Susan Kuretski
skure2@illinois.edu
CS410 - Fall 2020

Course Project Proposal

Team Members

Team of one: Susan Kuretski (de facto captain) - skure2@illinois.edu

Description - Option 5: Free Topic

Task

The objective of this project is to do sentiment analysis on images, also known as memes, from Reddit. Using Google Vision OCR, the characters from the image will be processed to tokenized strings/words, essentially this transforms the text from the meme into a bag of words. Using a probabilistic approach to sentiment analysis, I would ideally like to use naive Bayes with Laplace estimation to avoid the assignment of zero probabilities. I have never implemented an application using naive Bayes, so if this ends up being a rabbit hole, I may have to switch gears to a probabilistic semantic analysis (PLSA).

Importance/Interesting

In general, sentiment analysis is useful in determining users' feelings and attitudes towards certain items, whether it is a review, comment, or product. In regards to Reddit, it would be interesting to see which memes gather positive or negative sentiment and whether it correlates to upvotes or downvotes.

Planned Approach (Tools, systems, datasets, evaluation) with Time Estimates

Task	Time Estimate
1. Get dataset from Kaggle here: https://www.kaggle.com/sayangoswami/reddit-memes-dataset	-
2. Do a data cleaning pass to ensure all URLs of images are seemingly correct in terms of structure, downvotes and upvotes are integers >= 0.	0.5 hour

3. The dataset has 3327 images to be downloaded and/or stored in the cloud. Iterate through each entry in the dataset to fetch the image and store. Discard invalid images.	3 hour
4. Run images through Google Vision OCR. https://cloud.google.com/vision/docs/ocr	3 hour
5. Evaluate accuracy	
6. Store results of OCR. I want to minimize the need for repeat OCR processing since it can become expensive.	1 hour
7. Transform data from OCR to usable dataset for naive Bayes (bag of words, maybe try n-grams).	4 hour
8. Define and fit the model. Use scikit-learn for Python.	10 hour ** No previous experience with naive Bayes or scikit-learn
9. Evaluate model using scikit-learn metrics and comparing with upvote/downvotes.	4 hour ** No previous experience with evaluation of this
10. Attempt to make the model better (iterate steps 6 - 8).	10+ hours?
	Total: 35.5 hrs

Expected Outcomes

The expected outcomes are:

1. Have a final percentage of accuracy from using naive Bayes with hopes of it being greater than 50%
2. Implement improvements to accuracy if initial accuracy <= 60%
3. Possible factors affecting results: slang and intentional misspellings in memes, inaccuracy of OCR

Programming Languages and Systems

Python with scikit-learn
AWS s3 storage or Google Cloud Storage
Google Vision API