STAT 428: Individual Project Journal

Fu, Xingyu, xingyuf2

December 13

- Group meetings
- Reading Materials
- Research Thoughts
- Time spent working on code In general
- Group Report Writing
- Feelings of accomplishment or frustration you have as a result of your project making progress or not and why.

Group meetings

1. For Proposal 1: November 11.

First group meeting. Familiarized teamates. Finalized topic on LDA. Drafted the first proposal together.

11.13-11.18: Did the readings below for LDA.

2. For Proposal 2: November 25

Finalized what to do for the project and allocated work of coding.

11.25-12.2: Coding Work.

3. For Coding Work Integration: December 2

12.2 - 12.5 Exported LDA results and did evaluation.

4. For Final Report: December 9

I explained the output results. We wrote our final report.

Reading Materials

1. Mixture Language Models, PLSA, LDA, and EM Algorithm slides from CS 510 lecture notes: http://times.cs.uiuc.edu/course/510f18/schedule.html (http://times.cs.uiuc.edu/course/510f18/schedule.html)
2. Latent Dirichlet Allocation: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)
3. EM Algorithm: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm (https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm)
4. The EM Algorithm: An Optimization View. http://times.cs.uiuc.edu/course/510f18/notes/em-algorithm.pdf (http://times.cs.uiuc.edu/course/510f18/notes/em-algorithm.pdf).

Research Thoughts

We started the idea of lda because we're interested in text mining area using statistical methods. We first got to know about the principles of mixture language models, which extends naturally to principle of PLSA model. Converting it into the Bayesian version, we get our lda model. Then we looked into the EM algorithm to see how statistical methods e.g. optimization, introduced in 428 could help in lda. We also looked into corporating Random Number Generation into the whole lda model.

Time spent working on code In general

1. I did data collection on the raw json data (https://www.yelp.com/dataset/challenge (https://www.yelp.com/dataset/challenge)).
2. I did data cleaning, data filtering, data extraction, along with data processing for lda model in the code. — about a week
3. It successfully extracts and cleans reviews from yelp data, and transforms it into input to the lda model.
4. It needs to be fixed on the category extraction part.
5. I also helped in looking for lda reference resources and exporting the lda result for result presentation. — 2-3 days

Group Report Writing

I mainly worked on LDA and EM's introductions along with the experiment and data-presentation part.

Feelings of accomplishment or frustration you have as a result of your project making progress or not and why.

I had feelings of frustration since the our group proposed to use a lda model that was hard to understand at first.

When we had a good output, I felt accomplished since our group finally worked the project out.