky

Tato kapitola analyzuje autorství jednotlivců. Zaměřuje se na vícečetné autorství, a to jak v rámci jednoho ročníku, tak i meziročně.

a) Vícečetné autorství v rámci jednoho ročníku

Následující tabulka 17 ukazuje počty autorů a počty článků, které v jednotlivých ročnících daní autoři publikovali. Nejčastěji se jedno jméno vyskytovalo jako autor u čtyř článků v ročníku 2016.

	Rok				
	2016	**2017**	**2018**	**2019**	**2020**
Počet autorů s **jedním** publikovaným článkem	124	135	101	118	99
Počet autorů se **dvěma** publikovanými články	28	14	12	6	5
Počet autorů se **třemi** publikovanými články	3	3	4	0	0
Počet autorů se **čtyřmi** publikovanými články	4	1	0	0	0

Tabulka 17: Vícečetné autorství v rámci jednoho ročníku. Zdroj: vlastní zpracování.

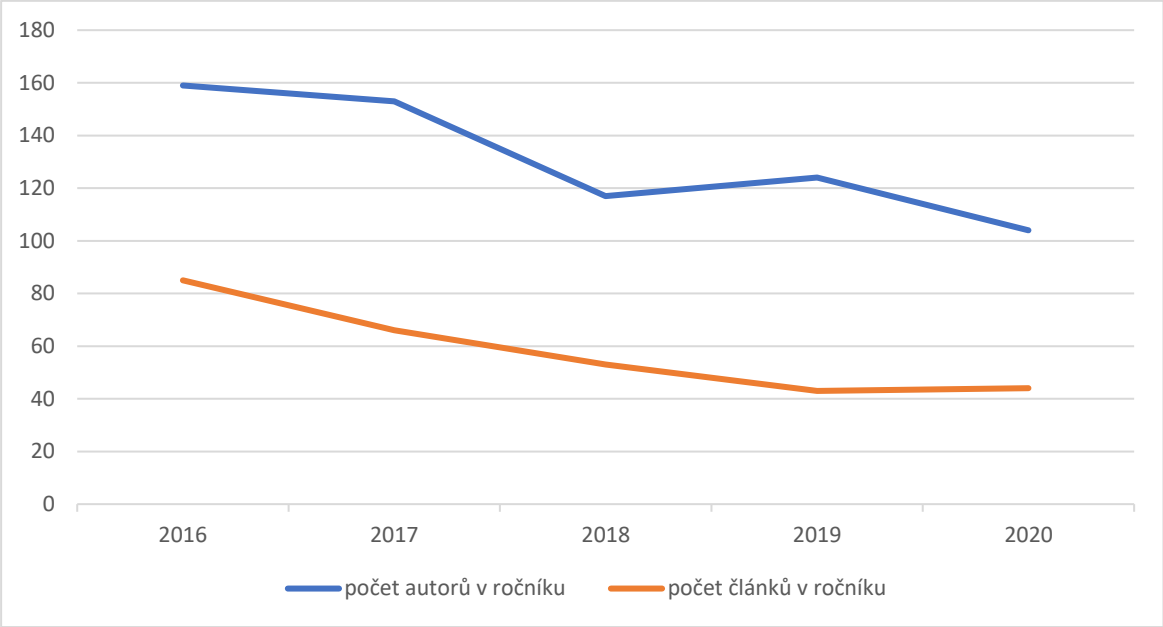
Do následující analýzy je každé jméno započítáno jako autor pouze jednou a všechna opakování jsou zanedbána. Podílem mezi počtem autorů a počtem článků v ročníku je bezrozměrné číslo sloužící k porovnání mezi ročníky. Čím vyšší číslo, tím méně se jména autorů v rámci ročníků opakují.

Nejčastěji autoři publikovali více než jeden článek v prvním ročníku konference v roce 2016, nejméně v ročníku 2019. Výsledky jsou patrné z tabulky 18.

	Rok				
	2016	**2017**	**2018**	**2019**	**2020**
počet autorů v ročníku	159	153	117	124	104
počet článků v ročníku	85	66	53	43	44
počet autorů/počet článků	1,9	2,3	2,2	2,9	2,4

Tabulka 18: Podíl mezi počtem autorů a počtem článků v ročníku. Zdroj: vlastní zpracování.

