Exercise 3

Information Retrieval

5. Scoring, Term Weighting and the Vector Space Model

Warm-Up

Exercise 5.1

- Are the following statements true or false? Give reasons for your answer.
  a) Ranking documents is especially important for small document collections. f small
  b) The relevance of a document does not depend on the query. f 和q一起计算
  c) The Jaccard coefficient is a measure for set similarity. t
  d) The Jaccard coefficient works well for ranking documents. f 用tf idf
  e) Rare terms are less informative than frequent terms. f rare are very informassiv
  f) The inverted document frequency (idf) has no effect on the ranking for one-term queries. t will not use global info
  g) The idea of the vector space model is to (i) represent documents and queries as vectors and (ii) calculate the relevance of a document as a vector similarity. t
  h) There is exactly one way to calculate the tf-idf weights. f 在ppt最后

Exercise 5.2

- What minimal and maximal values can the following variables have?

$\boxed{? \le tf_{t,d} \le ?}$ $\boxed{? \le df_t \le ?}$ $\boxed{? \le idf_t \le ?}$

ok

ppt上应该有

看照片

Fun with Calculations I

Exercise 5.3

tf 1,2.95 ? 1/11 1/3

- Compute the Jaccard matching score and the tf matching score for the following query-document pairs

  - Q_1: `information on cars`
    D_1: `all you have ever wanted to know about cars`

  - Q_2: `information on cars`
    D_2: `information on trucks, information on planes, information on trains`

- How well do these metrics reflect the relevance of the documents?

The Vector Space Model

Exercise 5.4

- Assume we have a corpus of $N = \mathbf{50\,000}$ documents
- Find below some information regarding 3 terms and 2 documents

Term t	df_t	$tf_{t,d1}$	$tf_{t,d2}$
`car`	500	0	10
`health`	5	10	100
`insurance`	50	1	100

a) Which of the documents d_1 and d_2 is more relevant to the query

q	**`health insurance`**

according to the vector space model? 看图，ok

ok

Use the weighting scheme `ltn.bnn` for creating the vectors and the cosine similarity for scoring.

SMART notation: **`ddd.qqq`**	
Term frequency	
`b` (boolean)	$\begin{cases} 1 \; if \; tf_{t,d} > 0 \\ 0 \; otherwise \end{cases}$
`l` (logarithm)	$\begin{cases} 1 + \log(tf_{t,d}) > 0 \\ 0 \; otherwise \end{cases}$
Document frequency	
`n` (no)	1
`t` (idf)	$\log \frac{N}{df_t}$
Normalization	
`n` (none)	1
`c` (cosine)	$\frac{1}{\sqrt{\sum_i w_i^2}}$

The Vector Space Model

Exercise 5.4

- Assume we have a corpus of $N = \mathbf{50\,000}$ documents
- Find below some information regarding 3 terms and 2 documents

Term t	df_t	$tf_{t,d1}$	$tf_{t,d2}$
`car`	500	0	10
`health`	5	10	100
`insurance`	50	1	100

b) Can we to save computations and still produce the same ranking (for any collection and query)? If so, how?

使用累加器，可以不用算d

Hint: Imagine (i) we want to answer only one query, and (ii) we want to answer many queries.

? -

SMART notation: `ddd.qqq`	
Term frequency	
`b` (boolean)	$\begin{cases} 1\ if\ tf_{t,d} > 0 \\ 0\ otherwise \end{cases}$
`l` (logarithm)	$\begin{cases} 1 + \log(tf_{t,d}) > 0 \\ 0\ otherwise \end{cases}$
Document frequency	
`n` (no)	1
`t` (idf)	$\log \frac{N}{df_t}$
Normalization	
`n` (none)	1
`c` (cosine)	$\frac{1}{\sqrt{\sum_i w_i^2}}$

Some Closing Questions

Exercise 5.5

Exercise 0.86

- If we were to stem `jealous` and `jealousy` to a common stem before setting up the vector space, detail how the definitions of tf and idf should be modified

their tf ' s and their df ' s would be added together

Exercise 5.6

Exercise 0.81

- What is the idf of a term that occurs in every document?
- Compare this to the use of stop word lists

the same effect as idf weighting: the word is ignored.

Exercise 5.7

Exercise 0.94

- Consider the case of a query term that is not in the set of indexed terms
- Thus, the query vector is not in the vector space created from the collection
- How would one adapt the vector space representation to handle this case?

Omit this term from the query and proceed