VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

SHLUKOVÁNÍ SLOV PODLE VÝZNAMU

WORD SENSE CLUSTERING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — MAREK JANKECH
AUTHOR

VEDOUCÍ PRÁCE — Doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2019

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání bakalářské práce

21483

Student: **Jankech Marek**
Program: Informační technologie
Název: **Shlukování slov podle významu**
Word Sense Clustering
Kategorie: Web
Zadání:

1. Prostudujte metody měření sémantické podobnosti slov.
2. Seznamte se s existujícími jazykovými nástroji, které mohou být použity ke zkvalitnění odhadu sémantického zařazení slova.
3. Shromážděte data potřebná pro průběžné testování jednotlivých fází řešení problému.
4. Na základě získaných poznatků navrhněte a realizujte systém, který dokáže k zadanému slovu automaticky najít a zobrazit slova sémanticky příbuzná.
5. Vyhodnoťte realizované řešení a porovnejte zvolený přístup s jinými metodami.
6. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 15. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Sémantickú podobnosť slov je možné kódovať pomocou vektorovej reprezentácie – vnorenia slov. Známymi predstaviteľmi typov modelov vytvárajúcich tieto vnorenia slov sú Word2Vec, FastText a Glove. V tejto práci je predstavený novší typ modelov s názvom Dict2Vec. Jedná sa o rozšírenie Word2Vec, ktoré využíva lexikálne slovníky. Práca opisuje prípravu dát z rôznych zdrojov korpusov a slovníkov a porovnáva presnosti jednotlivých typov modelov. Taktiež oboznamuje s implementovanou webovou aplikáciou využívajúcou vnorenia slov.

Abstract

Semantic similarity of words can be encoded using vector representation – word embedding. Known representatives of model types that produce these embeddings are Word2Vec, FastText, and Glove. In this thesis, a newer type of model named Dict2Vec is introduced. It is a Word2Vec extension that leverages lexical dictionaries. The thesis describes the preparation of data from various corpus and dictionary sources and compares accuracy of each model type. It also introduces a web application that uses word embedding.

Kľúčové slová

korpus, slovník, definície, lematizácia, Dict2Vec, Word2Vec, FastText, Glove, spracovanie prirodzeného jazyka

Keywords

corpus, dictionary, definitions, lemmatization, Dict2Vec, Word2Vec, FastText, Glove, natural language processing

Citácia

JANKECH, Marek. *Shlukování slov podle významu.* Brno, 2019. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Doc. RNDr. Pavel Smrž, Ph.D.

Shlukování slov podle významu

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána doc. RNDr. Pavla Smrža, Ph.D. Ďalšie informácie mi poskytol Ing. Martin Fajčík. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

. . . . . .
Marek Jankech
16. mája 2019

Poďakovanie

Rád by som poďakoval pánovi doc. RNDr. Smržovi, Ph.D. za vedenie práce a odbornú pomoc, podobne aj pánovi Ing. Fajčíkovi za rady a usmernenia.

Obsah

Kapitola 1

Úvod

Táto práca nadväzuje na existujúcu prácu od pána Hoštáka[2], ktorej cieľom bolo vytvoriť systém, ktorý by vedel k zadanému slovu nájsť slová sémanticky príbuzné. S úspechom k tomu využil distribučné sémantické modely Word2Vec, FastText a Glove a vytvoril kompletný systém na predspracovanie dát, trénovanie a vyhodnocovanie modelov.

Práca má za cieľ existujúci systém rozšíriť o ďalší typ modelov s názvom Dict2Vec. Dict2Vec rozširuje model Word2Vec Skip-gram s negatívnym vzorkovaním o použitie ďalších dát pri trénovaní, konkrétne definícií z lexikálnych slovníkov. Okrem toho, cieľom je natrénovať existujúce modely spolu s novým typom modelu na dátach z ďalších dostupných korpusov, vyhodnotiť ich úspešnosť a porovnať. V neposlednom rade je práca cielená aj na implementáciu webovej služby využívajúcej existujúce sémantické modely na určenie sémantickej príbuznosti zadaných slov.

Kapitola 2 predstavuje princíp nového typu modelu Dict2Vec. V kapitole 3 sú predstavené jednotlivé jazyky, knižnice, nástroje a všetky prostriedky použité v implementačnej časti práce. Kapitola 4 popisuje architektúru a spôsob implementácie nového systému na prípravu dát zo slovníkov a korpusov, trénovanie a vyhodnocovanie modelov. Taktiež popisuje návrh a implementáciu webovej aplikácie využívajúcej natrénované modely. V kapitole 5 je uvedená konfigurácia pri trénovaní a vyhodnocovaní jednotlivých typov modelov a zhodnotené sú ich úspešnosti na testovacej sade krycie mená. Záver 6 rekapituluje, čo bolo dosiahnuté v tejto práci a aké ponúka východiská do budúcna.

Kapitola 2

Dict2Vec

V minulej práci pána Hoštáka[2] boli s úspechom nasadené na český jazyk distribučné sémantické modely Word2Vec, FastText a Glove. V tejto kapitole je popísaný princíp fungovania nového typu modelu Dict2Vec[9], ktorý rozširuje model Word2Vec Skip-gram negative sampling.

Princíp je založený na predpoklade, že definície zo slovníkov obsahujú informácie, ktoré môžu vylepšiť reprezentáciu jazyka. Môžeme rozlíšiť slovné páry, kde sa každé slovo nachádza v definícií toho druhého a páry, kde sa iba jedno slovo nachádza v definícii toho druhého. Tie prvé môžeme nazvať silnými pármi a tie druhé slabými. Niektoré slabé páry môžu byť povýšené na silné, ak sa nazvájom nachádzajú v K najbližších slovách definície toho druhého.

Na základe silných a slabých párov sa vykonáva pozitívne vzorkovanie, kedy vektory slov slabých alebo silných párov sú posunuté bližšie k sebe. Negatívne vzorkovanie sa zase vykonáva vybratím náhodných slov zo slovníka a posunutím ich vektorov ďalej od seba. Existuje však pravdepodobnosť, že náhodne vybrané slová spolu sémanticky súvisia. Avšak, ak náhodne vybraté slová tvoria silný alebo slabý pár, negatívne vzorkovanie sa nevykonáva. Silné a slabé páry tak zabezpečujú akýsi dozor.

Kapitola 3

Použité technológie a prostriedky

Mieru rýchlosti, jednoduchosti a celkovo možnosti vývoja systému významne ovplyvňuje hardvérové a softvérové vybavenie. V tejto kapitole je uvedené, v akom prostredí sa vyvíjalo, aké programovacie jazyky, knižnice a nástroje boli použité pri implementácii systému a za akým účelom.

3.1 Servery

Spustené výpočty pomocou vlastných skriptov i prevzatých programov bežali na sprístupnených školských serveroch athena, knot a minerva. Z dôvodu vysokej náročnosti na pamäť i výkon procesora niektorých procesov bolo mnohokrát nutné voliť najvybavenejšie stroje. Napr. trénovanie modelu Glove na mohutnom korpuse seznam2017 dosiahlo v špičke spotrebu operačnej pamäte až 168 GB.

Na serveroch bežal operačný systém Ubuntu. Vďaka tomu, že sa jedná o distribúciu systému Linux[1], s veľkou výhodou boli využité poskytnuté nástroje a utility. Najčastejšie využívanými boli `less` na zobrazenie obsahu dát najmä z robustných korpusov, `head` a `tail` na získanie častí dát najmä na testovacie účely. Na filtrovanie obsahu sa zúžitkovali nástroje ako `grep`, `sed`, `awk`, `cut` a iné. Osoh z nástroja `screen` sa ukázal pri púšťaní dlhých, i niekoľkodňových výpočtov. Dĺžka výpočtov a spotreba operačnej pamäte sa merala pomocou príkazu `\time`.

3.2 Python

Najčastejšie využívaným programovacím jazykom bol Python vo verzii 3.5. Jednalo sa o prirodzenú voľbu, pretože mnoho knižníc, či nástrojov bolo vyvinutých s podporou práve pre tento jazyk. V tejto práci boli s miernymi úpravami využité aj niektoré existujúce skripty napísané pánom Hoštákom práve v tomto jazyku, najmä na trénovanie a vyhodnocovanie modelov Word2Vec a Glove[2]. Python je interpretovaný skriptovací jazyk[3], má pomerne zrozumiteľnú syntax a dynamickú správu pamäte, vďaka čomu je programátor uchránený od potenciálnych chýb pri prístupe k nepridelenej pamäti a umožňuje to tak rýchlejší vývoj. Na druhej strane, nevýhodou je vyššia pamäťová náročnosť a pomalší beh programov. Viacero modulov pre Python však bolo napísaných v jazyku C a je možné ich importovať

[1] www.ubuntu.com

vďaka kompilovanému jazyku Cython[2], vytvárajúcim rozhranie medzi zdrojovým kódom v Pythone a kompilovanými knižnicami v C.

3.3 virtualenv

Behom práce sa vyskytla potreba aktualizovať alebo doinštalovať určité knižnice pre Python. virtualenv[3] je nástroj, ktorý umožňuje vytvoriť izolované prostredie s vlastnou štruktúrou inštalačných adresárov s možnosťou inštalácie ľubovoľných Python knižníc aj bez práv administrátora operačného systému.

3.4 MorphoDiTa

Jednou z knižníc použitých v Python skriptoch je MorphoDiTa[4][6]. Osvedčila sa pri lematizácii slov českého jazyka už v práci pána Hoštáka a je znovu zužitkovaná aj v tejto práci. Využíva však novšie jazykové modely verzie 16115[7]. Ukázalo sa, že v tejto práci nebola potrebná kompletná morfologická analýza so všetkými gramatickými kategóriami. Pri filtrovaní lematizovaných slov vystačilo určenie slovného druhu a jeho bližšia špecifikácia. Za týmto účelom bol využitý k tomu určený jazykový model pre značkovanie lém s autormi uvádzanou rýchlosťou až 250 tisíc slov za sekundu, čo je asi šestnásťkrát väčšia rýchlosť ako pri použití hlavných modelov s kompletnou analýzou. Model pri analýze priraďuje POS značky (Part of Speech) českého morfologického systému navrhnutého pánom Janom Hajičom[1] a využité sú len prvé dve POS značky z celkových pätnástich.

3.5 UDPipe

Ako alternatíva ku knižnici MorphoDiTa sa ponúkal nástroj UDPipe[5]. Výsledky lematizácie však boli neuspokojivé. UDPipe síce interne využíva MorphoDiTu, avšak jeho modely neobsahujú morfologický slovník, a ak trénovacie dáta neobsahovali lemmu pre dané slovo, využije sa tzv. „hádač“(„guesser“)[5] na odhadnutie lemmy slova. Vzhľadom k častému priraďovaniu nesprávnej lemmy k slovám z korpusov sa v tejto práci od tohoto nástroja upustilo.

3.6 Gensim

Veľmi dôležitou a často využívanou knižnicou pri téme spracovania prirodzeného jazyka je Gensim[6]. V porovnaní s prácou pána Hoštáka[2], v tejto práci sa Gensim využíva okrem trénovania modelov Word2Vec aj na trénovanie modelov FastText, nakoľko od verzie 3.2.0[7] v sebe natívne podporuje aj implementáciu FastText. Je používaná aj pri vyhodnocovaní presnosti jednotlivých modelov pomocou upravených skriptov od pána Hoštáka.

[2] https://github.com/cython/cython
[3] https://virtualenv.pypa.io
[4] http://ufal.mff.cuni.cz/morphodita
[5] http://ufal.mff.cuni.cz/udpipe
[6] https://radimrehurek.com/gensim/
[7] https://github.com/RaRe-Technologies/gensim/releases/tag/3.2.0

3.7 Glove-python

Na vytvorenie matice spoluvýskytu a následné trénovanie modelov typu Glove bola znovu využitá knižnica Glove-python[8], tak ako aj u pána Hoštáka. Do niektorých súborov implementovaných v jazyku v Cython však bolo nutné zasiahnuť, pre možnosť trénovania aj na rozsiahlych korpusových dátach.

3.8 Beautiful Soup

Beautiful Soup[9] je knižnica na analýzu a extrakciu dát z HTML a XML dokumentov. Pre svoju zhovievavosť aj k nevalidným dokumentom bola použitá v tejto práci pri extrakcií definícií z lokálnych slovníkov v XML formáte. Vďaka využitiu lxml[10] syntaktického analyzátora napísaného v jazyku C je získavanie dát urýchlené.

3.9 Flask, Jinja2, jQuery

Flask[11] je webový framework. Využíva Jinja2[12] šablonovací systém. V tejto práci bol použitý na vytvorenie jednoduchej webovej aplikácie na hádanie príbuzných slov. V JavaScriptovej knižnici jQuery[13] bola implementovaná komunikáciu vo formáte JSON pomocou techniky AJAX.

3.10 Vertikalizátor

Korpusy feeds a bubing predstavujú súbory so zlúčeným obsahom z databázy viacerých webových stránok vo formáte Web Archive. Okrem iného obsahujú HTML či XML zdrojové kódy webových stránok. Pre získanie užitočného obsahu z týchto súborov bol v tejto práci zosožený nástroj vertikalizátor projektu corpproc[14] výskumnej skupiny znalostných technológií na FIT VUT. Je to nástroj, ktorý prevádza Web Archive súbory do vertikálneho formátu.

3.11 Dict2Vec

Dict2Vec[15] je framwerok na učenie vnorenia slov s použitím lexikálnych slovníkov. Jedná sa o rozšírenie a optimalizáciu originálneho Word2Vec frameworku. Zdrojový kód autorov príspevku[9] implementuje sťahovanie a extrakciu informácií z anglických slovníkov, generovanie silných a slabých párov, trénovanie modelov Dict2Vec a ich vyhodnocovanie na poskytnutých dátových sadách. V tejto práci bol s aplikovanými modifikáciami využitý len program na trénovanie modelov a generovanie silných a slabých párov z poskytnutých definícií. Získavanie definícií z českých lexikálnych slovníkov bolo znovu implementované mnou, nakoľko autori frameworku sa zamerali na konkrétne anglické slovníky.

[8] https://github.com/maciejkula/glove-python
[9] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[10] https://lxml.de/
[11] http://flask.pocoo.org/
[12] http://jinja.pocoo.org/
[13] https://jquery.com/
[14] https://knot.fit.vutbr.cz/wiki/index.php/Corpproc# 2._Vertikalizace
[15] https://github.com/tca19/dict2vec

Kapitola 4

Návrh a implementácia

Celú prácu je možné rozdeliť na 3 hlavné časti: príprava dát, trénovanie a vyhodnocovanie presnosti modelov. V tejto kapitole je podrobne opísané, čo sa deje v každej časti, je tu znázornená architektúra celého systému, použité korpusy a slovníky na trénovanie a spôsoby vyhodnocovania modelov.

