

Extrakcia kľúčových slov v slovenčine

Adam Rafajdus

Vedúci: Márius Šajgalík

20.10. 2015

Problém a motivácia

- Problém = Extrakcia kľúčových slov
- Motivácia – Existencia efektívnej metódy na vyhodnocovanie kvality kľúčových slov, ktorá je postavená na kategorizácii dokumentov, z ktorých sú slová extrahované
- Tým pádom, nemusím ručne vyhodnocovať kvalitu extrahovaných kľúčových slov cez dotazníky či iné pomalé kvalitatívne testy
 - > viem to urobiť kvantitatívne
- **Vieme sa učiť s učiteľom**

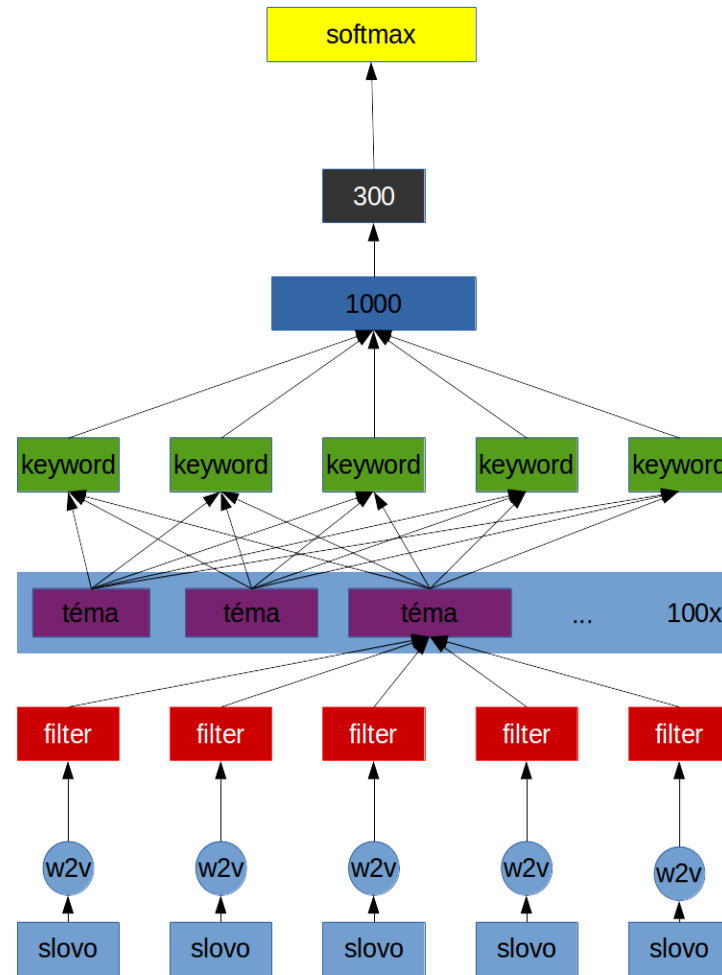
Cieľ

- **Navrhnuť novú architektúru neurónovej siete**
 - úspešne natrénovať na úlohu kategorizácie dokumentov
 - na základe malého počtu slov (5-10)
 - úzke hrdlo

Riešenie problému

- Dôrazná architektúra neurónovej siete
- Žiaden monolitický blackbox
- Modularizácia – hlavný bod záujmu
- Vyhodnocovanie – softmax

Experimentálny návrh



Dáta na spracovanie

- Zameriavame sa na slovenčinu ako jazyk
- Datasetsy SNK a SkWiki
- Nie som však jazykovo obmedzený
- Existuje veľa dostupných datasetov pre anglický jazyk, na ktorých môžem vykonávať dodatočné porovnávacie testy
- Porovnávať budem s LDA – populárny model tém

Zhrnutie

- Návrh novej architektúry neurónovej siete
- Kvantitatívne vyhodnotenie -> učenie sa z dát

Zdroje

- <https://code.google.com/p/word2vec/>
- Efficient estimation of word representations in vector space - Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean - 2013/1/16
- <http://torch.ch/>