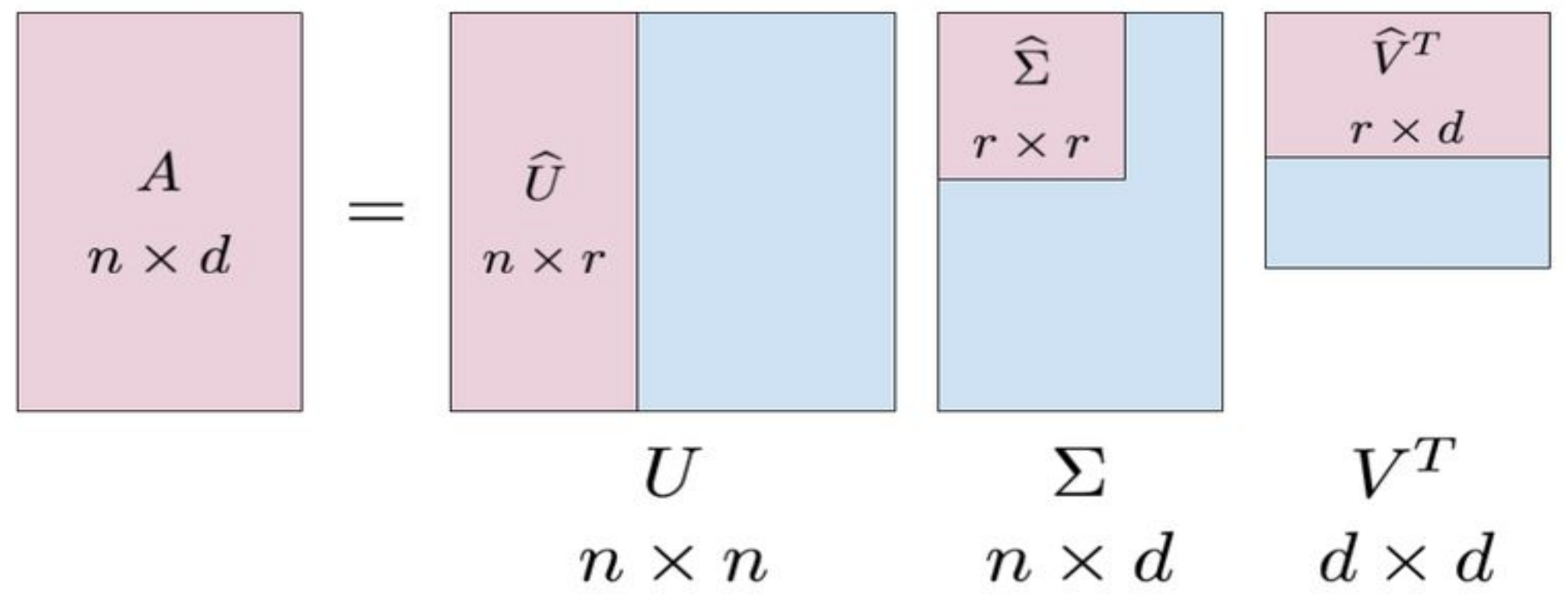
SVD and NMF: An Exploration in Text Mining

Madison Everett and Brian Tonnies
Department of Mathematics and Computer Science

Background

SVD (Singular Value Decomposition) and NMF (Nonnegative Matrix Factorization) are matrix decomposition methods that are commonly used within text mining to summarize large collections of text data.