Graf 2 je grafickým znázorněním tabulky 18. Zobrazuje závislost počtu článků v ročníku na počtu autorů v ročníku.



Graf 2: závislost počtu článků v ročníku na počtu autorů v ročníku.

b) Vícečetné autorství mezi ročníky

Z celkového počtu opakování se autorů v jednotlivých článcích a tabulky 17 bylo vypočítáno, že 92 autorů publikovalo článek více než v jednom ročníku.

Zhodnocení výsledků

Celkem bylo v 291 článcích identifikováno 757 autorů. Někteří autoři publikovali opakovaně. Po odstranění duplicitních jmen bylo identifikováno dohromady 403 jedinečných jmen autorů. Nejvíce se jeden autor podílel na devíti článcích. Nejvíce autorů (231) publikovalo pouze jeden článek.

Průměrně se na jednom článku podílel autorský kolektiv o velikosti 2,6 člena. Nejmenší analyzovaný autorský kolektiv má 1 člena, největší má pět členů. Nejmenší počet autorů na jeden článek byl zaznamenán z roce 2016 (2,4), největší počet v roce 2019 (3,0).

V nezměněné sestavě ve více ročnících publikovalo 17 autorských kolektivů. Společně ve dvou ročnících publikovalo 11 autorských kolektivů, ve třech ročnících 3 autorské kolektivy a ve všech pěti ročnících publikovaly v nezměněné sestavě 3 kolektivy. Bylo zjištěno, že vědecké kolektivy zkoumající podobná témata se v autorských kolektivech prolínají.

Co se týče vícečetného autorství v rámci jednotlivých ročníků, bylo zjištěno, že nejvíce autoři publikovali více než jeden článek v prvním ročníku konference v roce 2016, nejméně

v ročníku 2019. Absolutní hodnota počtu autorů (bez opakování jmen) v ročníku klesá společně s počtem článků v ročníku.

Autorství u více než jednoho článku bylo v rámci jednotlivých ročníků nalezeno 80krát, mezi ročníky 92krát. Z těchto výsledků lze usuzovat, že dlouholetými studiemi se zabývá jen o něco více autorů než více krátkodobými studiemi současně.

5. Závěr

Na základě provedených frekvenčních analýz bylo zvoleno 7 nejčastěji se v příspěvcích opakujících témat, které vytvořily kategorie, do kterých se jednotlivé příspěvky roztřídily. Bylo zjištěno, že celkově relativně nejvíce je zastoupena kategorie Teorie a metodologie pedagogiky a vzdělávání a nejméně zastoupenou je kategorie Řízení znalostí a znalostní inženýrství. Z nejobsáhlejší kategorie Teorie a metodologie pedagogiky a vzdělávání by se pro další analýzy tematicky daly vyčlenit dílčí kategorie Zkoušky a výsledky, Gamifikace jako učební metoda, Sociální media a Etika a morálka. Kategorií, která se v článcích nejčastěji vyskytuje spolu s ostatními kategoriemi je kategorie s nejširším tematickým záběrem.

Konstantní trend výskytu počtu článků v čase vykazují kategorie, které se zabývají velmi obecným nebo naopak velmi úzce specializovaným tématem. Rostoucí trend u některých kategorií kopíruje obecné rostoucí trendy jako větší generalizace nebo zaměření pozornosti spíše na lidi, jejich motivaci a životní komfort než na technologie.

Geografickou analýzou bylo zjištěno, že nejčastěji se v článcích vyskytovala entita Česká republika, druhou nejčastější byla Evropská unie, třetí v pořadí byla Jižní Korea. Analyzované články mají široký geografický záběr, obsahovaly geografické údaje z šesti kontinentů a 90% evropských zemí. Nejčastěji se v článcích vyskytoval kontinent Evropa, stát Česká republika a město Praha. Převažující výskyt České republiky v entitě stát lze vysvětlit tím, že články jsou psané převážně českými autory na území České republiky a samotná konference se koná na území České republiky. Tematickými kategoriemi silně závislými na geografii jsou Teorie a metodologie pedagogiky a vzdělávání, Lidské zdroje a pracovní uplatnění, Kvantitativní metody pro vzdělávání a / nebo vědu a Chování a motivace studentů. Nejméně geograficky závislými tématy jsou témata Řízení znalostí a znalostní inženýrství a Systémové inženýrství a informační inženýrství, software.

