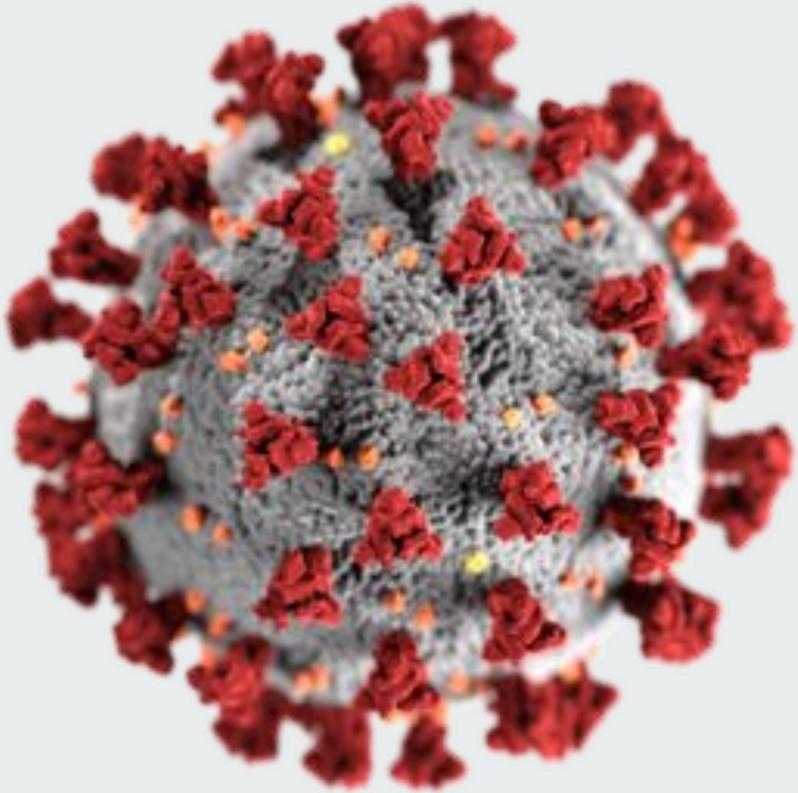
CORD-19

Information Retrieval 2021-2022

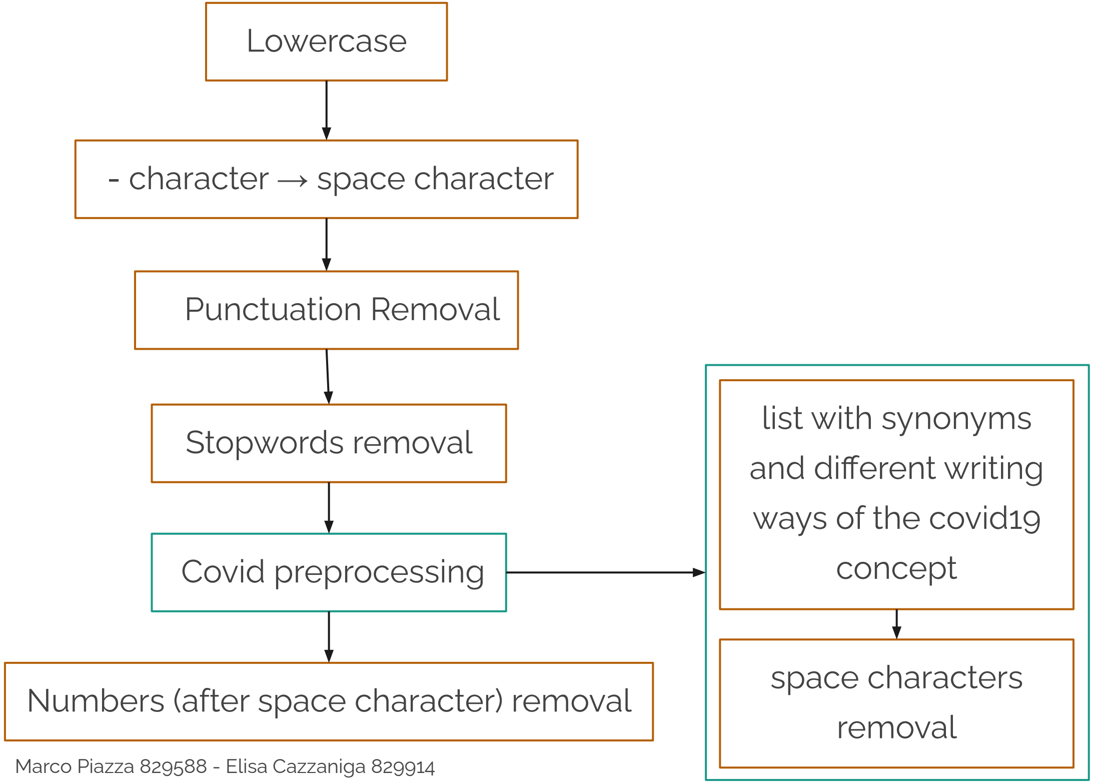


Marco Piazza 829588 - Elisa Cazzaniga 829914

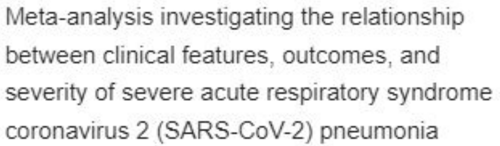
Project structure

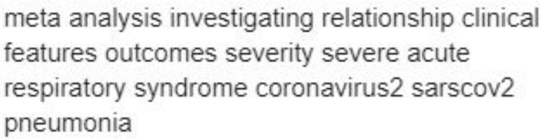
- First part
  - Analysis documents and queries.
  - Comparison between whole collection and documents related to covid19
- Second part
  - Analysis of the performance of *pyterrier* using different types of index configurations, pre-processing steps and models
- Third part:
  - Improvement of the effectiveness

Analysis of Queries and Documents

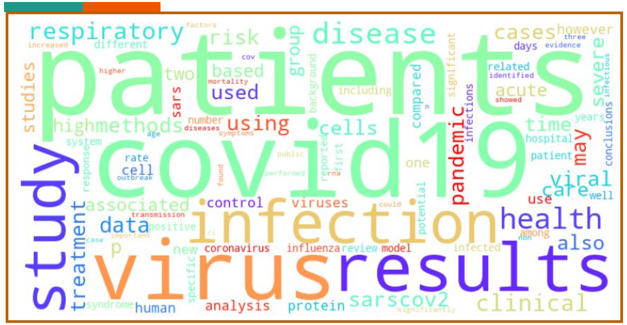

Lowercase
- character → space character