4.1 Ciele

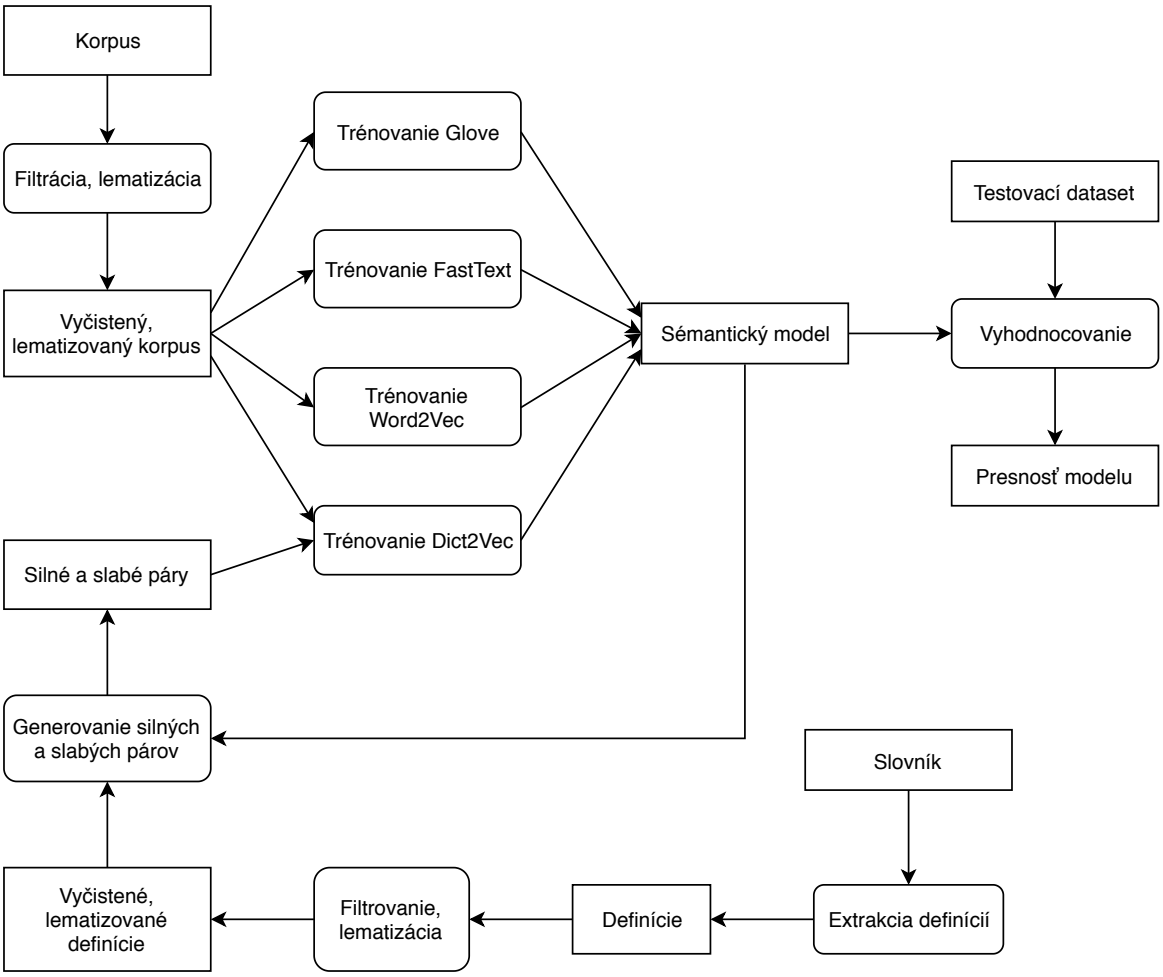
Ako pokračovanie práce pána Hoštáka je v tejto práci snaha o vytvorenie modelov trénovaných na ďalších dostupných korpusoch. Vychádzajúc z výsledkov práce pána Hoštáka, kde sa ukázal výrazný nárast presnosti modelov pri trénovaní na lematizovaných korpusoch[2], aj na tomto mieste v rámci prípravy dát sa korpusy lematizujú. Známe typy modelov ako Word2Vec, FastText a Glove použité pánom Hoštákom sú v tejto práci rozšírené o ďalší typ modelu – Dict2Vec, trénovaný na korpuse s využitím definícií z lexikálnych slovníkov. Ďalšie ciele zostávajú nezmenené – zostavenie systému schopného určovať mieru sémantickej podobnosti slov s čo najvyššou presnosťou. Okrem toho, pribudla aj požiadavka na webovú službu využívajúcu existujúce sémantické modely na určenie sémantickej vzdialenosti medzi zvolenými kandidátnymi slovami a nápovedou.

4.2 Architektúra

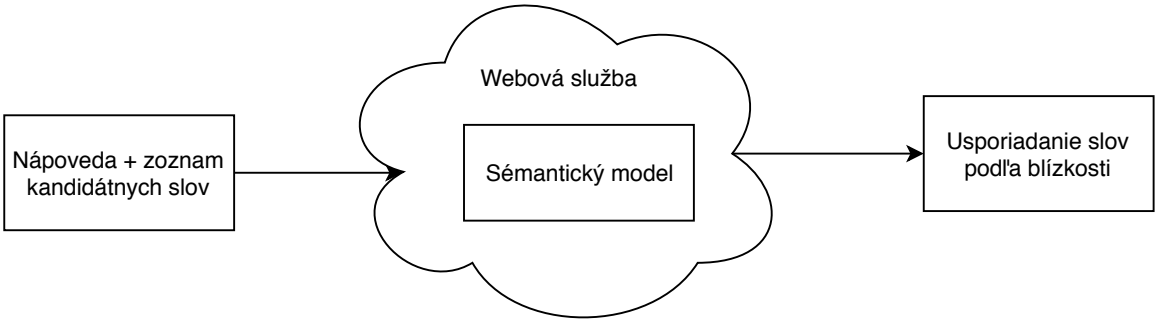
V prvej fáze bolo potrebné nachystať dáta na trénovanie modelov. Pre modely Word2Vec, FastText a Glove to predstavuje prevedenie slov z textu korpusov do vhodného formátu. Pri modeloch typu Dict2Vec však navyše bolo potrebné aj nachystať silné a slabé páry zo slovníkových definícií. To predstavuje celý rad krokov od extrakcie definícií zo slovníkov cez filtrovanie a lematizáciu až po generovanie samotných párov, pri ktorom je nutné využiť existujúceho modelu vnorenia slov.

V druhej fáze sa na pripravených dátach spustilo trénovanie vybraných typov modelov s rôznymi parametrami. Výstupom je model s vektorovou reprezentáciou sémantickej príbuznosti slov.

Poslednou fázou je vyhodnotenie presnosti vytvoreného modelu na testovacích dátových sadách použitých aj v práci pána Hoštáka. Model s najlepšou presnosťou bol potom využitý v back-end logike webovej služby. Webová služba očakáva na vstupe 1 slovo ako nápovedu a zoznam slov, ktoré na výstupe usporiada zostupne podľa sémantickej príbuznosti k danej nápovede.



Obr. 4.1: Architektúra systému spracovania dát, trénovania a vyhodnocovania modelov



Obr. 4.2: Architektúra webovej služby

4.3 Korpusy

Operovalo sa so 4 druhmi textových korpusov v rôznych formátoch. Výrazne najväčší zo všetkých je korpus s názvom seznam2017 tvorený rôznorodým obsahom z portálu seznam.cz. Korpusy feeds a bubing sú tvorené stiahnutým obsahom z rôznych webov a sú mnohonásobne menšej veľkosti, než seznam2017. Najmenší z korpusov, v tejto práci nazvaný ebooks, je tvorený zbierkou e-kníh. Všetky korpusy boli k dispozícii na školských serveroch.

Korpus	Počet slov	Veľkosť
Seznam2017	26 520 791 137	178 GB
Feeds	238 418 750	1,479 GB
Bubing	169 163 132	1,108 GB
E-books	108 064 999	713,7 MB

Tabuľka 4.1: Veľkosti korpusov a počty slov

4.3.1 Seznam2017

Jedná sa o najobjemnejší korpus. Po lematizácii a filtrovaní stop slov mal veľkosť 178 GB s počtom slov 26 520 791 137. Korpus sa nachádzal v dvoch verziách, pričom v druhej verzii, ktorá bola dvojnásobnej veľkosti, sa nachádzali za každým slovom aj ich lemmy oddelené znakom zvislice. Nakoľko v niektorých prípadoch dochádzalo k priradeniu nepresnej lemmy, rozhodol som sa pre použitie prvej verzie korpusu s následnou lematizáciou pomocou nástroja MorphoDiTa. O to, spolu s filtrovaním pomocných značiek a interpunkcie, sa stará skript *seznamPreprocessor.py*. Korpus obsahuje značky ako <d>, <t>, <p> a <s>. Pre nás je podstatná značka <s>, ktorá oddeľuje jednotlivé vety. Po spracovaní je korpus v požadovanom formáte s jednou vetou na každom riadku so slovami v lematizovanom tvare. Použitie skriptu:

- `-i, -input <file>`

  Cesta k vstupnému korpusu.

- `-o, -output_dir <dir>`

  Výstupný adresár na uloženie výstupného korpusu.

- `-t, -tagger <file>`

  Cesta k modelu morfologického taggeru kompatibilného s knižnicou MorphoDiTa.

- `-sw, -stop_words <file>`

  Cesta k súboru zo stop-slovami.

- `-l, -lemma {raw, full, none}`

  Určuje typ výstupu. `raw` – len lemmy (predvolene), `full` – lemmy doplnené o značky, `none` – pôvodné slovo bez lematizácie

4.3.2 Bubing a Feeds

Korpus feeds mal po lematizácii veľkosť 1,48 GB s počtom slov 238 418 750 a korpus bubing 1,1 GB s 169 163 132 slovami. Oba boli tvorené súbormi typu Web Archive. V prípade korpusu feeds sa jednalo o celý zoznam súborov sťahovaného obsahu z internetu, každý súbor pomenovaný podľa dátumu sťahovania. Pri korpuse bubing sa jednalo o 1 súbor s celým obsahom. Vďaka nástroju vertikalizátor a jeho úprávám od pánov Švaňa[8] a Matějku[4] sa ich úspešne podarilo previesť do vertikálneho formátu, čo odfiltrovalo nerelevantný obsah. Vertikál obsahuje na každom riadku buď značku alebo jedno slovo. Pre jednoduchšiu manipuláciu s množstvom súborov som vytvoril shell skript *preprocessWarc.sh*, ktorému sa zadá adresár a spustí vertikalizátor pre každý súbor v adresári. Po tomto kroku nasledovalo filtrovanie pomocných značiek vertikálu a prevod do horizontálneho formátu. Na tento účel bol využitý vlastný skript *vert2Horiz.py*. Znovu sa využila značka <s> na detekciu hraníc jednotlivých viet a výstup mal tvar jedna veta na každý riadok. Aj tu sa využil iný pomocný shell skript *preprocessVerth.sh* na hromadný prevod súborov v určenom adresári, ktorý obaľuje skript *vert2Horiz.py*. Posledným krokom bola lematizácia. Tá s vykonala pomocou skriptu *lemmHoriz.py*, ktorý má rovnaké parametre, ako *seznamPreprocessor.py*. Použitie skriptu *vert2Horiz*:

- `-i, -input <file>`

  Cesta k vstupnému korpusu vo vertikálnom formáte.

- `-o, -output_dir <dir>`

  Výstupný adresár na uloženie korpusu v horizontálnom formáte.

4.3.3 E-books

Posledný použitý korpus v tejto práci je tvorený obsahom e-kníh. S veľkosťou 713 MB a 108 064 999 slovami sa jedná o najmenší korpus. Je tu však predpoklad, že kvalitou by mal predčiť ostatné korpusy. Bol dodaný so slovami v už lematizovanom tvare, bez akýchkoľvek dodatočných značiek, nebolo teda už potrebné vykonávať ďalšie predspracovanie a bol tak pripravený priamo na použitie pri trénovaní modelov.

4.4 Slovníky

Modely typu Dict2Vec využívajú pri trénovaní okrem korpusov aj definície z lexikálnych slovníkov. Bolo teda potrebné zvoliť vhodné české slovníky. S týmto úmyslom boli vybrané slovníky s názvami CZCZ, CETY (Český etymologický slovník), NEO (Slovník neologizmov), PSJČ (Přiruční slovník jazyka českého), SCS (Slovník cizích slov), SSČ (Slovník spisovné češtiny) a SSJČ (Slovník spisovného jazyka českého). Výber bol uskutočnený na základe prieskumu obsahu každého slovníka s kritériom, aby bolo zo záznamov možné strojovo odlíšiť priamo definíciu daného slova od príkladov užitia, spôsobu skloňovania a iných pomocných informácií. Cieľom bolo totiž vygenerovať z definícií silné a slabé páry.

V definíciách sa vyskytovali pomerne často aj slová ako napríklad „mající“, „obsahující“, „týkající“, „vlastnost“, „vyjadřuje“a podobne. Takéto slová podľa môjho názoru nepredstavujú pridanú hodnotu pri určovaní sémantickej príbuznosti a tak boli z obsahu definícií

odfiltrované. Zoznam stop-slov bol zhotovený na základe analýzy najčastejšie vyskytovaných slov v slovníkoch a následného osobného posúdenia. Naopak, niektoré časté slová ako „človek“, či „žena“do stop-listu zaradené neboli, pretože v určitých kontextoch majú svoje sémantické opodstatnenie, ako napríklad pri názve funkcie či povolania. Ďalej boli vynechávané všetky skratky, nakoľko pri nasledovnej lematizácii jazykový model vo väčšine prípadov nedokázal určiť celý tvar slova. Problém s rozlíšením skratiek však nastával pri slovníku PSJČ, kde sa v definiciách vyskytovali vety ukončené bodkou. Riešilo sa to analýzou najčastejších slov na konci vety, a na základe toho bol zostavený ďalší stop-list. Ak sa slovo na konci vety nachádzalo v stop-liste, bolo považované za skratku a vynechalo sa, inak sa slovo ponechalo a bodka sa odstrihla. Z kľúčových slov sa odrezávali zvratné zámená („se“, „si“) a rôzne znaky, ktoré nepatria samotnému slovu (zátvorky, bodky, dvojbodky, čiarky,...). V slovníku CETY sa niekedy vyskytovali kľúčové slová ako napr. „fix(a)“, čo predstavovalo dva tvary slova: „fix“a „fixa“. V tomto prípade sa uložil rovnaký záznam s definíciou pre kľúčové slovo fix aj fixa. Podobne sa postupovalo aj v iných slovníkoch v prípade viacerých kľúčových slov oddelených čiarkou spárovaných s jednou definíciou. Aj tu sa uložil ten istý záznam zvlášť pre každé kľúčové slovo.

Zdrojové súbory slovníkov boli rôznych formátov, v najjednoduchšom prípade sa jednalo o formát – jeden riadok slovo, ďalší riadok definícia atď. – inokedy slovo od definície oddelené pomocou bieleho znaku. V takomto prípade na extrakciu definícií postačovali regulárne výrazy. Pri niektorých slovníkoch však bolo nutné vytiahnuť definície zo súborov obsahujúcich viaceré dokumenty typu XML. V tomto prípade pomohla knižnica Beautiful Soup, najmä ak sa jednalo o nevalidné XML.

Najväčším orieškom bol zdrojový súbor s XML záznamami pre SSČ, kde nebolo možné triviálnym spôsobom rozpoznať, či sa jednalo o kľúčové slovo alebo definíciu, bol preto naimplementovaný pomerne komplikovaný algoritmus. Hoci existovala aj varianta s kľúčovým slovom na každom nepárnom riadku a definíciou na každom párnom, definície obsahovali aj pomocné informácie o skloňovaní, slovnom druhu a pod. Tieto pomocné informácie nebolo možné jednoznačne strojovo odlíšiť od relevantného obsahu definície. Preto sa pracovalo s originálnym XML zdrojom.

Extrakciu definícií zo slovníkov mal na starosti skript *local_def_extractor.py*. Pre každý slovník sa očakáva súbor v konkrétnom formáte. Nasledujúce názvy súborov sú z adresára slovníkov projektu MA - morfologická analýza a sú v očakávanom formáte pre skript: *czcz.def, cety.def*, *neologizmy1.definitions*, *psjc.entries*, *scs.def*, *ssc.xml.lines*, *ssjc_snad_lepsi.entries*. Skript vygeneruje do aktuálneho adresára pre každý slovník výstupný súbor s definíciami na každom riadku v tvare: `kľúčové_slovo TAB názov_slovníka TAB definícia`. Parametre skriptu:

- `-dicts [czcz, cety, neo, psjc, scs, ssc, ssjc]`

  Zoznam názvov slovníkov, z ktorých sa majú extrahovať definície

- `-czcz <file>`

  Cesta k slovníku CZCZ.

