WORD EMBEDDING

Datum konání: 20.11.2020

Místo konání: online (prostredníctvom MS Teams)

Název přednášky: Word Embedding

Přednášející: RNDr. Radovan Garabík

Počet účastníků: <=30

Zpracoval: Ptáček Dušan

RNDr. Radovan Garabík

Vyštudoval študijný program Jadrová a subjadrová fyzika na Matematicko-fyzikálnej fakulte Univerzity Komenského v Bratislave, kde v roku 1998 získal titul RNDr. Postupne sa viac začal venovať programovaniu i počítačovej lingvistike a od roku 2002 pôsobí na Jazykovednom ústave Ľudovíta Štúra SAV v Bratislave na oddelení Slovenského národného korpusu. Zaoberá sa predovšetkým korpusovou lingvistikou a počítačovým spracovaním prirodzeného jazyka. Ako člen kolektívu získal v roku 2005 významné ocenenie Cena Slovenskej akadémie vied za budovanie infraštruktúry pre vedu [1].

Obr.1: Pozvánka na prednášku Dr. Garabíka [https://cestina.phil.muni.cz/aktualne/kalendar-akci-a-udalosti/radovan-garabik-word-embedding]

Word Embedding

Slovné spojenie „word embedding“ pochádza z angličtiny a voľne by sme ho mohli preložiť ako „zapúzdrenie slov“ príp. v češtine „vnoření slov“. Je to metóda používaná v oblasti počítačového spracovania jazyka (NLP – Natural Language Processing), pomocou ktorej možno priradiť vektor ku každému slovu z dostatočne rozsiahleho (konečného) súboru lingvistických dát. Tieto lingvistické dáta sú texty, ktoré nemusia byť anotované, postačí „rozumný výber“ textov a „dobrá“ tokenizácia týchto dát. Na týchto dátach je následne model neurónovej siete natrénovaný.

Vektory priradené k jednotlivým slovám generujú vektorový priestor. Veľkosť, či dimenzia vektorového priestoru sa obyčajne pohybuje rádovo v stovkách či tisíckach. V prezentovanom modeli sa jednalo o 200-rozmerný vektorový priestor. Súradnice vektorov sú usporiadané n-tice reálnych čísel. V našom prípade teda každý vektor obsahuje 200 reálnych čísel.

Jednotlivé bázové vektory (osi) nášho vektorového priestoru zodpovedajú sémantickej kategórii alebo kombinácii aspoň dvoch sémantických kategórií. Treba však zdôrazniť, že nie je to celkom presný popis skutočnosti a nejde o nejaké nami vybrané a pomenované sémantické kategórie. Skôr sa jedná o kategórie, ktoré vznikajú automaticky na základe nastaveného modelu a vyjadrujú kontextové vzťahy medzi slovami v danom súbore textov.

Využité algoritmy v modeli spadajú do softvérového balíčka word2vec [3], ktoré fungujú na princípe extrakcie spomenutých „sémantických kategórií“ a na princípe učenia sa neurónových sietí. Stručne dodajme, že word2vec ponúka dva algoritmy. Je to CBOW (continuous bag-of-words) a skip-gram. Algoritmus CBOW sa snaží odvodiť aktuálne slovo od slov okolitých, zatiaľ čo skip--gram pracuje procesom opačným, teda snaží sa odhadnúť kontext daného slova. Podrobnejšie technické či matematické konštrukcie týkajúce sa metódy word embedding a algoritmov word2vec, sú zrejme vysoko nad rámec prednášky a tejto správy, preto môžem záujemcom len odporučiť text tu [3] alebo tu [4].

Konceptuálne tak len uveďme, že v prípade oboch algoritmov sú slová incializované ako náhodné vektory. Tieto vektory sa postupne upravujú, aby maximalizovali podmienenú pravdepodobnosť výskytu slovo <–> kontext. Čísla súradníc sa tzv. učením sa neurónovej siete postupne upravujú a v závislosti na kvalite a istej vyváženosti našich dát, dostaneme lepšie výsledky. Pod lepšími výsledkami budeme mať na mysli presnejšie vyjadrené sémantické podobnosti medzi slovami.

