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

Abstract – This is a fast track tutorial on vector space calculations. A linear algebra approach is used. The tutorial covers term-document and term-query matrices, matrix transposition, dot products, cosine similarities, and local and global weights.

Keywords: vector space model, linear algebra, term-document, term-query matrices, dot products, cosine similarities, local weights, global weights
Published: 11-03-2006; Updated: 03-19-2016

Note: This article is part of a legacy series that the author published circa 2006 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated. The original articles can be found referenced in online research publications on IR and elsewhere.

Problem

A collection of five "documents" ($D = 5$) is searched with the query, *q*, *latent semantic indexing*.

d_1 = LSI tutorials and fast tracks.
d_2 = Books on semantic analysis.
d_3 = Learning latent semantic indexing.
d_4 = Advances in structures and advances in indexing.
d_5 = Analysis of latent structures.

Document terms are not reduced to *word roots*. However, the documents are

1. **linearized**, by removing markup tags, comments, style instructions, and scripts.
2. **tokenized**, by removing punctuation and lowercasing terms.
3. **filtered**, by removing stopwords; i.e., frequently used words like *and, of, in…*

Survival terms are arranged as an index of terms with frequency data. See Table 1.

Table 1. Index terms frequency data.

Index terms	q		d_1	d_2	d_3	d_4	d_5
advances	0		0	0	0	2	0
analysis	0		0	1	0	0	1
books	0		0	1	0	0	0
fast	0		1	0	0	0	0
indexing	1		0	0	1	1	0
latent	1		0	0	1	0	1
learning	0		0	0	1	0	0
lsi	0		1	0	0	0	0
semantic	1		0	1	1	0	0
structures	0		0	0	0	1	1
tracks	0		1	0	0	0	0
tutorials	0		1	0	0	0	0

Depending on the nature of documents and queries, index terms can be weighted with a given weighting scheme or a combination of these (Chisholm & Kolda, 1999). For instance, Salton tried a total of 1800 combinations of which 287 were found to be distinct (Salton & Buckley, 1987).

To start familiarizing yourself with two of the several models out there, in this tutorial query terms are scored with the Term Count Model and document terms with the TF-IDF Model.

Query: Term Count Model (FREQ Model): $w_{i,q} = L_{i,q} = f_{i,q}$

- $w_{i,q}$ = weight of index term i in query q.
- $L_{i,q}$ = local weight defined as the frequency of index term i in query q.

Documents: TF-IDF Model: $w_{i,j} = L_{i,j}\, G_i = f_{i,j}\, log(D/d_i)$

- $w_{i,j}$ = weight of index term i in document j.
- $L_{i,j}$ = local weight defined as the frequency of index term i in document j.
- $G_i = log(D/d_i)$ = global weight, defined as *Inverse Document Frequency (IDF)* where D is the collection size and d_i the number of documents that mention index term i.

Solution

We first construct the query and term-document matrices (**q**, **A**) and populate them with term weights. Vectors lengths are also computed as Euclidean Distances; i.e., as L_2-norms. Some times a vector is denoted with an arrow ($\rightarrow$). You may omit it as long as you are consistent.

Index terms		**q**		**d_1**	**d_2**	**d_3**	**d_4**	**d_5**
advances		0		0.00	0.00	0.00	1.40	0.00
analysis		0		0.00	0.40	0.00	0.00	0.40
books		0		0.00	0.40	0.00	0.00	0.00
fast		0		0.70	0.00	0.00	0.00	0.00
indexing		1		0.00	0.00	0.40	0.40	0.00
latent	**q =**	1	**A =**	0.00	0.00	0.40	0.00	0.40
learning		0		0.00	0.00	0.70	0.00	0.00
lsi		0		0.70	0.00	0.00	0.00	0.00
semantic		1		0.00	0.40	0.40	0.00	0.00
structures		0		0.00	0.00	0.00	0.40	0.40
tracks		0		0.70	0.00	0.00	0.00	0.00
tutorials		0		0.70	0.00	0.00	0.00	0.00
Lengths		1.73		1.40	0.90	0.98	1.51	0.70

Next, vectors are transformed into unit vectors, denoted with a hat (^), by normalizing vector elements with their lengths (Wikipedia, 2016; Abhimanyu, 2012). The **q** and **A** matrices become

Index terms		**$\hat{q}$**		**$\hat{d_1}$**	**$\hat{d_2}$**	**$\hat{d_3}$**	**$\hat{d_4}$**	**$\hat{d_5}$**
advances		0.00		0.00	0.00	0.00	0.93	0.00
analysis		0.00		0.00	0.44	0.00	0.00	0.58
books		0.00		0.00	0.78	0.00	0.00	0.00
fast		0.00		0.50	0.00	0.00	0.00	0.00
indexing		0.58		0.00	0.00	0.41	0.26	0.00
latent	**q =**	0.58	**A =**	0.00	0.00	0.41	0.00	0.58
learning		0.00		0.00	0.00	0.71	0.00	0.00
lsi		0.00		0.50	0.00	0.00	0.00	0.00
semantic		0.58		0.00	0.44	0.41	0.00	0.00
structures		0.00		0.00	0.00	0.00	0.26	0.58
tracks		0.00		0.50	0.00	0.00	0.00	0.00
tutorials		0.00		0.50	0.00	0.00	0.00	0.00

