A Comparison of SVD and NMF for Unsupervised Dimensionality Reduction

Chelsea Boling, Dr. Das

Mathematics Department

Lamar University

Outline

- Introduction to Text Mining
- Singular Value Decomposition
- Non-negative Matrix Factorization
- Methods
- Results
- Conclusion

Text Mining

- What is Text Mining?
  - How is it different from Data Mining?
- Why is it hard?
  - Unstructured Texts
- What would we like to do with derived information?
  - Discover new details.
  - Form new facts.

Singular Value Decomposition

$$A_{n\times m} = U_{n\times r} S_{r\times r} (V_{m\times r})^T$$

$A_{n\times m}$: n documents, m terms

$U_{n\times r}$: n documents, r concepts

$S_{r\times r}$: r rank ("strength" of each concept)

$V_{m\times r}$: m terms, r concepts

Non-negative Matrix Factorization

- Given a nonnegative target matrix *A* of dimension *m x n*, NMF algorithms aim at finding a rank *k* approximation of the form:

$$A_{m\times n} \approx W_{m\times k} \times H_{k\times n}$$

  - where W and H are nonnegative matrices of dimensions *m x k* and *k x n,* respectively.
  - *W* is the basis matrix, whose columns are the basis components. *H* is the mixture coefficient matrix, whose columns contain the contribution of each basis component to the corresponding column of X.
- How is the rank chosen for NMF?

$$\min_{W,H} f(W,H) = \left\| A - (WH) \right\|^{\frac{1}{2}}$$

Methods

- **Gather Documents**
- **Structure Text (Preprocessing)**
- **Implement the Techniques**
- **Evaluate Performance**