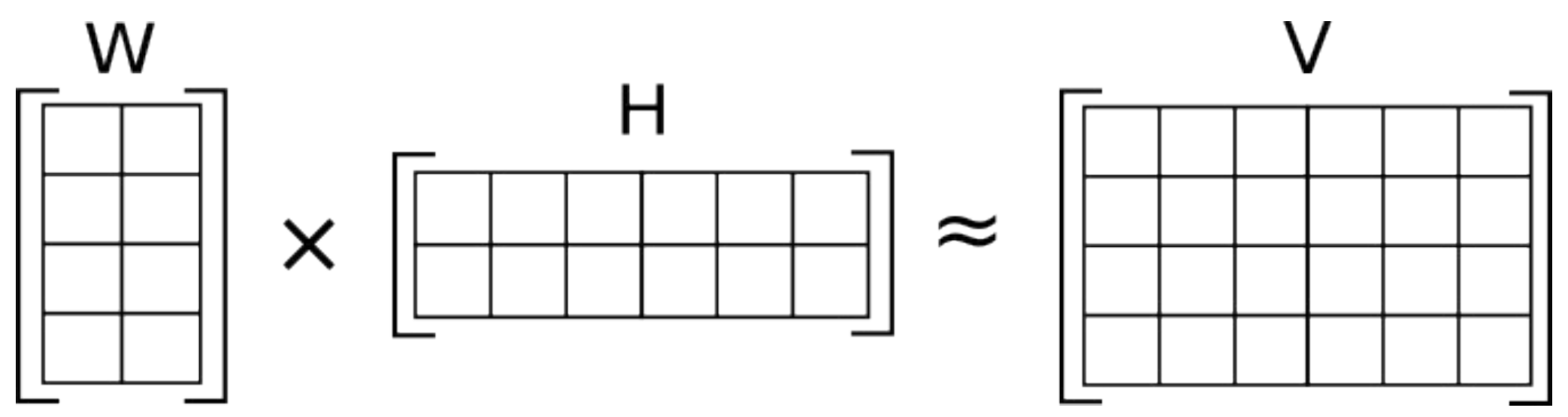
SVD: Singular Value Decomposition



Example of SVD using Matlab

$$\underset{A}{\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}} = \underset{U}{\begin{bmatrix} -0.5 & -0.9 \\ -0.9 & 0.5 \end{bmatrix}} \quad \underset{\Sigma}{\begin{bmatrix} 7.3 & 0 \\ 0 & 0.3 \end{bmatrix}} \quad \underset{V}{\begin{bmatrix} -0.6 & 0.8 \\ -0.8 & -0.6 \end{bmatrix}}$$

NMF: Nonnegative Matrix Factorization



Example of NMF using Matlab

$$\underset{V}{\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}} \approx \underset{W}{\begin{bmatrix} 1.2 & 1.0 \\ 2.4 & 0 \end{bmatrix}} \quad \underset{H}{\begin{bmatrix} 1.7 & 2.1 \\ 0 & 0.5 \end{bmatrix}}$$

This summarization occurs by using these processes on term-document frequency matrices, which are matrices that contain the frequency of terms within a collection of documents, as seen below:

Terms	d1	d2
Science	3	0
Football	0	4

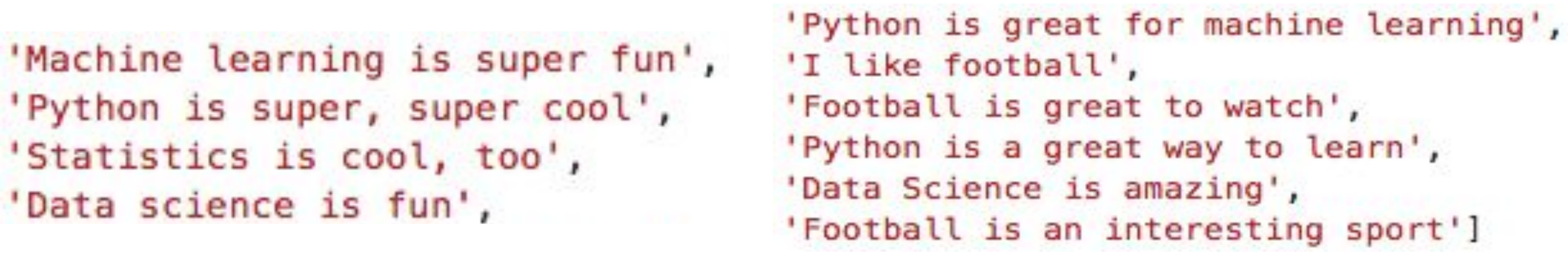
Motivation

Although both of these methods work extremely well, there has been a constant debate on which one of these methods performs better on real world data. Because of this, we have completed a comparison study on a toy dataset and on a large collection of documents in order to evaluate which method performs better.