Punctuation Removal
Stopwords removal
Covid preprocessing
Numbers (after space character) removal
list with synonyms and different writing ways of the covid19 concept
space characters removal

Example:

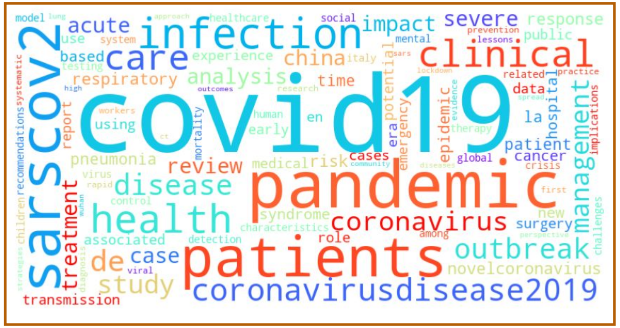

Meta-analysis investigating the relationship between clinical features, outcomes, and severity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pneumonia


meta analysis investigating relationship clinical features outcomes severity severe acute respiratory syndrome coronavirus2 sarscov2 pneumonia

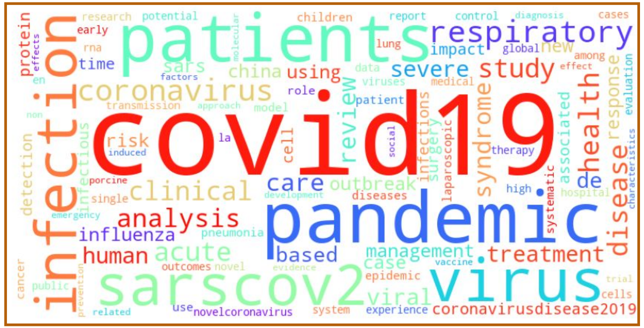
Documents



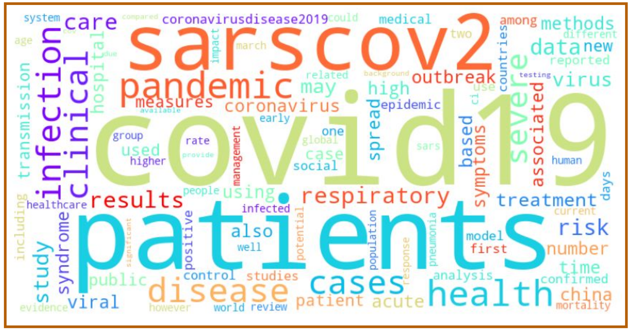
Most common words titles of covid19 documents