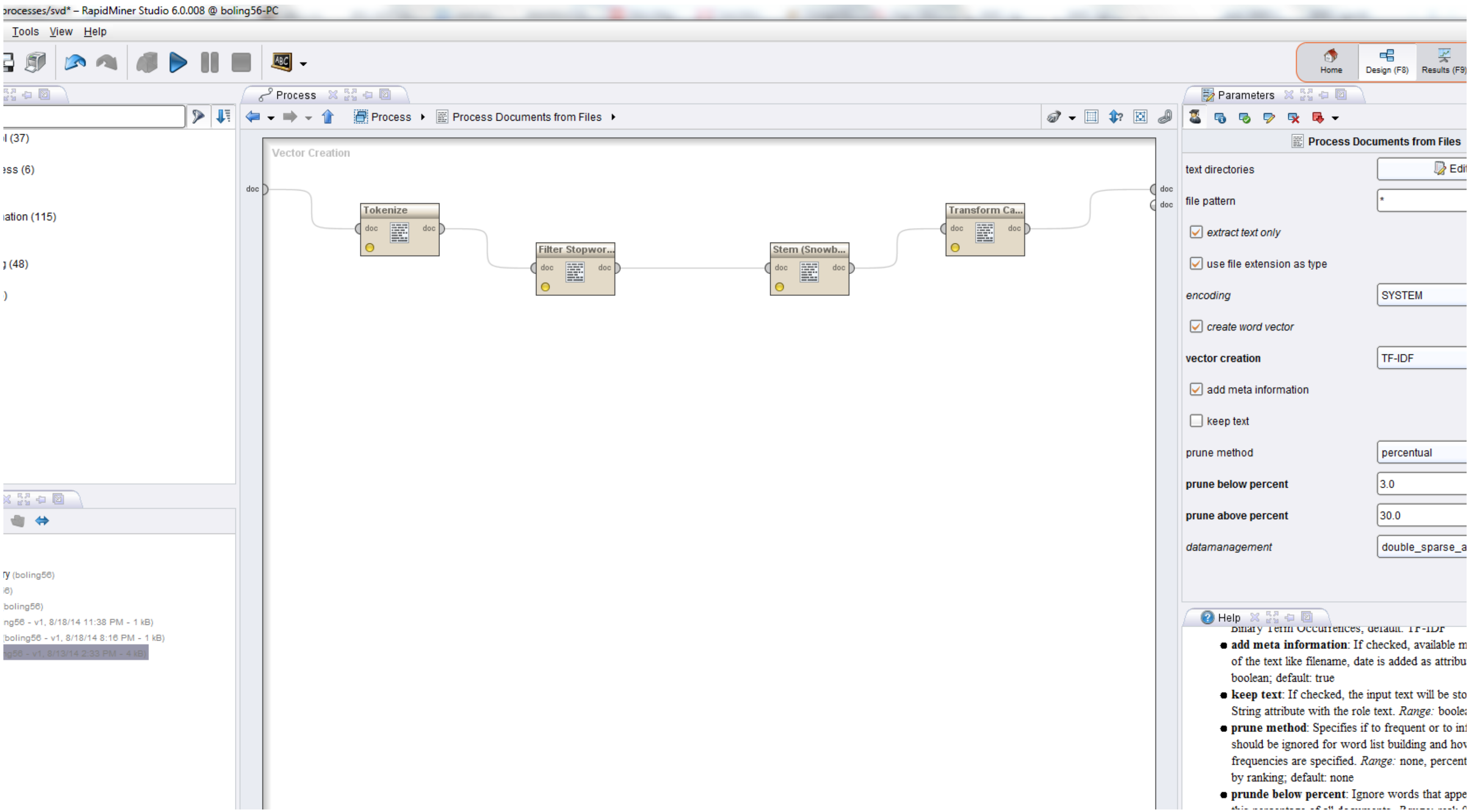
Methods

- We used the Pubmed Central Open Access Subset, which consists of 800,000+ full-text articles.
- Due to memory and space limitations, we did a keyword search on the data and found that 3,398 articles had the term “herbicides”.
- We preprocessed our dataset using RapidMiner and exported our preprocessed data to R.

Methods: Preprocessing Data

- Data Preparation
  - Tokenization
  - Filtering Stopwords and Length
  - Stemming (Snowball Stemming Algorithm)
  - Transform Cases (Lowercase letters)
- Pruning
  - Prune below 30%
  - Prune above 70%

Methods: Preprocessing Data



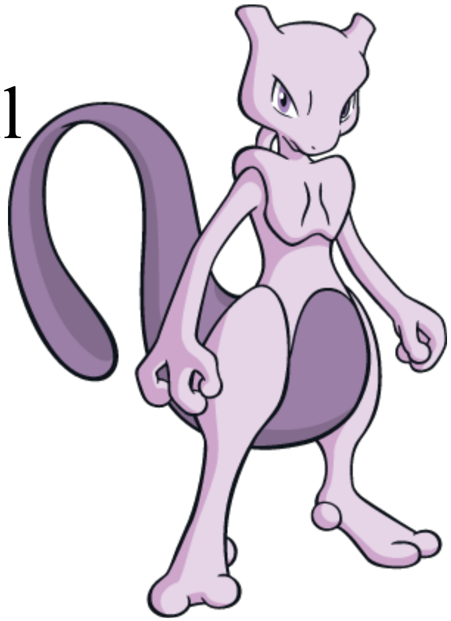
Methods: Term Weighting

- Term Frequency
  - It is the number of times that term *t* occurs in document *d*
  - Should we really use this?
- IDF Weighting
- Term Frequency-Inverse Document Frequency
  - Why Should We Use TF-IDF instead of TF?

Methods: Term Weighting

Assign to term *t* **a weight** in document *d* that is:

- highest when *t* occurs several times within a small number of documents.
  - We like to see rare things!
- lower when the term occurs fewer times in a document, or occurs in many documents.
- lowest when the term occurs in virtually all documents.



Reference: http://nlp.stanford.edu/IR-book/

Mewtwo image: http://images5.fanpop.com/image/polls/987000/987481_1333090091691_full.png

