OH SNAP!
REDDIT PROJECT

ALEX NGUYEN
DSI-ATX

DATA SCIENCE PROBLEM?

How can we scrape reddit data and use NLP classification to train a model that is able to differentiate between two subreddits.?

Subreddits chosen:

r/thanosdidnothingwrong and r/inthesoulstone.

Using push shift api, I was able to scrape 36983 reddit posts.

thanosdidnothingwrong	Inthesoulstone
20, 000 posts (54%)	16983 posts (46%)

THANOS

WHY THESE TWO SUBREDDITS?

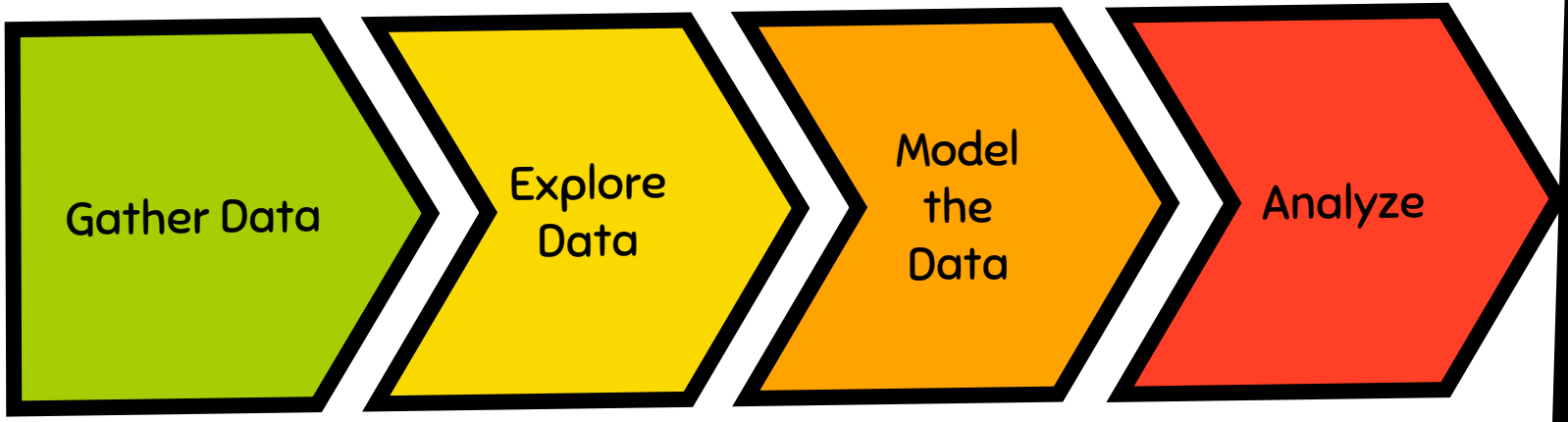
- On July 9th the subreddit thanosdidnothingwrong held an event to mimic and reenact a scene from the movie Avengers: Infinity War
- In the movie, Thanos brought balance to the universe and wiped out half of the population.
- To try and recreate this event moderators and Reddit admins allowed a mass ban to take place. (The largest ban in Reddit history)

350,000 REDDITORS BANNED

"THE END IS NEAR, AND WHEN I'M DONE, HALF OF HUMANITY WILL STILL EXIST. PERFECTLY BALANCED, AS ALL THINGS SHOULD BE..."

- THANOS

DATA WORKFLOW



Train Test Split > Setting up the models > Pipeline and used Gridsearch

MODEL

Types of Models used:

Count Vectorizer & TFIDF	Random Forest
Logistic Regression	AdaBoost

Pipeline with a gridsearch to optimize hyperparameters

Scores:

Logistic Regression training score = 78.73%
Logistic Regression training score = 65.73%

Random Forest training score = 76.99%
Random Forest testing score = 65.78%

Ada Boost training score = 65.78%

Ada Boost testing score = 64.71%

Ex: min_df=1, ngram_range=(1,2)
stop_words=None, strip_accents='unicode'

LOGISTIC REGRESSION MODEL PREDICTIONS

Subreddit	Predicted inthesoulstone	Predicted thanosdidnothingwrong
Actual inthesoulstone	2852	2753
Actual thanosdidnothingwrong	1431	5169