Most common words texts of covid19 documents

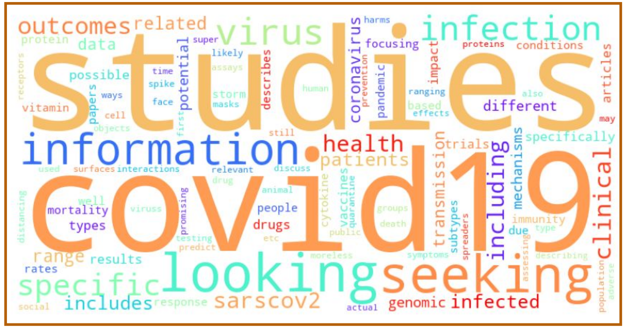


Most common words titles of general documents

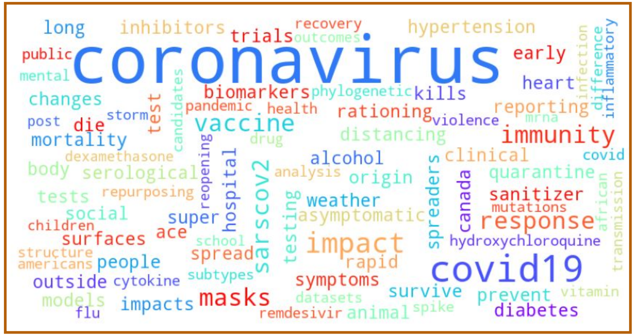


Most common words texts of general documents

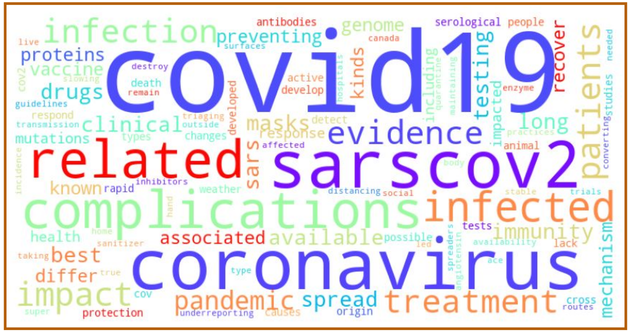
Queries



Most common words ad-hoc queries



Most common words desc queries



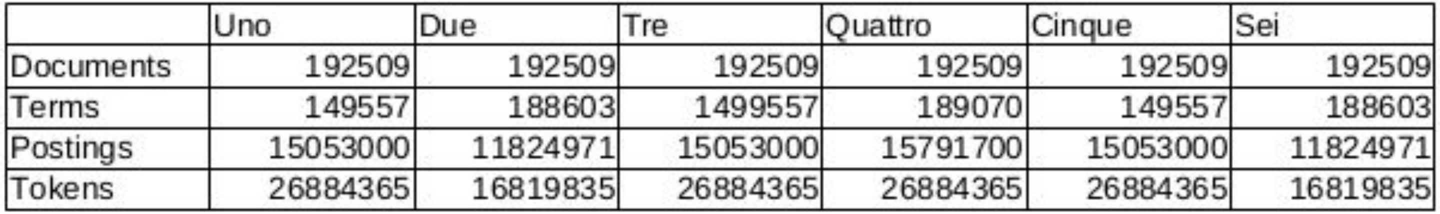
Most common words nar queries

Search Engines - Basic Search

- Different indexes - preprocessing approaches
- Different language models
- Performances of same queries

