Project 3: NLP & Reddit

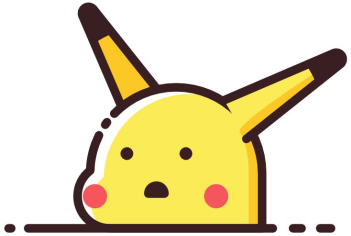
Vonn Johnson

Scenario

Reddit servers have gone down and in the process subreddit post have gotten mixed up!

Their developers were able to fix the problem, but it took weeks to organize posts back into their proper subreddit.

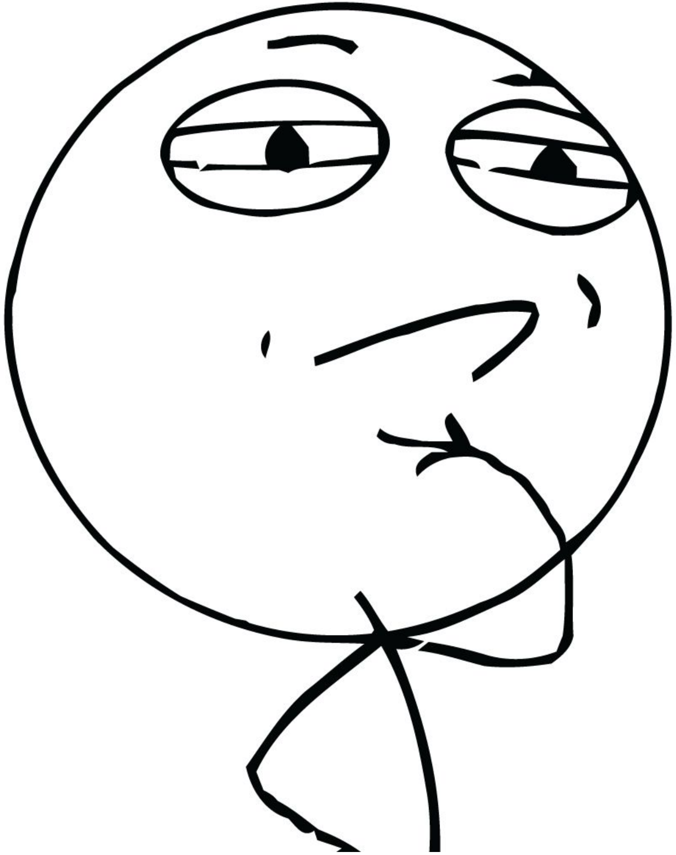
A team of data scientist are called to prevent this disorganization from ever happening again.



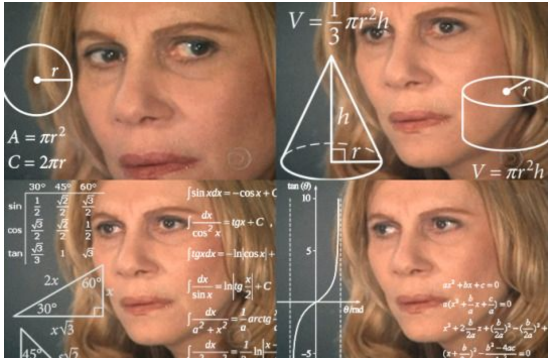
Objective

Based on post titles, can we predict where a post comes from using Natural Language Processing and Classification models?

