Assignment 3 - Information Retrieval and Text Mining

Stefan Truong	3338287
Suhas Sangolli	3437641
Tushar Balihalli	3437638

TASK 1.

d1 : pens write on paper paper
d2 : pencils write on envelope
d3 : ballpens write on paper
q: “ballpens envelope”.

Subtask 1:

tf: #terms in doc d

tf_pens_d1 = 1	tf_pens_d2 = 0	tf_pens_d3 = 0
tf_write_d1 = 1	tf_write_d2 = 1	tf_write_d3 = 1
tf_on_d1 = 1	tf_on_d2 = 1	tf_on_d3 = 1
tf_paper_d1 = 2	tf_paper_d2 = 0	tf_paper_d3 = 1
tf_pencils_d1 = 0	tf_pencils_d2 = 1	tf_pencils_d3 = 0
tf_envelope_d1 = 0	tf_envelope_d2 = 1	tf_envelope_d3 = 0
tf_ballpens_d1 = 0	tf_ballpens_d2 = 0	tf_ballpens_d3 = 1

df: # docs that t occurs in

df_pens = 1
df_write = 3

df_on = 3
df_paper = 2
df_pencils = 1
df_envelope =1
df_ballpens = 1

idf: log 3/df

df_pens = 0.477
df_write = 0
df_on = 0
df_paper = 0.176
df_pencils = 0.477
df_envelope = 0.477
df_ballpens = 0.477

tf_weight: (1+log tf) // 0 if tf is zero and 10er log

w_pens_d1 = 1	w_pens_d2 = 0	w_pens_d3 = 0
w_write_d1 = 1	w_write_d2 = 1	w_write_d3 = 1
w_on_d1 = 1	w_on_d2 = 1	w_on_d3 = 1
w_paper_d1 = 1,30103	w_paper_d2 = 0	w_paper_d3 = 1
w_pencils_d1 = 0	w_pencils_d2 = 1	w_pencils_d3 = 0
w_envelope_d1 = 0	w_envelope_d2 = 1	w_envelope_d3 = 0
w_ballpens_d1 = 0	w_ballpens_d2 = 0	w_ballpens_d3 = 1

tf-idf-weights & corresponding document vector: (1+log tf) log N/df

	d1	d2	d3
pens	0.477	0	0
write	0	0	0
on	0	0	0
paper	0.22898	0	0.176
pencils	0	0.477	0
envelope	0	0.477	0
ballpens	0	0	0.477

query -vector

	q
pens	0
write	0
on	0
paper	0
pencils	0
envelope	0.477
ballpens	0.477

Subtask2:

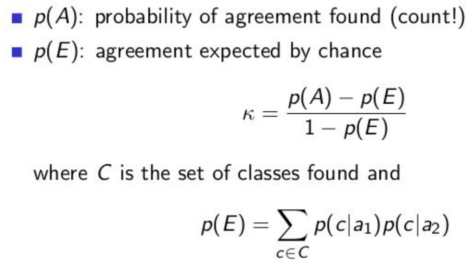
$$\cos(q,d) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q^2{}_i \sqrt{\sum_{i=1}^{|V|} d^2{}_i}}}$$

cos(d1,d2) = [(0.477 0 0 0.22898 0 0 0)^T (0 0 0 0 0.477 0.477 0)]/[(0.477² + 0.22898²)^(½)(0.477² + 0.477²)^(½)] = 0
cos(d1,d3) = $\frac{0.22898 \times 0.176}{\sqrt{0.477^2 + 0.22898^2}\ \sqrt{0.176^2 + 0.477^2}}$ *= 0.1498*
cos(d2,d3) = 0
cos(q,d1) = **0**
cos(q,d2) = $\frac{0.477^2}{\sqrt{0.477^2 + 0.477^2}\ \sqrt{0.477^2 + 0.477^2}}$ = **0.5**
cos(q,d3) = $\frac{0.477^2}{\sqrt{0.477^2 + 0.477^2}\ \sqrt{0.176^{22} + 0.477^2}}$ = **0.707**

Ranking according to cosine order: d3, d2, d1

TASK 2.