- `-cety <file>`

  Cesta k slovníku CETY.

- `-neo <file>`

  Cesta k slovníku NEO.

- ...

  Cesty k ďalším slovníkom zo zoznamu.

4.4.1 Definície

Pre ďalšie spracovanie je potrebné definície z rôznych slovníkov spojiť do jedného súboru. K tomu poslúžil UNIX nástroj `cat`. Nasleduje ponechanie len tých definícií, ktorých kľúčové výrazy sa nachádzajú v slovníku daného korpusu. Pre každý korpus tak bol vygenerovaný zvlášť súbor s definíciami. Keďže definície z českých slovníkov som vytváral vo vlastnom formáte, bolo potrebné aj upraviť pôvodný skript od autorov. Parametre upraveného skriptu *clean_definitions_new.py*:

- `-d, -definitions <FILE>`

  Cesta k súboru s definíciami.

- `-v, -vocab <FILE>`

  Cesta k slovníku unikátnych slov, každé slovo na samostatnom riadku.

- `-o, -output_fn <FILE>`

  Cesta k výstupnému súboru s ponechanými definíciami, bez názvov slovníkov.

Autori generovali silné a slabé páry z definícií iba pre slová, ktoré mali vo vstupnom korpuse viac ako 5 výskytov. Rozhodol som sa postupovať rovnako a pred spúšťaním skriptu *clean_definitions_new.py* som pre všetky korpusy vygeneroval súbory obsahujúce na každom riadku unikátne slovo, ktoré sa aspoň 5krát vyskytlo v korpuse. Vytvoril som na ten účel skript *wordOccur.py*, ktorý je schopný spočítať výskyty jednotlivých slov v korpuse, alebo priamo vytiahnuť tento slovník z existujúcich modelov. Výstup generuje do súboru vo formáte unikátne slovo s počtom výskytov oddelených tabulátorom na každom riadku. Slová sú usporiadané zostupne podľa počtu výskytov. Parametre skriptu pri spúšťaní s modelom na vstupe:

```
python wordOccur.py model
```

- `-i, -input <FILE>`

  Cesta k natrénovanému modelu.

- `-mt, -model_type {w2v, w2v_obj, ft_orig, ft_gensim}`

  Typ modelu. `w2v` – textový formát, `w2v_obj` – w2v objekt knižnice gensim, ft_orig – binárny formát modelov natrénovaných pomocou knižnice fastText, ft_gensim – fastText objekt vytvorený pomocou knižnice gensim.

Parametre skriptu pri spúšťaní s korpusom na vstupe:

```
python wordOccur.py corpus
```

- `-i, -input <FILE>`

  Cesta k vstupnému korpusu určenému na trénovanie.

- `-m, -model_saving`

  Ak je vybratá táto možnosť, nebude sa načítavať celý korpus naraz do pamäte.

Zo súboru som potom vybral riadky so slovami o 5 a viac výskytov pomocou nástroja `head` a nástroj `cut` poslúžil na vybratie samotných slov bez čísel.

Keďže pracujeme s lematizovanými korpusmi, bolo potrebné ešte lematizovať aj definície. Pri tomto kroku sa zároveň uskutočňovala ďalšia fáza filtrovania pomocou POS značiek. Z definícií sa odstránili priraďovacie a podraďovacie spojky, niektoré predložky, čísla, interpunkcia a väčšina zámen. K tomu sa využil podpríkaz `lemma` z vlastného skriptu *utils.py*, kotrý obsahuje aj ďalšie funkcie. Syntax skriptu pre spúšťanie lematizácie:

```
utils.py lemma -i FILE [-o FILE] [-sw STOPWORDS]
[-t TAGGER] [-l raw,full,none]
```

Parametre sú rovnaké ako pri skripte *seznamPreprocessor.py*.

Posledným krokom je generovanie silných a slabých párov. K tomu sa využil upravený skript od autorov frameworku Dict2Vec nazvaný *generate_pairs_new.py*. Úprava spočívala v spôsobe načítania prvého riadku existujúceho natrénovaného modelu v textovom formáte, pretože všetky mnou v minulosti natrénované modely obsahovali na prvom riadku počet slov a dimenzii, s čím pôvodný skript nepočítal. Skript teda očakáva na svojom vstupe okrem súboru s definíciami aj natrénovaný model, aby mohol vygenerovať silné a slabé páry. Preto sa najprv vytvárali modely typu Word2Vec, FastText a Glove a až potom modely Dict2Vec. Parametre skriptu:

- `-d, -definitions <FILE>`

  Cesta k súboru s definíciami.

- `-e, -embedding <FILE>`

  Cesta k natrénovanému modelu.

- `-o, -output_fn <FILE>`

  Cesta k výstupnému súboru s ponechanými definíciami, bez názvov slovníkov.

- `-sf, -strong-file <FILE>`

  Cesta k výstupnému súboru na uloženie silných párov.

- `-wf, -weak-file <FILE>`

  Cesta k výstupnému súboru na uloženie slabých párov.

- `-K <NUMBER>`

  K najbližších slov v okolí slova, ktoré môžu byť povýšené na silné páry.

Pre neskoršie porovnávanie presnosti modelov, bolo žiaduce vygenerovať silné a slabé páry s pomocou všetkých natrénovaných modelov Word2Vec, FastText a Glove. Aby sa tento proces zjednodušil a urýchlil, bol na ten účel vytvorený pomocný shell skript *pairs.sh*. Skript využíva fakt, že adresár projektu tejto práce má určitú pevnú adresárovú štruktúru. Je teda schopný automaticky vygenerovať všetky súbory silných a slabých párov pre každý existujúci model s odpovedajúcim korpusom. Skript postupne spúšťa obalený *generate_pairs_new.py* vždy s inými argumentami. Bez zadania parametrov skript predvolene generuje páry pre všetky 3 typy modelov natrénovaných na všetkých typoch korpusov. Generovanie párov je možné obmedziť pomocou parametrov:

- `-W`

  Zahrnúť modely Word2Vec do procesu generovania párov.

- `-F`

  Zahrnúť modely FastText do procesu generovania párov.

- `-G`

  Zahrnúť modely Glove do procesu generovania párov.

- `-e`

  Zahrnúť modely natrénované na korpusoch e-books do procesu generovania párov.

- `-b`

  Zahrnúť modely natrénované na korpusoch bubing do procesu generovania párov.

- `-f`

  Zahrnúť modely natrénované na korpusoch feeds do procesu generovania párov.

4.5 Trénovanie

Tak ako systém pána Hoštáka[2], aj tento systém podporuje trénovanie modelov typu Word2Vec, FastText a Glove, no doplnený je o model typu Dict2Vec.

4.5.1 Word2Vec a FastText

Rozhranie pôvodného skriptu pána Hoštáka *trainWord2Vec.py* na trénovanie modelu Word2Vec zostalo nezmenené, no, špecifická implementácia korpusu ako objektu bola nahradená štandardnými triedami z knižnice gensim *LineSentence* v prípade korpusu ako jediného súboru a *PathLineSentence* v prípade adresára s viacerými súbormi korpusu.

Čo sa týka modelu FastText, v knižnici gensim pribudla aj jeho implementácia, a preto som sa rozhodol použiť v tejto práci práve tú na trénovanie nových modelov. Inak je užívateľské rozhranie skriptu *trainFasttext.py* kompatibilné s predošlou implementáciou. Spoločné rozhranie skriptov *trainWord2Vec.py* a *trainFasttext.py*:

- `-i, -input <FILE/DIR>`

  Cesta k pripravenému korpusu alebo adresáru s korpusmi.

- `-o, -output_dir <DIR>`

  Cesta k adresáru, kde bude uložený model.

- `-e, -epochs <NUMBER`

  Počet iterácií.

- `-d, -dimensions <NUMBER>`

  Počet dimenzií vektorov.

- `-c, -count <NUMBER>`

  Pod touto hranicou výskytov sú slová v modeli ignorované

- `-m, -model {cbow, sg}`

  Trénovací algoritmus CBOW alebo SG.

- `-a, -algorithm {ns, hs}`

  Aproximácia negative sampling alebo hierarchical softmax.

- `-s, -samples <NUMBER>`

  Počet negatívnych vzoriek.

- `-t, -threads <NUMBER>`

  Počet vlákien.

- `-w, -window <NUMBER>`

  Šírka kontextového okna.

4.5.2 Glove

Aj v tomto prípade zostalo rozhranie skriptu *trainGlove.py* zachované. Nevyhnutný bol však zásah do použitej knižnice glove-python. Pri trénovaní na menších korpusoch ako feeds, bubing, či e-books nenastal žiadny problém. Avšak pri použití korpusu seznam2017 už problém nastal a to konkrétne pretečenie pôvodne použitých celočíselných dátových typov v cython súbore. Toto sa prejavilo pri trénovaní modelu tým, že celý proces bol

násilne ukončený operačným systém kvôli nedovolenému prístupu do pamäti. Nutná bola zmena z typov `int` na `long long` pri počtoch spoluvýskytov slov na niekoľkých miestach zdojového súboru *glove_cython.pyx*. Upravenú knižnicu bolo teda potrebné znovu preložiť príkazom `python setup.py cythonize` a nainštalovať cez nástroj pip v editačnom móde, teda `pip install -e <PATH>`. Parametre skriptu *trainGlove.py*:

- `-i, -input <FILE>`

  Cesta k pripravenému korpusu alebo matici spoluvýskytu.

- `-o, -output_dir <DIR>`

  Cesta k adresáru, kde bude uložený model.

- `-om, -output_mdir <DIR>`

  Cesta k adresáru, kde bude uložená matica spoluvýskytu.

- `-e, -epochs <NUMBER`

  Počet iterácií.

- `-d, -dimensions <NUMBER>`

  Počet dimenzií vektorov.

- `-c, -count <NUMBER>`

  Pod touto hranicou výskytov sú slová v modeli ignorované.

- `-m, -mode {corpus, matrix}`

  Určenie vstupu – korpus alebo matica spoluvýskytu.

4.5.3 Dict2Vec

Trénovanie modelu Dict2Vec je implementované v programe napísanom v jazyku C s názvom *dict2.c* od autorov frameworku. Keďže sa jedná o kompilovaný jazyk, je potrebné program preložiť do spustiteľného formátu pomocou priloženého Makefile súboru. Trénovanie modelov s použitím všetkých korpusov okrem seznam2017 prebiehalo v poriadku. Znova sa však zopakoval problém neoprávneného prístupu do pamäti pri použití korpusu seznam2017, ako pri modeli Glove. Ukázalo sa, že framework nebol prispôsobený na trénovanie na takých rozsiahlych korpusoch. Siahlo sa teda po úprave originálneho zdrojového kódu od iného autora, ktorý previedol niektoré konštrukcie kódu do ich varianty v jazyku C++[1]. Čiastočne to odstránilo problém, no program aj tak spadol na inom mieste. Ak existoval zámer docieliť možnosť trénovania aj na robustnom korpuse seznam2017, či iných rozsiahlych korpusoch do budúcna, vlastný zásah do kódu bol nevyhnutný. Identifikoval som kritické miesta, kde znovu pretiekol rozsah dátových typov `int`, zmenil na `long long` a program fungoval.

[1] https://github.com/w4-anthony/dict2vec/commit/4fc56ea9f590e33c0c00b7a601e1504d2b548a3c

Ďalšia modifikácia na urýchlenie trénovania spočívala v inicializácii slovníka na základe vstupného korpusu. Toto je prvá fáza práce programu. Ak trénujeme viacero modelov Dict2Vec s rôznymi vstupnými parametrami, avšak s rovnakým vstupným korpusom, dátová štruktúra, kde sa inicializuje slovník, bude vždy na konci tejto fázy rovnaká. Je teda zbytočné zakaždým opakovať tento krok prechádzania vstupného korpusu a počítania výskytov jednotlivých slov. Použil som teda súbory s výskytmi jednotlivých slov vygenerovaných pomocou skriptu *wordOccur.py* pri fáze čistenia definícii.

Vyrátal som, že inicializáciou slovníka pomocou špeciálneho súboru s unikátnymi slovami a ich výskytmi v korpuse je možné ušetriť na celom procese trénovania modelov Dict2Vec aj desiatky hodín času v prípade robustných korpusoch ako je seznam2017. Upravený program v jazyku C++ na trénovanie som nazval *dict2vec_with_vocab.cc*.

Na zjednodušenie práce som si znova vytvoril pomocný shell skript s názvom *trainD2V.sh*. Tomuto skriptu stačí v parametroch zadať cestu k vstupnému korpusu a súboru so silnými pármi, ak nám parametre trénovania modelov postačujú ponechané na predvolených hodnotách. Cestu k súboru so slabými pármi si vie sám odvodiť, ak je rovnaká ako cesta k silným párom, podobne aj cestu k výstupnému súboru. Znova sa spolieha na pevne danú adresárovú štruktúru projektu. Všetky parametre je však možné zadať aj ručne. Parametru skriptu:

- `-p <FILE>`

  Cesta k spustiteľnému programu na trénovanie modelov Dict2Vec.

- `-i <FILE>`

  Cesta ku korpusu na trénovanie.

- `-o <DIR>`

  Výstupný adresár na uloženie modelu.

- `-v <FILE>`

  Voliteľne cesta k existujúcemu slovníku s výskytmi slov.

- `-s <FILE>`

  Cesta k súboru so silnými pármi.

- `-w <FILE>`

  Cesta k súboru so slabými pármi.

- `-D <NUMBER>`

  Počet dimenzií.

- `-W <NUMBER>`

  Šírka kontextového okna.

- `-C <NUMBER>`

  Min. počet výskytov slov, na ktorých bude model trénovaný.

- `-S <NUMBER>`

  Počet silných párov na pozitívne vzorkovanie.

- `-W <NUMBER>`

  Počet slabých párov na pozitívne vzorkovanie.

- `-T <NUMBER>`

  Počet vlákien.

- `-E <NUMBER>`

  Počet trénovacích epoch.

- `-S <BOOL>`

  Ak je 1, ukladať model pri každej epoche. Ak 0, neukladať.

- `-1 <FLOAT>`

  Alpha hyperparameter rýchlosti učenia.

- `-2 <FLOAT>`

  Prah podvzorkovania slov s častými výskytmi.

- `-3 <FLOAT>`

  Koeficient pre silné páry.

- `-4 <FLOAT>`

  Koeficient pre slabé páry.

4.6 Hodnotenie presnosti

Pána Hošták[2] využíval vo svojej práci ako metriku na vyhodnocovanie modelov kosínusová vzdialenosť medzi vektormi slov. Využíval tri vyhodnocovacie sady: krycie mená, slovné analógie a slovník synoným. Výslednú presnosť previedol na percentá. Tento spôsob vyhodnocovania bol spolu s vyhodnocovacími skriptami prevzatý aj v tejto práci. Operujeme však s novšou sadou krycích mien. Všetky vyhodnocovacie skripty však boli rozšírené o možnosť výberu typu modelu, ako je v skripte *wordOccur.py* 4.4.1. Predvolene sa však ráta s textovým formátom modelu vnorenia slov.

Priemerná presnosť na krycích menách je však počítaná prísnejšie. Do výpočtu sú zahrnuté aj slová mimo slovníka („out of vocabulary“), pre kompatibilitu s vyhodnovacím spôsobom Ing. Fajčíka.

Na zjednodušenie a paralelizáciu vyhodnocovania viacerých modelov som znovu vytvoril pomocný skript *evalMultiModels.py*. Skript umožňuje naraz vyhodnocovať modely rovnakého formátu na zvolenej testovacej sade. Obecné parametre skriptu:

- `-m, -input <FILE>`

  Cesty k jednotlivým modelom oddelených čiarkami. Možnosť použitia aj shell vzorov („wildcards“), nutno však obaliť do úvodzoviek.