Methodology

To implement SVD, NMF, and term-document frequency matrices, we used the Scikit-learn and Sklearn framework that contained classes for each.

To compare these two methods, we first created a small toy dataset that consists of 10 sentences that can be summarized into the categories of football, machine learning and data science, as seen below:



```
'Machine learning is super fun',
'Python is super, super cool',
'Statistics is cool, too',
'Data science is fun',
```

```
'Python is great for machine learning',
'I like football',
'Football is great to watch',
'Python is a great way to learn',
'Data Science is amazing',
'Football is an interesting sport']
```

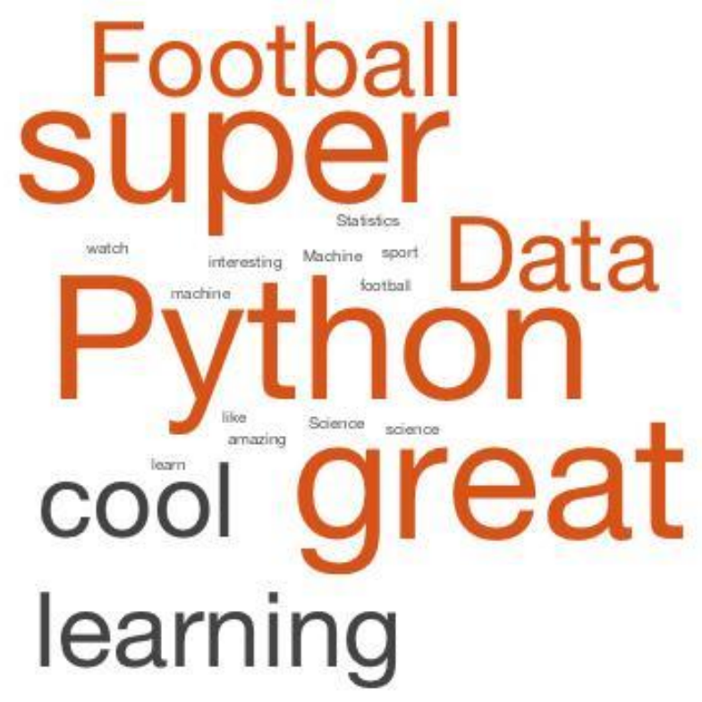
After creating the toy dataset, we then used the cryptography collection of documents from the 20 newsgroup dataset from Sklearn. This collection of documents contained 1000 text files that circulated around the general topic of cryptography.

In order to evaluate these methods, we gathered the general topic extraction results and displayed them and the unmodified document within multiple Word Clouds by using MATLAB.

