Inverse Document Frequency weighting

Document Frequency

- **Rare terms are more informative** than frequent terms
  - Recall stop words
- Consider a term in the query that is rare in the collection
  - A document containing this term is very likely to be relevant to the query → We want **a high weight for rare terms** like arachnocentric
- **Frequent terms are less informative** than rare terms
- Consider a query term that is frequent in the collection (e.g., high, increase)
- But it's not a sure indicator of relevance

  → For frequent terms, we want positive weights for words like high, increase,

  but lower weight than for rare terms
- We will use **document frequency (DF) to capture this** in the score.

IDF weight

- $df_t (\leq N)$ is the document frequency of t : the number of document
  - df_t is an inverse measure of the informativeness of t
  - $df_t \leq N.$
  - N : No. of documents
- We define the inverse document frequency (idf) of t by $idf_t = log_{10}(N/df_t)$
  - We use $log_{10}(N/df_t)$ instead of (N/df_t) to "dampen" the effect of idf

IDF example, suppose N=1 million

term	df_t	idf_t
calpurnia	1	6
animal	100	5
sunday	1,000	4
fly	10,000	3
under	100,000	2
the	1,000,000	1

- "the"라는 term t는 모든 문서에 나타난 단어. $idf_t = 0$
- tf와 달리 idf는 **모든 문서내에서 각 term에 해당하는 값이 같음**

Effect of idf on ranking

- Question : Does idf have an effect on ranking for one-term queires, like
  - iPhone
- idf has no effect on ranking one term queries
  - idf affects the ranking of documents for queries with at least two terms
  - For the query "capricious person", idf weighting makes occurrences of "capricious" count for much more in the final document ranking than occurrences of "person"
  - **df가 적은** term에 더 **많은 가중치**를 부여. 의미있는 단어에 가중치를 주게 됨 → 즉, **단어 자체에 가중치 부여**

Collection VS. Document Frequency

- The collection frequency of t is the number of occurrences of t in the collection, counting multiple occurrences.
- example :

Word	Collection Frequency	Document Frequency
insurance	10440	3997
try	10422	8760

  - Collection Frequency : collection에서 단어 word가 사용된 횟수를 의미하기에, 몇몇 문서내에서 "insurance" 단어가 많이 사용된다면 **"insurance"와 "try" 중 정보성은 "insurance"가 많더**라도 **score은 비슷**할 수 있음
  - Document Frequency : 한 문서내에서 단어 word를 포함하는지 안하는지를 (not multiple occurrences) 나타내기에, **문서에 자주 출현하지 않은 단어가 문서에 자주 출현한 단어보다 높은 score**를 받음

출처

- stanford IR 강의 (https://www.youtube.com/watch?v=7nWlI_TVid0&list=PLaZQkZp6WhWwoDuD6pQCmgVyDbUWl_ZUi&index=10)