[redacted] Final Report

Jesse Huang

May 2, 2018

Introduction

Reddit is a news aggregation site comprised of subreddits focused on specific topics. r/confession is a subreddit for people to confess things where they cannot otherwise do so. Consequently, the comments of this subreddit discuss very personal subjects of a very wide range. These comments, available via Google's BigQuery, can be used to examine the subreddit as a whole [1]. This project models these comments using Latent Dirichlet Allocation (LDA) and Latent Semantic Analisis (LSA), and compares their resulting comment visualizations.

Setup

Everything was done in Python, with sqlite3 used to store the comments, scikit-learn and NumPy for data analysis, and Bokeh for visualization[2]. The procedure outlined below was done for LDA and LSA[3].

- Google's BigQuery was queried for 12000 comments- 1000 randomly selected from each month of 2017.
- The topic model was used to generate 20 topics.
- t-distributed stochastic neighbor embedding (t-SNE) was used to reduce the dimensionality to 2.
- Each comment was plotted on a scatter plot, colored by the topic that it was most strongly associated with. These plots were generated for

each topic model plotting all comments, and plotting only comments with $Pr(Comment \in Topic_c) > 0.5$, where $Topic_c$ is the topic most strongly associated with a given comment. Each topic was represented by its top 8 most likely words.

Results

The above procedure generated the following plots:

Figure 1: Comment Scatterplot with All Comments using LDA, 20 Topics

Figure 2: Comment Scatterplot with All Comments using LSA, 20 Topics

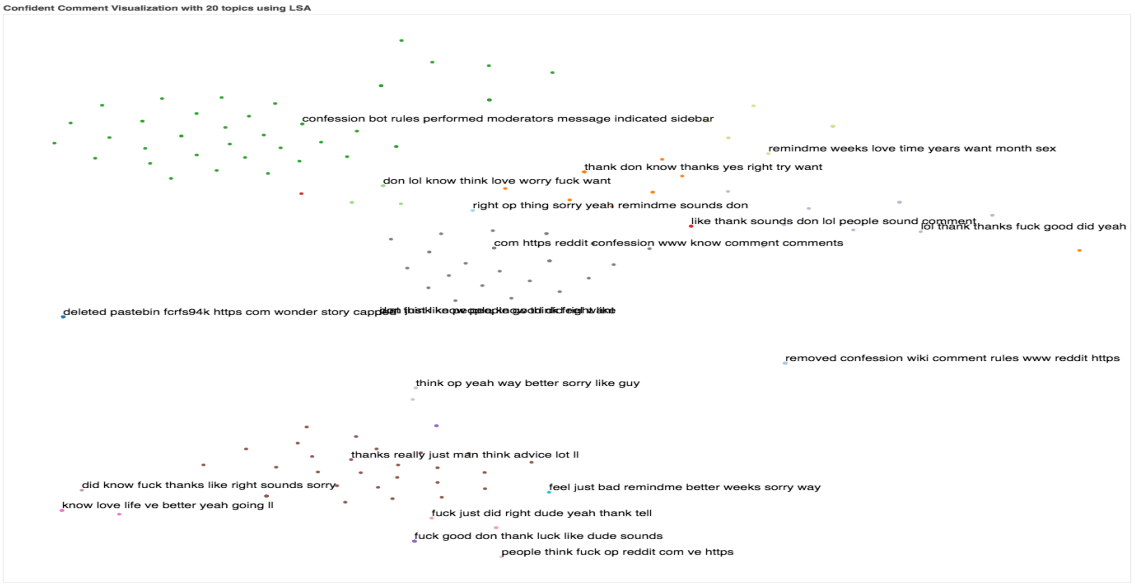

Figure 3: Scatterplot with Comments: $Pr(Comment \in Topic_c) > 0.5$ using LSA, 20 Topics

