Foundations of Information Retrieval

Exercise 3

Issue date: 01.11.2012
Exercise session on: 08.11.2012, 15.11.2012
Questions to: Elena Demidova: demidova@l3s.de

TF/IDF weighting and Vector Space Model

Given is a document collection with two documents:

Document 1:
A good cook could cook as much cookies as a good cook who could cook cookies.
Document 2:
Best chocolate chip cookies recipe.

a. Index the both documents using an inverted index.
Apply case folding, use stop words list: {a, as}.

good(DF=1)	*D1(2)*
cook (DF=1)	*D1(4)*
could(DF=1)	*D1(2)*
much(DF=1)	*D1(1)*
cookies(DF=2)	*D1(2), D2(1)*
who (DF=1)	*D1(1)*
best(DF=1)	*D2(1)*
chocolate(DF=1)	*D2(1)*
chip (DF=1)	*D2(1)*
recipe (DF=1)	*D2(1)*

Compute the cosine similarity between these documents and the following queries.
Apply the **ltc** weighting.
Q_1= *cookies*
Q_2= *how to cook chocolate cake*

	DF	idf= log(N/DF)	D_1: tf-raw	D_1: tf-wght (see slide 21)	D_1: tf-idf weight	D_1: n'lized (lengthD_1=*0.84*)
good	1	log(2/1)=0.3	2	1+log(2)=1.3	1.3*0.3=0.39	0.39/0.84=**0.46**
cook	1	0.3	4	1.6	0.48	**0.56**
could	1	0.3	2	1.3	0.39	**0.46**
much	1	0.3	1	1	0.3	**0.35**
cookies	2	0	2	1.3	0	**0**
who	1	0.3	1	1	0.3	**0.35**
best	1	0.3	0	0	0	**0**
chocolate	1	0.3	0	0	0	**0**
chip	1	0.3	0	0	0	**0**
recipe	1	0.3	0	0	0	**0**

length D_1 = *sqrt(*0.39^2+0.48^2+0.39^2+0.3^2+0.3^2*)= 0.84*

	DF	idf = log(N/DF)	D_2: tf-raw	D_2: tf-wght (see slide 21)	D_2: tf-idf weight	D_2: n'lized (lengthD_2=0.6)
good	1	log(2/1) = 0.3	0	0	0	**0**
cook	1	0.3	0	0	0	**0**
could	1	0.3	0	0	0	**0**
much	1	0.3	0	0	0	**0**
cookies	2	log(2/2) = 0	1	1	0	**0**
who	1	0.3	0	0	0	**0**
best	1	0.3	1	1	0.3	**0.3/0.6=0.5**
chocolate	1	0.3	1	1	0.3	**0.5**
chip	1	0.3	1	1	0.3	**0.5**
recipe	1	0.3	1	1	0.3	**0.5**

length D_2 = sqrt(4*0.3^2)=0.6

	Q1: tf-raw , tf-wght	Q1: tf-idf, n.length	Q2: tf-raw, tf-wght	Q2: tf-idf weight	Q2: n'lized (length=0.42)	D1: n'lized (see above)	D2: n'lized (see above)
good	0	**0**	0	**0**	**0**	**0.46**	**0**
cook	0	**0**	1	**0.3*1=0.3**	**0.3/0.42= 0.71**	**0.56**	**0**
could	0	**0**	0	**0**	**0**	**0.46**	**0**
much	0	**0**	0	**0**	**0**	**0.35**	**0**
cookies	1	**0**	0	**0**	**0**	**0**	**0**
who	0	**0**	0	**0**	**0**	**0.35**	**0**
best	0	**0**	0	**0**	**0**	**0**	**0.5**
chocolate	0	**0**	1	**0.3**	**0.71**	**0**	**0.5**
chip	0	**0**	0	**0**	**0**	**0**	**0.5**
recipe	0	**0**	0	**0**	**0**	**0**	**0.5**

length Q_2 = sqrt(2*0.3^2)=0.42

cos (d1, q1)=0
cos (d2, q1)=0

➔ The both documents are not relevant for Q1 under ltc weighting (as IDF of Q1 is 0).

cos(d1, q2)=0.71*0.56=0.3976
cos(d2, q2)=0.71*0.5=0.355

➔ D1 is more relevant for Q2 under ltc weighting.

Explain the similarity scores! How would the result change with the lnc weighting?
As Q_1 contains only one term whose IDF is 0, there will be no relevant results under the ltc weighting. Lnc does not take document frequency into account, such that the both documents would have non-zero scores for Q_1. D_1 would be ranked ahead of D_2, as it has a higher TF. No change for Q_2.