Celkem bylo v 291 článcích identifikováno 757 autorů. Někteří autoři publikovali opakovaně. Po odstranění duplicitních jmen bylo identifikováno dohromady 403 jedinečných jmen autorů. Průměrně se na jednom článku podílel autorský kolektiv o velikosti 2,6 člena. V nezměněné sestavě ve více ročnících publikovalo 17 autorských kolektivů. Bylo zjištěno, že vědecké kolektivy zkoumající podobná témata se v autorských kolektivech prolínají. Absolutní hodnota počtu autorů (bez opakování jmen) v ročníku klesá společně s počtem článků v ročníku. Bylo zjištěno, že dlouholetými studiemi se zabývá jen o něco více autorů než více krátkodobými studiemi současně.

V příštích ročnících lze předpokládat, že se bude zvyšovat podíl článků s tématem Chování a motivace studentů, počet přijatých konferenčních příspěvků se bude pohybovat mezi 40 – 50. Lze také předpokládat, že bude klesat počet publikujících autorů. Geografický rozptyl publikovaných článků lze předpokládat opět široký, stejně jako ve sledovaných ročnících.

Seznam použitých zdrojů

1. About the conference [online]. c2018 [cit.2020-11-25]. Prague: Czech University of Life Sciences Prague. Dostupné z: https://erie.v2.czu.cz/en/r-13310-about-the-conference

2. ALLAHYARI, M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D.; GUTIERREZ, J. B.; KOCHUT, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques [online]. 2017. Dostupné z: http://cobweb.cs.uga.edu/~safaei/A%20Brief%20Survey%20of%20Text%20Mining%20Classification,%20Clustering%20and%20Extraction%20Techniques.pdf

3. AMBUEHL, S.; SHENGWU. L. Belief updating and the demand for information [online]. Elsevier, 2008. Dostupné z: https://scholar.harvard.edu/files/shengwu_li/files/infodemand.pdf

4. BAEZA-YATES, R., RIBEIRO-NETO, B. Modern information retrieval [online]. Harlow: Pearson, 1999 [cit. 2020-11-21]. ISBN 0-201-39829-X.

5. BERRY, M. W. et al.; WILEY, J. & Sons. Text Mining : Applications and Theory [online]. *Incorporated*, 2010. Dostupné z: http://ebookcentral.proquest.com/lib/techlib-ebooks/detail.action?docID=496056.

6. DIAZ, G. Stopwords-iso [online] c2016 [cit.2021-01-24]. Dostupné z: https://github.com/stopwords-iso/stopwords-en/blob/master/stopwords-en.txt

7. DÖMEOVÁ, L., HOUŠKA, M., HOUŠKOVÁ BERÁNKOVÁ, M. Systems Approach to Knowledge Modelling. 1st ed. Prague: Graphical Studio, 2008. 282 s. ISBN 978-80-86703-30-5.

8. FELDMAN, R., SANGER, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, New York, 2007.

9. FRYDRYCHOVÁ, A. Praktické možnosti analýzy textových dat se zaměřením na analýzu sentimentu a vizualizaci: diplomová práce. Brno: Masarykova univerzita v Brně, Filozofická fakulta, 2020. Dostupné z: https://is.muni.cz/th/n38t2/diplomova_prace.pdf

10. GARRETA, R. Text mining [online]. c2020 [cit. 2020-12-03]. Dostupné z: https://monkeylearn.com/text-mining/

11. HAN, J.; SANGIORGI, F. Searching for information [online]. Elsevier, 2018. Dostupné z: https://www.researchgate.net/publication/322947957_Searching_for_Information

12. HEER, J. Data visualization, Text Visualization [online]. University of Washington, [cit.2021-02-17]. Dostupné z: https://courses.cs.washington.edu/courses/cse512/15sp/lectures/CSE512-Text.pdf

13. INDURKHYA, N.; DAMERAU, F. J. Handbook of Natural Language Processing [online]. CRC Press LLC, 2010. Dostupné z: http://ebookcentral.proquest.com/lib/techlib-ebooks/detail.action?docID=565922.