For the grand finale, matrix $\mathbf{q^TA}$ is computed where $\mathbf{q^T}$ is the transpose of $\mathbf{q}$.

$$\mathbf{q^TA} = \begin{bmatrix} d_1 & d_2 & d_3 & d_4 & d_5 \\ 0.00 & 0.26 & 0.70 & 0.15 & 0.33 \end{bmatrix}$$

The reason for computing unit vectors is now justified. When we multiply any two unit vectors, their dot product equals the cosine of the angle between them. So $\mathbf{q^TA}$ is a matrix of cosines equal to dot products. Any geometric-based differences between cosines and dot products relevant to the retrieval problem (Jones & Furnas, 1987) disappear. Looking at the $\mathbf{q^TA}$ matrix, documents are ranked in decreasing order of cosine similarities and as follows:

$d_3 > d_5 > d_2 > d_4 > d_1$

Therefore, the third document (d_3 = Learning latent semantic indexing.) is more similar to the query (q = latent semantic indexing) than the others.

Conclusion

Linear algebra provides a clean-cut approach to the vector space calculations used in IR. In recent years, Dirac Notation and Quantum Theory have been applied to vectors space models and, in general, to IR (Wang, 2007; Rijsbergen, 2004; Kantor, 2007; Baeza-Yates & Ribeiro-Neto, 1999; Grossman & Frieder, 2004).

If you are not familiar with linear algebra, you may want to read our new tutorial on vector space calculations without linear algebra (Garcia, 2016).

Exercises

1. Rework this tutorial exercise, this time
   a. including all stopwords during indexing.
   b. defining G_i in the TF-IDF model as $log((D - d_i)/d_i)$.
2. Rework this tutorial exercise, this time using the Term Count Model to score both document and query terms.

References

Abhimanyu, P. S. (2012). The Representation of Matrices in Unit Vector Notation. Journal of Mathematics Research, Vol 4, No. 4. Retrieved from http://www.ccsenet.org/journal/index.php/jmr/article/view/18102/12560

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Adisson Wesley. Book Review. Retrieved from http://www.amazon.com/gp/customer-reviews/R2HC8ULDSMXKZQ/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=020139829X

Chisholm, E. and Kolda, T. G. (1999). New Term Weighting Formulas for the Vector Space Method in Information Retrieval. Oak Ridge National Laboratory. Retrieved from http://www.sandia.gov/~tgkolda/pubs/pubfiles/ornl-tm-13756.pdf

Garcia, E. (2016). Vector Space Calculations Without Linear Algebra. Retrieved from http://www.minerazzi.com/tutorials/vector-space-calculations.pdf

Grossman, D. A., Frieder, O. (2004). Information Retrieval: Algorithms and Heuristics. Springer. Book Review. Retrieved from http://www.amazon.com/review/RACNGPXD2GNE7/ref=cm_cr_dp_title?ie=UTF8&ASIN=1402030045&channel=detail-glance&nodeID=283155&store=books

Jones, W. P. & Furnas, G. W. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. JASIS, 38(6), 420-442. Retrieved from http://furnas.people.si.umich.edu/Papers/PicturesOfRelevance.pdf

Kantor, P. B. (2007). Keith van Rijsbergen, The Geometry of Information Retrieval. Inf. Retrieval, 2007. 10:485–489. DOI 10.1007/s10791-007-9026-8. Retrieved from http://comminfo.rutgers.edu/~kantor/CURRIC.VITAE/CV%20PDFs/vanRijsbergen.pdf

Rijsbergen, K. (2004). The Geometry of Information Retrieval. Cambridge University Press, UK. Book Review. Retrieved from http://www.amazon.com/review/R3FM04FS4ZDHGC/ref=cm_cr_dp_title?ie=UTF8&ASIN=0521838053&channel=detail-glance&nodeID=283155&store=books

Salton, G. & Buckley, C. (1987). Term Weighting Approaches in Automatic Text Retrieval.87-881. Cornell University. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/6721/87-881.pdf?sequence=1&isAllowed=y See also http://www.cs.odu.edu/~jbollen/IR04/readings/article1-29-03.pdf

Wang, X. M. (2007). Dirac Notation, Fock Space and Riemann Metric Tensor in Information Retrieval Models. Retrieved from http://arxiv.org/ftp/cs/papers/0701/0701143.pdf

Wikipedia (2016). Vector Notation. Retrieved from https://en.wikipedia.org/wiki/Vector_notation