Low Rank Approximation in SVD

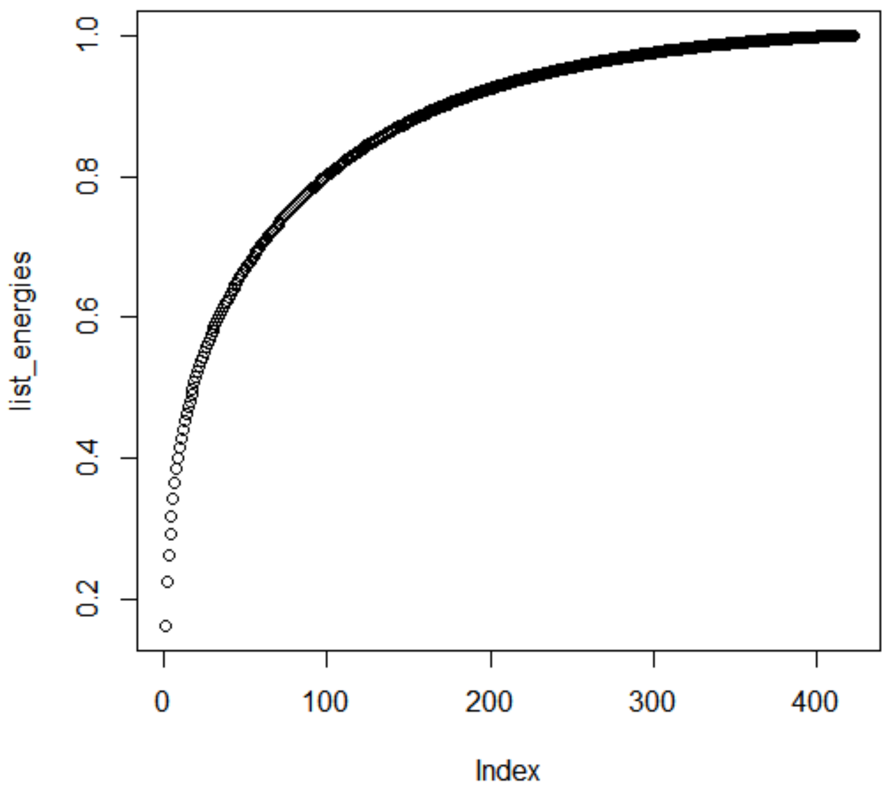
We retain k dimensions of matrix by computing the energy in . In order to retain 90% of the energy in , we compute and divide it by the total energy. This is defined as

$$E_k = \sum_{i=1}^{k} \sigma_{ii}^2$$

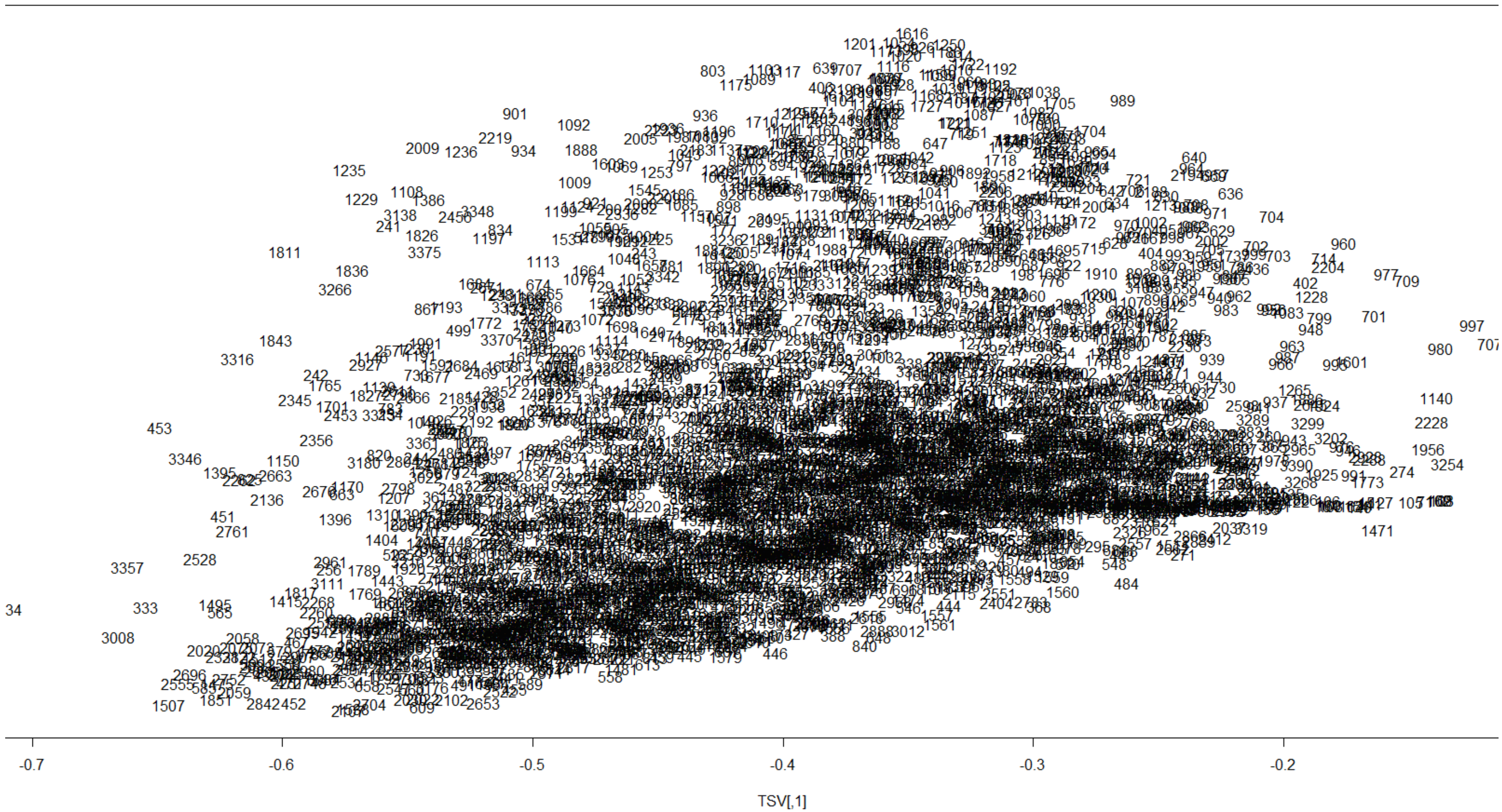
where k denotes the number of reduced dimensions and - σ_{ii} represents the singular values of Σ. By looking at all values of k, the retained energy is 90.0% at least $k = 172$.