14. KEMPE, S. The Data – Information – Knowledge Cycle [online]. c2013 [cit.2021-01-25]. Dostupné z: https://www.dataversity.net/the-data-information-knowledge-cycle/#

15. KUBIŠ, L. Kontextová analýza textu: diplomová práce. Ostrava: VŠB – Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, 2019. Dostupné z: https://dspace.vsb.cz/bitstream/handle/10084/138718/KUB0355_FEI_N2647_2612T025_2019.pdf?sequence=1&isAllowed=y

16. MANNING, CH. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval [online]. Cambridge University Press, 2008. Dostupné z: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

17. MUDRYCHOVÁ, K. Atributy znalostí a jejich hodnocení: disertační práce. Praha: Česká zemědělská univerzita, Provozně ekonomická fakulta, 2020

18. PARALIČ, J. a kol. Dolovanie znalostí z textov [online]. Technická univerzita v Košiciach, 2010 [cit. 2020- 11-31]. ISBN 978-80-89284-62-7. Dostupné z: http://people.tuke.sk/jan.paralic/knihy/DolovanieZnalostizTextov.pdf

19. RUSINKOVÁ, J. Aplikácie text miningových metod. Žilina: Žilinská univerzita v Žiline, 2015. Dostupné z: https://fhv.uniza.sk/mkd_revue/01_2015/01_2015_rusinkova.pdf

20. TAEHO, J. Text Mining : Concepts, Implementation, and Big Data Challenge. Springer International Publishing AG, 2018.

21. WACHSMUTH, H. Text Analysis Pipelines: Towards Ad-hoc Large-Scale Text Mining. Springer. ISBN 9783319257419.

Příloha 1

Slovo	Počet výskytů	Počet dokumentů	Slovo	Počet výskytů	Počet dokumentů	Slovo	Počet výskytů	Počet dokumentů
ability	528	164	experience	720	169	population	122	58
absolute	98	42	experts	85	42	position	195	76
abstract	312	290	explain	290	127	positive	651	170
academic	716	134	express	202	104	possibility	284	129
accepted	83	41	extent	112	63	potential	215	112
access	139	71	external	150	62	practice	906	198
account	122	82	factor	827	172	prefer	222	74
acquire	191	89	faculty	826	140	prepare	357	130
active	1369	219	failure	60	40	presentation	169	56
actual	94	60	family	161	41	previous	360	153
adapt	61	40	features	73	42	primary	472	123
addition	336	147	feedback	165	68	principles	131	65
adequate	80	48	feel	90	52	private	175	49
administration	136	60	female	241	52	probability	151	42
advantage	92	64	field	700	160	problem,	62	42
affect	125	69	figure	631	159	problem.	302	125
age	479	113	final	512	146	procedure	276	92
agree	162	75	financial	508	67	process	1551	252
achieve	482	159	finding	502	174	professional	700	119
aim	608	219	first,	151	103	program	963	181
allowed	71	52	focus	###	256	project	669	107
also,	69	44	follow	197	116	proper	84	44
alternative	105	61	foreign	246	62	proportion	121	52
analyse	2056	276	form	498	184	proposed	132	73
answer	823	181	formal	101	56	proved	110	64
apply	702	202	formed	241	114	provide	424	180
approach	692	199	formulated	80	51	psychological	89	48
article	255	110	framework	168	86	public	259	90
aspect	270	127	frequency	534	149	published	96	62
assess	864	162	full-time	296	46	pupils	704	59
assigned	117	65	function	216	79	purpose	252	143
assume	199	110	fundamental	67	46	quality	918	191
attempt	103	44	future	534	168	quantitative	125	72
attended	66	41	gain	255	108	question	2506	244
attention	290	116	gender	562	84	range	131	92
attitude	270	99	general,	56	48	rate	455	100
authors	472	154	global	94	48	ratio	85	46
average	756	139	goal	296	127	reach	145	81
aware	225	113	grade	529	78	real	212	106
background	131	57	graduates	252	60	realized	76	47

