STAT 428: Individual Report

Fu, Xingyu, xingyuf2

Group 15 – Yelp Review Classification with Latent Dirichlet Allocation Model

- 1. group member names and netids, identifying the group leader.
- 2. Individual contribution section:
  - My contribution to the group project.
  - The outcome of the project.
- 3. Evaluation of My team members section
  - My evaluation of each group members
- My ProjectJournal files
  - 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.

1. group member names and netids, identifying the group leader.

Leader: Yifan Zhang (yifan8)

Members: Xingyu Fu(xingyuf2), Haonan Wang(haonan3), Yixiang Yu(yyu50), Qiaoge Zhu(qzhu18)

2. Individual contribution section:

My contribution to the group project.

For the coding part, I took charge of data processing and helped in experiments. Specifically, I did data collection on the raw json data (https://www.yelp.com/dataset/challenge (https://www.yelp.com/dataset/challenge)), data cleaning, data filtering, and data extraction. These took me about a week since the yelp data was origanaly json file and was pretty messy. I also implemented data processing to transform data into Bag-of-words as input to our lda model. For our experiments, I extracted a business category list from the yelp data as a supplement file to compare with our LDA model's output. I helped to run the lda model and studies the result. I did the result explanation in the final report. I also helped in evaluation the model output. For the writing parts, I wrote the lda and em's principle parts.

The outcome of the project.

I got the following 20 extracted business categories from the whole yelp dataset: Restaurants; Shopping; Beauty & Spas; Bars; Home Services; Health & Medical; Local Services; Automotive; Event Planning & Services; Active Life; Coffee & Tea; Hair Salons; Home & Garden; Auto Repair; Hotels & Travel; Arts & Entertainment; Professional Services; Doctors; Real Estate; Pets

Here, this 20 business categories are the ones that we think could best describe business categories of the yelp dataset. Then, we applied LDA model on an sampled corpus and ran until convergence. Then we manually evaluate the sampled results for sentences like: [Restaurant "… if you want a great atmosphere no attitude mixed crowd fantastic service and reasonably priced food bev taste phoenix's newest cherry and tell me what you think …"]. Based on our study, the LDA output could sucessfully classify the review into the correct business category.

3. Evaluation of My team members section

Our group generally works on the project together evenly. We wrote group proposals and group report and held project meetings together. We all participated and offered ideas actively. Our group leader Yifan takes more charge in final modifications before submission.

Haonan searched tutorials of LDA and EM， helped in code writing for data processing and for em implementation.

Yifan worked on the implementation of the whole lda model and the em algorith (optimization methods).

Qiaoge and Yixiang helped in implementing RNG method and optimization methods.

All five of us evaluated 60-90 LDA model output on yelp reviews manually for the final evaluation. We did the final report together.

My evaluation of each group members

Yifan Zhang (yifan8): 10

Haonan Wang (haonan3): 10

Yixiang Yu (yyu50): 5

Qiaoge Zhu (qzhu18): 10

My ProjectJournal files

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.