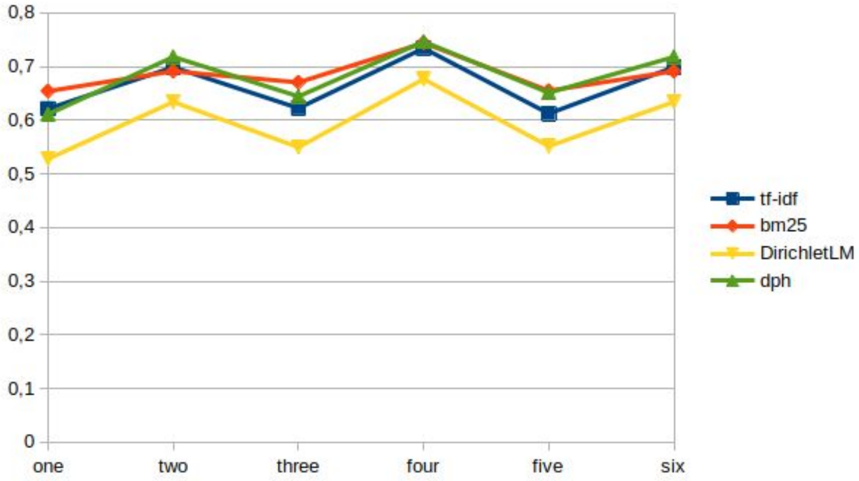
Different preprocessing steps

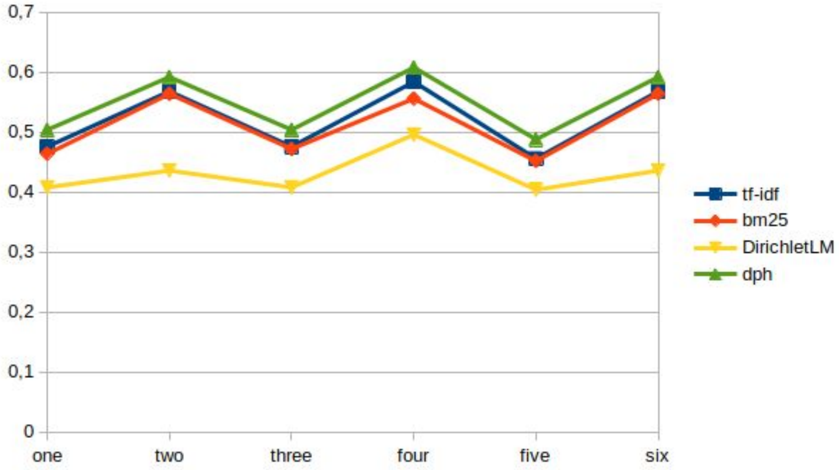
- One → stopwords removal and Porter stemmer for both documents and queries.
- Two → Stopwords removal for both categories.
- Three → Porter stemmer for both categories.
- Four → Neither stopwords removal nor stemmer.
- Five → Stopwords removal on queries and Porter Stemmer on documents.
- Six → Stopwords removal only on documents.



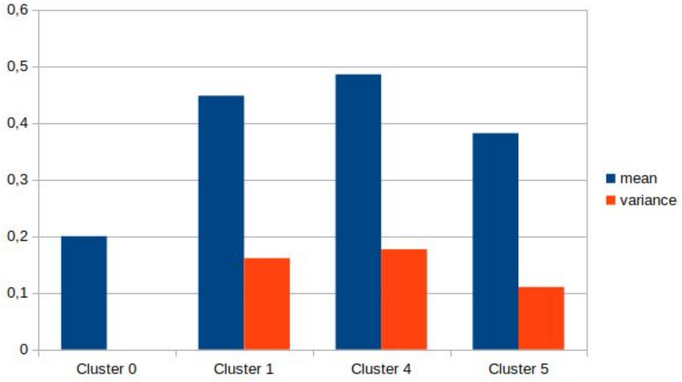
	Uno	Due	Tre	Quattro	Cinque	Sei
Documents	192509	192509	192509	192509	192509	192509
Terms	149557	188603	1499557	189070	149557	188603
Postings	15053000	11824971	15053000	15791700	15053000	11824971
Tokens	26884365	16819835	26884365	26884365	26884365	16819835