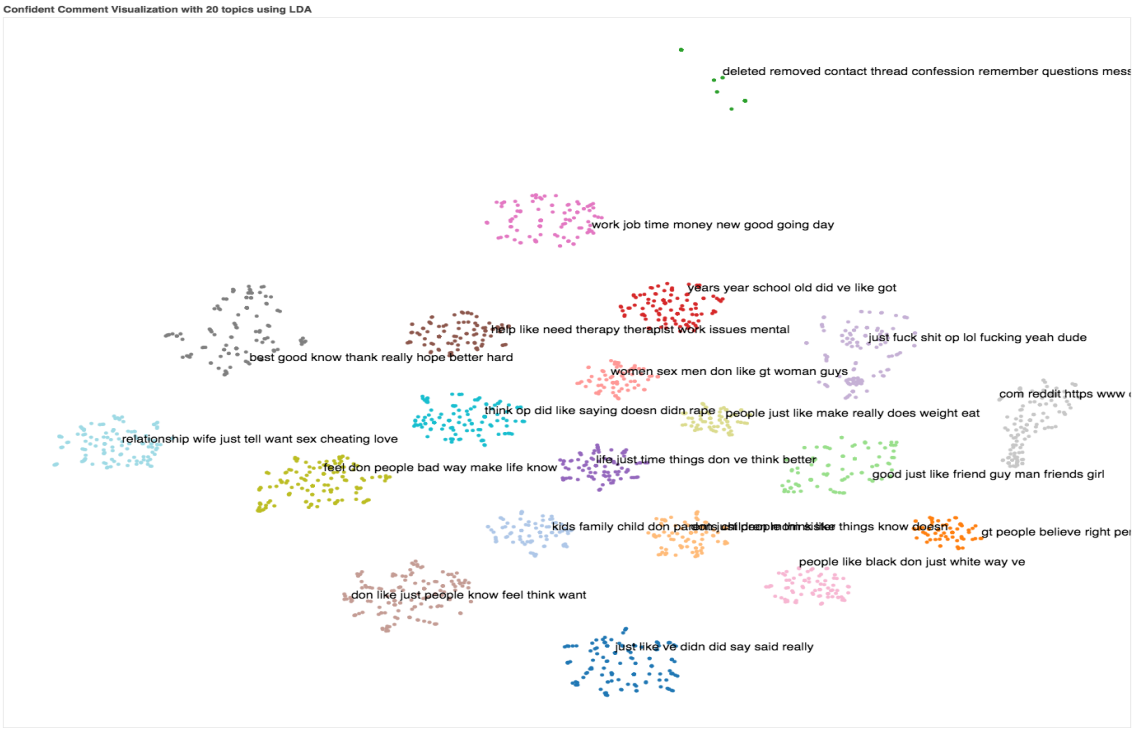

Figure 4: Scatterplot with Comments: $Pr(Comment \in Topic_c) > 0.5$ using LDA, 20 Topics

For each topic distribution, the following Kullback-Liebler divergences and divergence errors were obtained for each run through t-SNE:

	Kullback-Liebler Divergence	Divergence Error
LSA All Comments	81.078	1.228
LSA $Pr(Comment \in Topic_c) > 0.5$	-9.186	-9.336
LDA All Comments	88.055	1.639
LDA $Pr(Comment \in Topic_c) > 0.5$	47.960	0.192

It is worth noting that for the two topic distributions generated using LDA (Figure 1 and Figure 4), the negative log likelihood of LDA is extraordinarily high, at 1328844. Thus, the model is not very confident in the topics generated.

Interpretation

0.1 General

Although reasoning about topics using their most frequent words is often challenging and unreliable,[4] one can clearly intuit certain subjects revealed by the topic models. For example, in the upper-middle left side of Figure 4, the topic "best good know thank really hope better hard" seems to reflect comments that are comforting, or at least positive; perhaps words of encouragement for someone who is confessing a difficult experience they went through. Others such as "relationship wife just tell want sex cheating love" and "work job time money new good going day", middle left and center top of Figure 4, respectively, seem to reference a confession's subject matter.

0.2 Model Comparison

The results show that the topics generated by LDA and the "topics" generated by the document term matrix decomposition of LSA cannot be directly compared when a confidence threshold is applied. This is clearly shown by the nonsensical, negative error and Kullback-Liebler divergences calculated for LSA $Pr(Comment \in Topic_c) > 0.5$. However, imposing a confidence threshold of 0.5 to LDA significantly boosts its performance; we can see this improvement through the lower KL divergence and error of this model over regular LDA and LSA.

0.2.1 Confidence

The results imply that LDA is more confident in classifying Reddit comments. Figures 1 and 2 both plot all the comments, but far fewer points in Figure 3 cross the 0.5 threshold than do so in Figure 4.

0.2.2 Topic Distribution

As shown in Figure 2, LSA created a far more skewed topic distribution. Most of the comments in Figure 2 were classified under the orange topic, whereas no 1 topic dominates the distribution of Figure 1. Additionally, it can be seen in Figure 4 that the less skewed distribution of comments is maintained even when the confidence threshold is imposed. Therefore, not

only does LDA classify most comments under a wider range of topics, it also classifies them more strongly.

Final Thoughts

Latent Dirichlet Allocation seems preferable to Latent Semantic Analysis in modeling the topics of r/confession. It also models topics relatively well, as Figure 4 holds useful insights for the types of discussions that occur in r/confession.

References

[1] Felipe Hoffa. 1.7 billion reddit comments loaded on bigquery. https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery.

[2] Shuai Wang. Topic modeling and t-sne visualization. https://shuaiw.github.io/2016/12/22/topic-modeling-and-tsne-visualzation.html.

[3] Mike Bernico. Latent semantic analysis. https://github.com/mbernico/cs570/blob/master/module_1/lsa%20text.ipynb.

[4] Martin Wattenberg, Fernanda Vigas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.