Proposal

Team GGWP

Q: What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Yueyi Gao (yueyig2); Ted Chen (jundong2) as captain

Q: What is the type of your project: Is it Data Set Creation or Leaderboard Competition Creation?

Dataset creation.

Q: If your project is Data Set Creation, what is the novelty of your data set as compared with all the existing data sets? Which of the existing data sets is the closest to yours? What new task can your new data set be used to evaluate? How do you plan to create the data set?

Our group aims to create a new dataset including major news that can be used for text mining tasks such as semantics analysis and search engine evaluation. This dataset includes most of the recent news from some major news sites and we, as the human judges, will annotate the topics of each article piece from the dataset. There are some existing datasets focusing on the collection of research articles and social media posts. However, for a more comprehensive evaluation of the text system, a dataset including all the aspects from technology, finance, fashion, and lifestyle should be included as part of the evaluation system. This dataset is mainly focused on the evaluation of the developing text mining system though it can also be used for semantic analysis, sentiment analysis, and topic mining analysis. We plan to develop some web crawlers to fetch the news from some major news sites [See the lists below]. We may also consider building a data pipeline for stream processing the data if time permits.

https://www.cnn.com/ ; https://www.bbc.com/ ; https://www.bloomberg.com/
https://www.foxnews.com/ ; https://www.newyorker.com/ ; https://news.yahoo.com/
https://www.npr.org/ ; https://www.theatlantic.com/ ; https://www.nbcnews.com/
https://www.theguardian.com/us ; https://www.reuters.com/ ; https://www.vox.com/
https://www.washingtonpost.com/ ; https://www.nationalreview.com/
https://www.nytimes.com/;