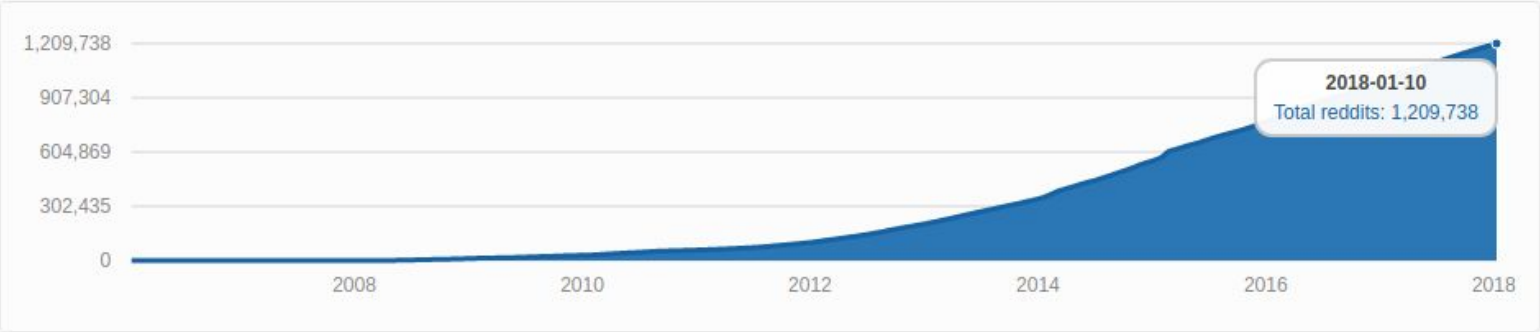
Reddit2Vec: Text Analysis and Recommendation

Connor Gouge

Data Scientist

Context on Reddit
"The Front Page of the Internet"

- Content Aggregator
  - 6th on the Alexa top list
- Content is divided into "subreddits"
  - Divided by subject: r/nfl , r/math, r/gaming etc.
  - There are over a million subreddits and more are created every day



graph by redditmetrics.com

The Problem to Solve:

- It is hard to find engaging content on a new social network or content aggregator

My Solution:

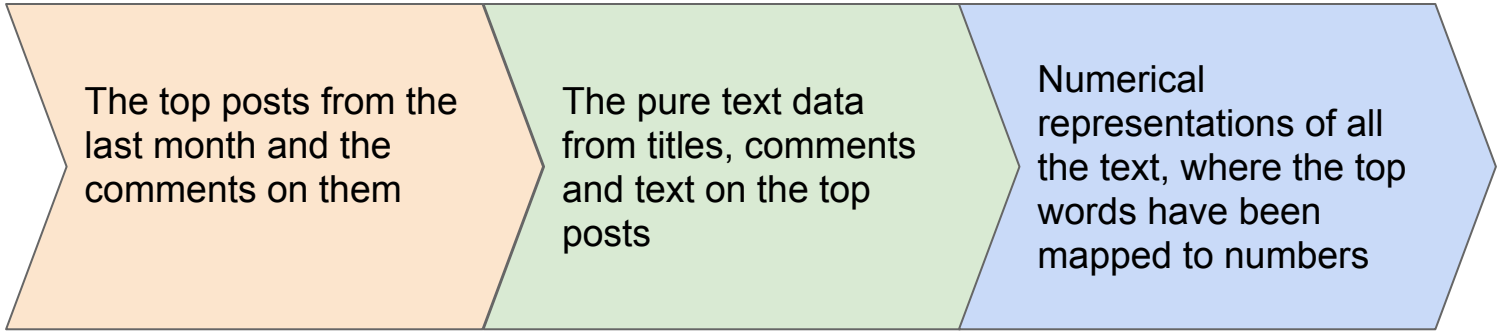
- Topic modeling on text data with NLP then classification of text into appropriate subreddits

www.findsubreddits.net

An interface to get recommendations based on your Twitter feed or some text that you provide

The Data Pipeline

- Around 3 million comments, titles and text posts from reddit
  - Scraped from up to 100 of the top posts in the last month on about 600 subreddits



