An information-theoretic perspective of tf-idf measures
(paper review)

Kholkin I., Kovrigin A., Shibaev E.

Faculty of Computer Scince
Constructor University Bremen

October 2023

Table of Contents

Table of Contents

Intro

Information retrieval abstraction

We have a set of N documents $D = \{d_1, \ldots, d_N\}$ and a set of M terms (words) from these documents $W = \{w_1, \ldots, w_M\}$.

tf-idf

The **term frequency–inverse document frequency (tf–idf)** is a common measure to weigh words in information retrieval systems:
For a term w_i and a document d_j, its score is $tf_{i,j} \cdot \log(idf_i)$, where

- $tf_{i,j}$ is a frequency of w_i in d_j
- idf_i is the inverse fraction of documents the word is present in.

It has numerous variations and is often considered as an empirical method.
We'll derive the tf-idf from a information-theoretical perspective and try generalizing it.

Table of Contents

Basic formulae

Let x_i and y_j be two distinct events from finite event spaces X and Y. A probability distribution $P(x_i, y_j)$ is given for $x_i \in X$ and $y_j \in Y$.

Marginal probability

The probability $P(x_i)$ is then calculated as $\sum_{y_j \in Y} P(x_i, y_j)$

Amount of information

The **amount of information** of x_i is then: $-\log P(x_i)$.

Let's $\mathcal{X}$ and $\mathcal{Y}$ be random variables representing distinct events from X and Y respectively. Then:

Self-entropy

The expected amount of information or **self-entropy** is defined as $\mathcal{H}(\mathcal{X}) = -\sum_{x_i \in X} P(x_i) \log P(x_i)$.

Basic formulae

Pairwise mutual information

The **pairwise mutual information** between x_i and y_j is the difference between the amounts of information based on (i) the actual joint probability, $P(x_i, y_j)$, and (ii) the expected probability when the independence of the two events are assumed, $P(x_i)P(y_j)$:
$\mathcal{M}(x_i, y_j) = \log \frac{P(x_i,y_j)}{P(x_i)(y_j)}$

Expected mutual information

Then, the **expected mutual information**, or mutual information between $\mathcal{X}$ and $\mathcal{Y}$ is defined as: $\mathcal{F}(\mathcal{X}, \mathcal{Y}) = \sum\limits_{x_i, y_j} P(x_i, y_j)\mathcal{M}(x_i, y_j) =$

$$\sum_{x_i, y_j} P(x_i, y_j) \log \frac{P(x_i,y_j)}{P(x_i)(y_j)} = \mathcal{H}(\mathcal{X}) - \mathcal{H}(\mathcal{X}|\mathcal{Y}) = \mathcal{H}(\mathcal{X}) + \mathcal{H}(\mathcal{Y}) - \mathcal{H}(\mathcal{X}\mathcal{Y})$$

Basic formulae

The reduction of uncertainty $\mathcal{Y}$ after observing a specific x_i can be expressed using

Kullback-Leibler information

Basically the Kullback-Leibler divergence between $P(\mathcal{Y}|x_i)$ and $P(\mathcal{Y})$:
$\mathcal{K}(P(\mathcal{Y}|x_i)||P(\mathcal{Y})) = \sum_{y_j \in Y} P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)}$.

As you might have noticed, our mutual information thing looks a lot like it. In fact:
$$\mathcal{F}(\mathcal{X},\mathcal{Y}) = \sum_{x_i \in X} P(x_i)\mathcal{K}(P(\mathcal{Y}|x_i)||P(\mathcal{Y})) = \sum_{y_j \in Y} P(y_j)\mathcal{K}(P(\mathcal{X}|y_j)||P(\mathcal{X}))$$

Information-theoretic view on tf-idf

Back to our problem.

We have a set of N documents $D = \{d_1, \ldots, d_N\}$ and a set of M terms from these documents $W = \{w_1, \ldots, w_M\}$. Let's say d_i is also an event of picking a document d_i (similarly with w_i).
Now, let $\mathcal{D}$ and $\mathcal{W}$ be random variables defined over D and W respectively. Our objective is to calculate the expected mutual information between $\mathcal{D}$ and $\mathcal{W}$.

Information-theoretic view on tf-idf

Let's assume that $P(d_j|w_j) = \frac{1}{N_j}$ and $P(d_j) = \frac{1}{N}$. Meaning that, in the first case, we know the word and that it comes from one of the N_j documents equally likely. And, in the second case, we don't have any word as a query. Hence, let's say that all the documents are equally likely to fit. With these assumptions:

Self-entropy for $\mathcal{D}$

$$\mathcal{H}(\mathcal{D}) = -\sum_{d_j \in D} P(d_j) \log P(d_j) = -N \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N}$$

Self-entropy for $\mathcal{D}|w_i$

$$\mathcal{H}(\mathcal{D}|w_i) = -\sum_{d_j \in D} P(d_j|w_i) \log P(d_j|w_i) = -N_i \frac{1}{N_i} \log \frac{1}{N_i} = -\log \frac{1}{N_i}$$

Information-theoretic view on tf-idf

Finally, let's assume that we randomly select a query term w_i from the whole document set. Denoting the frequency of w_i with in d_j as f_{ij},the frequency of w_i in the whole document set as f_{w_i} and the total frequency of all terms appearing in the whole document set as F,the probability that a specific w_i is selected is f_{w_i}/F. Then, the expected mutual information is calculated as

$$\mathcal{F}(\mathcal{D},\mathcal{W}) = \mathcal{H}(\mathcal{D}) - \mathcal{H}(\mathcal{D}|\mathcal{W}) = \sum_{w_i \in W} P(w_i)(\mathcal{H}(\mathcal{D}) - \mathcal{H}(\mathcal{D}|w_i)) =$$

$$\sum_{w_i \in W} P(w_i)(-\log \tfrac{1}{N} + \log \tfrac{1}{N_i}) = \sum_{w_i \in W} \tfrac{f_{w_i}}{F} \log \tfrac{N}{N_i} = \sum_{w_i \in W} \sum_{d_j \in D} \tfrac{f_{ij}}{F} \log \tfrac{N}{N_i}$$

Information-theoretic view on tf-idf

tf-idf in expected mutual information

$$\mathcal{F}(\mathcal{D}, \mathcal{W}) = \sum_{w_i \in W} \frac{f_{w_i}}{F} \log \frac{N}{N_i} = \sum_{w_i \in W} \sum_{d_j \in D} \frac{f_{ij}}{F} \log \frac{N}{N_i}$$

Hence, tf–idf can be interpreted as the quantity required for the calculation of the expected mutual information. The idf factor expresses the change in the amount of information after observing a specific term, and the tf factor expresses the probability estimation that the term is actually observed.

Assumptions

- The distribution of query terms is basically the same as the distribution of the terms in documents.
- We assumed that: $P(d_j) = \sum_{w_i \in d_j} \frac{f_{w_j}}{F} \frac{1}{N_i} \approx \frac{1}{N}$
- And that $P(w_i, d_j) = \frac{f_{w_j}}{F} \frac{1}{N_i} \approx \frac{f_{ij}}{F}$