

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Getting Users to Engage

A Look At Comments on Reddit Posts

Christopher Messier

General Assembly

Washington, D.C.

November 3, 2017

Overview

Engagement

Messier

Reddit

Methodology

Models

Conclusion

1 Reddit

2 Methodology

3 Models

4 Conclusion

What is Reddit?

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Reddit is a large link sharing website

- "The front page of the internet"
- Types of posts
 - Content - links, pictures, videos, articles
 - Self Posts - text posts by the user
- Users share links with communities known as subreddits

Graph of Subreddits by User Cross Posting

Engagement

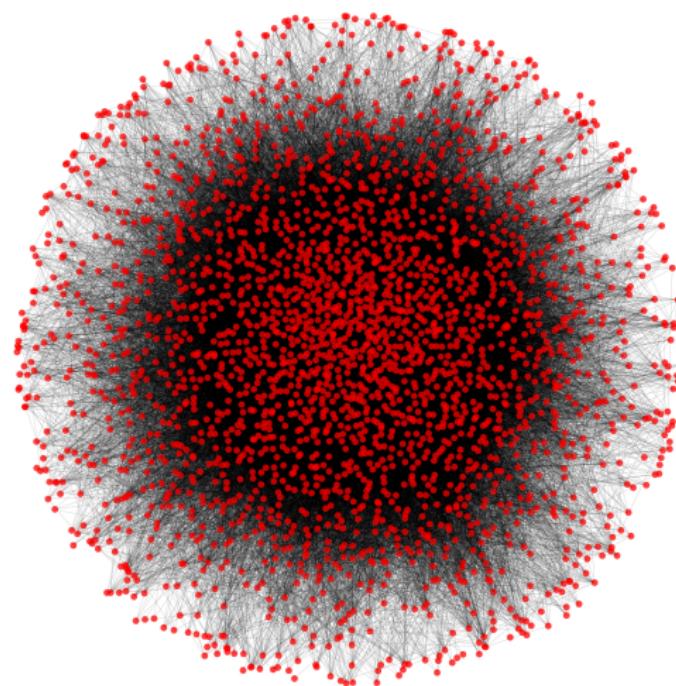
Messier

Reddit

Methodology

Models

Conclusion



Measuring Engagement

Engagement

Messier

Reddit

Methodology

Models

Conclusion

High Engagement posts will be defined as:

- Posts that have the top 10% most comments
- Binary Classification
 - $High\ Engagement = 1$
 - Everything Else = 0
- Determined for each subreddit

Why Comments?

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Reddit has several intrinsic post metrics:

- Upvotes/Downvotes
 - Count of other users choice to like or dislike a post
- Rank
 - The page position of a post, determined by its popularity
 - Comments capture active engagement with the site, interaction

Data Collection

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Data was collected during the week of 10/30/2017

- 2,923 subreddits
- 1,049,005 users
- 2,590,769 unique posts
- 42,374,242 comments

Data Collection (cont.)

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Collected using a proprietary python web scraper

- Attempted to get the most representative sample possible
- Used `r/random` to generate a random sample of subreddits
- Captured up to 2,000 most recent posts per subreddit

Omitted Data

Engagement

Messier

Reddit

Methodology

Models

Conclusion

The following data was removed/omitted:

- Duplicate posts
- Posts from bots, such as AutoModerator
- Non-english/non-ascii posts
- Adult Material

Omitted Data (cont.)

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Analysis was performed on a subsample, $n= 362,658$:

- Computational Reasons
 - Limited time
 - Limited processing power
 - Effect on model selection

Engagement

Messier

Reddit

Methodology

Models

Conclusion

TFIDF Vectorization of the title

- Text Frequency
- Inverse Document Frequency
- Approx. 2,538 word vectors

Sentiment Analysis

- Attempts to determine sentiment of titles from content

Modeling was performed using the python package Scikit-Learn

Models used include:

- Decision Trees
- k -Nearest Neighbors
- Random Forest

Scoring

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Models are evaluated using the following metrics:

- Precision:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

- Recall:

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

- F1:

$$F1 = \frac{2(P \times R)}{P + R} \quad (3)$$

Measuring the effects of the content of a post

- Used a decision tree model
 - Minutes since posting
 - Whether it's a question
 - Features outside content
 - Has an emoji in the title

Content Decision Tree Results

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Content Decision Tree Results

	Predicted True	Predicted False
True	89122	9352
False	9352	972

- Precision: 0.83
- Recall: 0.83
- F1: 0.83

Measuring the sentiment of a post's title, using nltk

- Used a decision tree model
 - Positive
 - Negative
 - Neutral scores

Sentiment Decision Tree Results

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Confusion Matrix

	Predicted True	Predicted False
True	98334	140
False	10308	16

- Precision: 0.83
- Recall: 0.90
- F1: 0.86

k-Nearest Neighbors

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Measuring the sentiment of the post title

- Full complement of features
- Lazy model
 - Computationally Inefficient
 - Keeps all observations in memory

k-Nearest Neighbors Results

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Confusion Matrix:

	Predicted True	Predicted False
True	97912	562
False	9927	397

- Precision: 0.86
- Recall: 0.90
- F1: 0.87

Measuring the sentiment of the post title

- Full complement of features
- Computationally Brilliant
 - Distributable
 - Efficient
 - *Accurate*

Random Forest Results

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Confusion Matrix:

	Predicted True	Predicted False
True	96340	2134
False	8382	1942

- Precision: 0.88
- Recall: 0.90
- F1: 0.88

As our ability to analyze and gain understandings from data grows, it leads us to some difficult problems

- Accuracy vs. Interpretability
- Difficulty distilling information into general consumption
- The article should be aimed at highlighting this gap, and help readers understand these difficulties

Engagement

Messier

Reddit

Methodology

Models

Conclusion

Thank You