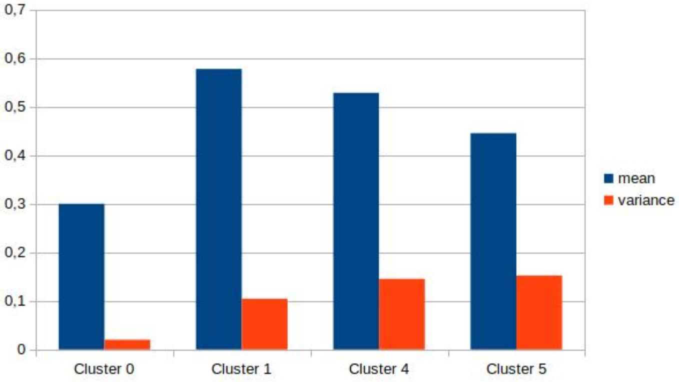
Different models

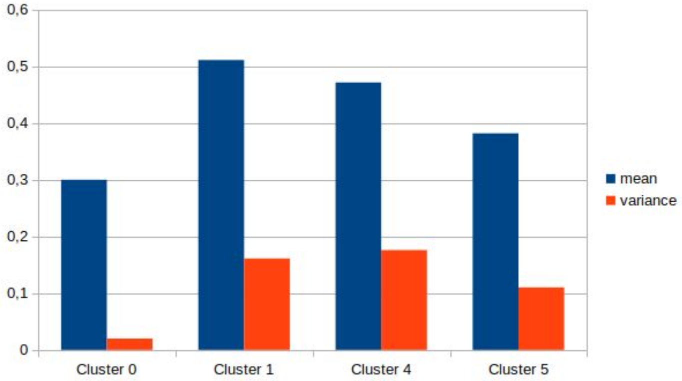


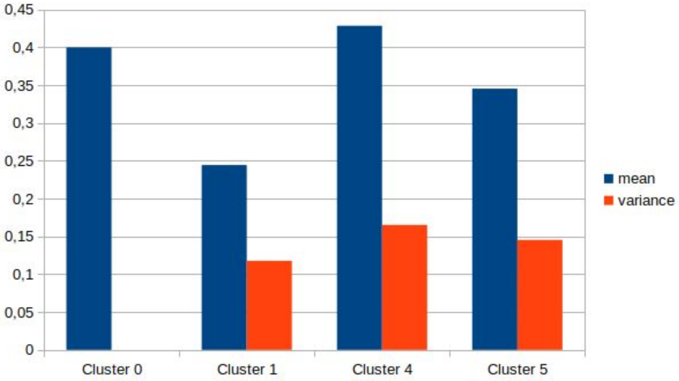


Same queries performance







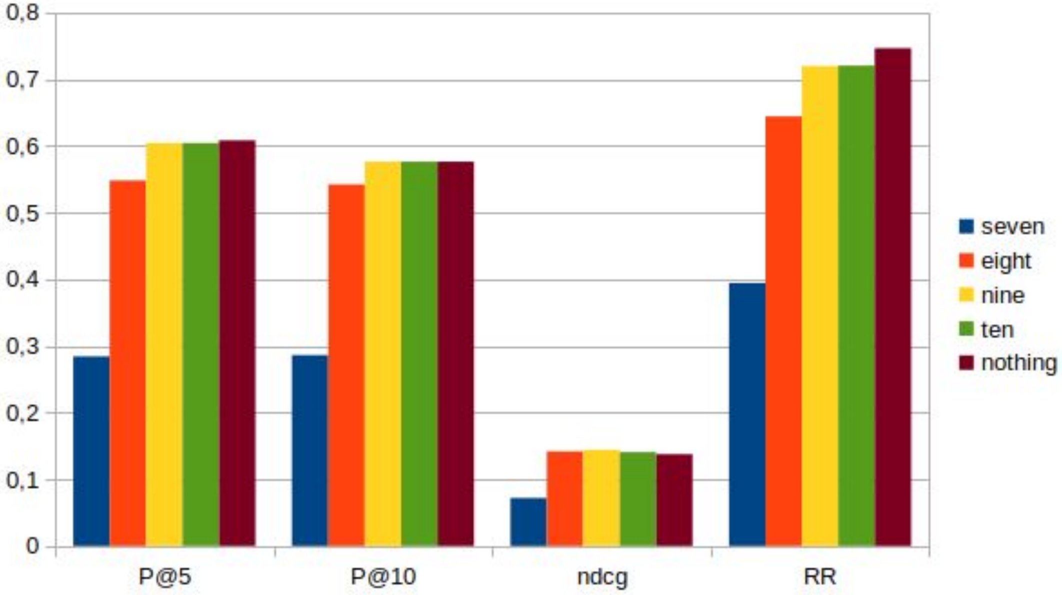


Search Engine - Advanced Search

- Improvement of the effectiveness:
  - Queries expansion
  - Queries reduction
  - Re-ranking by date
  - Re-ranking using neural approaches

Query expansion

- Seven → Glove Gigaword
- Eight → RM3 with BERT reranking
- Nine → RM3
- Ten → RM3 with linear combination

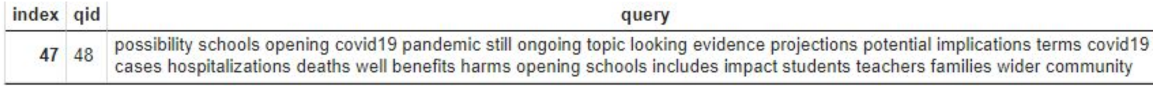


Query reduction

Example of query reduction:

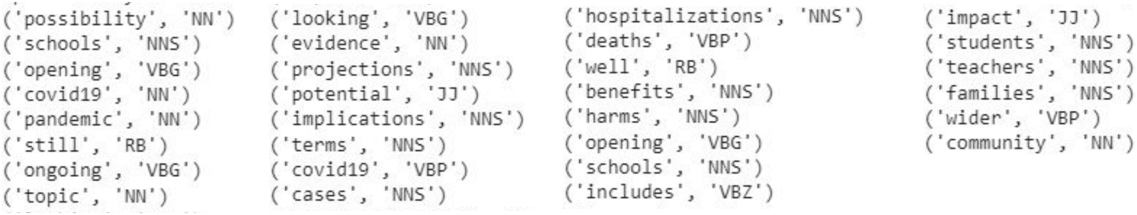
(Narrative)