Keďže využívame presne zadefinované matematické pojmy z oblasti lineárnej algebry a geometrie, nič nám tiež nebráni vyjadriť podobnosť slov matematicky presným vzťahom. A teda podobnosť slov budeme vnímať ako podobnosť vektorov. A na podobnosť vektorov budeme nahliadať ako na

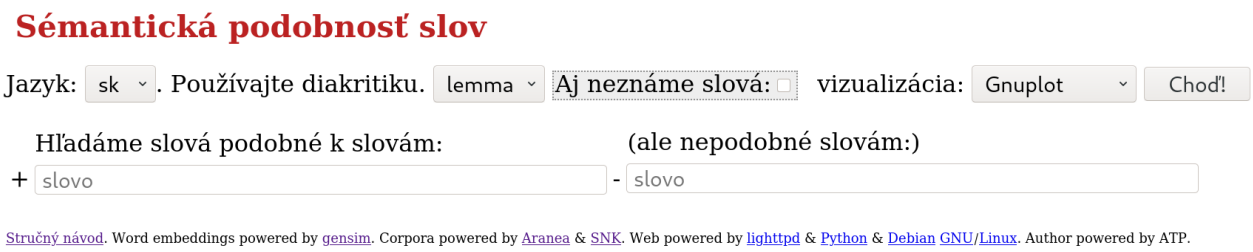
veľkosť uhla, ktorý vektory v a w spolu zvierajú. Tento vzťah môžeme teda vyjadriť ako:

cos(v, w) = v.w / |v|.|w|

pričom *v* a *w* sú vektory, *v.w* je skalárny súčin vektorov *v* a *w*, veľkosť vektora *v*, resp. *w* značíme ako |v|, resp. |w|. Súčin veľkostí dvoch vektorov *v* a *w* značíme |v|.|w|.

Sémantická podobnosť slov a užívateľské rozhranie

Na webe [2] bolo sprístupnené užívateľské rozhranie, pomocou ktorého môžeme vyhľadávať sémanticky podobné slová.

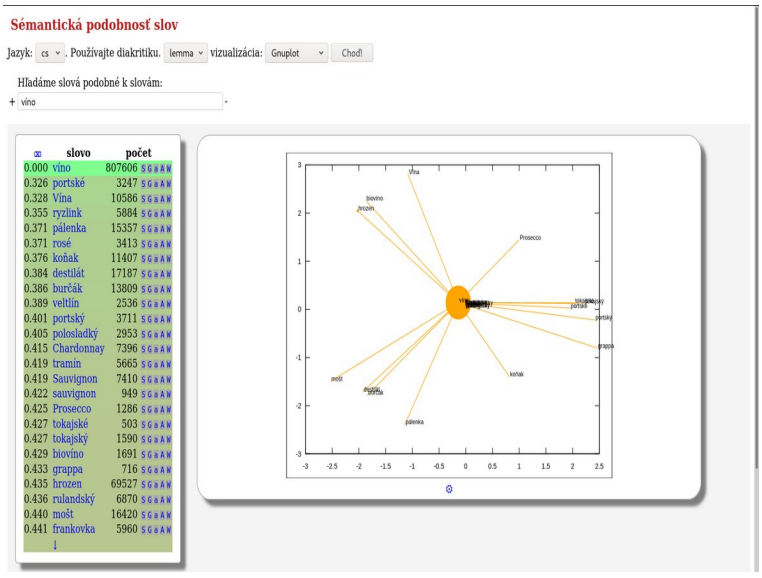


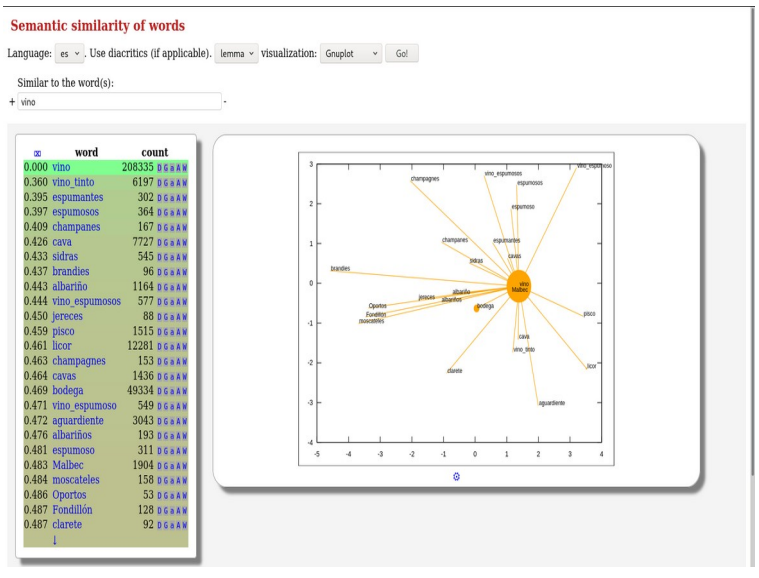
Obr.2: Užívateľské rozhranie na vyhľadávanie podobných slov v slovenčine