- `-d, -dictionary`

  Cesta k morfologickému slovníku knižnice MorphoDiTa.

- `-mt, -model_type {w2v, w2v_obj, ft_orig, ft_gensim}`

  Typ modelu. `w2v` – textový formát, `w2v_obj` – w2v objekt knižnice gensim, ft_orig – binárny formát modelov natrénovaných pomocou knižnice fastText, ft_gensim – fastText objekt vytvorený pomocou knižnice gensim.

Parametre skriptu pri použití vyhodnocovej sady krycie mená:

```
python evalMultiModels.py cn
```

- `-i, -input <FILE>`

  Cesta k vyhodnocovacej dátovej sade krycie mená.

- `-m, -model`

  Cesta k výstupnému súboru.

- `-t, -topn`

  Počet vyhľadaných najviac podobných slov.

- `-ns, -no-sort`

  Ak je uvedené, nezoraďovať záznamy podľa priemernej presnosti.

- `-v, -verbose`

  Zapnutie rozšíreného módu.

- `-c, -co_matrix`

  Cesta k matici spoluvýskytu.

Parametre skriptu pri použití vyhodnocovej sady slovných analógií:

```
python evalMultiModels.py analogy
```

- `-i, -input <FILE>`

  Cesta k vyhodnocovacej dátovej sade slovných analógií.

- `-t, -topn`

  Počet vyhľadaných najviac podobných slov.

- `-p, -phrases`

  Zahrnúť aj frázy.

Parametre skriptu pri použití vyhodnocovej sady slovníku synoným:

```
python evalMultiModels.py syn
```

- `-i, -input <FILE>`

  Cesta k vyhodnocovacej dátovej sade krycie mená.

- `-o, -output`

  Cesta k výstupnému súboru.

- `-t, -topn`

  Počet vyhľadaných najviac podobných slov.

- `-ns, -no-sort`

  Ak je uvedené, nezoraďovať záznamy podľa priemernej presnosti.

- `-p, -phrases`

  Zahrnúť aj frázy.

4.7 Webová služba

K experimentálnym účelom som vytvoril webovú služba pomocou frameworku Flask. Aplikácia očakáva na vstupe 1 slovo ako nápovedu a zoznam kandidátnych slov. Tieto slová najprv lematizuje a potom k nim vypočíta kosínusovú vzdialenosť od nápovedy a zašle na výstup. K tomuto potrebuje mať načítaný v pamäti natrénovaný sémantický model a morfologický slovník využívaný knižnicou MorphoDiTa. Komunikácia s aplikáciou funguje vo formáte JSON. K tomu sa dá využiť napr. nástroj `curl`. Aplikácia využíva dva existujúce skripty od pána Hoštáka. Skript *lemmatization.py* bol využitý na lematizáciu vstupných slov a skript *loggingConfig.py* na logovacie výpisy. Príklad možnej HTTP komunikácie cez terminál:

Klient zašle HTTP žiadosť:
```
curl -H Content-type: application/json" -X POST 127.0.0.1:8084 -d
'{"hint": "hmyz", "candidates": "moucha osa okno"}'
```
Server odpovedá:
```
"{\"osa\": \"0.9994203\", \"moucha\": \"0.9864203\", \"okno\":
\"0.2538827\"}"
```

Webová služba sa na serveri spúšťa príkazom `python3 webapp.py`
Parametre:

- `-ht, -host <NAME>`

  Názov hostiteľa.

- `-p, -port`

  Číslo portu.

- `-D, -debug`

  Zapnúť debug mód.

- `-m, -model <FILE>`

  Cesta k natrénovanému modelu.

- `-mt, -model_type {w2v, w2v_obj, ft_orig, ft_gensim}`

  Typ modelu. `w2v` – textový formát, `w2v_obj` – w2v objekt knižnice gensim, ft_orig – binárny formát modelov natrénovaných pomocou knižnice fastText, ft_gensim – fastText objekt vytvorený pomocou knižnice gensim.

- `-d, -dictionary`

  Morfologický slovník kompatibilný s knižnicou MorphoDiTa.

- `-s, -sort`

  Zoradiť záznamy podľa sémantickej blízkosti.

Kapitola 5

Výsledky

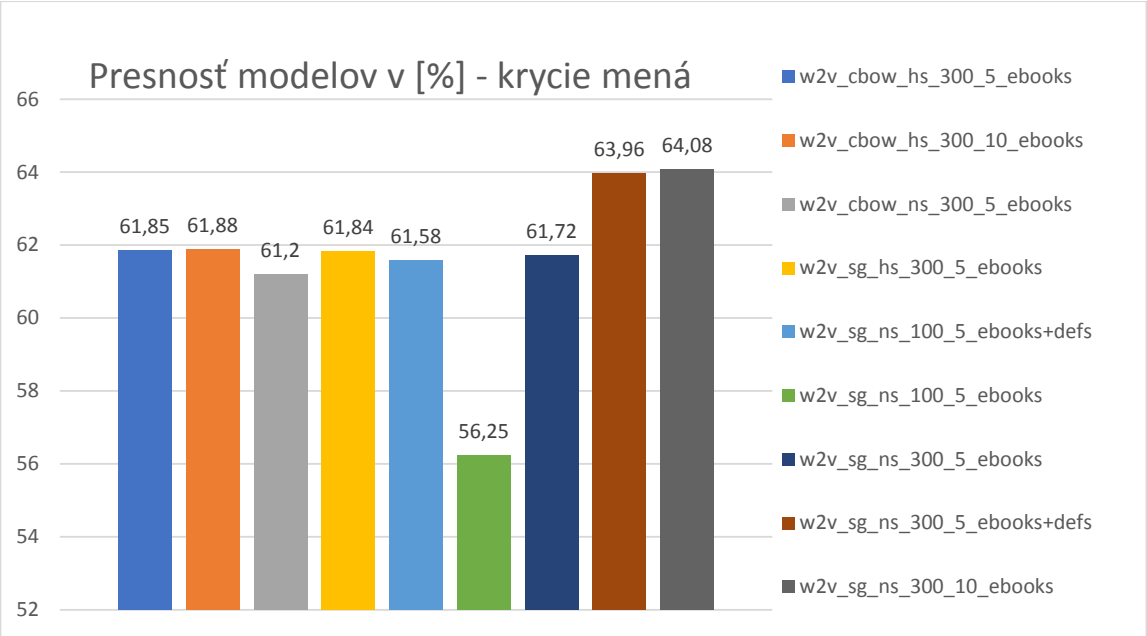
V tejto kapitole sú analyzované presnosti jednotlivých typov modelov. Porovnáva sa úspešnosť vzhľadom na rôzne parametre a vstupné korpusy. Zameriavame sa na výsledky na testovacej sade krycie mená, preto sú tu zobrazené grafy presností jednotlivých modelov na krycích menách. Kompletné výsledky zo všetkých dátových sád sú uvedené v tabuľkách v prílohe A. Spolu bolo vyhodnotených 93 natrénovaných modelov.

Označenie pre modely Word2Vec a FastText je vo formáte `<typ modelu>_<trénovací algoritmus>_<softmax aproximácia>_<počet dimenzií>_<počet epóch>_<korpus>`. Modely Glove sú značené vo formáte `<typ modelu>_<počet epóch>_<korpus>`. Dict2Vec modely majú za dvojitým podčiarkovníkom uvedený názov modelu, ktorý bol použitý pri generovaní silných a slabých párov. Prvá časť ich názvu je v tvare `<typ modelu><počet dimenzií><počet silných párov na pozitívne vzorkovanie><počet slabých párov na pozitívne vzorkovanie><koeficient pre silné páry><koeficient pre slabé páry>`.

5.1 Konfigurácia

Ako optimálna hodnota pre počet dimenzií bolo zvolené číslo 300, na základe výsledkov práce pána Hoštáka. Experimentovalo sa však aj so 100 dimenziami. Takéto modely boli trénované na 300 MB časti zo začiatku každého korpusu. Sledujeme tým hlavne, ako sa pri tom prejaví model Dict2Vec so svojimi silnými a slabými pármi, keďže podľa výsledkov autorov Dict2Vec sa najvýraznejšie prejavuje prevaha ich modelov na malých korpusoch[9]. Keďže modely Dict2Vec využívajú aj tieto dáta z definícií, pre spravodlivé porovnanie aspoň s modelmi Word2Vec sa aj tie natrénovali so špeciálnou variantou korpusu s pripojenými definíciami na konci. Samotné definície mali veľkosť iba 26 MB. Obdobne postupovali aj autori Dict2Vec. Pre všetky modeli bolo konštantne nastavený prah pre povdzorkovanie na hodnotu 0,0001. Pri aproximácii negative sampling bol nastavený počet negatívnych vzoriek na hodnotu 5. FastText mal min. počet ngramov nastavený na hodnotu 3 a max. na 6.

Pre modely Dict2Vec boli zvolené hodnoty parametrov 4 silné a 5 slabých párov na pozitívne vzorkovanie. Pri trénovaní na korpuse seznamu sa však testovali aj varianty s 2 silnými a 3 slabými pármi. Tieto parametre nastavujú vplyv silných a slabých párov pri trénovaní na korpuse. Autori odporúčali tento vplyv zmenšiť už pri korpusoch, ktoré majú desiatky gigabajtov. Koeficienty pre silné, resp. slabé páry boli konštatne nastavené na hodnoty 0.8, resp. 0.45. Modely Dict2Vec boli vždy trénované na korpuse, ktorý bol použitý



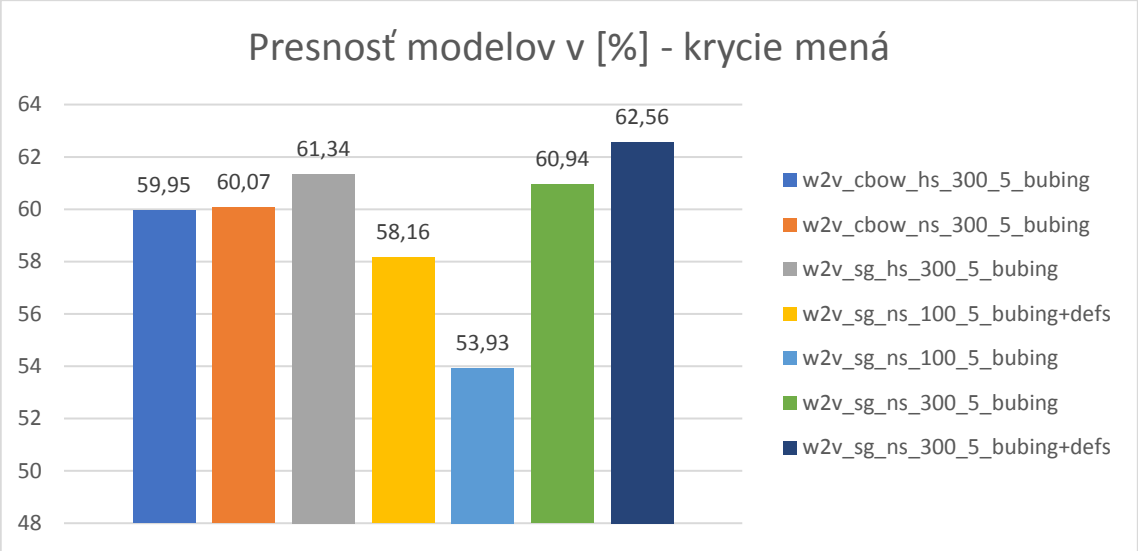
Obr. 5.1: Presnosť modelov Word2Vec trénovaných na korpuse E-books

aj na natrénovanie existujúceho modelu iného typu, a ktorý bol použitý pri generovaní silných a slabých párov pre konkrétny model Dict2Vec.

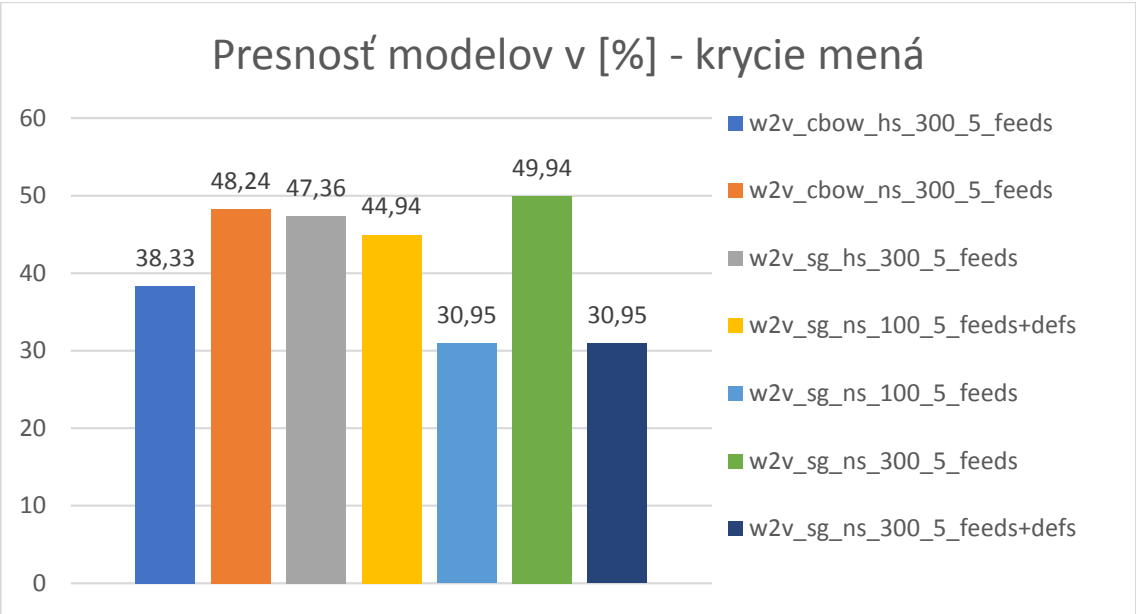
Modely boli vyhodnocované na slovníku synoným iba s prísnejšou variantou top 10 najbližších slov. Pri slovných analógiách bolo pri každej otázke očakávané správne slovo na prvom mieste.

