TF-IDF
Term Frequency – Inverse Document Frequency

Seminar By:
SUDHARSAN R

Introduction

- In information retrieval, tf–idf (also TF*IDF, TFIDF, TF–IDF, or Tf–idf), short for **term frequency – inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. tf–idf can be successfully used for stop-words filtering in various subject fields, including text summarization and classification.

Definition

- The tf–idf is the product of two statistics, *term frequency* and *inverse document frequency*. There are various ways for determining the exact values of both statistics.

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

- Where tf(t,d) – term frequency

  idf(t,D) - inverse document frequency

  t - relative frequency of term

  d – document

  D – total number of documents

Term frequency

- E.g. Query = "the brown cow“

  A set of documents

1. Eliminate Documents that doesn’t contain all three words
2. To further distinguish them, we might count the number of times each term occurs in each document.
3. The number of times a term occurs in a document is called its **term frequency**.

Term frequency

- Term frequency, *tf(t,d)*, is the relative frequency of term t within document d,

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

- where *ft,d* is the raw count of a term in a document, i.e., the number of times that term t occurs in document d.
- denominator is simply the total number of terms in document *d*.

Inverse document frequency

- E.g. Query = "the brown cow“

  A set of documents

1. The term "the" is so common, term frequency will tend to incorrectly emphasize documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "brown" and "cow".
2. The term "the" is not a good keyword unlike the less-common words "brown" and "cow".
3. To decrease the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely , Inverse document frequency is used

Inverse document frequency

- The **inverse document frequency** is a measure of how much information the word provides, i.e., if it is common or rare across all documents.
- Obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- Where,

  N - total number of documents in the corpus N = | D |

  $|\{d \in D : t \in d\}|$ - number of documents where the term t appears.

  If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$

Term frequency – inverse document frequency

- Then *tf–idf* is calculated as,

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

- A high weight in *tf–idf* is reached by a high term frequency (in the given document) and a low inverse document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

Example

- Suppose that we have term count tables of a corpus consisting of only two documents,

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

- The calculation of tf–idf for the term "this" is performed as follows…

Example

- Term Frequency :

$$\mathrm{tf}(\text{"this"}, d_1) = \frac{1}{5} = 0.2$$
$$\mathrm{tf}(\text{"this"}, d_2) = \frac{1}{7} \approx 0.14$$

- Inverse Document Frequency :

$$\mathrm{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

- So tf–idf is zero for the word "this", which implies that the word is not very informative as it appears in all documents.

$$\mathrm{tfidf}(\text{"this"}, d_1, D) = 0.2 \times 0 = 0$$
$$\mathrm{tfidf}(\text{"this"}, d_2, D) = 0.14 \times 0 = 0$$

Example

- The calculation of *tf–idf* for the term "example" is performed as follows:

$$\mathrm{tf}(\text{"example"}, d_1) = \frac{0}{5} = 0$$
$$\mathrm{tf}(\text{"example"}, d_2) = \frac{3}{7} \approx 0.429$$
$$\mathrm{idf}(\text{"example"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

- Finaly,

$$\mathrm{tfidf}(\text{"example"}, d_1, D) = \mathrm{tf}(\text{"example"}, d_1) \times \mathrm{idf}(\text{"example"}, D) = 0 \times 0.301 = 0$$
$$\mathrm{tfidf}(\text{"example"}, d_2, D) = \mathrm{tf}(\text{"example"}, d_2) \times \mathrm{idf}(\text{"example"}, D) = 0.429 \times 0.301 \approx 0.129$$

Thank you..

LoneWolf