Hodnota textových dát

Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV
3. slovenský ELRC workshop

2021-05-25, Bratislava

Veľkosti korpusov

- 1 milión – morfologická/syntaktická anotácia
- 10 miliónov – terminológia
- 100 miliónov – vreckový slovník (korpus textov zákonov 43M, slovenská wikipédia 50M)
- 1 miliarda – veľký slovník, word embeddings, špičkové jazykové modely (SNK 1.6G)
- 10 miliárd – webový korpus (ARANEA)
- 100 miliárd – anglické webové korpusy

Veľkosti korpusov

- 1 milión – morfologická/syntaktická anotácia
- 10 miliónov – terminológia
- 100 miliónov – vreckový slovník (korpus textov zákonov 43M, slovenská wikipédia 50M)
- 1 miliarda – veľký slovník, word embeddings, špičkové jazykové modely (SNK 1.6G)
- 10 miliárd – webový korpus (ARANEA)
- 100 miliárd – anglické webové korpusy
- koľko slov človek „spracuje“ za celý život?

Veľkosti korpusov

- 1 milión – morfologická/syntaktická anotácia
- 10 miliónov – terminológia
- 100 miliónov – vreckový slovník (korpus textov zákonov 43M, slovenská wikipédia 50M)
- 1 miliarda – veľký slovník, word embeddings, špičkové jazykové modely (SNK 1.6G)
- 10 miliárd – webový korpus (ARANEA)
- 100 miliárd – anglické webové korpusy
- koľko slov človek „spracuje“ za celý život?
- 400 miliónov?

Dostupnosť korpusov

- Big Data
- okrem výnimiek a starých textov nie sú dostupné...

Dostupnosť korpusov

- Big Data
- okrem výnimiek a starých textov nie sú dostupné...
- GFDL, CC

Dostupnosť korpusov

- Big Data
- okrem výnimiek a starých textov nie sú dostupné...
- GFDL, CC
- výnimky zo zákona (právne predpisy, technické normy, náboženské obrady)

Dostupnosť korpusov

- Big Data
- okrem výnimiek a starých textov nie sú dostupné...
- GFDL, CC
- výnimky zo zákona (právne predpisy, technické normy, náboženské obrady)
- “the Google way”

Korpus textov zákonov

- texty zákonov Slovenskej republiky
- 43 miliónov slov
- 1955 – 2020
- `https://www.juls.savba.sk/legalcorp.html`

korpusy ARANEA

- rodina porovnateľných webových korpusov
- veľké
- `http://aranea.juls.savba.sk/`

korpusy ARANEA

- rodina porovnateľných webových korpusov
- veľké
- `http://aranea.juls.savba.sk/`
- voľne dostupné po registrácii

Word Embedding (vektorová reprezentácia slov)

- uhádnuť slovo na základe kontextu
- uhádnuť kontext na základe slova

Word Embedding (vektorová reprezentácia slov)

- uhádnuť slovo na základe kontextu
- uhádnuť kontext na základe slova
- hidden layer je veľká matica, ktorá násobí vstup (vektor) $\rightarrow$ výstup (vektor)
- word2vec
- neodmysliteľný základ počítačového spracovania jazyka
- zachytáva „významové“ rozdiely medzi slovami – uhol medzi vektormi
  - $\vec{A} \cdot \vec{B} = |\vec{A}||\vec{B}| \cos\theta$
  - $\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$

Implementácia

- `https://www.juls.savba.sk/semä.html`

Implementácia

- `https://www.juls.savba.sk/semä.html`
- Aranea korpusy + SNK
- linky do korpusov, Google, slovníkov, Wiktionary, iné...
- vizualizácia

Použitie

- tezaurus
- významovo blízke slová
- nazeráme do myšlienkového modelu „priemerného človeka“

Použitie

- tezaurus
- významovo blízke slová
- nazeráme do myšlienkového modelu „priemerného človeka“
- neovplyvnené lingvistickými teóriami

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$
- $|\text{kľudne,pokojne}| = 0.392$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$
- $|\text{kľudne,pokojne}| = 0.392$
- $|\text{parfum,parfém}| = 0.192$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$
- $|\text{kľudne,pokojne}| = 0.392$
- $|\text{parfum,parfém}| = 0.192$
- $|\text{Clinton,Bush}| = 0.191$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$
- $|\text{kľudne,pokojne}| = 0.392$
- $|\text{parfum,parfém}| = 0.192$
- $|\text{Clinton,Bush}| = 0.191$
- $|\text{Lotyšsko,Litva}| = 0.107$

Sémantická blízkosť slov

- $|\text{vodič,šofér}| = 0.323$
- $|\text{vodička,šoférka}| = 0.549$
- $|\text{kľudne,pokojne}| = 0.392$
- $|\text{parfum,parfém}| = 0.192$
- $|\text{Clinton,Bush}| = 0.191$
- $|\text{Lotyšsko,Litva}| = 0.107$
- $|\text{debil,idiot}| = 0.173$

Vektorová aritmetika

- operácie $+$ a $-$

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie
- kráľ - muž $+$ žena

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie
- kráľ - muž $+$ žena
- premiér - muž $+$ žena

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie
- kráľ - muž $+$ žena
- premiér - muž $+$ žena
- pes - on $+$ ona

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie
- kráľ - muž $+$ žena
- premiér - muž $+$ žena
- pes - on $+$ ona
- letuška - žena $+$ muž

Vektorová aritmetika

- operácie + a −
- vyváženie
- kráľ - muž + žena
- premiér - muž + žena
- pes - on + ona
- letuška - žena + muž
- doctor - man + woman

Vektorová aritmetika

- operácie $+$ a $-$
- vyváženie
- kráľ - muž $+$ žena
- premiér - muž $+$ žena
- pes - on $+$ ona
- letuška - žena $+$ muž
- doctor - man $+$ woman
- 007 - 009

Vektorová aritmetika

- operácie + a −
- vyváženie
- kráľ - muž + žena
- premiér - muž + žena
- pes - on + ona
- letuška - žena + muž
- doctor - man + woman
- 007 - 009
- 666 - 555
- Bratislava - Slovensko + Francúzsko
- Paris - France + Slovaquie

Ďakujem za pozornosť