Vector Space Models

Kirill Semenov

NPFL103, December 2022

Architecture: Brief Information

- Language: Python
- Main data structures:
  - For collecting data (from documents, queries): dict
  - For all computations: sparse matrices
- Libraries:
  - Preprocessing, tokenization: re, string, nltk, ufal.morphodita
  - Computations: scipy, numpy

Experiments

- Approximate time: 3-6 minutes per run (English ≈1.5 times slower than Czech)
- Overall number of experiments: 15
- Parameters for experiment:
  - Tokenization (+lemmatization)
  - String preprocessing
  - Idf weighting
  - Tf weighting
  - Query construction
  - ~~Vector normalization~~

Experiments: Tokenization, Lemmatization, Preprocessing

Tokenization: default (spaces+punctuation); nltk; morphodita (tokenization + lemmatization); nltk+clear stopwords; nltk+clear punctuation

Preprocessing: case-insensitive

run	explanation	map	
		en	cs
run-0	**all default values**	**0.0781**	**0.1459**
	preprocessing: case-insensitive	0.1714	0.1879
	tokenization: nltk	0.1029	0.1461
run-1	**tokenization: nltk; preprocessing: case-insensitive**	**0.1984**	**0.1882**
	tokenization(+lemmatization): morphodita; preprocessing: case_insensitive	0.0859	0.0518
	tokenization(+lemmatization): morphodita	0.0456	0.0761
	tokenization+preprocessing: clear_stopwords	0.1027	0.1462
	tokenization: nltk; preprocessing: clear punctuation	0.1028	0.1459
	tokenization: nltk; preprocessing: clear punctuation, case_insensitive	0.1714	0.1879

Experiments: TF, IDF Weighting

TF weighting: default (logarithm), boolean, augmented

IDF weighting: default (idf), none, prob idf

run	explanation	map	
		en	cs
run-0	**all default values**	**0.0781**	**0.1459**
	idf weighting: none	0.0689	0.124
	idf weighting: prob	0.0782	0.1459
	tf weighting: bool	0.0627	0.0962
	tf weighting: augmented	0.0628	0.0968

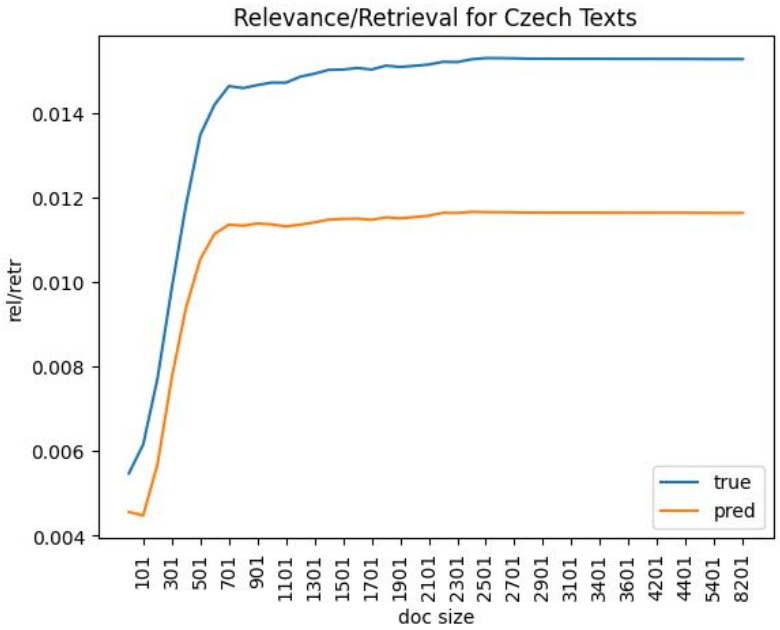
Experiments: Query Construction

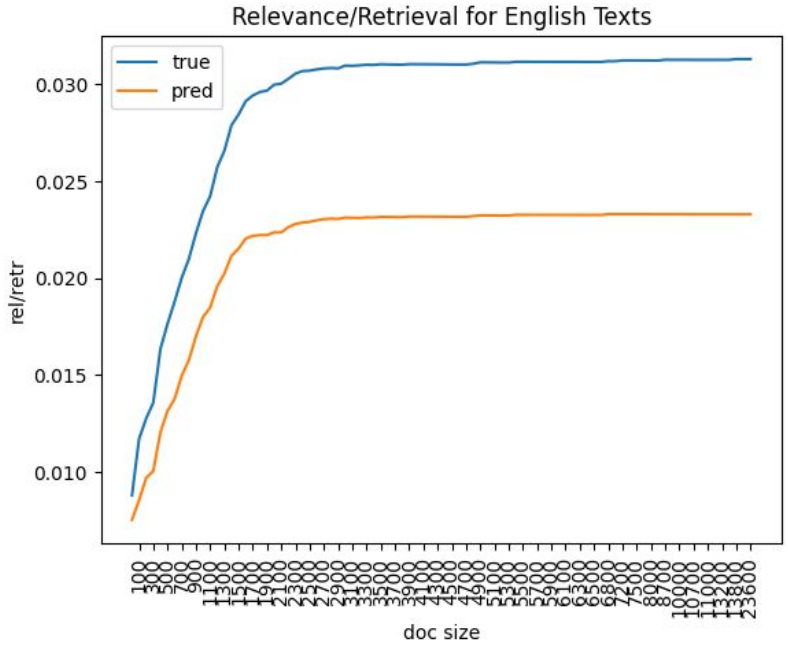
Query Construction: default (only titles), enhanced (titles + desc), enhanced_2 (titles + desc + narr)

run	explanation	map	
		en	cs
run-0	**all default values**	**0.0781**	**0.1459**
run-2	**query construction: enhanced; tokenization: nltk; preprocessing: case-insensitive**	0.2641	**0.229**
	query construction: enhanced_2; tokenization: nltk; preprocessing: case-insensitive	**0.2848**	0.2262

Experiments (Abandoned): Normalization

Pivot normalization: no intersection between the ground truth relevance and my results (best system)





Results

- Morphodita worse than nltk
  - Nltk more important for English: apostrophes?
- Filtering stop words does not help
- TF, IDF ablations do worse
- Description of title does worse (for Czech) - probably because too wordy and “misleads” the vector

run	explanation	map	
		en	cs
run-0	**all default values**	**0.0781**	**0.1459**
run-1	**tokenization: nltk; preprocessing: case-insensitive**	**0.1984**	**0.1882**
run-2	**query construction: enhanced; tokenization: nltk; preprocessing: case-insensitive**	0.2641	**0.229**

Thank you for your attention!