Depression and Suicidal Ideation Detection (NLP/Machine Learning)

Question: Can we detect depression or suicidal ideation with natural language processing?

Aim: To be able to read a personal text written by someone and predict with high accuracy if the writer is depressed/in danger.

Tools to Use:
(will narrow these down)

NLP Librarys:
CRF Suite Python.
PySpark
SparkSession
SciKit (Python)

Data We'll Start with:
Challenge - making sure control (happy) training data is personal enough to be as informative about the writer as the sad data

Testing Data:

- Reddit thread asking people to summarize their lives - link

Training Data:
(might need to parse through all the data we have and manually classify the more ambiguous texts to happy or sad - eg. texts with a neutral sentiment)

- subreddit r/casualconversations
- Subreddit r/suicidewatch

Timeline:
By Saturday (4/7)

- Create basic website - Rommel
- Set up flask app for backend - Rommel
- Scrape data from relevant subreddits and threads - Allen
- Create MongoDB with all the scraped - Allen
- Evaluate and choose the ML library to use - Jenny and Mehreen
- Research the process of putting data through and have some solid data to put it through - Jenny and Mehreen

By Tuesday (4/10)

- Clean and manually check data as much as possible
- Create model
- Put data through process and test accuracy
- Change + adjust as needed

By Thursday (4/12)

- Finish up final website which gives:
    - Summary of project
    - Examples of happy/sad texts
    - Takes input from user (eg. paragraph) and gives an alert if the words hint towards them being depressed

By Saturday (4/14)

- Present!

More Details

Dataset: Scrape Reddit r/SuicideWatch for posts. (We assume that people posting there are having suicidal ideation or at least are very dissatisfied about the trajectory of their life). (https://www.reddit.com/r/SuicideWatch/)

Training: Use at least 10,000 posts to accumulate the data required to make a training set.

Testing: Use posts from r/SuicideWatch as positives, and other texts (where the author is not sad or depressed) as negatives.

Metric: We would like to see a probability or score that the person who wrote the text is likely depressed or has suicidal ideation.

Presentation Website:

Main page: has a field that allows for text to be typed or pasted into it. Once the Enter/Submit button is clicked, it spits out a probability that the text is associated with someone who is either sad or has suicidal ideation.

Methodology Page: Explains the steps that we took and the data we used to train our model. We would also explain what the model actually does in order to come up with the score.

Technology Page: Lists and diagrams the different technologies that we used in order to make create the application.