Vector Space Models with PyTerrier

David Burian

January 2, 2023

Choice of IR framework

- Xapian
- Lucene
- Anserini (Pyserini)
- Solr
- ElasticSearch

Baselines

- **run-0**
  - *tokenizer*: regex tokenizer spliting at any of the following characters `-,.;;?!␣\t\n[]()'"`
  - *stop words*: -
  - *stemmer*: -
  - *weight model*: term frequency
- **run-0-tfidf**
  - *tokenizer*: regex tokenizer spliting at any of the following characters `-,.;;?!␣\t\n[]()'"`
  - *stop words*: -
  - *stemmer*: -
  - *weight model*: Robertson's TF-IDF

Baselines results

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0	0.0433	0.0680	0.1011	0.1360
run-0-tfidf	0.1905	0.1800	0.2359	0.2720

Table: Results of **run-0** and **run-0-tfidf**.

Tokenizers

- **run-0-pyterrier-tok**:
  - builtin PyTerrier tokenizers; 'english' for English dataset, 'utf' for Czech dataset
- **run-0-nltk-tok**:
  - nltk's tokenizers; 'english' for English dataset, 'czech' for Czech dataset

Tokenizers results

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0-pyterrier-tok	0.2540	0.2560	0.3472	0.3960
run-0-nltk-tok	0.1893	0.1800	0.2110	0.2560

Table: Results of **run-0-pyterrier-tok** and **run-0-nltk-tok**.

Stopwords

- **run-0-pyterrier-stop**:
  - builtin PyTerrier's stopwords list for English dataset only
- **run-0-nltk-stop**:
  - nltk English stopword list
- **run-0-kaggle-stop**:
  - Kaggle Czech stopword list

Stopwords results

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0-pyterrier-stop	-	-	0.3443	0.4080
run-0-nltk-stop	-	-	0.3442	0.4080
run-0-kaggle-stop	0.2576	0.2640	-	-

Table: Results of **run-0-pyterrier-stop**, **run-0-nltk-stop** and **run-0-kaggle-stop**.

Stemming\Lemmatization

- **run-0-porter-stemm**:
  - PyTerrier's Porter stemmer
- **run-0-snowball-stemm**:
  - PyTerrier's Snowball stemmer
- **run-0-udpipe-lemm**:
  - Lemmatization using UDPipe 2 with the 'czech' model
- **run-0-czech-stemm**:
  - Czech stemmer implemented by Prof. Jacques Savoy, 'light' version

Stemming\Lemmatization results

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0-porter-stemm	0.2573	0.2600	0.3929	0.4240
run-0-snowball-stemm	0.2576	0.2640	0.3960	0.4240
run-0-udpipe-lemm	0.0000	0.0000	-	-
run-0-czech-stemm	0.3117	0.3240	0.3430	0.3800

Table: Results of **run-0-porter-stemm**, **run-0-snowball-stem**, **run-0-udpipe-lemm** and **run-0-czech-stemm**.

Weighting models

- **run-0-tfidf-pivoted**:
  - TF-IDF with Pivoted length normalization
- **run-0-tfidf-pivoted-robertson**:
  - TF-IDF with Pivoted Robertson's normalization
- **run-0-bm25**:
  - BM25
- **run-0-pl2**:
  - PL2
- **run-0-lemur-tfidf**:
  - Lemur's version of TF-IDF

Weighting models results

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0-tfidf-pivoted-robertson	0.3121	0.3320	0.3925	0.4200
run-0-tfidf-pivoted	0.2799	0.2840	0.3738	0.4160
run-0-pl2	0.2901	0.3160	0.3821	0.3960
run-0-bm25	0.3082	0.3240	0.3806	0.3760
run-0-lemur-tfidf	0.2947	0.3000	0.3877	0.4280

Table: Results of **run-0-tfidf-pivoted**, **run-0-tfidf-pivoted-robertson**, **run-0-bm25**, **run-0-pl2** and **run-0-lemur-tfidf**.

Query expansion

- **run-2**:
  - divergence from Randomness query expansion model using PyTerrier's built in 'Bo1' model

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-2	0.3548	0.3520	0.3983	0.4360

Table: Results of **run-2**.

Best run

- **run-2**
  - *tokenizer*: builtin PyTerrier tokenizers; 'english' for English dataset, 'utf' for Czech dataset
  - *stop words*: Kaggle Czech stopword list for Czech data only
  - *stemmer*: Czech stemmer implemented by Prof. Jacques Savoy, 'light' version for Czech; Snowball stemmer for English
  - *weight model*: Robertson's TF-IDF
  - *query expansion*: divergence from Randomness query expansion model using PyTerrier's built in 'Bo1' model

	Czech		English	
run ID	MAP	P@10	MAP	P@10
run-0	0.0433	0.0680	0.1011	0.1360
run-2	0.3548	0.3520	0.3983	0.4360

Table: Results of **run-0** and **run-2**.

Thank you for your attention.