bachelor	189	68	group.	267	122	reason	438	147
based	1240	268	growth	160	83	receive	176	83
basic	392	146	hand	309	163	recommendations	68	40
basis	268	134	helps	80	50	recorded	107	54
behaviour	341	62	highly	163	83	reduce	56	40
belong	54	40	his/her	93	50	refer	165	105
benefit	252	75	human	336	104	reflect	163	105
bring	76	61	hypotheses	553	120	regard	135	69
build	57	41	chance	66	40	regular	77	50
business	456	99	change	369	148	relation	934	200
calculation	523	153	characteristics	251	119	relative	195	80
called	141	87	children	374	51	relevant	240	130
career	229	48	choice	580	184	remaining	59	41
carried	211	118	idea	190	113	report	236	96
case,	183	105	identify	618	187	represent	439	182
category	488	108	impact	411	148	require	540	185
caused	116	69	implement	523	136	research,	623	210
class	549	128	improve	637	199	resources	220	74
classroom	292	72	include	837	246	respect	162	81
close	78	50	increase	762	198	response	1790	199
coefficient	158	60	independent	272	87	rest	74	47
cognitive	205	63	indicators	100	44	result	711	230
collection	251	129	individual	866	211	revealed	114	68
combination	82	53	influence	616	178	role	344	147
common	305	156	information,	169	71	rules	103	41
communicate	811	153	initial	161	74	sample	470	165
companies	131	44	institution	343	87	satisfaction	156	44
compare	923	235	insufficient	66	41	scale	295	100
competence	634	96	integrated	69	42	science	493	157
competitive	83	42	interaction	272	89	score	576	89
complete	439	178	internal	199	74	search	91	48
comprehensive	69	42	international	307	88	secondary	608	127
concept	577	129	internet	173	51	select	615	186
concerned	48	40	interpreted	122	75	semester	261	65
conclude	632	291	interviews	152	56	sense	81	47
condition	266	123	introduction	518	291	serve	68	51
conducted	336	151	investigated	85	55	service	286	73
confirm	258	129	involvement	225	108	set	404	171
connected	357	146	issue	449	161	share	146	72
considered	352	170	item	390	75	short	63	42
consists	380	194	job	327	62	school	2141	196
content	722	193	kappa	47	42	significance	386	122
continue	182	88	key	309	137	similarly,	61	51

contrary,	58	42	keywords	324	289	simple	109	70
contribute	208	106	knowledge	###	231	single	86	59
control	197	59	labour	150	55	situation	579	161
cooperation	116	59	lack	296	125	size	131	70
core	59	41	language	576	86	skills	990	167
correct	308	92	larger	73	57	social	938	174
correlation	394	76	lead	327	156	software	335	85
correspond	136	90	learn	###	221	solution	1167	162
country	461	88	lesson	532	78	source	241	109
course,	756	141	level	###	270	space	81	51
create	724	203	life	234	110	special	159	81
criteria	182	60	limited	131	79	specific	478	181
critical	205	83	linear	107	40	staff	119	42
crucial	116	71	link	184	93	stage	119	47
cultural	136	42	list	89	57	standard	409	139
current	398	169	literature	101	63	start	252	129
curriculum	165	69	long-term	88	62	stated	448	137
czech	1120	188	male	218	53	statistic	1264	193
data	1786	262	manage	###	188	status	65	41
deal	302	151	market	308	88	step	78	50
decision	516	133	master	241	75	strategy	579	122
decrease	64	40	material	698	282	stress	179	47
deeper	74	52	mathematical	###	106	strong	150	72
define	373	144	matter	91	54	structure	398	143
degree	418	114	meaning	87	51	study	####	286
demand	181	91	measure	377	128	subject	1272	191
demonstrate	118	80	meet	94	57	subsequent	66	45
department	81	42	mentioned	284	142	success	718	171
dependent	346	129	method	###	291	sufficient	121	76
describes	358	165	minimum	84	43	suggests	164	93
design	342	121	ministry	75	43	suitable	162	97
detailed	101	73	missing	68	45	sum	109	54
determine	307	145	model	722	136	support	820	220
develop	1956	246	modern	156	68	survey	666	147
deviation	78	40	monitoring	59	43	system,	466	156
didactic	139	43	motivate	452	124	table	1545	234
difference	1076	197	multiple	191	80	takes	50	41
difficult	462	142	mutual	104	56	target	102	46
dimensions	125	41	national	244	82	task	547	123
direct	112	68	nature	246	113	teacher	5244	233
discuss	692	285	negative	342	108	team	224	61
distance	192	42	network	165	39	technology	853	168
distributed	359	124	objective	323	130	tend	182	82
divided	228	128	observe	290	123	term	560	187

