

Vector Space Models

David Burian

6.12. 2022

Implementation

- ▶ Python
- ▶ `use multiprocessing.pool.Pool.imap`
 - ▶ 120s for Czech
 - ▶ 320s for English

run-0

run ID	Czech				English			
	MAP	#relevant	#rel. & ret.	MRR	MAP	#relevant	#rel. & ret.	MRR
run-0	0.1114	382	153	0.3571	0.1228	782	296	0.3217

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

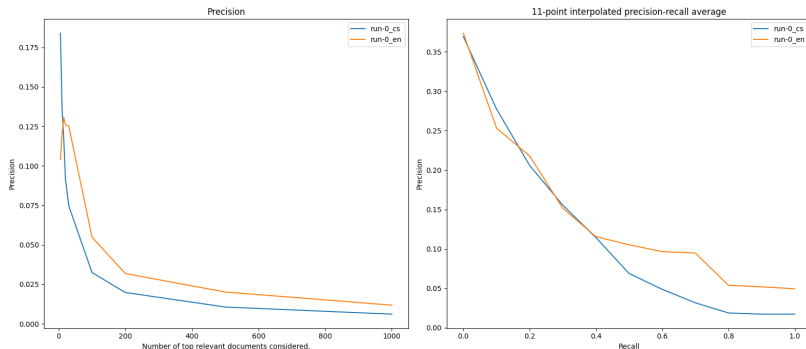


Figure: Precision and 11-point interpolated precision-recall curves for **run-0**.

run-0-tfidf

run ID	Czech				English			
	MAP	#relevant	#rel. & ret.	MRR	MAP	#relevant	#rel. & ret.	MRR
run-0	0.1114	382	153	0.3571	0.1228	782	296	0.3217
run-0-tfidf	0.1169	382	161	0.3989	0.1309	782	312	0.3856

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

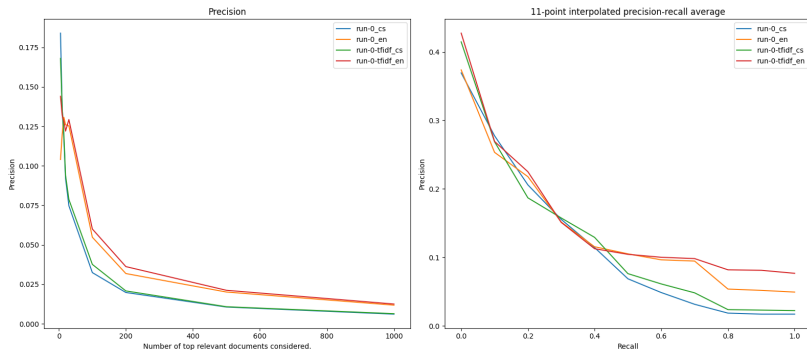


Figure: Precision and 11-point interpolated precision-recall curves for **run-0-tfidf** and **run-0**.

run-0-sep

run ID	Czech				English			
	MAP	#relevant	#rel. & ret.	MRR	MAP	#relevant	#rel. & ret.	MRR
run-0-tfidf	0.1169	382	161	0.3989	0.1309	782	312	0.3856
run-0-sep-par	0.1166	382	159	0.3971	0.1316	782	315	0.3856
run-0-sep-punc	0.1108	382	156	0.3890	0.1063	782	303	0.3222
run-0-sep-quot-par	0.1202	382	162	0.3965	0.1345	782	327	0.3583
run-0-sep-quot	0.1202	382	159	0.3983	0.1336	782	325	0.3583

Table: Results of run-0-sep-punc, run-0-sep-quot, run-0-seps-par, run-0-seps-quot-par and run-0-tfidf.

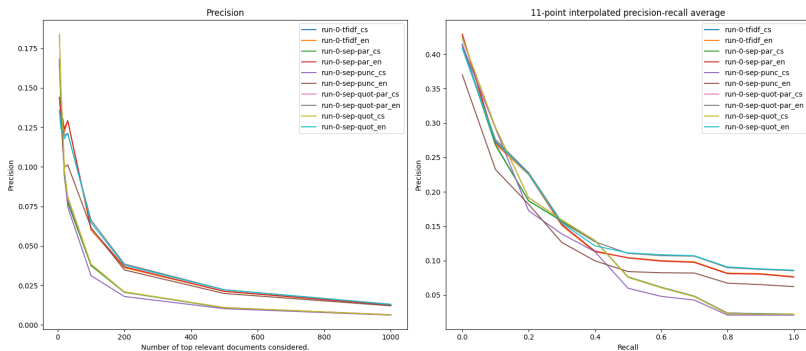


Figure: Precision and 11-point interpolated precision-recall curves for run-0-sep-punc, run-0-sep-quot, run-0-seps-par, run-0-seps-quot-par and run-0-tfidf.