Cohen's Kappa: measure of consistency in the agreement among judges



- $p(A)$: probability of agreement found (count!)
- $p(E)$: agreement expected by chance

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)}$$

where C is the set of classes found and

$$p(E) = \sum_{c \in C} p(c|a_1)p(c|a_2)$$

Step0:

Query: "corona beschränkungen baden-württemberg"
Information need: I want to know what i am allowed/ not allowed to do during a corona crisis in baden württemberg

Step1&2:

Search Engine of duckduckgo (Bing)
1: Is Relevant
2: Not Relevant

Annotator Agreement	Stefan	Tushar	Suhas
https://www.baden-wuerttemberg.de/de/service/aktuelle-infos-zu-corona/	1	2	1
https://www.swr3.de/aktuell/nachrichten/neue-corona-massnahmen-in-baden-wuerttemberg-100.html	1	1	1
https://www.baden-wuerttemberg.de/de/service/aktuelle-infos-zu-corona/aktuelle-corona-verordnung-des-landes-baden-wuerttemberg/	1	1	1
https://www.bw24.de/stuttgart/corona-regeln-baden-wuerttemberg-weihnachten-silvester-beschraenkungen-massnahmen-kontakte-verbote-feiern-feiertage-90110226.html	1	1	1
https://www.stuttgarter-nachrichten.de/inhalt.corona-beschraenkungen-in-baden-wuerttemberg-nachts-darf-niemand-mehr-aus-dem-haus.457220ec-fe36-4203-8879-b1637cab11fe.html	2	2	2

https://www.tz.de/welt/corona-baden-wuerttemberg-lockdown-regeln-ausgangsbeschraenkungen-mannheim-heilbronn-zr-90120240.html	2	2	2
https://www.swr.de/swraktuell/baden-wuerttemberg/coronavirus-liveblog-bw-100.html	2	2	2
https://sozialministerium.baden-wuerttemberg.de/de/gesundheit-pflege/gesundheitsschutz/infektionsschutz-hygiene/informationen-zu-coronavirus/	2	1	2
https://www.zdf.de/nachrichten/politik/coronavirus-kontaktbeschraenkung-ausgangsbeschraenkung-bundeslaender-100.html	1	2	1
https://www.bundesregierung.de/breg-de/themen/coronavirus/corona-massnahmen-1734724	1	1	1
https://coronavirus.stuttgart.de/	1	2	1
https://www.badische-zeitung.de/ministerpraesident-kretschmann-informiert-ueber-die-neuen-corona-beschraenkungen--198813285.html	2	2	1
https://www.stuttgarter-zeitung.de/inhalt.neue-corona-beschraenkungen-ministerpraesident-kretschmann-nimmt-im-livestream-stellung.254c06c0-44df-4a71-a5d1-b307c15fb1f6.html	2	2	2
https://www.rnz.de/politik/suedwest_artikel,-corona-ticker-baden-wuerttemberg-schul-Oeffnung-am-11-januar-noch-unklar-fast-27500-menschen-geimpft-_arid,501540.html	2	2	2
https://www.swr.de/swraktuell/baden-wuerttemberg/polizei-bw-will-an-silvester-staerker-kontrollieren-100.html	2	2	2
https://www.karlsruhe-insider.de/news/achtung-strengere-corona-einschraenkungen-fuer-baden-wuerttemberg-61419/	2	2	2
https://www.stuttgarter-zeitung.de/inhalt.corona-beschraenkungen-baden-wuerttemberg-lockerungen-ab-9-juni-feiern-mit-unter-100-gaesten-erlaubt.2beeb1fc-7f5d-464c-a32b-a6d99e3a7935.html	2	2	2
https://www.bw24.de/baden-wuerttemberg/baden-wuerttemberg-ruft-hoechste-corona-warnstufe-aus-neue-einschraenkungen-kommen-90072605.html	2	2	2

https://www.rnz.de/politik/suedwest_artikel,-corona-ticker-baden-wuerttemberg-schul-Oeffnung-am-11-januar-noch-unklar-fast-27500-menschen-geimpft-_arid,501540.html	2	2	2
https://www.mannheim.de/de/corona	1	2	2