Original query:



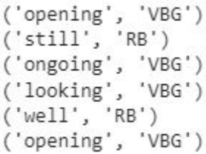
index	qid	query
47	48	possibility schools opening covid19 pandemic still ongoing topic looking evidence projections potential implications terms covid19 cases hospitalizations deaths well benefits harms opening schools includes impact students teachers families wider community

TAGs:



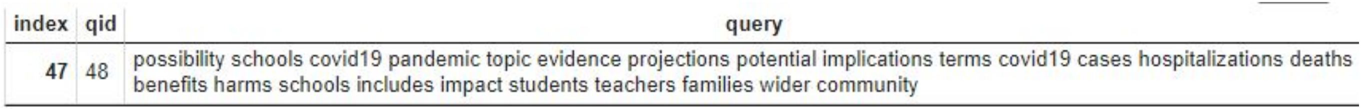
```
('possibility', 'NN')
('schools', 'NNS')
('opening', 'VBG')
('covid19', 'NN')
('pandemic', 'NN')
('still', 'RB')
('ongoing', 'VBG')
('topic', 'NN')
('looking', 'VBG')
('evidence', 'NN')
('projections', 'NNS')
('potential', 'JJ')
('implications', 'NNS')
('terms', 'NNS')
('covid19', 'VBP')
('cases', 'NNS')
('hospitalizations', 'NNS')
('deaths', 'VBP')
('well', 'RB')
('benefits', 'NNS')
('harms', 'NNS')
('opening', 'VBG')
('schools', 'NNS')
('includes', 'VBZ')
('impact', 'JJ')
('students', 'NNS')
('teachers', 'NNS')
('families', 'NNS')
('wider', 'VBP')
('community', 'NN')
```

TAGs that are removed:



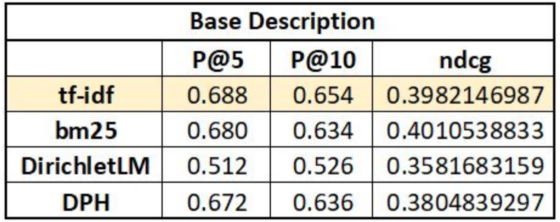
```
('opening', 'VBG')
('still', 'RB')
('ongoing', 'VBG')
('looking', 'VBG')
('well', 'RB')
('opening', 'VBG')
```

Final query:

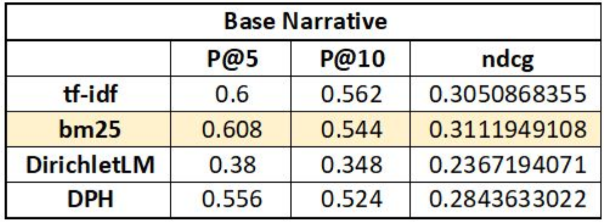


index	qid	query
47	48	possibility schools covid19 pandemic topic evidence projections potential implications terms covid19 cases hospitalizations deaths benefits harms schools includes impact students teachers families wider community

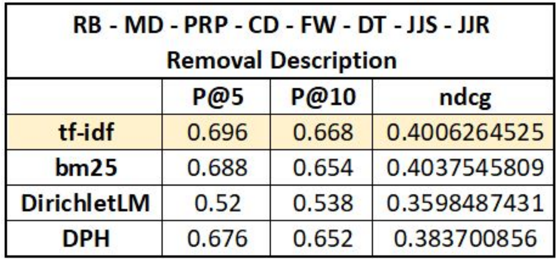
Query reduction - best results



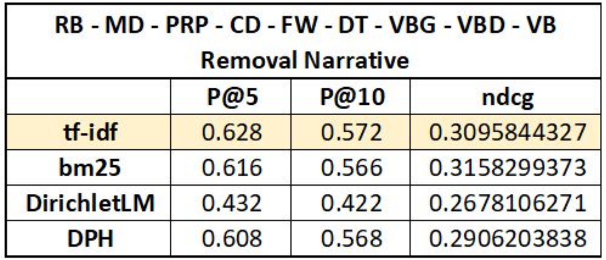
Base Description			
	P@5	**P@10**	**ndcg**
tf-idf	0.688	0.654	0.3982146987
bm25	0.680	0.634	0.4010538833
DirichletLM	0.512	0.526	0.3581683159
DPH	0.672	0.636	0.3804839297



Base Narrative			
	P@5	**P@10**	**ndcg**
tf-idf	0.6	0.562	0.3050868355
bm25	0.608	0.544	0.3111949108
DirichletLM	0.38	0.348	0.2367194071
DPH	0.556	0.524	0.2843633022