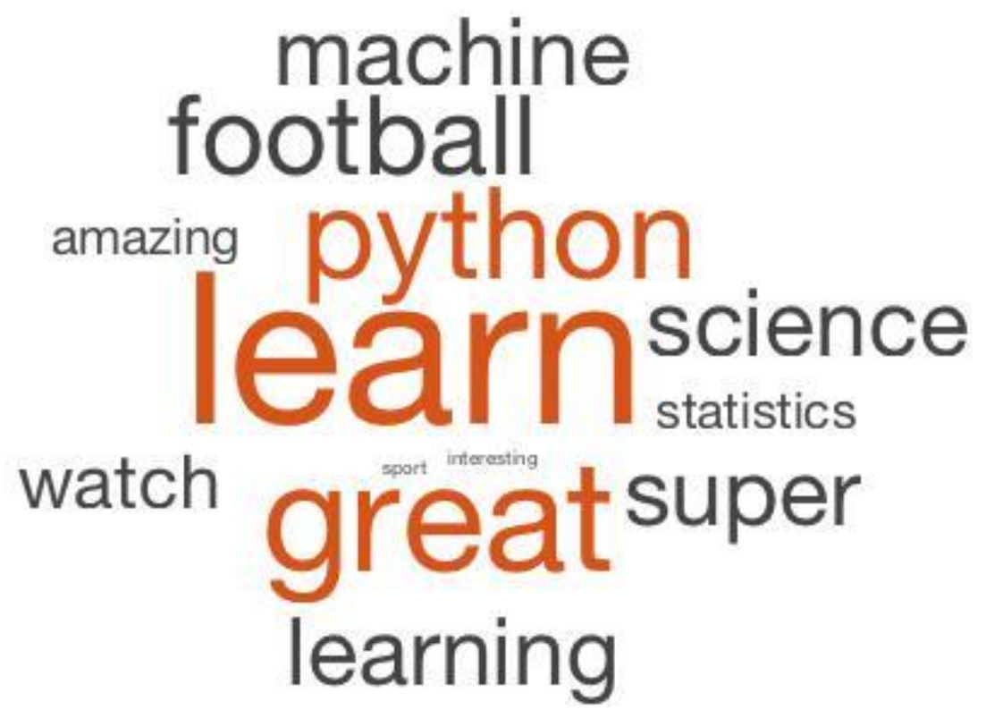
Results

From the small toy dataset, it appears that NMF and SVD perform almost the same compared to the original corpus of the ten sentences.

