The Ross Monica Continuum: Using Latent Semantic Indexing for Gender Classification

Anonymous ACL submission

1 Abstract

2 Introduction

Latent Semantic Indexing (LSI) is an applied matrix factorization method used for Natural Language Processing [1]. LSI operates over a term-document matrix, with each vector representing a given document and each ith row representing the occurrence ofword i in the given document. These counts are then discounted by their total occurrence in other documents, to decrease the importance of common words like "a", "the", "and" etc. We call this matrix a Term Frequency– Independent Document Frequency (TF-IDF) matrix(Jurafsky and Martin, 2009). LSI operates under the assumption that semantic meaning is preserved by a TF-IDF matrix A , and can be uncovered using the Singular Value Decomposition(SVD) of A.

SVD is a matrix factorization technique that constructs three matrices U, Σ, V^T. U is the matrix of right eigenvectors of A (representing the terms of A), V^T is the left eigenvectors of A (with rows representing the documents of A), and Σ is the Diagonal matrix of the eigenvalues of A.(Dumais et al., 1988) A rank-k approximation is often used in LSI to only condense the term vectors into a k dimensional space for efficient storage and calculation.

3 Methodology

In this project, I performed LSI over the Emory University's Friends Corpus (Emory University) to see if there were latent gender features encoded in the dataset. I pruned the dataset to include only lines from the six main characters, Ross, Chandler, Joey, Rachel, Phoebe and Monica. This particular corpus was selected because of its roughly equal gender breakdown, and manageable size for TF-IDF. I trained the model over the first nine seasons of the show, reserving season ten for testing purposes. I then fed the model a string containing every line from a given character in season ten (fit to the TF-IDF). The model then measured the Euclidean distance between the query vector and each character's document vector from V^T, and returned the closest match between the two gender groups (Ross, Chandler and Joey vs Monica Phoebe and Rachel). On top of this, I manually reviewed three episodes to identify characteristic features, and performed Principal Component Analysis on the TF-IDF matrix

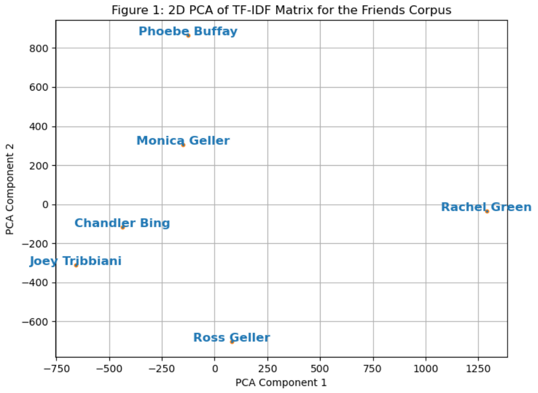

Figure 1: PCA of each speaker in the truncated Friends Corpus

4 Results

The PCA (principal component analysis) graph in Figure 1 shows a two-dimensional projection of each character's document vector. The most prominent grouping is in the bottom left corner, with Ross, Joey and Chandler (all male), with Joey and Chandler in the third quadrant of the graph and Ross barely in the fourth quadrant. While the women are more evenly distributed, they are all distinctly distant from the male cluster, with opposite parity values in at least one of the two principal components. The model's gender estimations for season ten reflect this, with the model correctly identifying all six character's gender identities based on the unseen testing data from season 10.

5 Discussion

After manually reviewing three episodes and the TF-IDF features, the male speakers demonstrate far more NG-coalescence (dropping the ng sound at the end of words) than the female speakers, which helps explain their similarity under the model. While the model's performance matches the PCA embeddings, it's results can not and should not be generalized to any population of people , since the model is trained on very narrow data from characters on a syndicated TV show, not real speakers. In general, this model demonstrates LSI's ability to perform binary classification, and I propose that LSI is a useable baseline for this task Furthermore, these results show that features (such as gendered features) preserved under a vector embedding will be captured by LSI.

5.1 Limitations

The Friends corpus is very limited when compared to the scope of many modern day training sets, which assists in my models performance. The scale which presents issues with both time complexity and data storage for LSI however. The larger the datasets get the less feasible LSI becomes, which is part of the reason that more modern techniques shy away from pure matrix representations.

5.2 Future Work

Possible modifications to this study may include changing the document classes (ie: two document matrix of male vs female speakers) or applying the same method to a different corpus. It would also be interesting to compare the accuracy of the LSI model with something like Naive Bayes to show this method's comparative strength for this task. Finally, this paper shows that LSI may be prone to the similar biases as TF-IDF, which presents a need for further research(Mahmood and Srinivasan, 2019).

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, page 281–285, New York, NY, USA. Association for Computing Machinery.

2015 Emory University. Corpus of lines from friends, liscensed under the apache liscense 2.0.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Asad Mahmood and Padmini Srinivasan. 2019. Twitter bots and gender detection using tf-idf. In *Conference and Labs of the Evaluation Forum*.