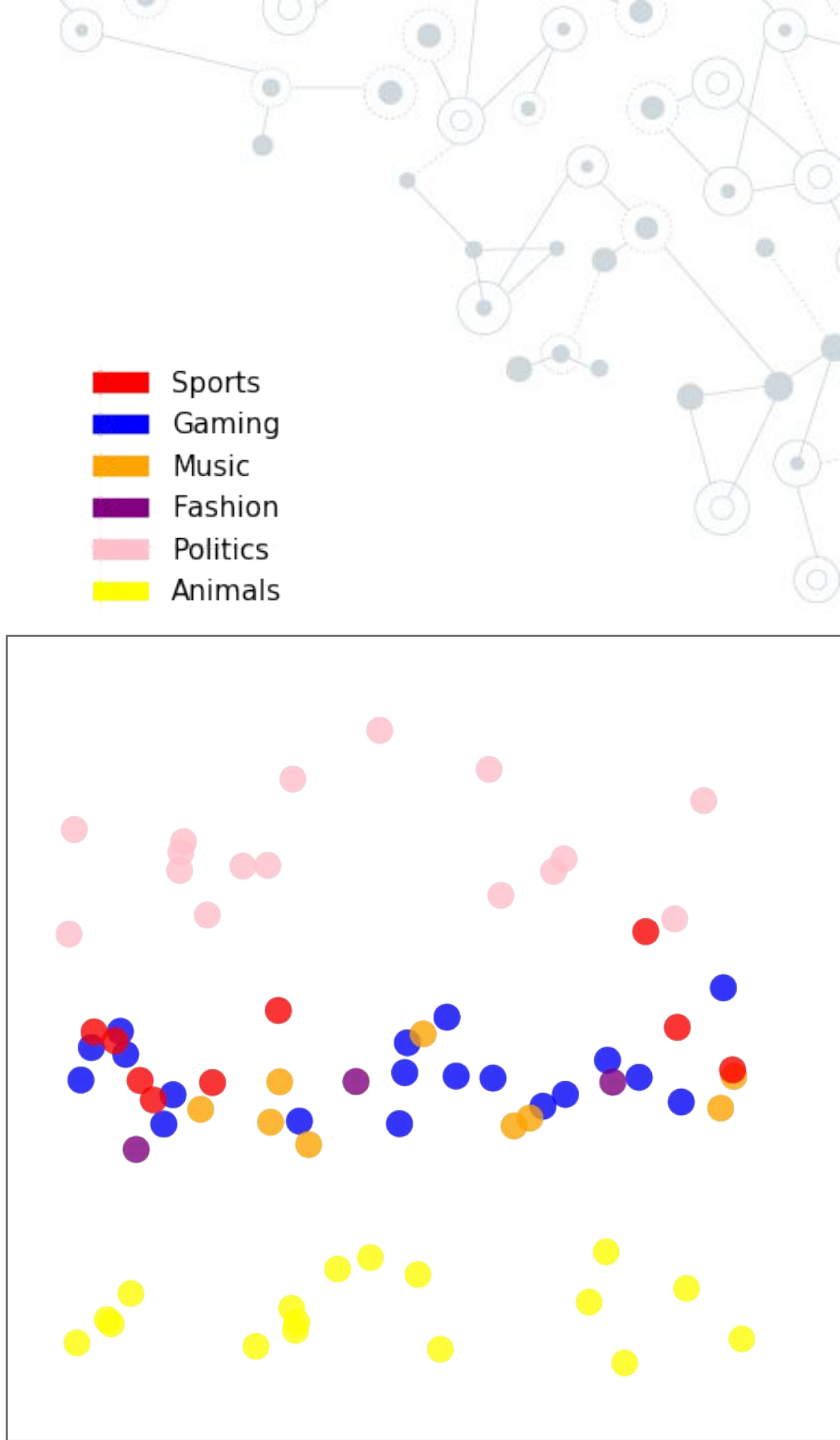
Building the Model

First Model: Clustering

- One vector for each subreddit, based on the average sentence in the subreddit

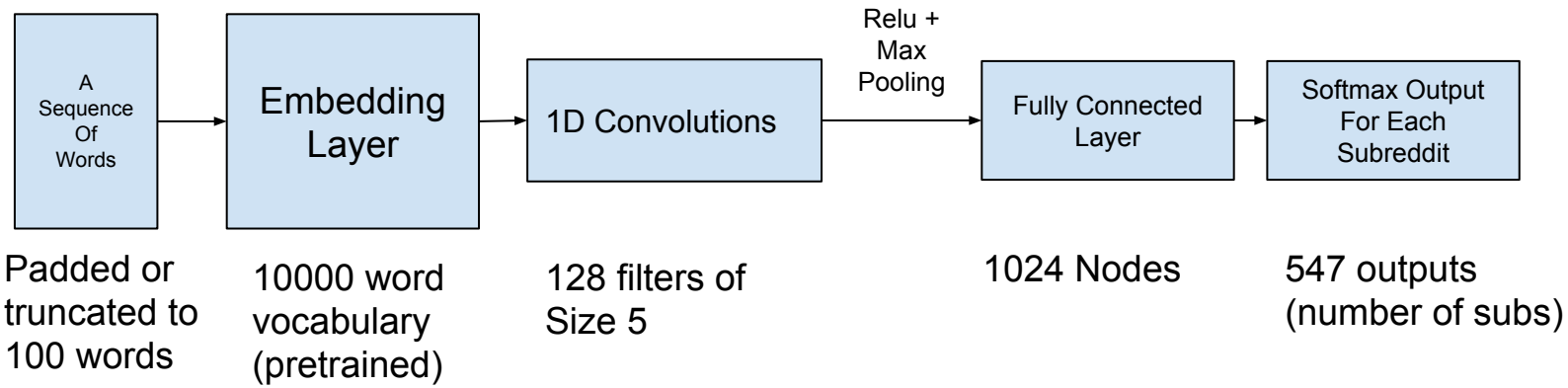
Cons:

- Slow to make predictions even with only my subset of ~600 subreddits
- Prediction quality had large variance depending on the type of content



The Current Model

- CNN Classifier
  - Makes predictions on which subreddit input content is most likely to be found in



Next Steps

- Expand and further curate subreddit selection
- Experiment with RNNs
- Experiment with CNN structure
- Further experiments with embeddings
- Bring in other types on content (images, articles, etc)

Tech Stack

python™

Keras

jupyter

amazon web services

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

NumPy

Flask

mongoDB

References and Open Source Libraries

- Tweepy
  - Twitter API wrapper for python
- pymongo
- PRAW
  - reddit API wrapper for python
- Gensim
- Beautiful Soup
- GloVe word embeddings
  - https://nlp.stanford.edu/projects/glove/
- The following articles and repositories:
  - https://github.com/adventuresinML/adventures-in-ml-code
  - https://github.com/tensorflow/models/blob/master/tutorials/embedding/word2vec.py
  - https://github.com/keras-team/keras/blob/master/examples/pretrained_word_embeddings.py

Connor Gouge

Data Scientist

www.findsubreddits.net

github.com/GougeC

linkedin.com/in/connorgouge

connorgouge@gmail.com