CSE515T Status Report

Jesse Huang

April 2, 2018

1 Progress

I have obtained and cleaned the data provided by Bigquery. My technology stack is Python with the Natural Learning Tool Kit, and a LDA package. I am storing my data localling with SQLite3. Currently, I am working on getting this data into document-term matrix form, so that I can feed it into LDA and LSA algorithms. 10 sample 'cleaned' comments are shown below:

2 Concerns

The reddit comment database is broken up into separate datasets by month. As such, I unsure how to sample uniformly from the database as a whole. Currently, I have been randomly sampled 200 comments from each month, but feedback would be appreciated on whether this is a solid choice to have made in building my local dataset. This problem is further complicated by BigQuery's limit of 30TB/month of free processing, of which I have used roughly 40%

3 Pivots

This limit has made my original proposal unfeasable, as to sample comments from all of Reddit would use up my processing limit to obtain relatively few samples. Because of this limit, and the extremely wide range of topics of the site as a whole, I have decided to focus on analysing the comments from a single subreddit, r/confession. I believe this choice also helped mitigate some of the difficulties that LDA has in modelling conversation, as most comments within this specific subreddit are centered around responding to confessions, as opposed to other comments.

4 Response to Feedback

Thank you very much for the feedback that was provided for my initial report. The questions posed are addressed below:

Why is feature selection involved in topic modeling, say with LDA?
It is not. I had initially proposed feature selection under with a misunderstanding of topic models.

How would I proceed to sentiment analysis based on topic modeling results?
One method I was considering is to look at how the topic models change overtime. However, I have not found much formal literature on established ways to proceed with sentiment analysis through topic models, only this somewhat informal article. As such, suggestions as to how to proceed would be welcome.