5.2 Hodnotenie Word2Vec

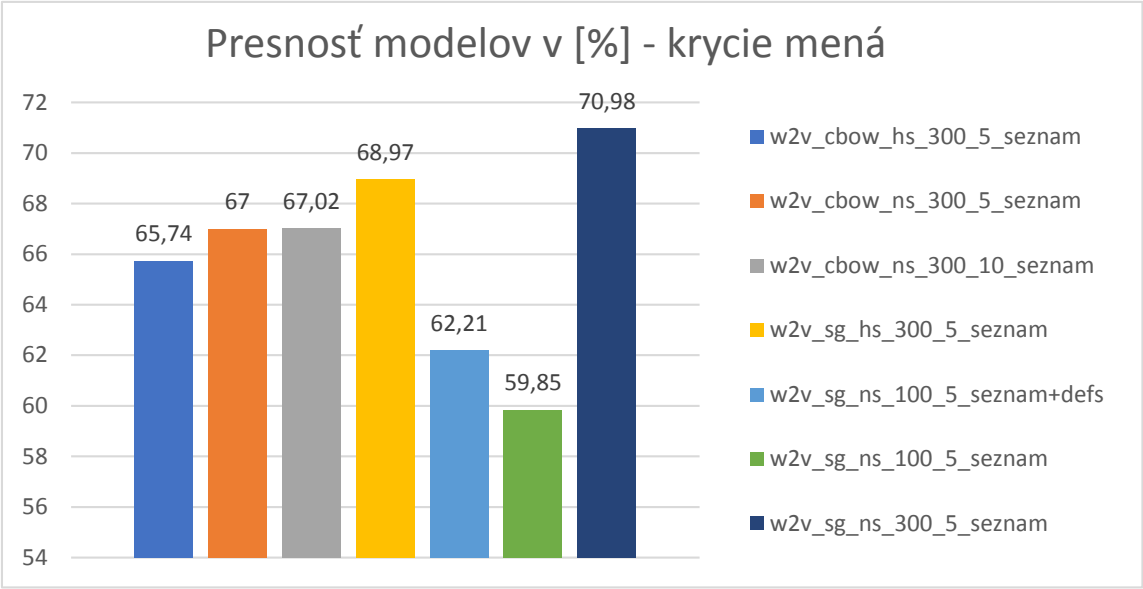
Všade, kde sa trénovalo aj s 10 trénovacími epochami možno vidieť takmer nulový nárast presnosti oproti variantám s 5 epochami. Presnosť modelov bola očakávane najnižšia pri skrátení každého korpusu na 300 MB a trénovaní 100-dimenzionálnych vektorov. Vždy však pomerne výrazne vzrástla po pridaní definícií na koniec korpusu. Najvýraznejší nárast možno badať pri korpuse feeds, kde sú celkovo presnosti modelov pomerne nízke a definície zvýšili presnosť 100-dimenzionálneho modelu až o 14 %. Zaujímavé však je, že pri tom istom korpuse s použitím definícií pri 300-dimenzionálnom modeli presnosť naopak veľmi výrazne poklesla. Aj tak vždy zvíťazil model SG NS. V prípade korpusu seznamu s presnosťou 70,98% sa dokonca jedná o absolútneho víťaza aj zo všetkých typov modelov. Medzi modelmi korpusu seznamu chýba model s 300 dimenziami s pripojenými definíciami na konci modelu. Tento model sa už bohužiaľ nepodarilo natrénovať, pretože na 2 rôznych serveroch jeho trénovanie spadlo s chybou bus error. Nie je však očakávaný výrazný nárast presnosti vzhľadom k robustnosti korpusu seznamu.



Obr. 5.2: Presnosť modelov Word2Vec trénovaných na korpuse Bubing



Obr. 5.3: Presnosť modelov Word2Vec trénovaných na korpuse Feeds



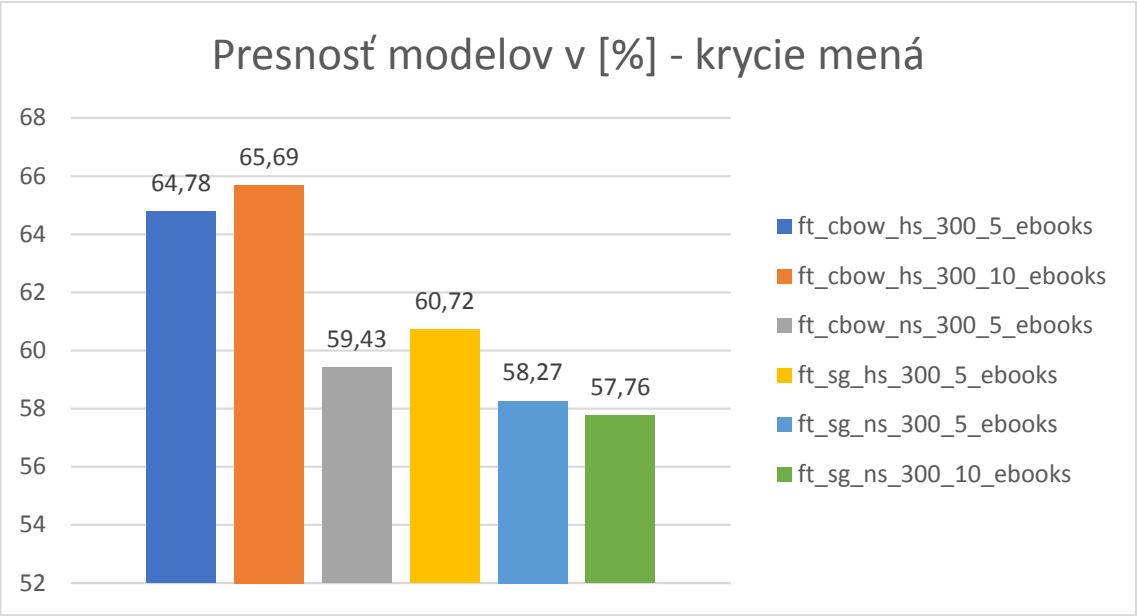
Obr. 5.4: Presnosť modelov Word2Vec trénovaných na korpuse Seznam2017

5.3 Hodnotenie FastText

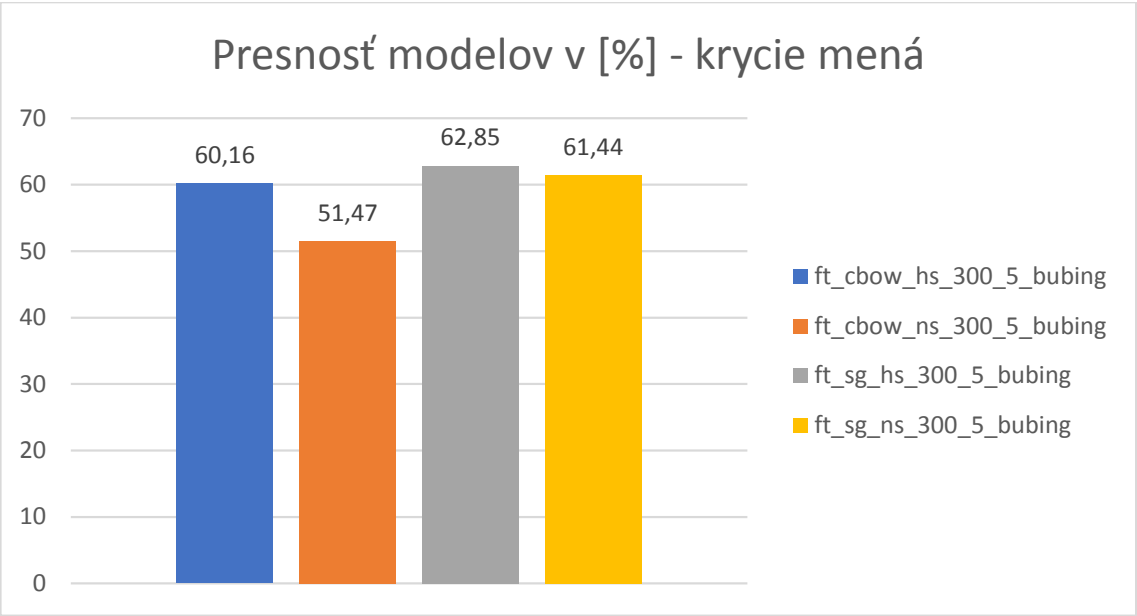
Modely FastText neobsahujú varianty modelov natrénovaných aj na pripojených definíciách, ani skrátené 300 MB korpusy, porovnávame teda iba vplyv parametrov. Nie je to však jednoduché, pretože takmer pri každom korpuse sú výsledky úplne odlišné. Zdá sa však, že pri veľkom korpuse seznamu CBOW modely zaostávajú, aj keď nie až tak výrazne. Pri najmenšom korpuse e-books zase CBOW HS kombinácia predbieha ostatné modely. Spomedzi modelov FastText znova však víťazí varianta SG NS na korpuse seznamu.

5.4 Hodnotenie Glove

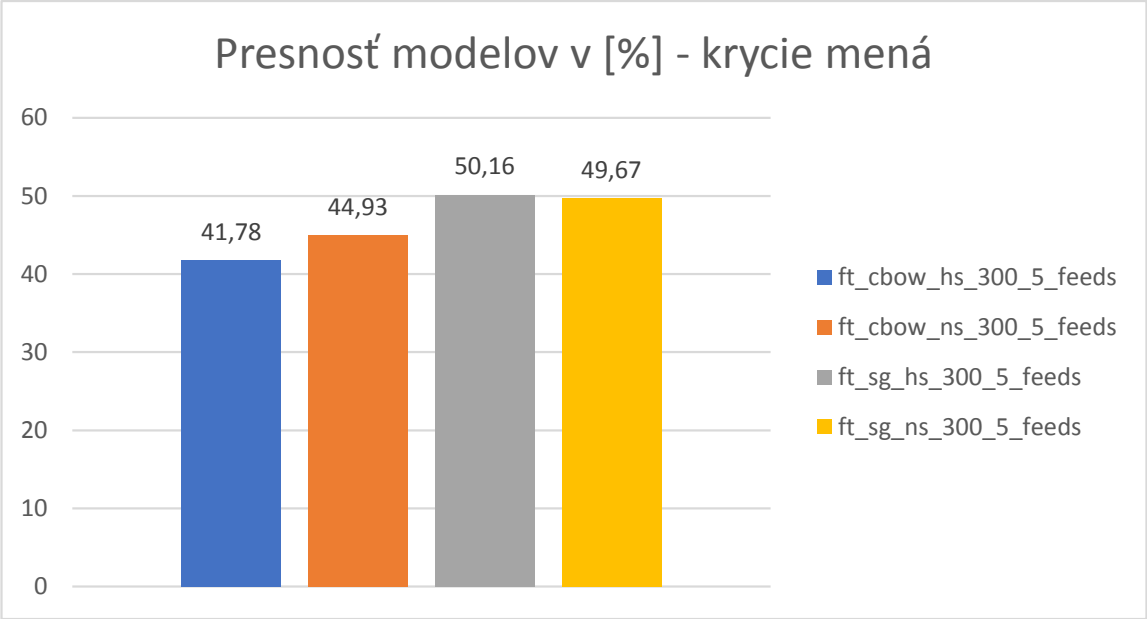
Modely Glove sa zmestili do jedného grafu aj so všetkými variantami korpusov. Líšia sa totiž len v počte trénovacích epóch. Ani tieto modely neobsahujú varianty s definíciami. Špeciálne však boli trénované aj na variantách s polovicou, štvrtinou a osminou korpusu seznamu a takisto aj s prehodeným poradím viet celého korpusu. Možno tu totiž pozorovať zaujímavý úkaz, kde modely trénované na polovici a celom korpuse seznamu dosahujú najnižšie presnosti. Prevrátením obsahu korpusu seznamu sa tak sledovalo, či sa náhodou nenachádza v neskorších partiách korpusu menej kvalitný obsah, ktorý by mal za následok zníženie presnosti. Toto sa nepotvrdilo, nakoľko aj modely s prevráteným obsahom dosahujú obdobné presnosti. Ukazuje sa teda, že modely Glove majú problémy s príliš obšírnymi korpusmi ako je seznam. Varianty so štvrtinou a osminou korpusu seznamu sa ukazujú ako optimálne na trénovanie. Korpusy bubing, feeds a e-books so svojimi veľkosťami sa zase zdajú byť príliš malé na trénovanie Glove. Víťazom sa stala varianta s 10 trénovacími epochami na štvrtine prevráteného korpusu seznamu.



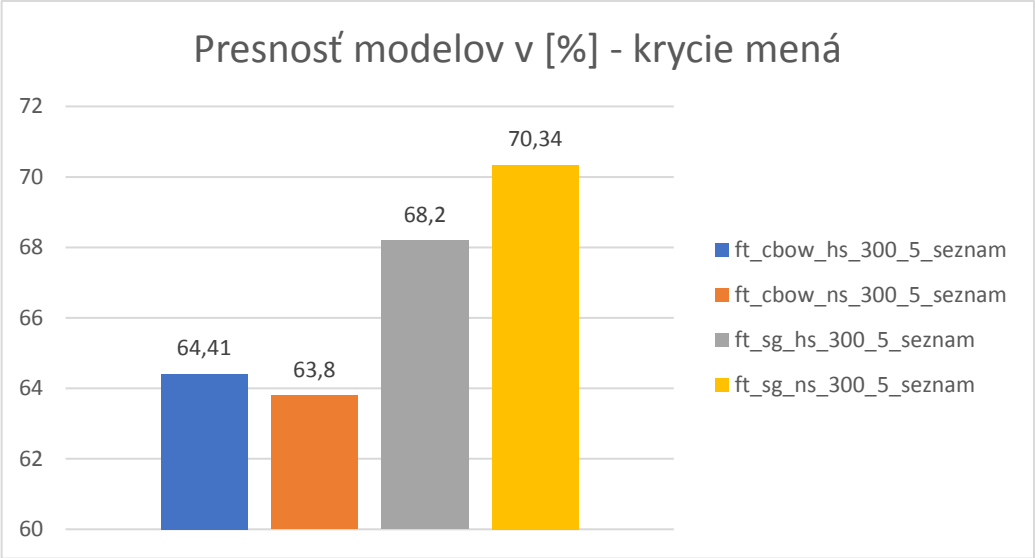
Obr. 5.5: Presnosť modelov FastText trénovaných na korpuse E-books



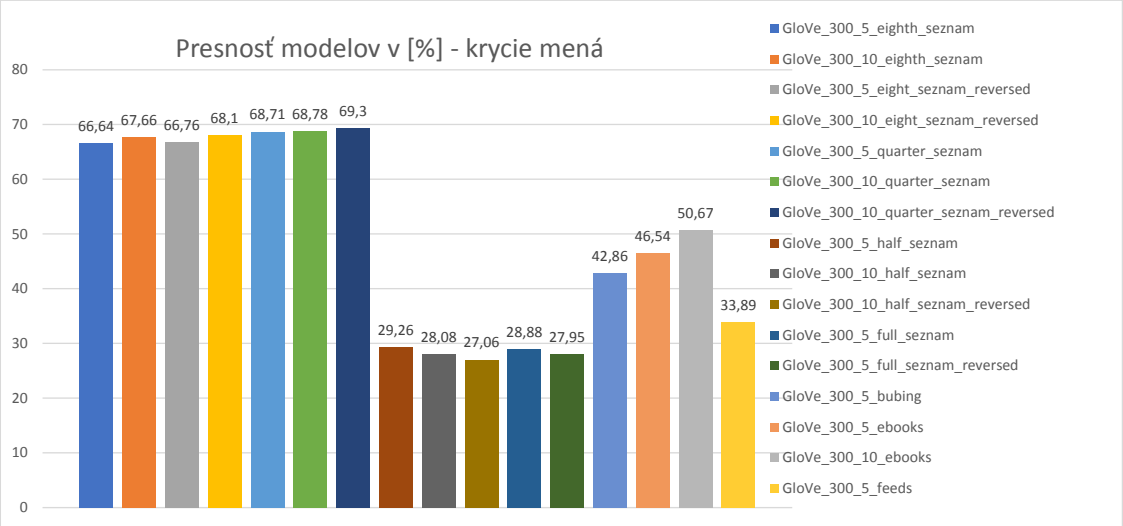
Obr. 5.6: Presnosť modelov FastText trénovaných na korpuse Bubing



Obr. 5.7: Presnosť modelov FastText trénovaných na korpuse Feeds



Obr. 5.8: Presnosť modelov FastText trénovaných na korpuse Seznam2017



Obr. 5.9: Presnosť modelov glove trénovaných na všetkých typoch korpusov

5.5 Hodnotenie Dict2Vec

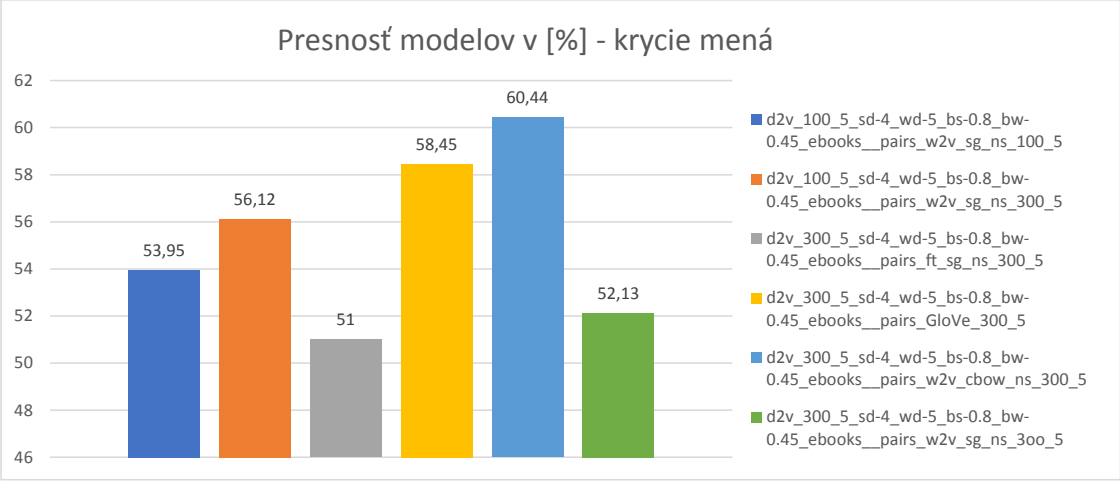
Podľa výsledkov testov autorov Dict2Vec boli tieto modely očakávanými favoritmi na víťazstvo, no nestalo sa tak. Práve naopak. Zo všetkých typov modelov dosahujú najnižšie presnosti. Koncept slabých a silných párov generovaných z definícií sa teda neujal. Svoju silu neprejavili dokonca ani pri 100-dimenzionálnych vektoroch, pretože v porovnaní s Dict2Vec, model Word2Vec ich väčšinou predbiehal aj v tejto kategórii. Avšak pri korpuse Feeds spomedzi všetkých typov modelov vyhral práve Dict2Vec.