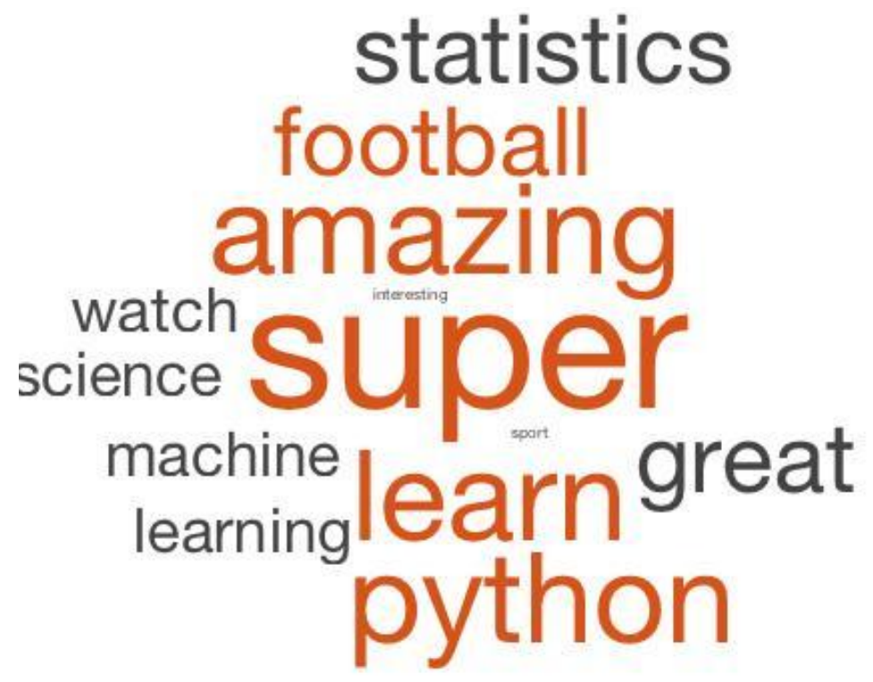
Original Corpus



NMF Results

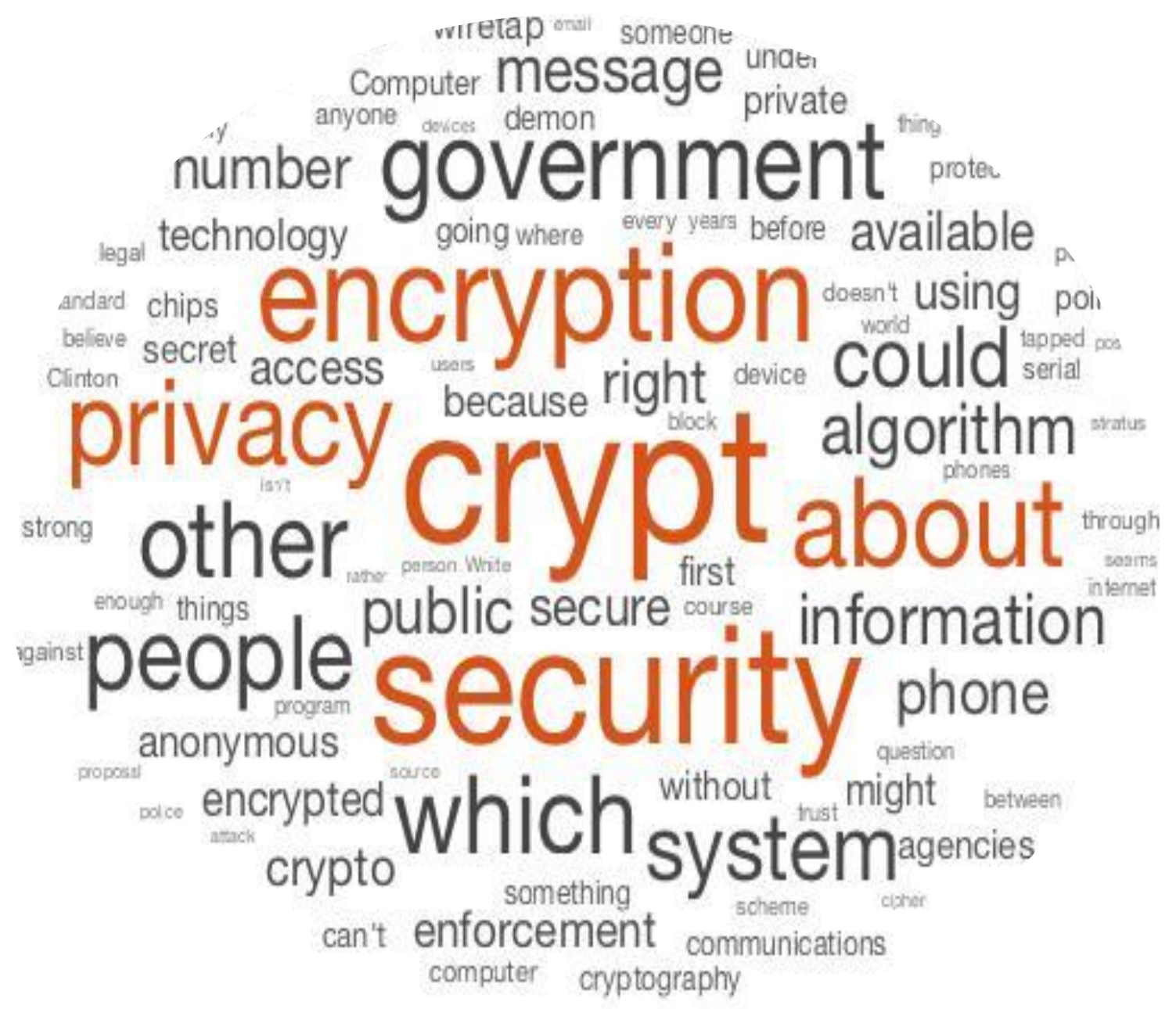


SVD Results