Precision = 66%

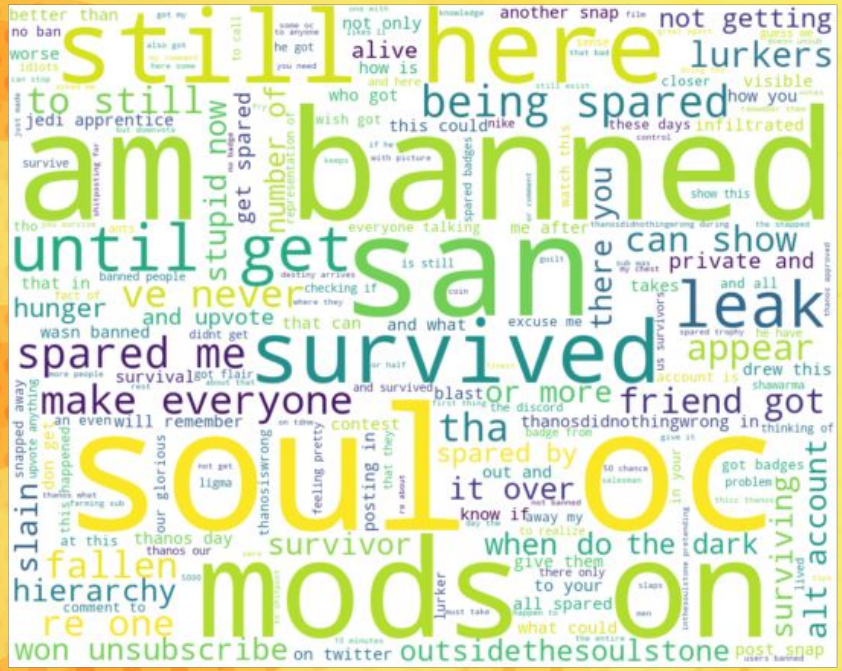
Sensitivity = 50%

Specificity = 78%

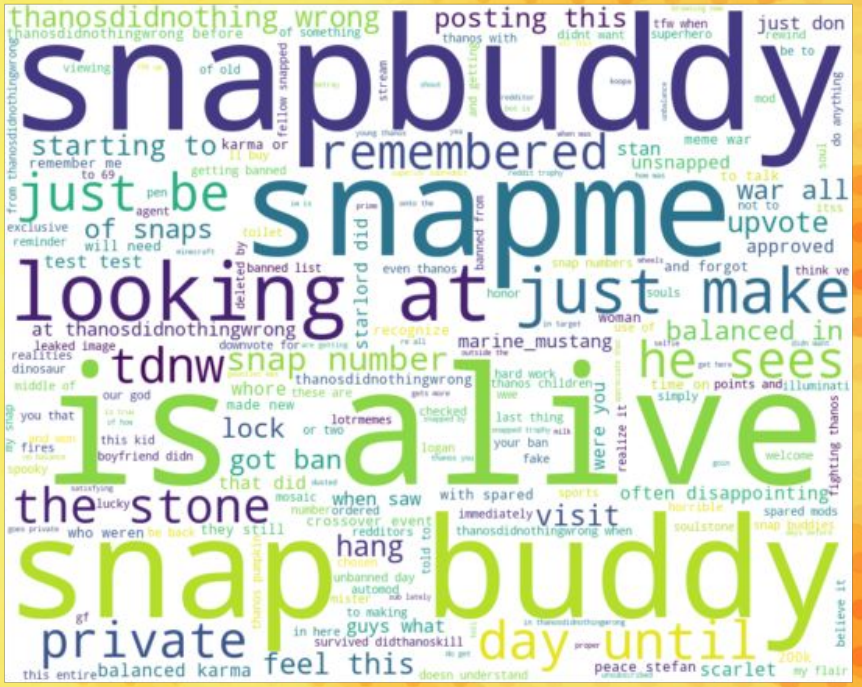
Accuracy = 65%

WORD CLOUDS

THANOSDIDNOTHINGWRONG



INTHESOULSTONE



FINAL THOUGHTS

- These two subreddits are very similar. Both are focused on the Marvel universe, a lot of the posts are memes!
- If I had more time I would have liked to apply NLTK to my model.
- Also I believe looking at different timeframes for the data being used heavily impacts these subreddits

THANKS!

ANY QUESTIONS?