5.6 Celkové hodnotenie

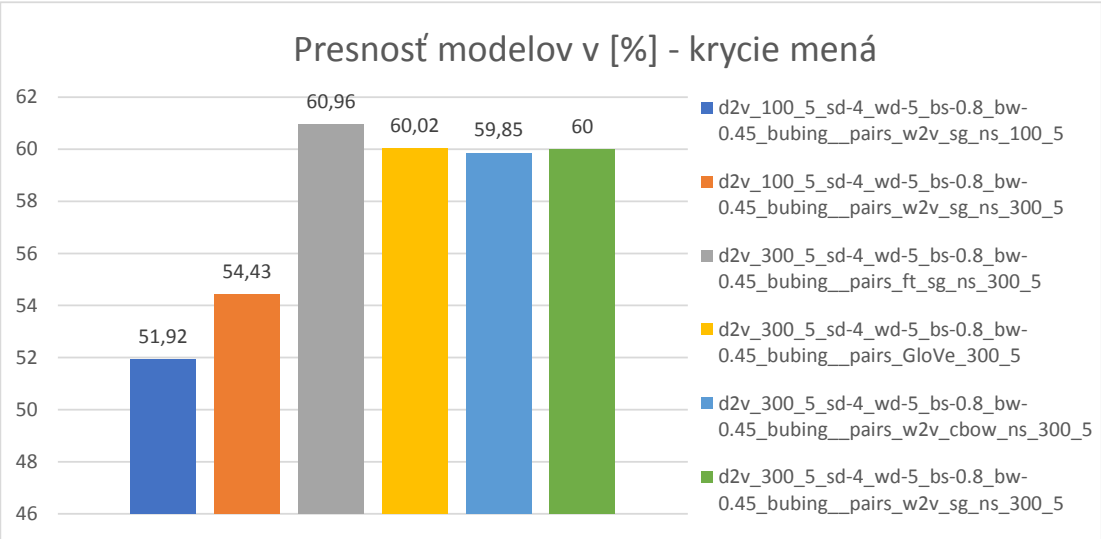
Víťazní zástupcovia každého typu modelu sa nachádzajú v grafe 5.14. Je tu vidno veľký rozdiel v presnosti medzi zástupcom Dict2Vec a ostatnými modelmi. Maximálna presnosť modelov Dict2Vec sa vyšplhala iba na 60,96 %. Čo je však zaujímavé, všetky ostatné typy modelov dosiahli najvyššiu presnosť na korpuse seznamu. V prípade modelu Glove, na štvrtine tohoto korpusu. Dict2Vec však najviac vyhovoval korpus Bubing.

V grafe 5.15 sa zase nachádzajú víťazné modely podľa typu korpusu. Prekvapivo sa do tohoto zoznamu dostal aj zástupca Dict2Vec. So svojou presnosťou 62.25 % predbehol ostatné modely na korpuse Feeds, ktorý sa zdal byť inak najmenej vyhovujúcim korpusom na trénovanie s najmenej kvalitným obsahom. Do tohoto zoznamu sa však nedostal žiadny model Glove, nakoľko nezvíťazil pri žiadnom korpuse.

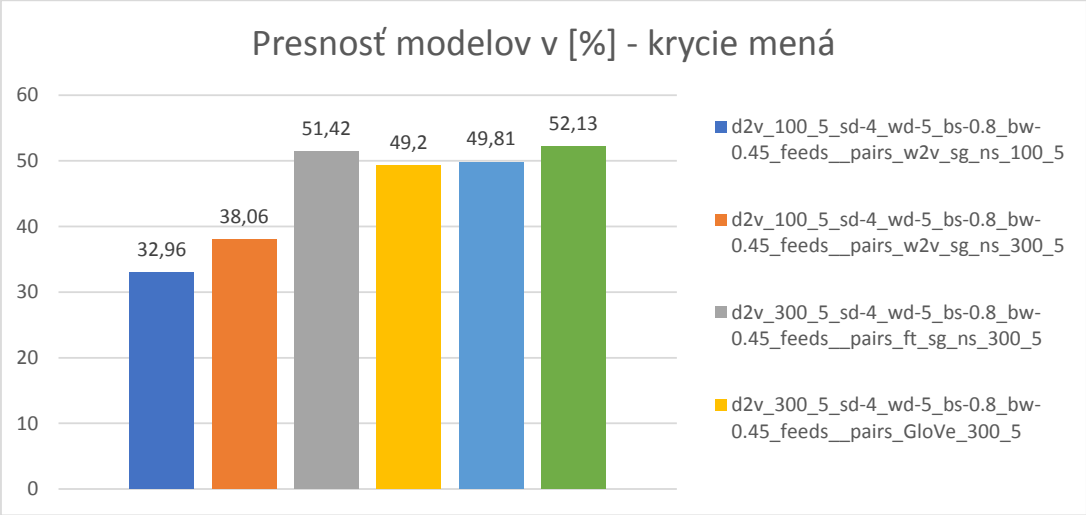
V grafoch sa nenachádzal ešte jeden model špecifický model. Jednalo sa o typ Word2Vec SG NS s 300 dimenziami a 5 epochami pri trénovaní na spojení všetkých 4 korpusov v poradí E-books, Bubing, Feeds a Seznam. Očakávalo sa, že vďaka spojeniu všetkých korpusov predbehne s presnosťou všetky ostatné modely. Na krycích menách však dosiahol presnosť 70,66 % a bol tak predbehnutý modelom natrénovaným iba na korpuse seznamu s rovnakými parametrami. Tento model s presnosťou 70,98 % sa tak stal absolútnym víťazom.



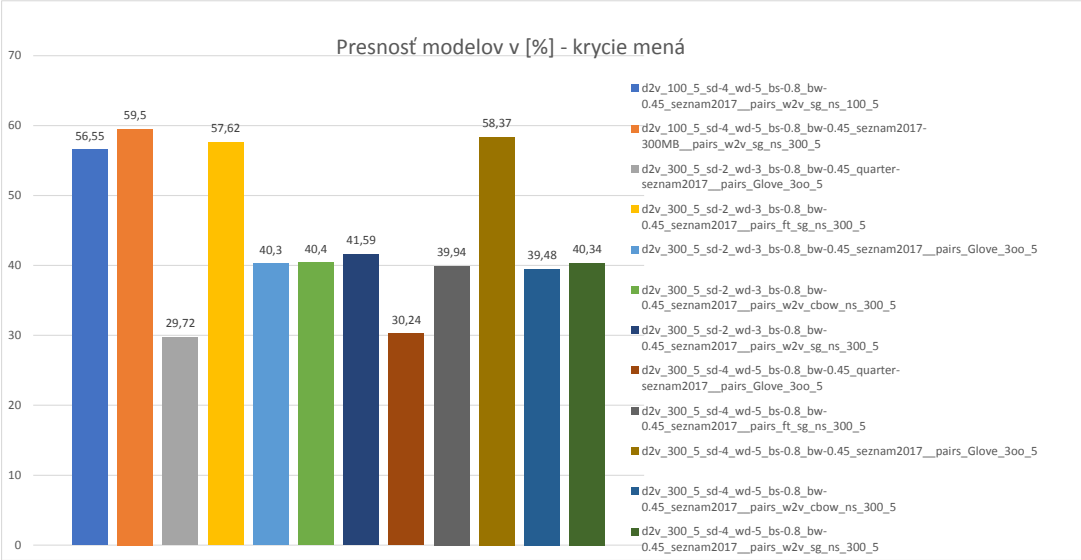
Obr. 5.10: Presnosť modelov Dict2Vec trénovaných na korpuse E-books



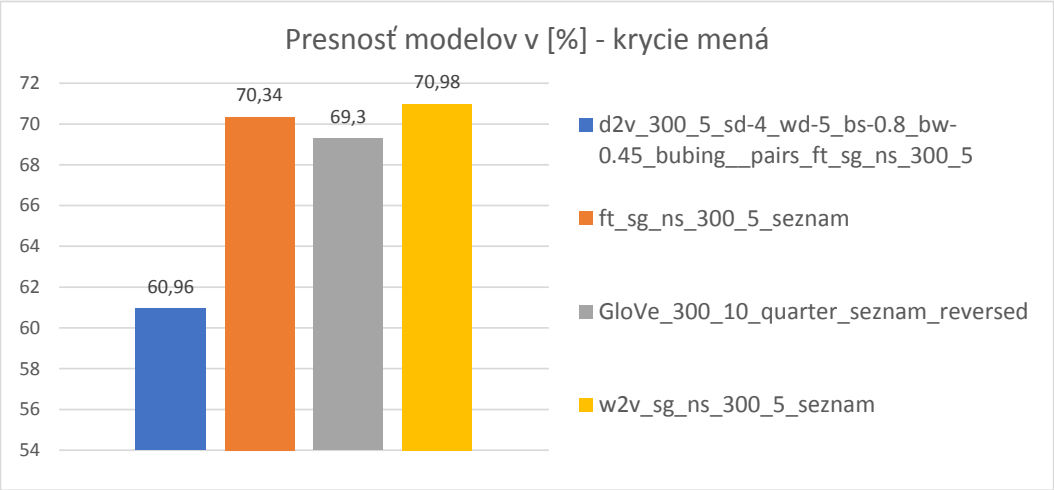
Obr. 5.11: Presnosť modelov Dict2Vec trénovaných na korpuse Bubing



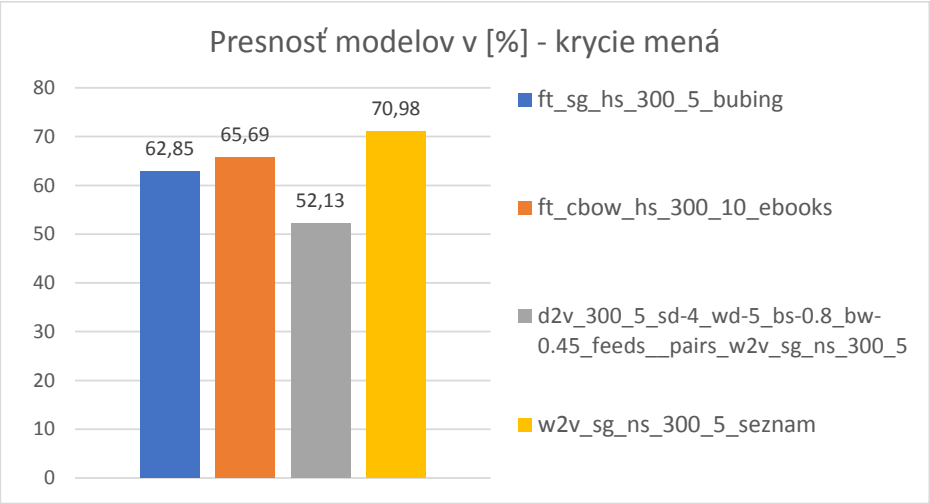
Obr. 5.12: Presnosť modelov Dict2Vec trénovaných na korpuse Feeds



Obr. 5.13: Presnosť modelov Dict2Vec trénovaných na korpuse Seznam2017



Obr. 5.14: Zástupca každého typu modelu s najvyššou dosiahnutou presnosťou



Obr. 5.15: Víťazný model podľa typu korpusu

Kapitola 6

Záver

Existujúci systém na trénovanie modelov typu Word2Vec, FastText a Glove bol v tejto práci rozšírený o modely typu Dict2Vec a prácu so slovníkmi. Z vybraných typov slovníkov je možné extrahovať definície, očistiť ich od menej významných slov, lematizovať a vygenerovať z nich silné a slabé páry pre modely Dict2Vec alebo použiť na iný účel. Boli natrénované a vyhodnotené modely všetkých 4 typov na testovacích sadách krycích mien, slovných analógií a slovníka synoným. Vďaka novým skriptom je možné predspracovať do vhodnej podoby na trénovanie sémantických modelov aj korpusy pochádzajúce zo súborov vo formáte Web Archive. Ďalšie skripty umožňujú aj rýchlejšie vyhodnocovanie existujúcich modelov.

Ďalej bola naimplementovaná webová služba, ktorá vie za pomoci sémantického modelu k zadanému zoznamu kandidátnych slov na základe slova-nápovedy priradiť kosínusovú vzdialenosť. Aplikácia umožňuje komunikáciu vo formáte JSON.

Úspešnosť nových modelov Dict2Vec nebola veľmi vysoká. V práci však nebolo experimentované so všetkými parametrami na trénovanie týchto modelov. Do budúcna by tak bolo možné vďaka existujúcemu systému vytvoriť aj ďalšie varianty týchto modelov a vyhodnotiť ich úspešnosť.

Literatúra

[1] Hajic, J.: *Disambiguation of Rich Inflection - Computational Morphology of Czech*. Charles University, 2004, ISBN 9788024602820.
URL https://books.google.sk/books?id=sB63AAAACAAJ

[2] Hošták, V. S.: *Shlukování slov podle významu*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2017.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=14835

[3] Keyser, C.: *Python Programming Language and Applications*. World Technologies, 2012, ISBN 9788132317586.

