Information Retrieval
Topic- Scoring, Term Weighting, The Vector Space Model (tf-idf weighting)
Lecture-24

Prepared By

Dr. Rasmita Rautray & Dr. Rasmita Dash
Associate Professor
Dept. of CSE

Content

- Collection frequency
- Document frequency
- tf-idf weighting

Frequency in document vs. frequency in collection

- In addition, to term frequency (the frequency of the term in the document) . . .
- . . .we also want to use the frequency of the term in the collection for weighting and ranking.

Desired weight for rare terms

- Rare terms are more informative than frequent terms.
- Consider a term in the query that is rare in the collection (e.g., arachnocentric).
- A document containing this term is very likely to be relevant.

→ We want high weights for rare terms like arachnocentric.

Desired weight for frequent terms

- Frequent terms are less informative than rare terms.
- Consider a term in the query that is frequent in the collection (e.g., GOOD, INCREASE, LINE).
- A document containing this term is more likely to be relevant than a document that doesn’t . . .
- . . . but words like GOOD, INCREASE and LINE are not sure indicators of relevance.
- → For frequent terms like GOOD, INCREASE, and LINE, we want positive weights . . .
- . . . but lower weights than for rare terms.

Document frequency

- We want high weights for rare terms like ARACHNOCENTRIC.
- We want low (positive) weights for frequent words like GOOD, INCREASE, and LINE.
- We will use document frequency to factor this into computing the matching score.
- The document frequency is the number of documents in the collection that the term occurs in.

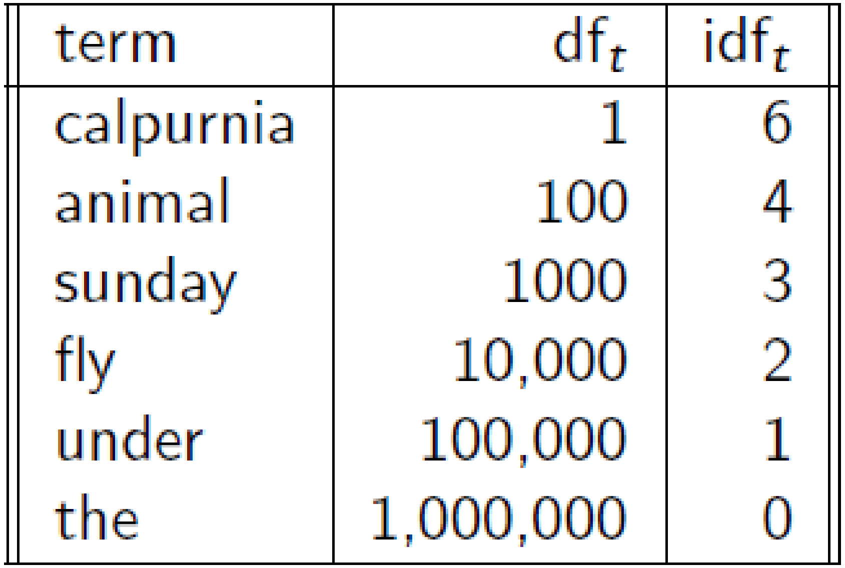
idf weight

- df_t is the document frequency, the number of documents that *t* occurs in.
- df_t is an inverse measure of the informativeness of term *t*.
- We define the idf weight of term *t* as follows:
- $\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$

  (N is the number of documents in the collection.)
- idf_t is a measure of the informativeness of the term.
- [log N/df_t] instead of [N/df_t] to “dampen” the effect of idf
- Note that we use the log transformation for both term frequency and document frequency.

Examples for idf

- Compute idf_t using the formula: $\mathrm{idf}_t = \log_{10} (1{,}000{,}000 / \mathrm{df}_t)$



term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

Effect of idf on ranking

- idf affects the ranking of documents for queries with at least two terms.
- For example, in the query “arachnocentric line”, idf weighting increases the relative weight of ARACHNOCENTRIC and decreases the relative weight of LINE.
- idf has little effect on ranking for one-term queries.

Collection frequency vs. Document frequency

- Collection frequency of t: number of tokens of t in the collection
- Document frequency of t: number of documents t occurs in

Word	cf	df
try	10422	8760
Insurance	10440	3997

tf-idf weighting

- The tf-idf weight of a term is the product of its tf weight and its idf weight.

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log (N / \mathrm{df}_t)$$

- tf-weight
- idf-weight
- Best known weighting scheme in information retrieval
- Alternative names: tf.idf is tf x idf

tf-idf

- Assign a tf-idf weight for each term t in each document d:

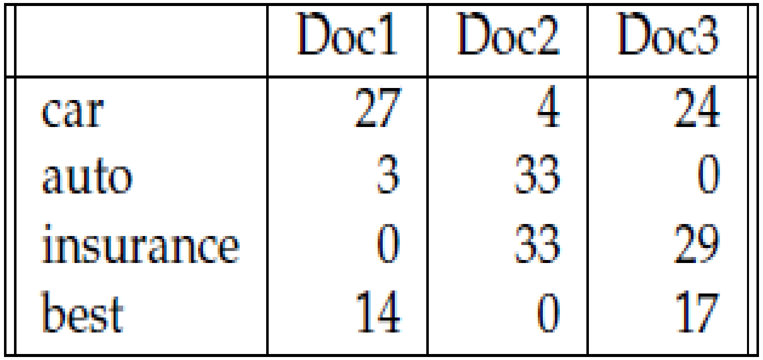
$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log (N / \mathrm{df}_t)$$

- The tf-idf weight . . .
  - . . . increases with the number of occurrences within a document. (term frequency)
  - . . . increases with the rarity of the term in the collection. (inverse document frequency)

Term, collection and document frequency

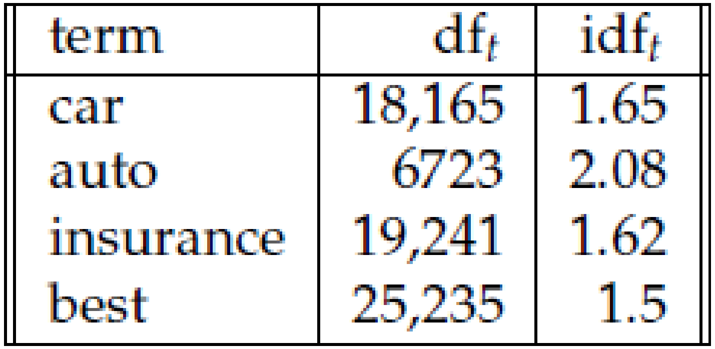
Quantity	Symbol	Definition
term frequency	$\mathrm{tf}_{t,d}$	number of occurrences of *t* in *d*
document frequency	df_{t}	number of documents in the collection that *t* occurs in
collection frequency	Cf_{t}	total number of occurrences of *t* in the collection

Consider the 1st table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3. Compute the tf-idf weights for the terms car, auto, insurance, best, for each document, using the idf values given below.



	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Term frequencies for 3 documents



term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

idf's of terms with various frequencies in the Reuters collection of 806,791 documents

Thank You