$k = 172$	$k = 173$	$k = 174$	$k = 175$	$k = 176$	$k = 177$	$k = 178$	$k = 179$	$k = 180$	$k = 181$	$k = 182$
0.9006	0.9016	0.9026	0.9035	0.9044	0.9053	0.9062	0.9071	0.9080	0.9089	0.9097

“Energy” Plot for SVD



Document Similarity by VΣ



Comparing Approximation Errors Using **Frobenius Norm**

- Singular Value Decomposition

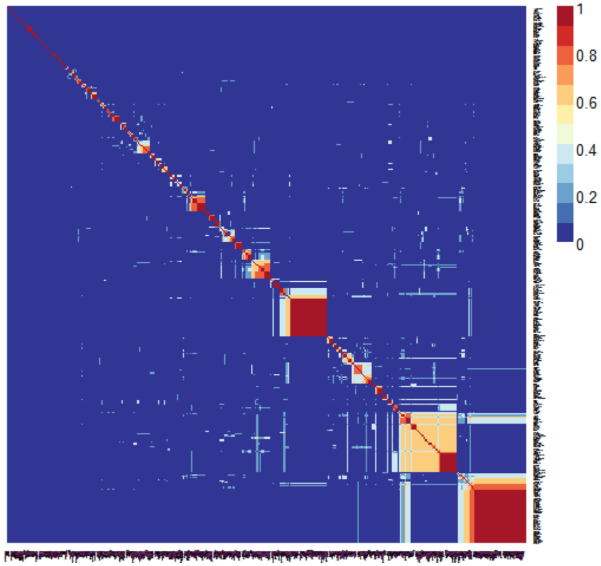
K= 170	K= 172	K= 175	K= 178	K= 182
18.54	18.37	18.11	17.84	17.50

- Non-negative Matrix Factorization

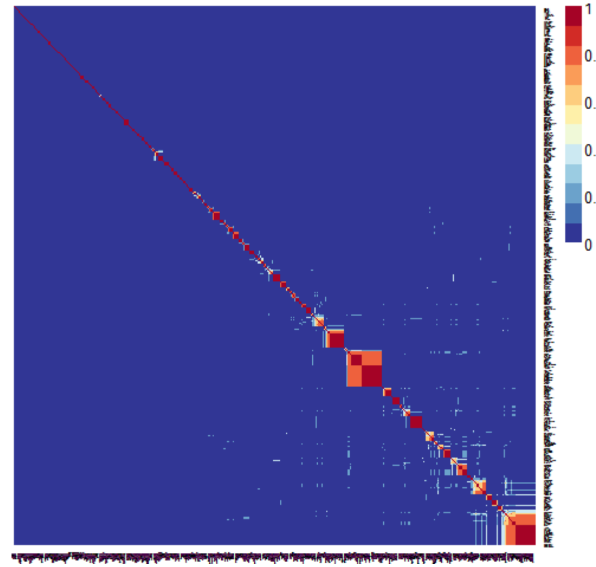
K=170	K=172	K=175	K=178	K=182
19.601	19.47	19.16	18.90	18.57