CHALLENGE CONSIDERED

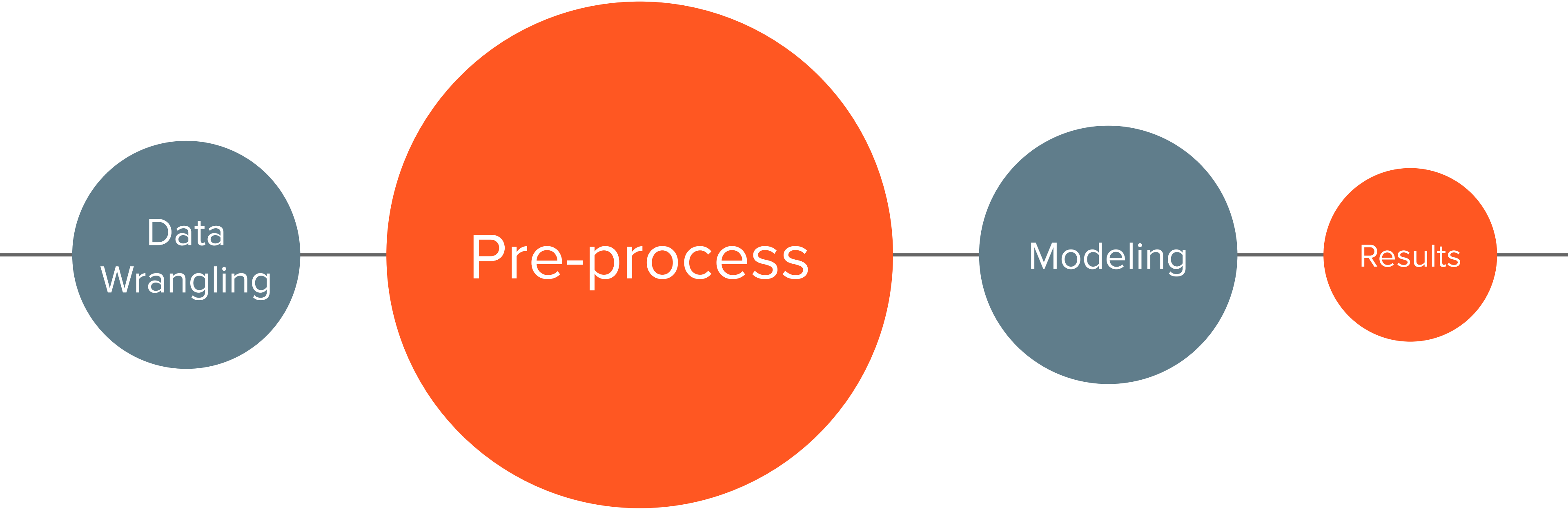


NLP?



r
A = πr²
C = 2πr
V = 1/3 πr²h
h
r
r
V = πr²h
30° 45° 60°
sin
cos
tan
2x
60°
30°
x√3
45°
∫sin xdx = −cos x + C
tan (θ)
10
5
−5
θ/rad
ax² + bx + c = 0

Methodology



Data Wrangling

Pre-process

Modeling

Results

Data Wrangling

- Extraction & Beautiful Soup
- Dataframes & Classify
- Test case: Politics v Stocks

Politics

Word	Word Count
Trump	383
white	129
Mueller	107
report	92
says	90

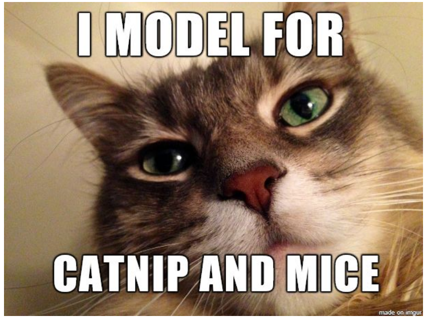
Stocks

Word	Word Count
stocks	156
stock	121
market	107
today	92
thoughts	90

Preprocessing

- Regex
- Stop Words
- Cleaning Round 1
- Cleaning Round 2
- Tokenizing

Modeling



- Binarize & Split
- Model Types
- Countvectorizor v. TFIDF
- Evaluation Method

Modeling: Deep Dive

Logistic Regression

- Similar to linear regression
- Predicts whether something is true or false, is or isn’t
- Classic classification model

Multinomial Naive Bayes

- Based off Bayes’ Theorem
- “Probability of an event occurring given the probability of another event(s) that has already occured”
- Naive: Assumed independence amongst features

Random Forests

- Based off Decision trees
- Uses bootstrapping to resample data set
- Uses Aggregation of multiple decision trees to draw conclusion.
- Aka Bagging

Results

(With a Baseline Accuracy score of 51%)

Logistic Regression

1st Place: .98 Accuracy on Train, .95 Accuracy on the Test

Multinomial Naive Bayes

2nd Place: .97 on the Train, .94 on the Test

Random Forests

3rd Place: .88 on the Train, .86 on the Test

After-Action Review

- Can be implemented
- Focus on subreddits w/ explicit relationship
- Try new models and evaluation methods
- Clean better

Thank You!