Validation of Text Similarity Methods

Matching of Grant Proposals to Reviewers: PostDoc Mobility 2021

Gabriel Okasa & Anne Jorstad

Validation Summary

We validate several TF-IDF models with various hyperparameter settings and data inputs for the task of semantic textual similarity.

We do not impose any restrictions for the similarity search and restrict the applications and referees' texts to English texts only. We keep only referees that have at least English 10 publications.

For matching the referees, we take into account the similarity average of 20 percent most similar publication of a given referee.

The research area distribution for the validated applications is as follows:

- LS: 123 (0.38)
- MINT: 147 (0.46)
- SSH: 50 (0.16)

We validate the following models:

- tfidf

and extract the text embeddings using:

- 3_gram; uni_gram

for the following type of texts:

- abstract; title; title_abstract

based on the following years of publications:

- 5; 10

For each validation scenario, we extract the text embeddings and compute the similarity between the applications and each publication of a referee based on the cosine similarity.

We measure the performance of the methods based on the mean average precision at $K = 2$ and $K = 5$.

Validation Results

Table 1: Validation results: Mean Average Precision

model	embedding	years	text	map_at_2	map_at_5
tfidf	3_gram	10	title_abstract	0.3094	0.4033
tfidf	3_gram	10	abstract	0.3023	0.4000
tfidf	3_gram	5	abstract	0.3117	0.3932
tfidf	3_gram	5	title_abstract	0.3102	0.3925
tfidf	uni_gram	10	abstract	0.3008	0.3900
tfidf	uni_gram	10	title_abstract	0.2906	0.3811
tfidf	uni_gram	5	abstract	0.3078	0.3792
tfidf	uni_gram	5	title_abstract	0.2914	0.3692
tfidf	3_gram	10	title	0.2055	0.2585
tfidf	uni_gram	10	title	0.1906	0.2504
tfidf	3_gram	5	title	0.1883	0.2316
tfidf	uni_gram	5	title	0.1797	0.2305

Validation Results by Research Area

Table 2: Validation results: Mean Average Precision by Research Area: LS

model	embedding	years	text	map_at_2	map_at_5
tfidf	3_gram	5	title_abstract	0.3008	0.3794
tfidf	3_gram	5	abstract	0.3008	0.3770
tfidf	uni_gram	5	abstract	0.3069	0.3752
tfidf	3_gram	10	title_abstract	0.2785	0.3752
tfidf	uni_gram	5	title_abstract	0.2907	0.3713
tfidf	3_gram	10	abstract	0.2622	0.3686
tfidf	uni_gram	10	title_abstract	0.2703	0.3622
tfidf	uni_gram	10	abstract	0.2663	0.3619
tfidf	3_gram	10	title	0.2175	0.2675
tfidf	uni_gram	10	title	0.2053	0.2558
tfidf	uni_gram	5	title	0.2012	0.2496
tfidf	3_gram	5	title	0.2114	0.2487

Table 3: Validation results: Mean Average Precision by Research Area: MINT

model	embedding	years	text	map_at_2	map_at_5
tfidf	3_gram	10	abstract	0.3656	0.4497
tfidf	3_gram	10	title_abstract	0.3605	0.4428
tfidf	uni_gram	10	abstract	0.3571	0.4329
tfidf	3_gram	5	abstract	0.3316	0.4204
tfidf	uni_gram	10	title_abstract	0.3350	0.4197
tfidf	3_gram	5	title_abstract	0.3265	0.4145
tfidf	uni_gram	5	abstract	0.3350	0.4078
tfidf	uni_gram	5	title_abstract	0.3146	0.3935
tfidf	3_gram	10	title	0.2177	0.2762
tfidf	uni_gram	10	title	0.2092	0.2752
tfidf	uni_gram	5	title	0.1939	0.2505
tfidf	3_gram	5	title	0.1956	0.2448

Table 4: Validation results: Mean Average Precision by Research Area: SSH

model	embedding	years	text	map_at_2	map_at_5
tfidf	3_gram	5	title_abstract	0.285	0.3603
tfidf	3_gram	10	title_abstract	0.235	0.3562
tfidf	3_gram	5	abstract	0.280	0.3533
tfidf	uni_gram	10	abstract	0.220	0.3328
tfidf	3_gram	10	abstract	0.215	0.3312
tfidf	uni_gram	10	title_abstract	0.210	0.3140
tfidf	uni_gram	5	abstract	0.230	0.3048
tfidf	uni_gram	5	title_abstract	0.225	0.2925
tfidf	3_gram	10	title	0.140	0.1840
tfidf	uni_gram	10	title	0.100	0.1643
tfidf	3_gram	5	title	0.110	0.1512
tfidf	uni_gram	5	title	0.085	0.1245