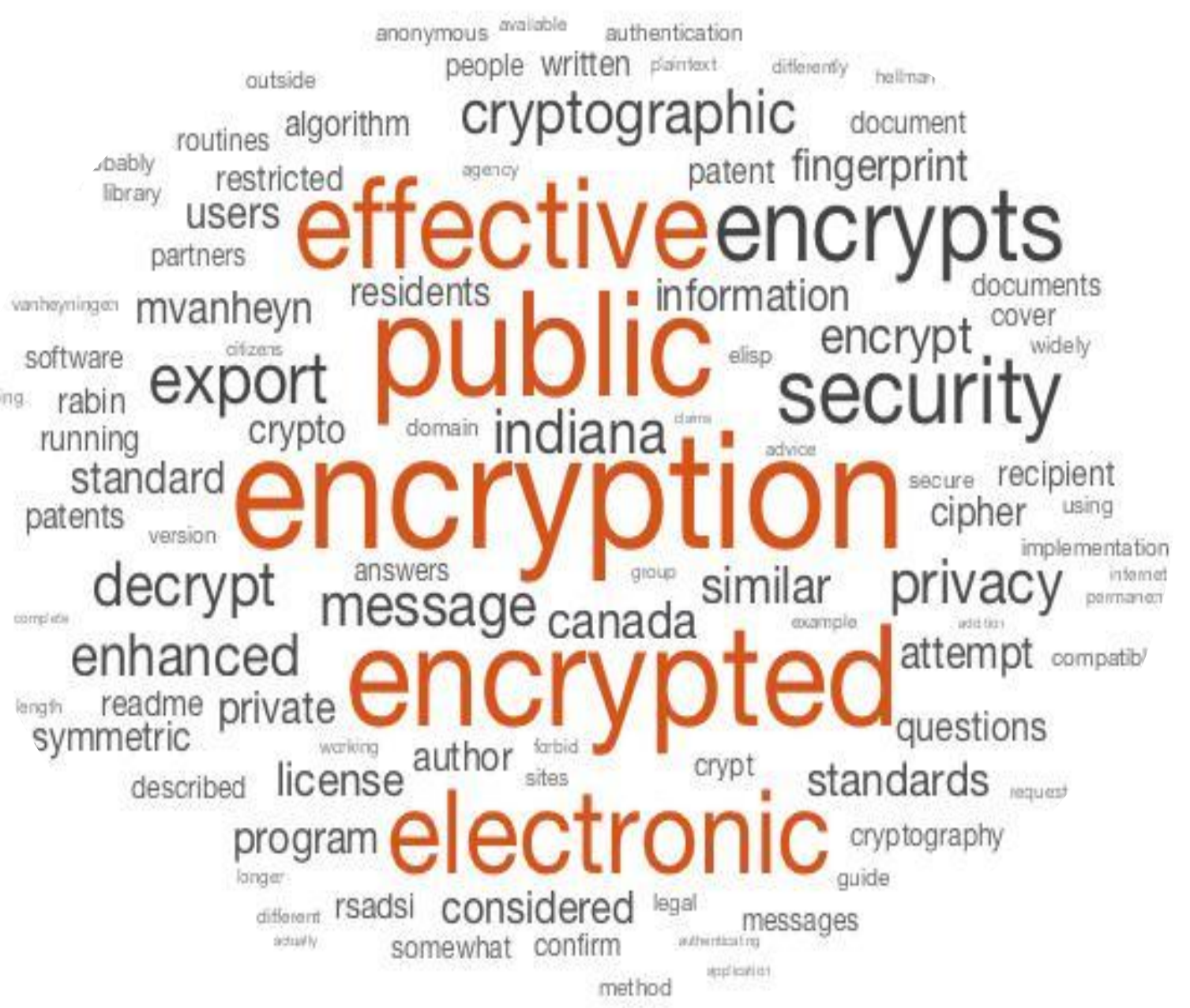


However, on the cryptography dataset, it appears that NMF has more similar results to the original corpus than SVD.