Consensus Maps for NMF

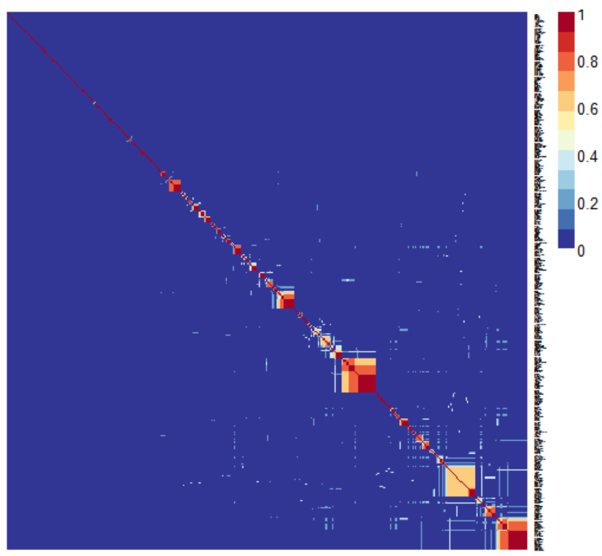
k = 100



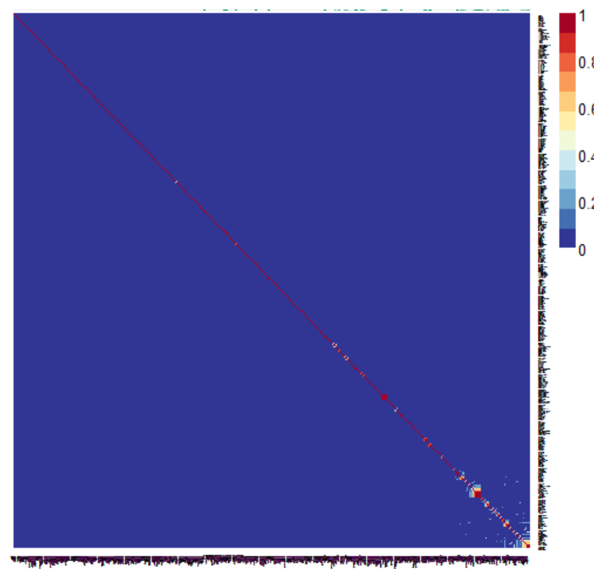
k = 175



k = 182



k = 375



Conclusion and Discussion

- Standard SVD produces a “deeper” factorization of the original data. Singular values come in handy when looking for valuable information.

- Standard NMF iteratively refines a solution, and one may choose to look at a lower rank, which is generally chosen so that $(n + m)r < nm$.

- NMF should be better in terms of its non-negativity constraints.

References

Berry, M. W., S.T. Dumais, and G.W. O'Brien. (1995). Using Linear Algebra for Intelligent Information Retrieval. SIAM Review, 37(4), 573–595.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Dumais, S. T., G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. (1988). Using Latent Semantic Analysis to Improve Access to Textual Information. CHI 88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press, 25(23), 281–285.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). From data mining to knowledge discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*,

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1-36.

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).

Fogel, P., Hawkins, D. M., Beecher, C., Luta, G., & Young, S. S. (2013). A Tale of Two Matrix Factorizations. *The American Statistician*, *67*(4), 207-218.

Kumar, A. C. (2009). Analysis of unsupervised dimensionality reduction techniques. *Computer Science and Information Systems/ComSIS*, *6*(2), 217-227.

Landauer, T.K., D. Laham, B. Rehder, and M.E. Schreiner. (1997). How Well Can Passage Meaning Be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. Proceedings of the 19th Annual Meeting of the Cognitive Science Society, 412–417.

Landauer, T. K., P. W. Foltz, and D. Laham. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25(23), 259–284.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556-562).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). ACM.

Peter, R., Shivapratap, G., Divya, G., & Soman, K. P. (2009). Evaluation of SVD and NMF methods for latent semantic analysis. *International Journal of Recent Trends in Engineering*, *1*(3).

PubMed Central Open Access Subset (2014). The National Center for Biotechnology Information. URL http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.