

Model distribúcie interpunkčných znamienok v rôznych textoch

Martin Hošek, 2mAIN

Školiteľka: doc. RNDr. Mária Markošová, PhD.

Cieľ diplomovej práce

- ▶ analýza distribúcie interpunkčných znamienok v rôznych textoch od rovnakého autora
 - ▶ ako prvého autora na testovanie sme vybrali autora Erich Maria Remarque a knihu All Quiet on the Western Front (Na západe nič nové) v angličtine
 - ▶ jeho druhá kniha: The Road Back (Cesta späť) v angličtine
- ▶ porovnanie výsledkov pre rôznych autorov a rôzne jazyky (zistenie „zvyklostí“ pre daný jazyk)
- ▶ vytvorenie matematického modelu distribúcie týchto interpunkčných znamienok, analýza grafov susedností, porovnanie textu bez týchto znamienok, ...

Dynamika sietí

- ▶ v bežnom svete existuje veľa vecí, ktoré sú v nejakých vzťahoch. Tieto vzťahy sa dajú zapísať pomocou sietí
- ▶ matematický model siete je **graf**
- ▶ uzly (objekty, vrcholy) a vzťahy medzi nimi (hrany) sa v čase prirodzene menia, pričom ich počet spravidla narastá, zatiaľ čo ich zánik je len minimálny
- ▶ rozširovanie (rast) sietí sa deje podľa rôznych modelov:
 - ▶ náhodné pripájanie
 - ▶ preferenčné pripájanie
 - ▶ zmiešané pripájanie, ...

Kulig a kol. - hlavný článok práce

- ▶ podľa tohto článku sa všetky nepodstatné interpunkčné znamienka (úvodzovky, zátvorky) ignorujú a bodky, čiarky, ... sa považujú za samostatné slovo a zároveň za jednoznačný oddeľovač novej vety
- ▶ Zipfova analýza, Zipfov zákon:
 - ▶ štatistické pozorovanie o frekvencii výskytu prvkov (najčastejšie slov) v jazyku
 - ▶ **frekvencia slova je nepriamo úmerná jeho poradovému číslu v zozname najčastejších slov**
 - ▶ inými slovami:
 - najčastejšie slovo v texte sa vyskytuje približne **dvakrát častejšie** ako druhé najčastejšie
 - **trikrát častejšie** ako tretie najčastejšie
 - **n-krát častejšie** ako n-té slovo
 - ▶ slová ktoré slúžia ako „lepidlo“, teda funkčné slová, sú v tomto rebríčku najvyššie
 - ▶ slová ktoré majú hlavne sémantický význam, sú v rebríčku nižšie
 - ▶ je vidno, že v grafe ozaj sú tieto slová najvyššie

Kulig a kol.

- ▶ **Heapsov zákon** (známy aj ako **Heapsova krivka**):
 - ▶ opisuje, ako rastie počet rôznych slov (**slovná zásoba**) v texte s rastúcim počtom všetkých slov (**tokeny**) - tento zákon dopĺňa **Zipfov zákon** a často sa využíva v lingvistike a spracovaní prirodzeného jazyka
 - ▶ počet rôznych slov $V(n)$ v texte s n slovami rastie približne podľa mocninovej funkcie $V(n) = K * n^\beta$, kde $V(n)$ je počet rôznych (unikátnych) slov pri všetkých slovách, ktorých je n , K je konštanta závislá o typu jazyka a typu textu a β je koeficient, typicky v rozsahu 0.4 – 0.6
 - ▶ porovnanie rastu dvojíc slov (zatiaľ iba pre prvú knihu)
- ▶ **záver**:
 - ▶ interpunkčné (základné) znamienka sa môžu používať v analýze ako štandardné slová

Knižnice, ktoré som zatiaľ využil

- ▶ programujem v jazyku Java (využívam programovacie prostredie IntelliJ)
- ▶ knižnica PDFBox (<https://pdfbox.apache.org/download.html>), využívam verziu 3.0.4
- ▶ knižnica umožňuje načítať text, ktorý sa nachádza v súbore, ktorý je vo formáte .pdf
- ▶ využívam čítanie čistého textu (t.j. po jednotlivých znakoch), tento typ textu (pred analýzou) sa zvykne označovať aj ako „rawText“.
- ▶ využil som knižnicu na vyrábanie grafov Xchart (<https://pdfbox.apache.org/download.html>), využívam verziu 3.8.8

Drozd et al. Quantifying principles of the narrative text formation

článok, ktorý hovorí o tom, čo sa spomína aj v prvej a hlavnej literatúre Kulig)

- ▶ interpunkčné znamienka sa môžu považovať za štandardné slová
- ▶ interpunkcia sa neobjavuje náhodne
- ▶ Interpunkčné znamienka určujú dĺžku a rytmus vety

Moja implementácia

- ▶ analýza vstupného súboru vo formáte .PDF (konkrétne EMR: All Quiet on the Western Front a The Road Back)
- ▶ nastavenie čítania od konkrétnej strany (preskočenie úvodných častí, obsah a pod.) => využívanie všeobecného konštruktora
- ▶ preskakovanie opakujúcich sa sprievodných textov (čísla strán, názvy kapitol, ...) aj v ďalších textoch
- ▶ rozobratie (analýza textu na jednotlivé typy (tokeny): slovo, bodka, dvojbodka, úvodzovka, bodkočiarka, otáznik, výkričník, ... => vyradené nepotrebné znaky
- ▶ spočítanie počtu výskytu jednotlivých slov (mapa slovo-počet), urobenie prehľadných výpisov, vyrobenie grafu frekvencie prvých n slov, zatiaľ nastavené na n=50.
- ▶ úprava počítadla slov, bez vyradených tokenov
- ▶ vyrobenie grafu susednosti (mapa „<slovo1>_<slovo2>-počet) a jeho prehľadný výpis => porovnanie na jednej knihe s heapsovým zákonom

Písanie DP

- ▶ rozpísaná teória k prvým dvom článkom
- ▶ popísaná implementácia SW
 - ▶ hlavné triedy a metódy
 - ▶ využívané knižnice
- ▶ zhruba analýza prvej knihy

Použitá a prečítaná literatúra

1. Dynamika sietí. Vladimír Kvasnička, Jozef Pospíchal, Jiří Navrátil, Bohumil Lacko, and Peter Trebatický, editors, Umelá inteligencia a kognitívna veda II, strany 321-377. STU, 2010.
2. Kulig et al. In narrative texts punctuation marks obey the same statistics as words. Information Sciences, 375:98-113, Január 2017.
3. Drozd et al. Quantifying principles of the narrative text formation, CoRR, abs/1412.8319, 2014
4. All Quiet on the Western Front, Erich Maria Remarque
https://www.glscott.org/uploads/2/1/3/3/21330938/aqwf-book_20size.pdf
5. The Road Back, Erich Maria Remarque
<https://www.fadedpage.com/showbook.php?pid=20241109>