Final Project (STAT 5870)

The goal of this project is to learn more about text analytics applications (Sentiment analysis, Text classification, Topic modeling, etc.). Pick a real text data set for which you believe there are interesting questions to answer. Choose one or more text analytics applications, and study the methods or tools people use and apply to your data set using Python. Feel free to do a Google Search to find interesting text analytics applications. Since text mining and text analytics are relatively new research area, you will find more information online than textbook. Of course you can't directly copy and paste what others published online, but you can still use those materials as a reference for your project.

Our final project will be an individual project but if you really want to work as a group then you can form a group with maximum three people. If you form a group then make sure to send me an email about information of your group members. If you work as a group then I am expecting more from you compared to working as an individual.

I don't want you to choose too ambitious project because then you might not able to finish. At the same time, I don't want you to choose too simple project. But don't worry since I will give you a feedback on your project proposal whether it is an appropriate size or not.

1. **Project Proposal** (Due by 11:59 pm, Jun. 20)

   (a) Description of the data set

   (b) Description of the application

   (c) Brief explanation about methods and tools you are trying to use.

   Maximum 1 page. The proposal should be concise and clear. You also need to include the citation for the data set of choice.

2. **Project Report** (Due by 11:59 pm, Jun. 30)

   (a) Introduction: Describe the data set and the problem you want to solve.

   (b) Methodology: Explain the methods and tools you used.

   (c) Result: Summarize the results.

   (d) Conclusion: Summarize your overall findings and potential direction of the future work.

   Maximum 5 pages. Tables, figures, or Python codes are not counted. The report must also include your Python code in appendix or submit separate .py file together with your report.

Grading

Final project is 25% of total grade.

- Project Proposal: 5%
- Project Report: 20%

Data Repositories

- https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38
- https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/
- Kaggle (https://www.kaggle.com/datasets)
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=text&sort=nameUp&view=table)