dynamic	58	41	occur	58	47	tertiary	203	48
easy	177	101	offer	257	129	test,	975	171
economic	764	168	online	558	70	that,	91	65
education	4816	272	opinion	230	100	theory	603	154
effective	428	153	opportunity	274	128	thinking	182	77
efficiency	457	116	option	65	43	time	1015	247
effort	92	64	organization	208	56	tool	670	162
electronic	140	52	oriented	127	56	topic	407	137
elementary	212	91	original	99	57	total	741	180
emphasis	87	47	output	229	99	traditional	174	81
empirical	93	48	overview	64	39	training	644	111
employees	385	87	paid	107	59	trend	100	55
enable	136	89	paper	916	233	type	780	208
english	295	50	partial	94	53	understand	674	192
enrolled	122	41	participate	695	162	units	85	40
ensure	87	63	pass	173	46	university	2135	234
entire	79	39	pay	120	50	users	245	119
environment	516	155	pedagogical	266	69	valid	133	78
equal	97	59	people	297	128	values	347	116
essential	130	85	perceive	412	119	variable	751	160
establish	171	80	percentage	159	79	verify	179	91
european	220	73	perception	182	61	version	91	54
evaluate	310	122	perform	681	167	view	170	98
evaluation	1316	203	period	220	88	way,	116	78
everyday	73	41	person	470	140	weak	67	45
evident	161	95	perspective	100	54	well.	86	62
exam	940	142	phase	87	41	wide	59	52
example,	297	139	phenomenon	61	46	women	150	45
exams	220	42	plan	212	86	work,	168	93
existing	189	96	play	173	96	written	269	87
expect	469	157	policy	84	41	year.	296	139

Příloha 2

Město	Počet článků	Město	Počet článků	Město	Počet článků	Město	Počet článků
praha	102	helsinki	2	paterson	1	zhenjiang	1
ústí nad labem	13	šumperk	2	edinburgh	1	linz	1
ostrava	11	tábor	2	komatsu	1	santos	1
české budějovice	11	bucharest	2	xinxiang	1	khān	1
milan	10	jihlava	2	libeň	1	paris	1
bologna	10	roma	2	brussels	1	mary	1
bratislava	9	irving	2	yūnis	1	grande	1
brno	8	new york	2	dudley	1	shanghai	1
constantine	7	chicago	2	durrës	1	murcia	1
jackson	6	austin	2	kawaguchi	1	utrecht	1
zhengzhou	6	rodriguez	2	london	1	xinzhou	1
pardubice	6	virginia	2	kara	1	logan	1
timişoara	5	montgomery	2	třinec	1	venice	1
hradec králové	5	omsk	2	seoul	1	alor	1
yekaterinburg	5	guadalupe	2	melaka	1	bangkok	1
pereira	4	poole	2	new	1	lausanne	1
fontana	4	samsun	2	tashkent	1	krupka	1
moscow	3	pavlodar	2	andijon	1	montréal	1
cangzhou	3	braga	2	budapest	1	campos	1
mosul	3	lille	1	sumy	1	coast	1
olomouc	3	brighton	1	kharkiv	1	opava	1
karlovy vary	3	donu	1	sousse	1	salerno	1
oxford	3	mino	1	boleslav	1	astana	1
liberec	3	ciudad de mexico	1	kyzyl	1	semey	1
saint gilbert	3	nice	1	henderson	1	portland	1
košice	3	žatec	1	klatovy	1	děčín	1
lárisa	3	hull	1	cheb	1	aydın	1
león	3	tachikawa	1	gent	1	hong kong	1
zlín	3	hong	1	wellington	1	novosibirsk	1
győr	3	tehrān	1	gresham	1	stuttgart	1
pulandian	3	taraz	1	vaughan	1	torrance	1
livingstone	2	almaty	1	jičín	1	braunschweig	1
monterrey	2	bāneh	1	litoměřice	1	bolton	1
qostanay	2	leiden	1	madrid	1	portsmouth	1
kokshetau	2	kingston	1	serra	1	escobedo	1
engel	2	karviná	1	lisbon	1	teplice	1
valley	2	sakai	1	mladá	1	boston	1
rio de janeiro	2	erlangen	1	kiev	1	juárez	1
salé	2	athens	1	bilimora	1	rostov	1
stockholm	2	berlin	1	samarqand	1	lyon	1
carolina	2	setar	1	kashi	1	dali	1
thessaloníki	2	tulsa	1	timon	1		

sofia	2	constanţa	1	okazaki	1		
plzeň	2	hassan	1	warren	1		