[4] Matějka, J.: *Rozšíření systému pro získávání, zpracování a analýzu rozsáhlých kolekcí textů z webu*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2018.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=20846

[5] Straka, M.; Hajic, J.; Straková, J.: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 05 2016.
URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/873.html

[6] Straka, M.; Straková, J.: MorphoDiTa: Morphological Dictionary and Tagger. 2014, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
URL http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0

[7] Straka, M.; Straková, J.: Czech Models (MorfFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115. 2016, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
URL http://hdl.handle.net/11234/1-1836

[8] Švaňa, M.: *Čištění, extrakce textu a převod webových stránek do vertikálního formátu*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2016.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=18729

[9] Tissier, J.; Gravier, C.; Habrard, A.: Dict2vec : Learning Word Embeddings using Lexical Dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational

Linguistics, September 2017, s. 254–263, doi:10.18653/v1/D17-1024.
URL https://www.aclweb.org/anthology/D17-1024

Príloha A

Tabuľky presností všetkých modelov

Tabuľka A.1: Prenosť v [%] modelov Dict2Vec

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_w2v_sg_ns_100_5	
Krycie mená	51,92
Synonymá	6,9
Antonymá (pods. mená)	6,9
Antonymá (príd. mená)	15,74
Antonymá (slovesá)	2,95
Stát-mesto	5,08
Rodinné vzťahy	19,29
Zamestnania	12,46
Národnosť	3,5

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_w2v_sg_ns_300_5	
Krycie mená	54,43
Synonymá	8,2
Antonymá (pods. mená)	10,24
Antonymá (príd. mená)	17,65
Antonymá (slovesá)	1,16
Stát-mesto	8,91
Rodinné vzťahy	22,22
Zamestnania	19,36
Národnosť	6,16

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_ft_sg_ns_300_5	
Krycie mená	60,96
Synonymá	9,24
Antonymá (pods. mená)	14,79
Antonymá (príd. mená)	13,12
Antonymá (slovesá)	2,41
Stát-mesto	16,58
Rodinné vzťahy	25,31
Zamestnania	33,92
Národnosť	10,98

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_GloVe_300_5	
Krycie mená	60,02
Synonymá	7,56
Antonymá (pods. mená)	13,51
Antonymá (príd. mená)	15,39
Antonymá (slovesá)	3,66
Stát-mesto	11,76
Rodinné vzťahy	27,62
Zamestnania	24,07
Národnosť	8,14

Tabuľka A.2: Prenosť v [%] modelov Dict2Vec

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_w2v_cbow_ns_300_5	
Krycie mená	59,85
Synonymá	8,91
Antonymá (pods. mená)	14,37
Antonymá (príd. mená)	16,03
Antonymá (slovesá)	3,48
Stát-mesto	14,71
Rodinné vzťahy	28,09
Zamestnania	29,12
Národnosť	7,95

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_bubing__pairs_w2v_sg_ns_300_5	
Krycie mená	60
Synonymá	8,16
Antonymá (pods. mená)	14,44
Antonymá (príd. mená)	12,14
Antonymá (slovesá)	3,3
Stát-mesto	15,06
Rodinné vzťahy	27,62
Zamestnania	27,78
Národnosť	7,77

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_w2v_sg_ns_100_5	
Krycie mená	53,95
Synonymá	18,53
Antonymá (pods. mená)	6,83
Antonymá (príd. mená)	19,98
Antonymá (slovesá)	2,5
Stát-mesto	4,72
Rodinné vzťahy	15,28
Zamestnania	5,72
Národnosť	7,29

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_w2v_sg_ns_300_5	
Krycie mená	56,12
Synonymá	18,86
Antonymá (pods. mená)	5,55
Antonymá (príd. mená)	13,41
Antonymá (slovesá)	2,59
Stát-mesto	14,88
Rodinné vzťahy	22,53
Zamestnania	13,64
Národnosť	9,56

Tabuľka A.3: Prenosť v [%] modelov Dict2Vec

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_ft_sg_ns_300_5	
Krycie mená	51
Synonymá	20,91
Antonymá (pods. mená)	11,74
Antonymá (príd. mená)	4,88
Antonymá (slovesá)	6,88
Stát-mesto	18,72
Rodinné vzťahy	29,01
Zamestnania	29,71
Národnosť	16

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_GloVe_300_5	
Krycie mená	58,45
Synonymá	18,26
Antonymá (pods. mená)	9,89
Antonymá (príd. mená)	13,01
Antonymá (slovesá)	10,45
Stát-mesto	9,8
Rodinné vzťahy	24,54
Zamestnania	12,12
Národnosť	8,62

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_w2v_cbow_ns_300_5	
Krycie mená	60,44
Synonymá	21,34
Antonymá (pods. mená)	12,87
Antonymá (príd. mená)	9,29
Antonymá (slovesá)	9,91
Stát-mesto	13,99
Rodinné vzťahy	20,68
Zamestnania	26,68
Národnosť	12,12

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_ebooks__pairs_w2v_sg_ns_3oo_5	
Krycie mená	52,13
Synonymá	20,69
Antonymá (pods. mená)	11,66
Antonymá (príd. mená)	10,16
Antonymá (slovesá)	4,02
Stát-mesto	15,24
Rodinné vzťahy	26,7
Zamestnania	21,89
Národnosť	12,41

Tabuľka A.4: Prenosť v [%] modelov Dict2Vec

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_w2v_sg_ns_100_5	
Krycie mená	32,96
Synonymá	2,57
Antonymá (pods. mená)	1,14
Antonymá (príd. mená)	1,34
Antonymá (slovesá)	0,09
Stát-mesto	0
Rodinné vzťahy	0,46
Zamestnania	0,42
Národnosť	1,04

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_w2v_sg_ns_300_5	
Krycie mená	38,06
Synonymá	2,91
Antonymá (pods. mená)	3,49
Antonymá (príd. mená)	4,7
Antonymá (slovesá)	0,09
Stát-mesto	1,34
Rodinné vzťahy	3,09
Zamestnania	2,86
Národnosť	1,23

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_ft_sg_ns_300_5	
Krycie mená	51,42
Synonymá	5,39
Antonymá (pods. mená)	11,31
Antonymá (príd. mená)	5,17
Antonymá (slovesá)	2,05
Stát-mesto	2,23
Rodinné vzťahy	7,1
Zamestnania	8,75
Národnosť	5,78

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_GloVe_300_5	
Krycie mená	49,2
Synonymá	4,39
Antonymá (pods. mená)	9,82
Antonymá (príd. mená)	5,11
Antonymá (slovesá)	0,54
Stát-mesto	0,89
Rodinné vzťahy	5,86
Zamestnania	8,25
Národnosť	4,64

Tabuľka A.5: Prenosť v [%] modelov Dict2Vec

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_w2v_cbow_ns_300_5	
Krycie mená	49,81
Synonymá	5,03
Antonymá (pods. mená)	11,02
Antonymá (príd. mená)	6,33
Antonymá (slovesá)	1,7
Stát-mesto	2,23
Rodinné vzťahy	6,79
Zamestnania	7,15
Národnosť	5,68

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_feeds__pairs_w2v_sg_ns_300_5	
Krycie mená	52,13
Synonymá	4,63
Antonymá (pods. mená)	10,67
Antonymá (príd. mená)	3,72
Antonymá (slovesá)	0
Stát-mesto	2,5
Rodinné vzťahy	7,1
Zamestnania	7,66
Národnosť	6,44

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_seznam2017__pairs_w2v_sg_ns_100_5	
Krycie mená	56,55
Synonymá	10,19
Antonymá (pods. mená)	9,53
Antonymá (príd. mená)	20,5
Antonymá (slovesá)	1,96
Stát-mesto	8,29
Rodinné vzťahy	17,13
Zamestnania	19,11
Národnosť	6,91

d2v_100_5_sd-4_wd-5_bs-0.8_bw-0.45_seznam2017-300mB__pairs_w2v_sg_ns_300_5	
Krycie mená	59,5
Synonymá	13,51
Antonymá (pods. mená)	5,9
Antonymá (príd. mená)	14,05
Antonymá (slovesá)	1,79
Stát-mesto	22,01
Rodinné vzťahy	17,59
Zamestnania	21,04
Národnosť	13,26

Tabuľka A.6: Prenosť v [%] modelov Dict2Vec

d2v_300_5_sd-2_wd-3_bs-0.8_bw-0.45_quarter-seznam2017__pairs_Glove_3oo_5	
Krycie mená	29,72
Synonymá	0,09
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

d2v_300_5_sd-2_wd-3_bs-0.8_bw-0.45_seznam2017__pairs_ft_sg_ns_300_5	
Krycie mená	57,62
Synonymá	20,16
Antonymá (pods. mená)	11,24
Antonymá (príd. mená)	12,14
Antonymá (slovesá)	4,55
Stát-mesto	50,71
Rodinné vzťahy	28,4
Zamestnania	66,92
Národnosť	41,1

d2v_300_5_sd-2_wd-3_bs-0.8_bw-0.45_seznam2017__pairs_Glove_3oo_5	
Krycie mená	40,3
Synonymá	9,71
Antonymá (pods. mená)	5,69
Antonymá (príd. mená)	2,96
Antonymá (slovesá)	1,88
Stát-mesto	30,57
Rodinné vzťahy	19,29
Zamestnania	41,84
Národnosť	29,64

d2v_300_5_sd-2_wd-3_bs-0.8_bw-0.45_seznam2017__pairs_w2v_cbow_ns_300_5	
Krycie mená	40,4
Synonymá	16,17
Antonymá (pods. mená)	7,89
Antonymá (príd. mená)	6,33
Antonymá (slovesá)	3,13
Stát-mesto	49,29
Rodinné vzťahy	20,22
Zamestnania	60,44
Národnosť	34,85

Tabuľka A.7: Prenosť v [%] modelov Dict2Vec

d2v_300_5_sd-2_wd-3_bs-0.8_bw-0.45_seznam2017__pairs_w2v_sg_ns_300_5	
Krycie mená	41,59
Synonymá	16,67
Antonymá (pods. mená)	8,61
Antonymá (príd. mená)	8,59
Antonymá (slovesá)	2,95
Stát-mesto	52,5
Rodinné vzťahy	21,45
Zamestnania	54,46
Národnosť	31,53

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_quarter-seznam2017__pairs_Glove_3oo_5	
Krycie mená	30,24
Synonymá	0,04
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_seznam2017__pairs_ft_sg_ns_300_5	
Krycie mená	39,94
Synonymá	15,41
Antonymá (pods. mená)	6,9
Antonymá (príd. mená)	6,1
Antonymá (slovesá)	3,04
Stát-mesto	48,57
Rodinné vzťahy	17,75
Zamestnania	55,22
Národnosť	30,78

d2v_300_5_sd-4_wd-5_bs-0.8_bw-0.45_seznam2017__pairs_Glove_3oo_5	
Krycie mená	58,37
Synonymá	14,11
Antonymá (pods. mená)	9,96
Antonymá (príd. mená)	6,16
Antonymá (slovesá)	2,32
Stát-mesto	35,03
Rodinné vzťahy	24,54
Zamestnania	45,37
Národnosť	36,93

Tabuľka A.8: Prenosť v [%] modelov Word2Vec

w2v_cbow_hs_300_5_bubing	
Krycie mená	59,95
Synonymá	8,9
Antonymá (pods. mená)	9,89
Antonymá (príd. mená)	23,93
Antonymá (slovesá)	2,14
Stát-mesto	4,1
Rodinné vzťahy	14,04
Zamestnania	21,13
Národnosť	1,04

w2v_cbow_ns_300_5_bubing	
Krycie mená	60,07
Synonymá	10,85
Antonymá (pods. mená)	19,2
Antonymá (príd. mená)	27,99
Antonymá (slovesá)	3,57
Stát-mesto	14,88
Rodinné vzťahy	29,94
Zamestnania	48,32
Národnosť	7,2

w2v_sg_hs_300_5_bubing	
Krycie mená	61,34
Synonymá	6,49
Antonymá (pods. mená)	11,81
Antonymá (príd. mená)	19,05
Antonymá (slovesá)	2,95
Stát-mesto	13,19
Rodinné vzťahy	29,01
Zamestnania	25
Národnosť	3,6

w2v_sg_ns_100_5_bubing+defs	
Krycie mená	58,16
Synonymá	13,42
Antonymá (pods. mená)	9,25
Antonymá (príd. mená)	15,74
Antonymá (slovesá)	2,05
Stát-mesto	7,4
Rodinné vzťahy	24,38
Zamestnania	24,66
Národnosť	3,5

w2v_sg_ns_100_5_bubing	
Krycie mená	53,93
Synonymá	3,42
Antonymá (pods. mená)	8,11
Antonymá (príd. mená)	13,65
Antonymá (slovesá)	1,88
Stát-mesto	7,31
Rodinné vzťahy	20,99
Zamestnania	17,42
Národnosť	2,18

w2v_sg_ns_300_5_bubing	
Krycie mená	60,94
Synonymá	4,21
Antonymá (pods. mená)	14,58
Antonymá (príd. mená)	17,13
Antonymá (slovesá)	2,59
Stát-mesto	12,48
Rodinné vzťahy	33,95
Zamestnania	31,65
Národnosť	3,31

w2v_sg_ns_300_5_bubing+defs	
Krycie mená	62,56
Synonymá	9,78
Antonymá (pods. mená)	15,65
Antonymá (príd. mená)	17,02
Antonymá (slovesá)	3,57
Stát-mesto	13,73
Rodinné vzťahy	34,57
Zamestnania	34,01
Národnosť	5,3

w2v_cbow_hs_300_5_ebooks	
Krycie mená	61,85
Synonymá	24,21
Antonymá (pods. mená)	10,17
Antonymá (príd. mená)	31,07
Antonymá (slovesá)	16,07
Stát-mesto	16,84
Rodinné vzťahy	29,32
Zamestnania	15,15
Národnosť	14,2

Tabuľka A.9: Prenosť v [%] modelov Word2Vec

w2v_cbow_hs_300_10_ebooks	
Krycie mená	61,88
Synonymá	25,17
Antonymá (pods. mená)	11,02
Antonymá (príd. mená)	30,26
Antonymá (slovesá)	13,04
Stát-mesto	15,33
Rodinné vzťahy	29,63
Zamestnania	15,07
Národnosť	13,16

w2v_cbow_ns_300_5_ebooks	
Krycie mená	61,2
Synonymá	22,54
Antonymá (pods. mená)	11,02
Antonymá (príd. mená)	25,96
Antonymá (slovesá)	11,43
Stát-mesto	21,66
Rodinné vzťahy	33,33
Zamestnania	21,72
Národnosť	15,72

w2v_sg_hs_300_5_ebooks	
Krycie mená	61,84
Synonymá	19,2
Antonymá (pods. mená)	9,67
Antonymá (príd. mená)	36,7
Antonymá (slovesá)	13,93
Stát-mesto	34,14
Rodinné vzťahy	33,64
Zamestnania	19,11
Národnosť	21,31

w2v_sg_ns_100_5_ebooks+defs	
Krycie mená	61,58
Synonymá	22,58
Antonymá (pods. mená)	6,4
Antonymá (príd. mená)	22,76
Antonymá (slovesá)	7,14
Stát-mesto	11,41
Rodinné vzťahy	23,3
Zamestnania	13,38
Národnosť	14,96

w2v_sg_ns_100_5_ebooks	
Krycie mená	56,25
Synonymá	13,07
Antonymá (pods. mená)	6,76
Antonymá (príd. mená)	20,73
Antonymá (slovesá)	7,77
Stát-mesto	12,57
Rodinné vzťahy	20,83
Zamestnania	7,49
Národnosť	11,84

w2v_sg_ns_300_5_ebooks	
Krycie mená	61,72
Synonymá	16,89
Antonymá (pods. mená)	11,24
Antonymá (príd. mená)	34,96
Antonymá (slovesá)	10,8
Stát-mesto	23,89
Rodinné vzťahy	36,57
Zamestnania	21,8
Národnosť	22,06

w2v_sg_ns_300_5_ebooks+defs	
Krycie mená	63,96
Synonymá	22,3
Antonymá (pods. mená)	10,24
Antonymá (príd. mená)	32,69
Antonymá (slovesá)	9,73
Stát-mesto	23,08
Rodinné vzťahy	36,11
Zamestnania	23,32
Národnosť	27,56

w2v_sg_ns_300_10_ebooks	
Krycie mená	64,08
Synonymá	18,7
Antonymá (pods. mená)	13,16
Antonymá (príd. mená)	33,86
Antonymá (slovesá)	11,7
Stát-mesto	27,45
Rodinné vzťahy	31,48
Zamestnania	18,86
Národnosť	24,91