RB - MD - PRP - CD - FW - DT - JJS - JJR Removal Description			
	P@5	**P@10**	**ndcg**
tf-idf	0.696	0.668	0.4006264525
bm25	0.688	0.654	0.4037545809
DirichletLM	0.52	0.538	0.3598487431
DPH	0.676	0.652	0.383700856



RB - MD - PRP - CD - FW - DT - VBG - VBD - VB Removal Narrative			
	P@5	**P@10**	**ndcg**
tf-idf	0.628	0.572	0.3095844327
bm25	0.616	0.566	0.3158299373
DirichletLM	0.432	0.422	0.2678106271
DPH	0.608	0.568	0.2906203838

RB → adverb

MD → modal

PRP → personal pronoun

CD → cardinal digit

FW → foreign word

DT → determiner

JJS → adjective, superlative

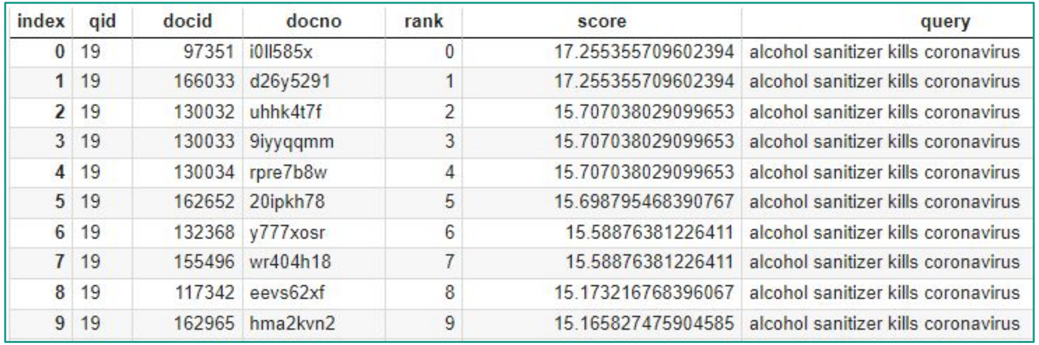
JJR → adjective, comparative

VBG → verb, gerund/present participle

VBD → verb, past tense

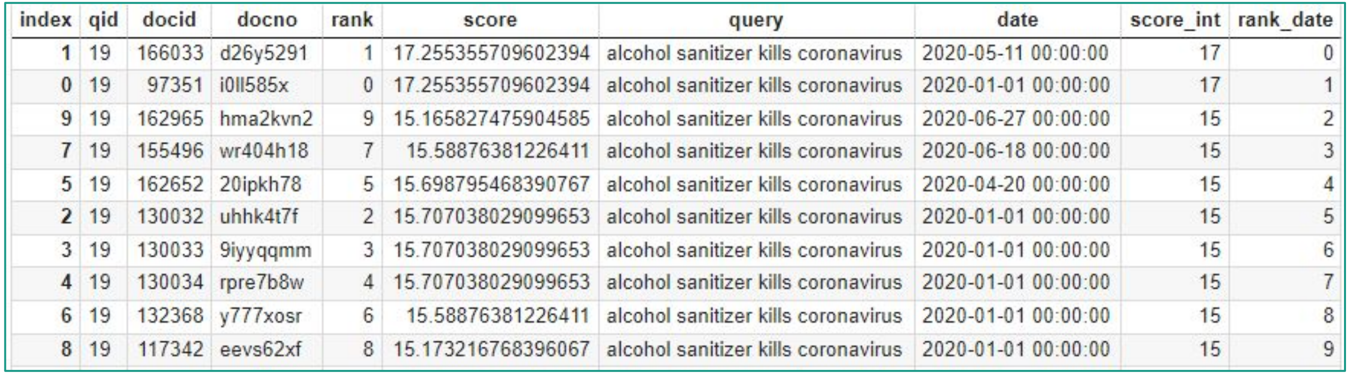
VB → verb, base form

Re-ranking documents by date



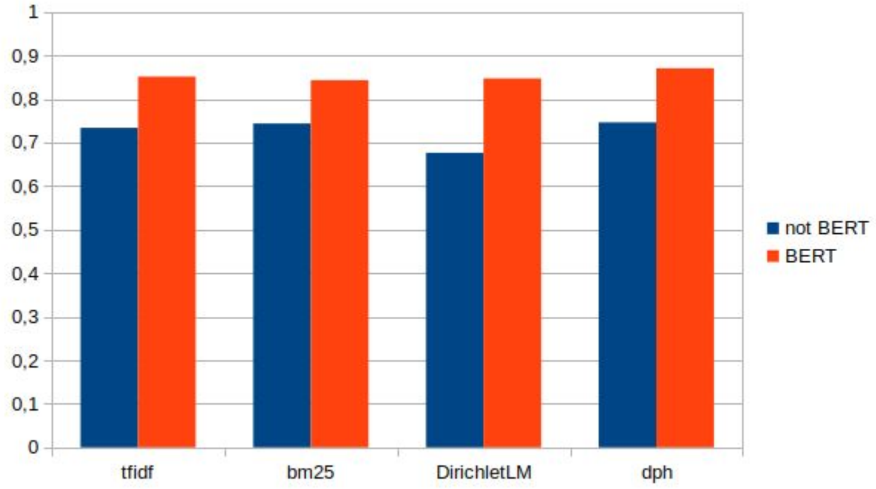
index	qid	docid	docno	rank	score	query
0	19	97351	i0ll585x	0	17.255355709602394	alcohol sanitizer kills coronavirus
1	19	166033	d26y5291	1	17.255355709602394	alcohol sanitizer kills coronavirus
2	19	130032	uhhk4t7f	2	15.707038029099653	alcohol sanitizer kills coronavirus
3	19	130033	9iyyqqmm	3	15.707038029099653	alcohol sanitizer kills coronavirus
4	19	130034	rpre7b8w	4	15.707038029099653	alcohol sanitizer kills coronavirus
5	19	162652	20ipkh78	5	15.698795468390767	alcohol sanitizer kills coronavirus
6	19	132368	y777xosr	6	15.58876381226411	alcohol sanitizer kills coronavirus
7	19	155496	wr404h18	7	15.58876381226411	alcohol sanitizer kills coronavirus