Step3:

Tushar/Stefan	Relevant	Not Relevant	Total
Relevant	4	1	5
Not Relevant	4	11	15
Total	8	12	20

P(A) = $\frac{4+11}{20} = \frac{3}{4}$
P(E) = $\frac{8}{20} * \frac{5}{20} + \frac{12}{20} * \frac{15}{20} = \frac{11}{20}$

K(Tushar, Stefan) = ($\frac{3}{4}$ - $\frac{11}{20}$)/ (1- $\frac{11}{20}$) = $\frac{4}{9}$ = 0.44

Suhas/Stefan	Relevant	Not Relevant	Total
Relevant	7	1	8
Not Relevant	1	11	12
Total	8	12	20

P(A) = $\frac{7+11}{20} = \frac{9}{10}$
P(E) = $\frac{8}{20} * \frac{8}{20} + \frac{12}{20} * \frac{12}{20} = \frac{13}{25}$

K(Suhas, Stefan) = ($\frac{9}{10}$ - $\frac{13}{25}$)/ (1- $\frac{13}{25}$) = $\frac{19}{24}$ = 0.79

Suhas/Tushar	Relevant	Not Relevant	Total
Relevant	4	4	8

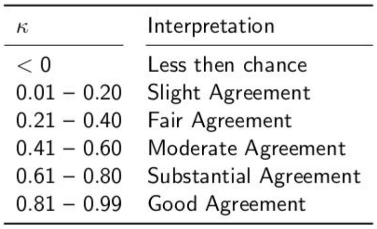
Not Relevant	1	11	12
Total	5	15	20

P(A) = $\frac{7+11}{20} = \frac{9}{10}$

P(E) = $\frac{5}{20} * \frac{8}{20} + \frac{15}{20} * \frac{12}{20} = \frac{11}{20}$

K(Suhas, Tushar) = ($\frac{9}{10}$ - $\frac{11}{20}$)/ (1- $\frac{11}{20}$) = $\frac{7}{9}$ = 0.77

Step4:

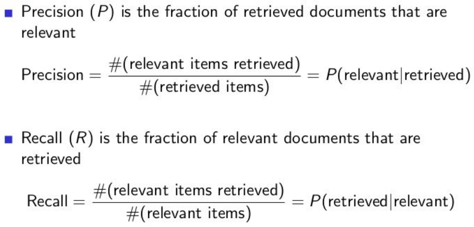


κ	Interpretation
< 0	Less then chance
0.01 − 0.20	Slight Agreement
0.21 − 0.40	Fair Agreement
0.41 − 0.60	Moderate Agreement
0.61 − 0.80	Substantial Agreement
0.81 − 0.99	Good Agreement

Tushar/Stefan: Moderate Agreement
Suhas/Stefan: Substantial Agreement
Tushar/Suhas: Substantial Agreement

The inner annotator agreement is substantial according to the Kappa index.
Qualitative Analysis: There are some differences in the expectations in the formulation in the information need. For example, when one expects a list immediately on the website, others might be content to read the infection rate etc. and then click on a link, what regulations have been passed. Further we didn’t expect that the information to be convoluted with other (rather useless) information like if a politician has embedded his speech on the website. Thus there are different tolerance levels for clutter amongst ourself.

TASK 3.



