A Linear Algebra Approach to the Vector Space Model
A Fast Track Tutorial

Dr. E. Garcia
admin@miislita.com

Last Update: October 21, 2009. Previous Update: February 10, 2008. Published: September 3, 2006.

Abstract

This is an improved (revised and updated) version of a fast track tutorial on vector space calculations. The tutorial covers term-document and term-query matrices, matrix transposition, dot products, and cosine similarities.
A modern linear algebra approach, based on reaction equations, is used. The virtues of this approach are described in http://www.miislita.com/information-retrieval-tutorial/association-scalar-clusters-tutorial-1.pdf.

Keywords: term vector calculations, cosine similarities, term-document, term-query matrices, matrix transposition, dot products, frobenius norm

Problem

A collection consisting of the following five documents is queried for *latent semantic indexing* (**q**):

d1 = LSI tutorials and fast tracks.
d2 = Books on semantic analysis.
d3 = Learning latent semantic indexing.
d4 = Advances in structures and advances in indexing.
d5 = Analysis of latent structures.

Rank documents in decreasing order of cosine similarities. Assume that:

1. Documents are linearized, tokenized, and their stopwords removed. Stemming is not used. Survival terms are used to construct a term-document matrix **A**. This matrix is populated with term weights a_{ij} which are the products of local (L_{ij}), global (G_i), and normalization (N_j) weights; i.e

   $a_{ij} = L_{ij}\, G_i\, N_j$

   In this equation terms are defined as follows:

   a. $L_{ij} = f_{ij}$, where f_{ij} is the frequency of term in document j. This is the so-called FREQ model.
   b. $G_i = log(D/d_i)$, where D is the collection size and d_i is the number of documents containing term i. This is the so-called IDF model. IDF stands for Inverse Document Frequency.
   c. N_j = 1/l; i.e. document lengths are normalized to 1/l. For N_j = 1,

   $a_{ij} = f_{ij}\, log(D/d_i)$

   N_j no need to be defined as 1/l and when it is done is called cosine normalization (see references 1 and 2). In general, l is the so called L2-norm or Frobenius length.

2. Query terms are scored using FREQ; i.e., $a_{iq} = L_{iq} = f_{iq}$, where f_{iq} is the frequency of term i in the query **q**.

Solution

1. Compute **A** and **q**.

2. Make the following “reaction” equation transformations

   $\mathbf{A} \rightarrow \mathbf{A}_{(u)}$

   $\mathbf{q} \rightarrow \mathbf{q}_{(u)}$

   $\mathbf{q}_{(u)} \rightarrow \mathbf{q}_{(u)}^{T}$

3. Compute $\mathbf{q}_{(u)}^{T}\,\mathbf{A}_{(u)}$.

where the u subscript denotes unit vectors. Unit vector means their elements are normalized so that vector lengths are equal to 1. Since unit vectors are used, **Cosine θ = Dot Products**.

Step 1. Compute **A** and **q**.

	d1	d2	d3	d4	d5
lsi	1*log(5/1)	0	0	0	0
tutorials	1*log(5/1)	0	0	0	0
fast	1*log(5/1)	0	0	0	0
tracks	1*log(5/1)	0	0	0	0
books	0	1*log(5/1)	0	0	0
semantic	0	1*log(5/2)	1*log(5/2)	0	0
analysis	0	1*log(5/2)	0	0	1*log(5/2)
learning	0	0	1*log(5/1)	0	0
latent	0	0	1*log(5/2)	0	1*log(5/2)
indexing	0	0	1*log(5/2)	1*log(5/2)	0
advances	0	0	0	2*log(5/1)	0
structures	0	0	0	1*log(5/2)	1*log(5/2)