8	19	117342	eevs62xf	8	15.173216768396067	alcohol sanitizer kills coronavirus
9	19	162965	hma2kvn2	9	15.165827475904585	alcohol sanitizer kills coronavirus

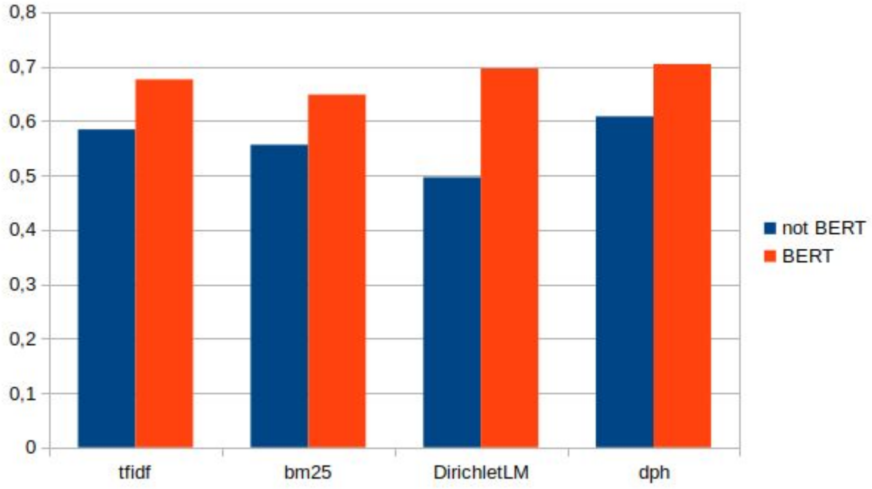
Grouping the documents by integer score and sort each group by descending date



index	qid	docid	docno	rank	score	query	date	score_int	rank_date
1	19	166033	d26y5291	1	17.255355709602394	alcohol sanitizer kills coronavirus	2020-05-11 00:00:00	17	0
0	19	97351	i0ll585x	0	17.255355709602394	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	17	1
9	19	162965	hma2kvn2	9	15.165827475904585	alcohol sanitizer kills coronavirus	2020-06-27 00:00:00	15	2
7	19	155496	wr404h18	7	15.58876381226411	alcohol sanitizer kills coronavirus	2020-06-18 00:00:00	15	3
5	19	162652	20ipkh78	5	15.698795468390767	alcohol sanitizer kills coronavirus	2020-04-20 00:00:00	15	4
2	19	130032	uhhk4t7f	2	15.707038029099653	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	15	5
3	19	130033	9iyyqqmm	3	15.707038029099653	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	15	6
4	19	130034	rpre7b8w	4	15.707038029099653	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	15	7
6	19	132368	y777xosr	6	15.58876381226411	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	15	8
8	19	117342	eevs62xf	8	15.173216768396067	alcohol sanitizer kills coronavirus	2020-01-01 00:00:00	15	9

Re-ranking using Neural Approach


tfidf
bm25
DirichletLM
dph
🟦 not BERT
🟥 BERT


tfidf
bm25
DirichletLM
dph
🟦 not BERT
🟥 BERT