Getting People to Engage

A Look At Reddit Posts

Christopher Messier

General Assembly

Washington, D.C.

November 3, 2017

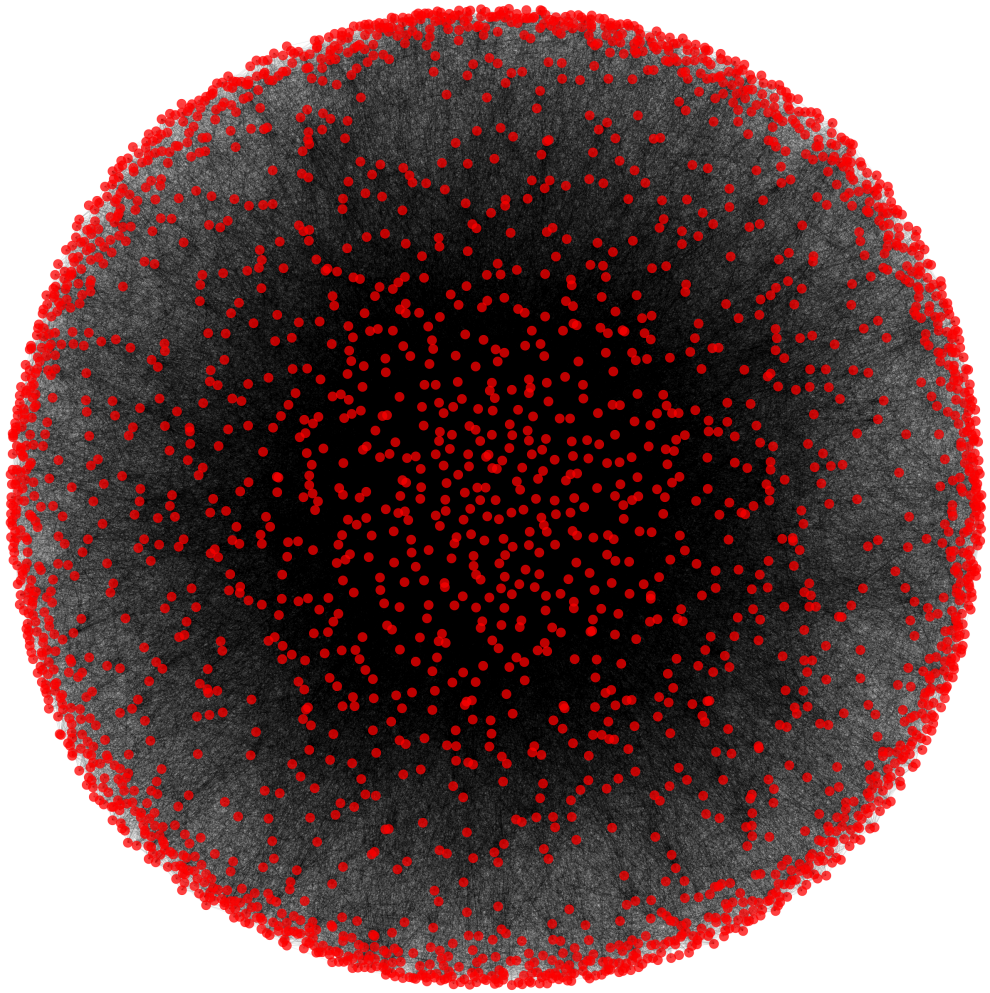
Overview

What is Reddit?

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



Measuring Engagement

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?

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

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.)

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

The following data was removed/omitted:

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

Omitted Data (cont.)

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

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

Natural Language Processing

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

Models

Modeling was performed using the `python` package `Scikit-Learn`

Models used include:

- Decision Trees
- k-Nearest Neighbors
- Random Forest

Scoring

Models are evaluated using the following metrics:

- Precision:

$$P = \frac{T_p}{T_p + F_p} \tag{1}$$

- Recall:

$$R = \frac{T_p}{T_p + F_n} \tag{2}$$

- F1:

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

Content

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

	Predicted True	Predicted False
True	27398	26740
False	27166	27494

- Precision: 0.50
- Recall: 0.50
- F1: 0.50

Sentiment

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

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

Sentiment Decision Tree Results

Confusion Matrix

	Predicted True	Predicted False
True	52569	1569
False	53061	1599

- Precision: 0.50
- Recall: 0.50
- F1: 0.36

k-Nearest Neighbors

Measuring the sentiment of the post title

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

k-Nearest Neighbors Results

Confusion Matrix:

	Predicted True	Predicted False
True	32203	21935
False	26670	27990

- Precision: 0.55
- Recall: 0.55
- F1: 0.55

Random Forest

Measuring the sentiment of the post title

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

Random Forest Results

Confusion Matrix:

	Predicted True	Predicted False
True	41136	13002
False	16322	38338

- Precision: 0.73
- Recall: 0.73
- F1: 0.73

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

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

Thank You