$$
= \begin{array}{c}
\begin{array}{ccccc} \mathbf{d1} & \mathbf{d2} & \mathbf{d3} & \mathbf{d4} & \mathbf{d5} \end{array} \\
\begin{bmatrix}
0.6990 & 0 & 0 & 0 & 0 \\
0.6990 & 0 & 0 & 0 & 0 \\
0.6990 & 0 & 0 & 0 & 0 \\
0.6990 & 0 & 0 & 0 & 0 \\
0 & 0.6990 & 0 & 0 & 0 \\
0 & 0.3979 & 0.3979 & 0 & 0 \\
0 & 0.3979 & 0 & 0 & 0.3979 \\
0 & 0 & 0.6990 & 0 & 0 \\
0 & 0 & 0.3979 & 0 & 0.3979 \\
0 & 0 & 0.3979 & 0.3979 & 0 \\
0 & 0 & 0 & 1.3979 & 0 \\
0 & 0 & 0 & 0.3979 & 0.3979
\end{bmatrix}
\end{array} = \mathbf{A}
\qquad
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \mathbf{q}
$$

Step 2. Compute $\mathbf{A}_{(u)}$, $\mathbf{q}_{(u)}$, and $\mathbf{q}_{(u)}^{T}$.

Unit vectors are obtained by dividing column vectors by their Frobenius norm (L_2-norms, Euclidean lengths). Essentially for a given vector, we square their elements, add them together, and square root the result.

The following Frobenius norms are obtained:

	d1	**d2**	**d3**	**d4**	**d5**		**q**
vector lengths	1.3980	0.8973	0.9816	1.5069	0.6891		1.7321

Each vector element is now divided by the corresponding length. The following matrices are then obtained:

$$
\begin{array}{c}
\begin{array}{ccccc} \mathbf{d1} & \mathbf{d2} & \mathbf{d3} & \mathbf{d4} & \mathbf{d5} \end{array} \\
\begin{bmatrix}
0.5000 & 0 & 0 & 0 & 0 \\
0.5000 & 0 & 0 & 0 & 0 \\
0.5000 & 0 & 0 & 0 & 0 \\
0.5000 & 0 & 0 & 0 & 0 \\
0 & 0.7790 & 0 & 0 & 0 \\
0 & 0.4434 & 0.4054 & 0 & 0 \\
0 & 0.4434 & 0 & 0 & 0.5774 \\
0 & 0 & 0.7121 & 0 & 0 \\
0 & 0 & 0.4054 & 0 & 0.5774 \\
0 & 0 & 0.4054 & 0.2640 & 0 \\
0 & 0 & 0 & 0.9277 & 0 \\
0 & 0 & 0 & 0.2640 & 0.5774
\end{bmatrix}
\end{array} = \mathbf{A}_{(u)}
\qquad
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5774 \\ 0 \\ 0 \\ 0.5774 \\ 0.5774 \\ 0 \\ 0 \end{bmatrix} = \mathbf{q}_{(u)}
\qquad
\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.5774 & 0 & 0 & 0.5774 & 0.5774 & 0 & 0 \end{bmatrix} = \mathbf{q}_{(u)}^{T}
$$

Step 3. Compute $\mathbf{q}_{(u)}^T \mathbf{A}_{(u)}$.

$$\begin{array}{cc} & \begin{array}{ccccc} d1 & d2 & d3 & d4 & d5 \end{array} \\ \mathbf{q}_{(u)}^T \mathbf{A}_{(u)} = & \begin{bmatrix} 0 & 0.2560 & 0.7022 & 0.1524 & 0.3334 \end{bmatrix} \end{array}$$

Thus, documents rank as follows:

d3 > d5 > d2 > d4 > d1

Exercises

1. Repeat the above calculations, this time including all stopwords. Explain any difference in computed results.

2. Repeat the above calculations, this time scoring global weights using IDF probabilistic (IDFP):

   $$G_i = \log((D - d_i)/d_i)$$

   Explain any difference in computed results.

References

1. http://sra.itc.it/people/polettini/PAPERS/Polettini_Information_Retrieval.pdf
2. http://csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/ornl-tm-13756.pdf.
3. http://www.miislita.com
4. http://www.miislita.com/term-vector/term-vector-1.html
5. http://www.miislita.com/term-vector/term-vector-4.html
6. http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html
7. http://www.miislita.com/information-retrieval-tutorial/association-scalar-clusters-tutorial-1.pdf