run-0-stopwords

run id	#Czech stop words	#English stop words
run-0-stopwords-kaggle	256	1298
run-0-stopwords-200	34	71
run-0-stopwords-600	8	21

Table: Stop word sets considered for given runs.

run-0-stopwords

run ID	Czech				English			
	MAP	#relevant	#rel. & ret.	MRR	MAP	#relevant	#rel. & ret.	MRR
run-0-sep-quot-par	0.1202	382	162	0.3965	0.1345	782	327	0.3583
run-0-stopwords-200	0.1225	382	168	0.4229	0.1340	782	320	0.3470
run-0-stopwords-600	0.1227	382	168	0.4232	0.1340	782	320	0.3470
run-0-stopwords-kaggle	0.1132	382	162	0.3758	0.1370	782	327	0.3505

Table: Results of **run-0-stopwords-kaggle**, **run-0-stopwords-200** and **run-0-stopwords-600**.

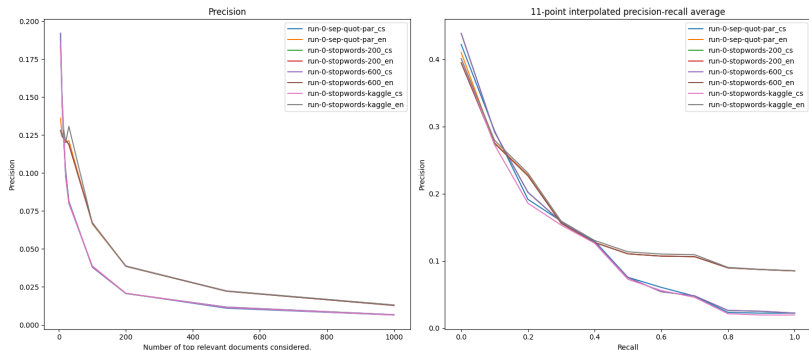


Figure: Precision and 11-point interpolated precision-recall curves for **run-0-stopwords-kaggle**, **run-0-stopwords-200** and **run-0-stopwords-600**.

run-0-tagblacklist

Left out tags:

- ▶ for Czech documents: DOCNO, DOCID,
- ▶ for English documents: DOCNO, DOCID, SN, PD, PN, PG, PP, WD, SM, SL, CB, IN, FN.

run-0-tagblacklist

run ID	Czech				English			
	MAP	#relevant	#rel. & ret.	MRR	MAP	#relevant	#rel. & ret.	MRR
run-0-stopwords-600	0.1227	382	168	0.4232	-	-	-	-
run-0-stopwords-kaggle	-	-	-	-	0.1370	782	327	0.3505
run-0-tagblacklist	0.1227	382	168	0.4232	0.1370	782	325	0.3505

Table: Results of **run-0-stopwords-600** on Czech, **run-0-stopwords-kaggle** on English and **run-0-tagblacklist** on both datasets.

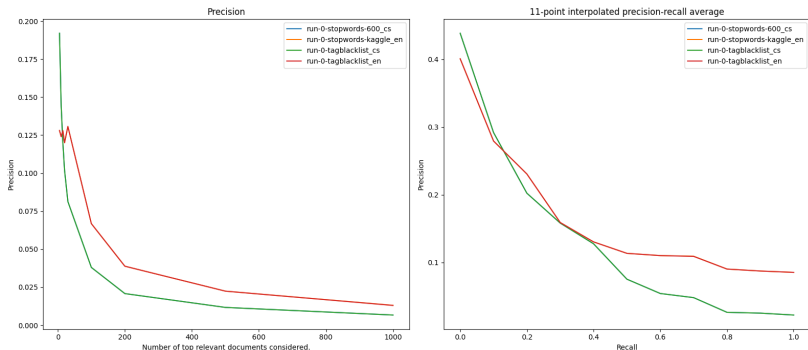


Figure: Precision and 11-point interpolated precision-recall curves for **run-0-stopwords-600** on Czech, **run-0-stopwords-kaggle** on English and **run-0-tagblacklist** on both datasets.

The end

Thank you for your attention