- Precision (P) is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

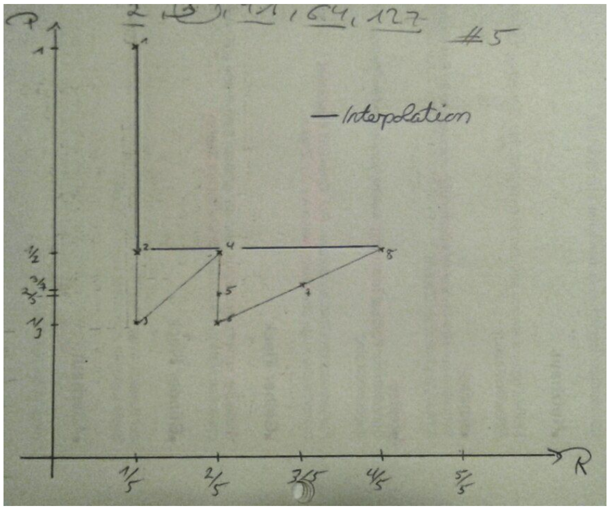
Ranked result: q1 : 127, 9, 10, 2, 35, 32, 41, 64

Correct result: q1 : 2, *34*, 41, 64, 127

Overall

Precision = 4/8 = 0.5

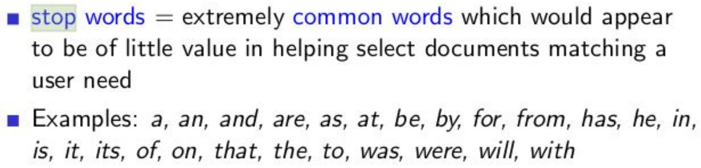
Recall = ⅘ = 0.8



k	Result Set	Precision	Recall
1	127	1/1	1/5
2	127, 9	1/2	1/5
3	0127, 9, 10	1/3	1/5
4	127, 9, 10, 2	2/4	2/5
5	127, 9, 10, 2, 35	2/5	2/5
6	127, 9, 10, 2, 35, 32	2/6	2/5
7	127, 9, 10, 2, 35, 32, 41	3/7	3/5
8	127, 9, 10, 2, 35, 32, 41, 64	4/8	4/5

The user is willing to look at more stuff as both precision and recall stays the same

TASK 4.



- stop words = extremely common words which would appear to be of little value in helping select documents matching a user need
- Examples: *a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with*

“Rare terms are more informative than frequent terms.”

- As stop words appear more common, their information value is not very high. With a tf-idf weight, the importance of a term increases how many times it appears in a document, offsetted how common the term appears across the collection. Thus stop-words should usually induce noise in the weights of the overall tf-idf weights. This increases or decreases the relative weight of other rare terms and could have a negative impact on the ranking.
- Using stop-word-elimination reduces the number of terms which has to be indexed → This might increase query processing time
- But stop words are useful for phrase queries and can increase the performance of the ranking as it better reflect the informativeness in the search results

TASK 5.

Term at a time: The index is organised by postings list so it minimises the disk seeks. In this method, a query is processed term-at-a-time and an accumulator stores the score of each term in the query. When all terms are processed, the accumulator contains the scores of the documents.

Query: “christopher movie hollywood”
Score function : 10*#(christopher)+1*#(movie)+5*#(hollywood)
Postings Lit:

Term	Postings List
christopher	(3,1)
movie	(3,1) (4,1)(5,1)
hollywood	(4,1)(5,1)

Term at a time Processing:

Term	Document1	Document2	Document3	Document4	Document5
	0	0	0	0	0
Christopher	0	0	10 (10*christopher)	0	0
Movie	0	0	11 (1*movie)	1 (1*movie)	1 (1*movie)
hollywood	0	0	11	6 (5*hollywood)	6 (5*hollywood)

Document at a time: Processing a query a document at a time requires several disk seeks. In this method, all the documents containing at least one term are scored. Each document is scored sequentially.

Query: “christopher movie hollywood”
Score function : 10*#(christopher)+1*#(movie)+5*#(hollywood)
Postings List:

Term	Postings List
christopher	(3,1)
movie	(3,1) (4,1)(5,1)
hollywood	(4,1)(5,1)

Document at a time processing:

Document ID	Score
D3	10*1+1*1 = 11
D4	1*1 +1 *5 = 6
D5	1*1 + 1*5 = 6