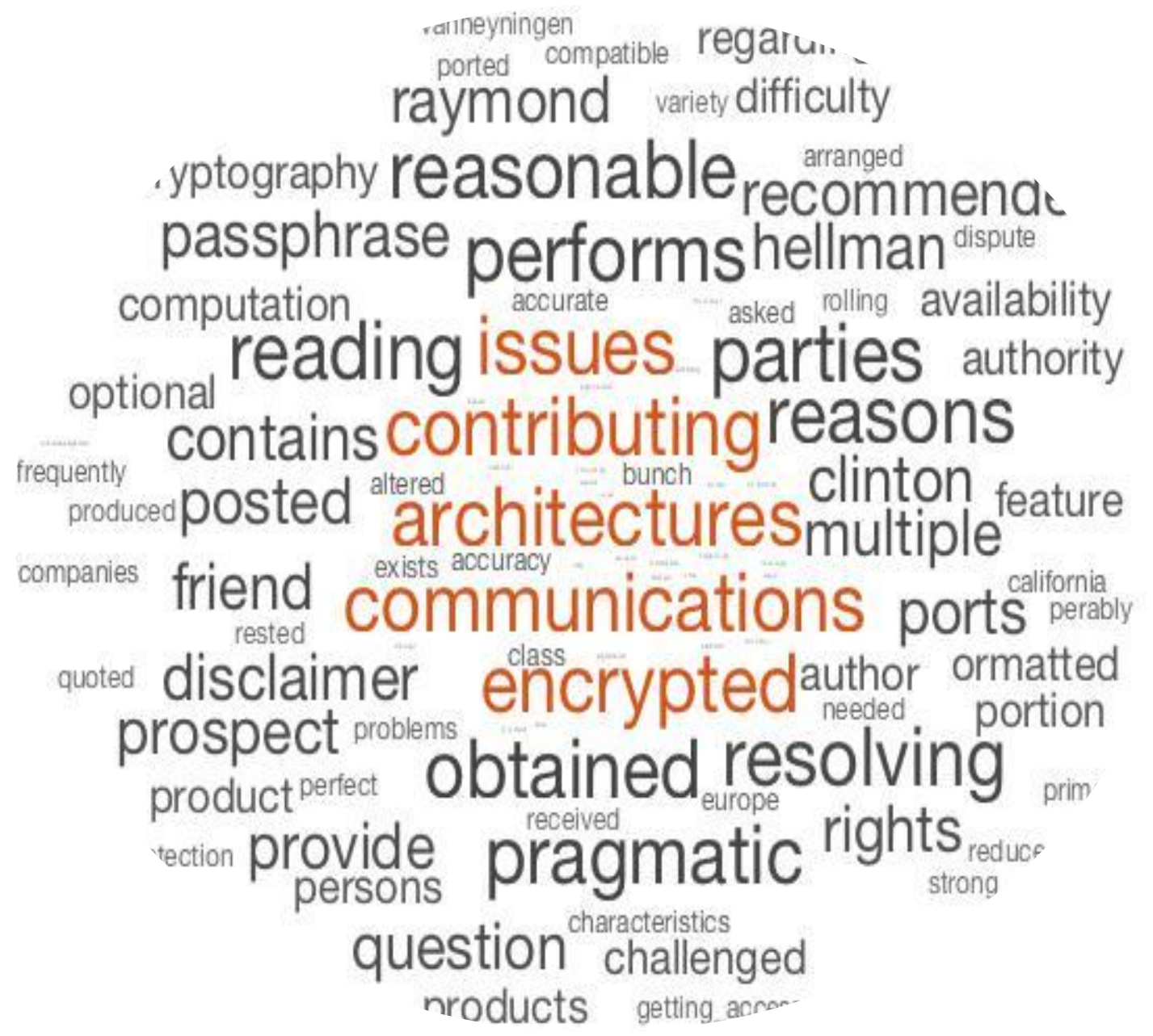
Original Corpus



NMF Results



SVD Results



Conclusion

From this exploration, we have concluded that NMF provides more logical results than SVD when performing general topic extraction from large datasets. This research allows us to further examine and evaluate these two methods and their performance with large real-life datasets.

Future Work

Despite getting intriguing results, some further explorations for this project are:

- Testing NMF and SVD on various inhomogenous datasets to see the comparison in performance and topic extraction.
- Comparing NMF and SVD in query-based topic extraction rather than general topic extraction.
- Exploring other matrix decompositions used for text mining and comparing the performance to NMF and SVD.

References

- Albright, Russ. 2004. Taming Text with the SVD. SAS Institute Inc, Cary, NC.
- Berry, Michael W., Browne, Murray., Langville, Amy N., Pauca, V. Paul.., Plemmons, Robert J. 2006. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, Volume 52, Issue 1, pp 153-173.
- Gao Jing, Jun Zhang. 2005. Clustered SVD Strategies in Latent Semantic Indexing. Information Processing &amp; Management, Volume 41, Issue 5, pg. 1051-1063.
- Ju Hong Lee, Sun Park, Chan-Min Ahn, Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. Information Processing &amp;Management, Volume 45, Issue 1, pg. 20-34.
- Lee, D., Seung, H. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788-791.
- Rakesh Peter, Shivapratap G, Divya G, Soman KP. 2009. Evaluation of SVD and NMF Methods for Latent Semantic Analysis, International Journal of Recent Trends in Engineering ,Vol 1, No. 3.
- Xu, S., Zhang, J., Han, D. et al. 2006. Singular value decomposition based data distortion strategy for privacy protection. Knowledge and Information Systems, Volume 10, Issue 3, pp 383–397.