Implementáciu algoritmov z word2vec obsahuje tiež knižnica Gensim (v programovacom jazyku Python) a bola využitá pri budovaní tohto webového rozhrania a jeho modelu.

Na natrénovanie modelu boli využité webové korpusy z projektu Aranea [5]. V užívateľskom rozhraní tak možno vyhľadávať slová z viac ako 20 jazykov. Súčasťou rozhrania je i vizualizácia daných hľadaných vzťahov.

Príklady sémanticky najpodobnejších slov napríklad k slovu „víno“ vo vybraných jazykoch:



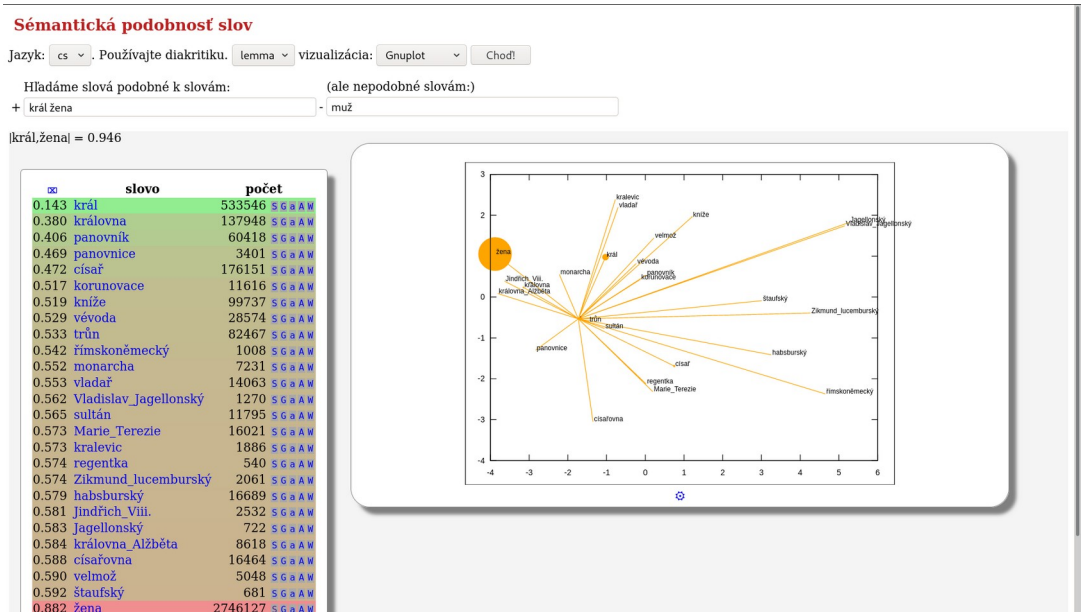


Obr.3 a 4: „Víno“ v češtine a španielčine

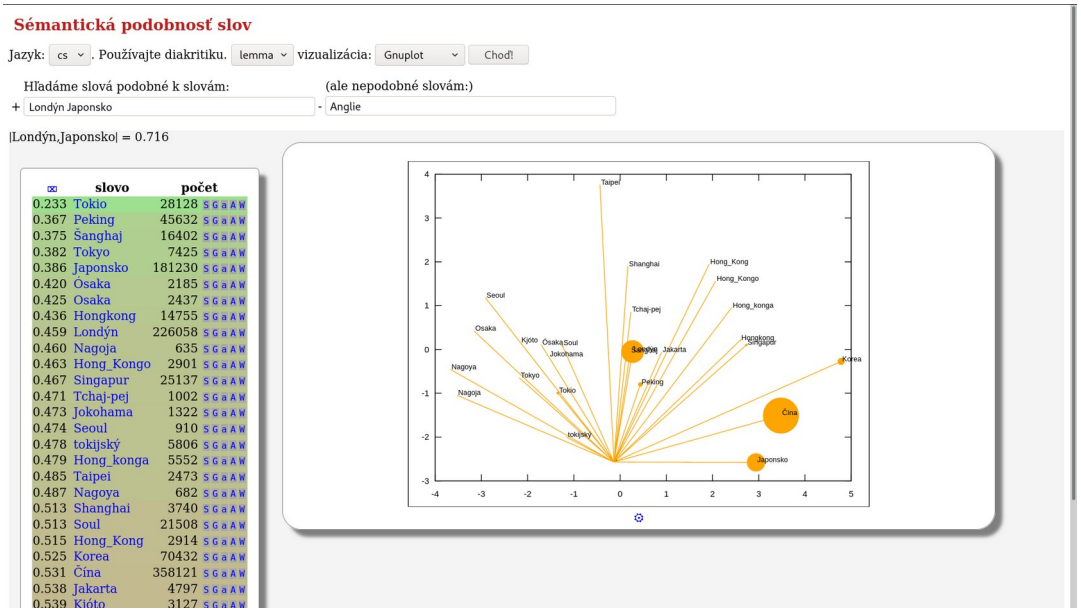
Vektorová aritmetika

Keďže pracujeme so slovami ako s vektormi z mnohorozmerného vektorového priestoru, nič nám nebráni v použití jednoduchých vektorových operácií – súčet či rozdiel ľubovoľných vektorov, teda slov.

Niekoľko zaujímavých výsledkov:



Obr.5: král + žena – muž = královna (očakávaný výsledok), v tomto prípade na druhom mieste



Obr.6: Londýn + Japonsko – Anglie = Tokio

Záverečné zhodnotenie

Metóda word embedding je jedným z ďalších pozoruhodných prepojení medzi lingvistikou a počítačovým spracovaním jazyka. Táto metóda by mohla byť nápomocná v lexikografickom výskume alebo napríklad pri analýze sentimentu.

Záverom treba tiež dodať, že Dr. Garabík zvládol prednášku na veľmi vysokej pedagogickej úrovni. Podarilo sa mu hovoriť zrozumiteľne o pomerne komplikovaných, matematicky zložitejších konceptoch.

