Lightning Talk

Project Concept and Supporting Data

Analyzing Text

With the large volume of text now available to us, there is significant opportunity to extract information from the text for analysis.

The goal of my final project will be to understand tone and context of a data set, including sentiment and meaning.

Option 1: Twitter Sentiment Analysis for Airlines

Problem: Many people take to Twitter to complain about airline service. What are they saying and what is the sentiment behind it?

Data: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Data fields: tweet_id, airline_sentiment, airline_sentiment_confidence, negativereason, negativereason_confidence, airline, airline_sentiment_gold, name, negativereason_gold, retweet_count, text, tweet_coord, tweet_created, tweet_location, user_timezone

Hypothesis: Negative sentiment will dominate. There will be clusters of complaints around particular incidents.

Related Research: Linguistic Features for Twitter Sentiment analysis
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251?height%3D90%%26iframe%3Dtrue%26width%3D90%

Option 2: Rumors and Sentiment

Problem: Too many rumors. Are they affected by sentiment?

Data: https://www.kaggle.com/arminehn/rumor-citation

Data Fields: mergent_page claim, claim_description, claim_label, tags, claim_source_domain, claim_course_url, date, body, page_domain, page_url, page_headline

Hypothesis: Rumors will have certain sentiment patterns

Related Research: https://www.media.mit.edu/cogmac/publications/Soroush_Vosoughi_PHD_thesis.pdf

Option 3: Repeated Topics Extraction on Quora

Problem: What types of topics are most frequently asked on Quora?

Data: https://www.kaggle.com/c/quora-question-pairs/data

Data Fields: id, qid1, qid2, question1, question2, is_duplicate

Hypothesis: The most popular types of questions will be non-technical questions since there are alternate forums for technical questions

Related Research: