1 Term Frequency and Inverted Document Frequency

Term Frequency $\mathrm{tf}_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.

Inverse Document Frequency Estimate the *rarity* of a term in the whole document collection. (If a term occurs in all the documents of the collection, its IDF is zero.)

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

with $|D|$: cardinality of D, or the total number of documents in the corpus $|\{j : t_i \in d_j\}|$: number of documents where the term t_i appears (viz. the document frequency) (that is $n_{i,j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1 + |\{j : t_i \in d_j\}|$

Example $|D| = 1,000,000$ $\mathrm{idf}_t = \log_{10} \frac{1,000,000}{\mathrm{df}_t}$

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

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

Normalized tf tf count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term ti within the particular document d_j.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $|d_j|$.

Alternative: $\frac{tf_{t,d}}{max \;\; tf_d}$

where max tf_d is the max frequency within the document.

Exercise

Given a document with the terms A, B and C with the following frequencies A: 3, B: 2, C: 1

The document belongs to a collection of 10,000 docs. The document frequencies are: A: 50, B:1300, C:250.

Compute the normalized tf and the tf-idf and compare them. You could also check the effects of using normalized tf measures. The idf are as below

A idf = log(10000/50) = 5.3;
B idf = log(10000/1300) = 2.0;
C idf = log(10000/250) = 3.7

Results:

A tf = 3/3; idf = log(10000/50) = 5.3; tf-idf=5.3
B tf = 2/3; idf = log(10000/1300) = 2.0; tf-idf=1.3
C tf = 1/3; idf = log(10000/250) = 3.7; tf-idf=1.2

Recall: The logarithm of a number y with respect to base b is the exponent to which b has to be raised in order to yield y. In other words, the logarithm of y to base b is the solution x of the equation

$b^x = y$

Exercise 1 Given the tables below

tf)

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

idf)

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

Compute the tf-idf weights for the terms in the tables for each document.

Tf normalization Take the values in the tf table above and replace them with normalized tf weights. Compute the tf-idf again. and compare the results.

2 Similarity Measures

- Jaccard

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Cosine Similarity

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$

Exercise 3 Given the tf for term for the three novels “Sense and Sensibility” (SaS), “Pride and Prejudice” (PaP) by Austin and “Wuthering Heights”

term	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6

Weighted terms:

term	SaS	PaP	WH
affection	0.996	0.993	0.847
jealous	0.087	0.120	0.466
gossip	0.17	0	0.254

compute the cosine similarity and the jaccard measure between SaS-PaP and SaS-WH.

Results: cos-sim(SaS,PaP) = 0.999 vs. cos-sim(SaS,WH)=0.888.

3 Evaluation Measures

Accuracy Percentage of documents correctly classified by the system.

Error Rate Inverse of accuracy. Percentage of documents wrongly classified by the system

Precision percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents *retrieved by the system* (TP + FP). (how many of the retrieved books are relevant?)

Recall : percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents *relevant for the human* (TP + FN). (how many of the relevant books have been retrieved?)

F-Measure : Combine in a single measure Precision (P) and Recall (R) giving a *global estimation of the performance* of an IR system

	Relevant	Not Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

Accuracy $\frac{\mathbf{TP + TN}}{TP+TN+FP+FN}$

Error Rate $\frac{\mathbf{FP+FN}}{TP+TN+FP+FN}$

Precision $\frac{TP}{TP+\mathbf{FP}}$

Recall $\frac{TP}{TP+\mathbf{FN}}$

F $\frac{2PR}{R+P}$

Exercise An IR system returns eight relevant documents and ten non-relevant documents. There are a total of twenty relevant documents in the collection. What is the precision of the system on this search, and what is its recall? Calculate the above measures for the following IR systems:

	Relevant	Not Relevant
Retrieved	40	0
Not retrieved	50	10

	Relevant	Not Relevant
Retrieved	40	50
Not retrieved	0	10

	Relevant	Not Relevant
Retrieved	40	25
Not retrieved	25	10