Tabuľka A.10: Prenosť v [%] modelov Word2Vec

w2v_cbow_hs_300_5_feeds	
Krycie mená	38,33
Synonymá	1,81
Antonymá (pods. mená)	1,42
Antonymá (príd. mená)	1,1
Antonymá (slovesá)	0
Stát-mesto	0,53
Rodinné vzťahy	0,15
Zamestnania	5,64
Národnosť	0,66

w2v_cbow_ns_300_5_feeds	
Krycie mená	48,24
Synonymá	4,27
Antonymá (pods. mená)	7,11
Antonymá (príd. mená)	12,2
Antonymá (slovesá)	2,5
Stát-mesto	4,1
Rodinné vzťahy	10,19
Zamestnania	14,81
Národnosť	4,36

w2v_sg_hs_300_5_feeds	
Krycie mená	47,36
Synonymá	2,1
Antonymá (pods. mená)	7,33
Antonymá (príd. mená)	12,43
Antonymá (slovesá)	0,18
Stát-mesto	3,57
Rodinné vzťahy	13,27
Zamestnania	9,76
Národnosť	2,27

w2v_sg_ns_100_5_feeds+defs	
Krycie mená	44,94
Synonymá	12,2
Antonymá (pods. mená)	2,42
Antonymá (príd. mená)	2,09
Antonymá (slovesá)	0,36
Stát-mesto	0,27
Rodinné vzťahy	5,4
Zamestnania	0,25
Národnosť	0,47

w2v_sg_ns_100_5_feeds	
Krycie mená	30,95
Synonymá	0,2
Antonymá (pods. mená)	1,35
Antonymá (príd. mená)	0,81
Antonymá (slovesá)	0,09
Stát-mesto	0,27
Rodinné vzťahy	0,31
Zamestnania	0,08
Národnosť	0,09

w2v_sg_ns_300_5_feeds	
Krycie mená	49,94
Synonymá	1,27
Antonymá (pods. mená)	6,97
Antonymá (príd. mená)	9,12
Antonymá (slovesá)	0
Stát-mesto	1,6
Rodinné vzťahy	4,01
Zamestnania	8,16
Národnosť	2,18

w2v_sg_ns_300_5_feeds+defs	
Krycie mená	30,95
Synonymá	0,23
Antonymá (pods. mená)	1,49
Antonymá (príd. mená)	0,06
Antonymá (slovesá)	0
Stát-mesto	0,36
Rodinné vzťahy	0,46
Zamestnania	0
Národnosť	0

w2v_cbow_hs_300_5_seznam	
Krycie mená	65,74
Synonymá	23,41
Antonymá (pods. mená)	7,04
Antonymá (príd. mená)	21,2
Antonymá (slovesá)	7,14
Stát-mesto	32,71
Rodinné vzťahy	29,63
Zamestnania	60,19
Národnosť	40,34

Tabuľka A.11: Prenosť v [%] modelov Word2Vec

w2v_cbow_ns_300_5_seznam	
Krycie mená	67
Synonymá	26,8
Antonymá (pods. mená)	19,06
Antonymá (príd. mená)	26,31
Antonymá (slovesá)	10,36
Stát-mesto	65,42
Rodinné vzťahy	52,01
Zamestnania	81,48
Národnosť	67,33

w2v_cbow_ns_300_10_seznam	
Krycie mená	67,02
Synonymá	26,9
Antonymá (pods. mená)	18,21
Antonymá (príd. mená)	26,02
Antonymá (slovesá)	10,18
Stát-mesto	65,33
Rodinné vzťahy	51,85
Zamestnania	81,4
Národnosť	67,23

w2v_sg_hs_300_5_seznam	
Krycie mená	68,97
Synonymá	18,97
Antonymá (pods. mená)	7,61
Antonymá (príd. mená)	15,85
Antonymá (slovesá)	4,29
Stát-mesto	46,97
Rodinné vzťahy	37,96
Zamestnania	55,56
Národnosť	44,22

w2v_sg_ns_100_5_seznam+defs	
Krycie mená	62,21
Synonymá	15,85
Antonymá (pods. mená)	8,18
Antonymá (príd. mená)	22,24
Antonymá (slovesá)	3,04
Stát-mesto	18,36
Rodinné vzťahy	33,02
Zamestnania	41,16
Národnosť	15,25

w2v_sg_ns_100_5_seznam	
Krycie mená	59,85
Synonymá	5,72
Antonymá (pods. mená)	7,89
Antonymá (príd. mená)	21,66
Antonymá (slovesá)	4,2
Stát-mesto	16,13
Rodinné vzťahy	30,86
Zamestnania	35,52
Národnosť	10,42

w2v_sg_ns_300_5_seznam	
Krycie mená	70,98
Synonymá	21,18
Antonymá (pods. mená)	9,25
Antonymá (príd. mená)	14,75
Antonymá (slovesá)	5,45
Stát-mesto	63,55
Rodinné vzťahy	45,83
Zamestnania	76,68
Národnosť	66,38

w2v/mix_books+bubing+feeds+seznam	
Krycie mená	70,66
Synonymá	21,2
Antonymá (pods. mená)	9,74
Antonymá (príd. mená)	14,63
Antonymá (slovesá)	5,54
Stát-mesto	64,08
Rodinné vzťahy	47,84
Zamestnania	78,03
Národnosť	67,05

Tabuľka A.12: Prenosť v [%] modelov FastText

ft_cbow_hs_300_5_bubing	
Krycie mená	60,16
Synonymá	8,19
Antonymá (pods. mená)	6,61
Antonymá (príd. mená)	13,07
Antonymá (slovesá)	3,75
Stát-mesto	1,52
Rodinné vzťahy	9,72
Zamestnania	60,69
Národnosť	7,29

ft_cbow_ns_300_5_bubing	
Krycie mená	51,47
Synonymá	5,94
Antonymá (pods. mená)	9,46
Antonymá (príd. mená)	3,14
Antonymá (slovesá)	4,29
Stát-mesto	0,98
Rodinné vzťahy	10,8
Zamestnania	63,05
Národnosť	22,92

ft_sg_hs_300_5_bubing	
Krycie mená	62,85
Synonymá	8,13
Antonymá (pods. mená)	10,1
Antonymá (príd. mená)	16,26
Antonymá (slovesá)	3,39
Stát-mesto	5,79
Rodinné vzťahy	30,09
Zamestnania	58,59
Národnosť	9,38

ft_sg_ns_300_5_bubing	
Krycie mená	61,44
Synonymá	6,58
Antonymá (pods. mená)	11,59
Antonymá (príd. mená)	9,58
Antonymá (slovesá)	2,77
Stát-mesto	2,67
Rodinné vzťahy	22,84
Zamestnania	65,66
Národnosť	15,25

ft_cbow_hs_300_5_ebooks	
Krycie mená	64,78
Synonymá	22,14
Antonymá (pods. mená)	13,73
Antonymá (príd. mená)	13,24
Antonymá (slovesá)	12,68
Stát-mesto	8,91
Rodinné vzťahy	31,17
Zamestnania	77,86
Národnosť	34,09

ft_cbow_hs_300_10_ebooks	
Krycie mená	65,69
Synonymá	25,51
Antonymá (pods. mená)	12,94
Antonymá (príd. mená)	18,58
Antonymá (slovesá)	14,82
Stát-mesto	11,76
Rodinné vzťahy	31,79
Zamestnania	77,19
Národnosť	32,2

ft_cbow_ns_300_5_ebooks	
Krycie mená	59,43
Synonymá	14,23
Antonymá (pods. mená)	13,02
Antonymá (príd. mená)	5,23
Antonymá (slovesá)	12,32
Stát-mesto	5,35
Rodinné vzťahy	17,75
Zamestnania	73,06
Národnosť	29,64

ft_sg_hs_300_5_ebooks	
Krycie mená	60,72
Synonymá	21,76
Antonymá (pods. mená)	12,87
Antonymá (príd. mená)	18,93
Antonymá (slovesá)	11,25
Stát-mesto	16,13
Rodinné vzťahy	37,65
Zamestnania	71,04
Národnosť	29,92

Tabuľka A.13: Prenosť v [%] modelov FastText

ft_sg_ns_300_5_ebooks	
Krycie mená	58,27
Synonymá	17,73
Antonymá (pods. mená)	13,51
Antonymá (príd. mená)	10,69
Antonymá (slovesá)	9,11
Stát-mesto	13,1
Rodinné vzťahy	33,8
Zamestnania	82,49
Národnosť	35,32

ft_sg_ns_300_10_ebooks	
Krycie mená	57,76
Synonymá	19,83
Antonymá (pods. mená)	12,52
Antonymá (príd. mená)	12,72
Antonymá (slovesá)	10,54
Stát-mesto	14,53
Rodinné vzťahy	35,03
Zamestnania	81,65
Národnosť	32,1

ft_cbow_hs_300_5_feeds	
Krycie mená	41,78
Synonymá	5,08
Antonymá (pods. mená)	5,05
Antonymá (príd. mená)	2,38
Antonymá (slovesá)	0,89
Stát-mesto	0
Rodinné vzťahy	7,72
Zamestnania	51,26
Národnosť	4,55

ft_cbow_ns_300_5_feeds	
Krycie mená	44,93
Synonymá	4,6
Antonymá (pods. mená)	5,26
Antonymá (príd. mená)	0,12
Antonymá (slovesá)	6,7
Stát-mesto	1,07
Rodinné vzťahy	7,87
Zamestnania	56,9
Národnosť	15,63

ft_sg_hs_300_5_feeds	
Krycie mená	50,16
Synonymá	4,98
Antonymá (pods. mená)	6,33
Antonymá (príd. mená)	12,78
Antonymá (slovesá)	0,54
Stát-mesto	1,34
Rodinné vzťahy	22,38
Zamestnania	48,57
Národnosť	7,01

ft_sg_ns_300_5_feeds	
Krycie mená	49,67
Synonymá	5
Antonymá (pods. mená)	8,25
Antonymá (príd. mená)	3,31
Antonymá (slovesá)	2,14
Stát-mesto	0,09
Rodinné vzťahy	14,81
Zamestnania	57,58
Národnosť	7,77

ft_cbow_hs_300_5_seznam	
Krycie mená	64,41
Synonymá	9,65
Antonymá (pods. mená)	2,2
Antonymá (príd. mená)	6,45
Antonymá (slovesá)	2,5
Stát-mesto	22,55
Rodinné vzťahy	28,7
Zamestnania	55,3
Národnosť	49,91

ft_cbow_ns_300_5_seznam	
Krycie mená	63,8
Synonymá	6,25
Antonymá (pods. mená)	2,28
Antonymá (príd. mená)	1,57
Antonymá (slovesá)	2,14
Stát-mesto	25,76
Rodinné vzťahy	21,14
Zamestnania	58,16
Národnosť	53,31

Tabuľka A.14: Prenosť v [%] modelov FastText

ft_sg_hs_300_5_seznam	
Krycie mená	68,2
Synonymá	15,2
Antonymá (pods. mená)	4,98
Antonymá (príd. mená)	16,49
Antonymá (slovesá)	1,96
Stát-mesto	47,59
Rodinné vzťahy	41,05
Zamestnania	56,14
Národnosť	51,99

ft_sg_ns_300_5_seznam	
Krycie mená	70,34
Synonymá	12,25
Antonymá (pods. mená)	5,83
Antonymá (príd. mená)	12,25
Antonymá (slovesá)	1,52
Stát-mesto	54,28
Rodinné vzťahy	47,22
Zamestnania	62,88
Národnosť	62,12

Tabuľka A.15: Prenosť v [%] modelov Glove

GloVe_300_5_bubing	
Krycie mená	42,86
Synonymá	1,93
Antonymá (pods. mená)	2,84
Antonymá (príd. mená)	4,07
Antonymá (slovesá)	1,52
Stát-mesto	3,74
Rodinné vzťahy	26,23
Zamestnania	2,78
Národnosť	0,28

GloVe_300_5_ebooks	
Krycie mená	46,54
Synonymá	9,88
Antonymá (pods. mená)	3,34
Antonymá (príd. mená)	23,93
Antonymá (slovesá)	9,46
Stát-mesto	5,53
Rodinné vzťahy	25,15
Zamestnania	4,29
Národnosť	3,31

GloVe_300_10_ebooks	
Krycie mená	50,67
Synonymá	13,05
Antonymá (pods. mená)	6,05
Antonymá (príd. mená)	22,71
Antonymá (slovesá)	10,36
Stát-mesto	7,84
Rodinné vzťahy	25,77
Zamestnania	5,22
Národnosť	6,82

GloVe_300_5_feeds	
Krycie mená	33,89
Synonymá	0,49
Antonymá (pods. mená)	0,36
Antonymá (príd. mená)	0,06
Antonymá (slovesá)	0
Stát-mesto	0,36
Rodinné vzťahy	0,46
Zamestnania	0,25
Národnosť	0,28

GloVe_300_5_quarter_seznam	
Krycie mená	68,71
Synonymá	15,03
Antonymá (pods. mená)	6,05
Antonymá (príd. mená)	5,57
Antonymá (slovesá)	8,75
Stát-mesto	45,99
Rodinné vzťahy	25,62
Zamestnania	37,88
Národnosť	31,91

GloVe_300_5_half_seznam	
Krycie mená	29,26
Synonymá	0
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

GloVe_300_5_full_seznam	
Krycie mená	28,88
Synonymá	0
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

GloVe_300_5_eighth_seznam	
Krycie mená	66,64
Synonymá	13,46
Antonymá (pods. mená)	7,04
Antonymá (príd. mená)	8,19
Antonymá (slovesá)	5,63
Stát-mesto	44,21
Rodinné vzťahy	22,69
Zamestnania	35,44
Národnosť	25,95

Tabuľka A.16: Prenosť v [%] modelov Glove

GloVe_300_10_quarter_seznam	
Krycie mená	68,78
Synonymá	16,59
Antonymá (pods. mená)	5,83
Antonymá (príd. mená)	6,91
Antonymá (slovesá)	7,41
Stát-mesto	47,06
Rodinné vzťahy	22,38
Zamestnania	31,65
Národnosť	28,5

GloVe_300_10_half_seznam	
Krycie mená	28,08
Synonymá	0
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

GloVe_300_10_eighth_seznam	
Krycie mená	67,66
Synonymá	15,01
Antonymá (pods. mená)	5,62
Antonymá (príd. mená)	9,99
Antonymá (slovesá)	5,63
Stát-mesto	47,24
Rodinné vzťahy	13,27
Zamestnania	31,4
Národnosť	26,23

GloVe_300_5_eight_seznam_reversed	
Krycie mená	66,76
Synonymá	13,54
Antonymá (pods. mená)	10,1
Antonymá (príd. mená)	8,13
Antonymá (slovesá)	7,23
Stát-mesto	48,04
Rodinné vzťahy	22,69
Zamestnania	37,04
Národnosť	25

GloVe_300_5_full_seznam_reversed	
Krycie mená	27,95
Synonymá	0
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0

GloVe_300_10_eight_seznam_reversed	
Krycie mená	68,1
Synonymá	15,18
Antonymá (pods. mená)	10,1
Antonymá (príd. mená)	10,74
Antonymá (slovesá)	6,25
Stát-mesto	50,8
Rodinné vzťahy	24,69
Zamestnania	29,8
Národnosť	23,3

GloVe_300_10_quarter_seznam_reversed	
Krycie mená	69,3
Synonymá	16,76
Antonymá (pods. mená)	7,97
Antonymá (príd. mená)	2,5
Antonymá (slovesá)	5,09
Stát-mesto	47,68
Rodinné vzťahy	26,7
Zamestnania	34,26
Národnosť	25,76

GloVe_300_10_half_seznam_reversed	
Krycie mená	27,06
Synonymá	0
Antonymá (pods. mená)	0
Antonymá (príd. mená)	0
Antonymá (slovesá)	0
Stát-mesto	0
Rodinné vzťahy	0
Zamestnania	0
Národnosť	0