THE POWER OF REDDIT

PREDICTING SUBREDDIT'S THROUGH NATURAL LANGUAGE PROCESSING

SAM LUNDBERG

THE QUESTION

- Pulling info from Reddit threads
- What information can be gained?
- What Data Science tools are used?
- What is the outcome?
- What are the benefits of this?

THE SUBREDDITS

THE PROCESS

- Pulled 600 posts from each subreddit
- Cleaned the data
  - Removed Stop Words
- Used Natural Language Processes (NLP)
- Created models to make predictions
- Built additional models

THE RESULTS

- Used words in the Title to determine if in Star Wars subreddit
  - Achieved 85% accuracy rate
    - Used TF-IDF with a Ridge Penalty for best results
    - Count Vectorizer showed similar results
- Using the mean number of comments:
  - Achieved 93% accuracy rate with Random Forest

THE RESULTS

- Used specific words in the Title to determine if in Star Wars subreddit
  - Using the word “episode”
    - Achieved 91% accuracy rate
  - Using the word “new”
    - Achieved 90% accuracy rate

Word	Count
star	226
trek	151
wars	83
new	44
just	41
series	38
tng	36
episode	35
discovery	31
like	31
time	26
ds9	26
picard	24
think	24
did	23
jedi	22
watching	21
enterprise	20
does	20
space	20

NEXT STEPS

- Applying NLP to Reddits can be successful
  - Does the heavy lifting
  - Further tuning of stop words
  - Show which posts are useful to follow up on
  - Additional models to look through descriptions of posts
- Other use cases to scrape Reddit
  - See how people react to certain events or products