Prednáška mi tiež pomohla lepšie pochopiť, že dimenzia vektorového priestoru sa nerovná počtu sémantických kategórií v pravom slova zmysle, že tieto sémantické kategórie nevyberá tvorca modelu, ale že tieto „sémantické kategórie“ sa vyabstrahujú automaticky na základe kontextu jednotlivých slov z dát, ktoré na natrénovanie modelu použijeme.

V neposlednom rade, prednáška opäť vo mne podnietila fascináciu matematickými pojmami a ich obecnosťou, vďaka ktorej tieto matematické pojmy a teórie možno aplikovať v tak rozličných oblastiach, ako je napr. fyzika, počítačová grafika, či v našom prípade lingvistika.

Zdroje

[1] Garabík, Radovan: P*rofil pracovníka*. Slovenský národný korpus JÚĽŠ SAV, Bratislava. [https://korpus.sk/rgarabik.html]

[2] Garabík, Radovan: *Sémantická podobnosť slov*. Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied v Bratislave, 2017-2020. [https://www.juls.savba.sk/semä.html]

[3] Mikolov, Tomas; Chen, Kai ; Corrado , Greg ; Dean , Jeffrey: *Efficient Estimation of Word Representations in Vector Space*. Proceedings of Workshop at ICLR. Université de Montreal. Scottsdale, 2013. [https://arxiv.org/pdf/1301.3781.pdf]

[4] Rong, Xin: *word2vec Parameter Learning Explained*, 2016. [https://arxiv.org/abs/1411.2738]

[5] Benko, Vladimír: Aranea: Yet Another Family of (Comparable) Web Corpora. Text, Speech and Dialogue. 17th International Conference, TSD 2014. Eds. Sojka, Petra et al. Springer International Publishing Switzerland. Brno. 257–264.