Proposal: Predicting Subreddits from Reddit Posts

Alexander Lamson <alamson@umass.edu>
Hari Hara Subramanian Krishnamoorthy <hkrishnamoor@umass.edu>

Problem statement

Reddit is a website where users can post about things that interest them on communities contained within reddit called “subreddits”. Between the subreddits, there are distinct words, phrases and manners of speech that distinguish themselves. Our goal is to predict the subreddit the particular post came from, given only the title and other text in the post. For feature engineering, we plan to try a bag-of-words representation, word2vec, and doc2vec. For models, we plan to test a fully connected neural network, a SVM classifier, nearest neighbors, a convolutional neural network, naive Bayes, and a random forest classifier. The first 3 models will be implemented by Alex and the last 3 will be implemented by Hari. In addition to simply creating models which perform with high accuracy, we want to find out what models perform really well on closely related as well as dissimilar classes.

Text Classification Challenges

1. We explicitly chose subreddits which were similar, which increases the difficulty of the classification task.
2. We are choosing not to exploit various domain specific characteristics (upvote score, post author, etc) to classify the data.
3. Posts can vary in length and use many words not found in common english, which makes our vocabulary large and our vectors sparse.
4. The vector representations of the documents do not capture semantic information thus making text classification hard.

Prior work

The paper [1] explores a variety of deep learning approaches such as CNN and RCNN to classify reddit posts to the correct subreddit in which RCNN performed very well with an accuracy of 54 percent on 20 categories. The problem with this paper is that they used 20 subreddits that are chosen based on how distinct they were from each other, thus making the classification task much easier. We propose to do similar experiments but with a very different dataset and analyze which models classify diverse classes and which models perform well when classes are closely related.

In paper [2] they used vector based classification techniques. They generated vectors with a bag-of-words with different order n-grams, as well as with Mikolov's Word2Vec [4] and Le's Paragraph Vector [5]. They used a naive bayes classifier using a bag-of-words with n-grams as their baseline. The classification models they used were regularized logistic regression, a SVM with a linear kernel and a boosted decision tree. We propose to use a SVM in a similar way as well as a random forest classifier.

The paper [3] discusses various techniques for text classification in general. The techniques proposed in the paper are Decision Trees, Naive Bayes, Neural Networks, Rule-based classifiers and other techniques like genetic algorithms. We decided to implement random forests and naive bayes because they makes decisions based on feature values like word counts. This makes sense for reddit posts because information about a particular word's occurrence can have an impact on accuracy. We decided to use neural networks because they are the current state of the art in methods for classification of textual data.

Dataset

The dataset is publicly available for download here: https://github.com/linanqiu/reddit-dataset

Total dataset contains threads scraped from 51 subreddits. We chose 10 subreddits as output classes.

Classes:

1. r/anime
2. r/comicbooks
3. r/dota2
4. r/leagueoflegends
5. r/conservative
6. r/libertarian
7. r/askscience
8. r/explainlikeimfive
9. r/gameofthrones
10. r/thewalkingdead

When looking at data for r/thewalkingdead, there were 73,408 posts, 450 of which had been deleted, 2,394 of which had no text. We expect similar counts for the other subreddits.
Our reasoning for choosing these particular subreddits was to choose pairs of subreddits which had some similarity to each other (r/anime being similar to r/comics for example), and very different from the subreddits chosen outside of the pair (r/anime is very different than r/explainlikeimfive).

Approach

We plan to split up the data into 70 percent training data, 10 percent validation data, 20 percent test data. To implement the models, we plan to use Scikit-Learn, Gensim and Tensorflow for certain parts of the code. As a preliminary experiment, we will establish a baseline by finding the maximum occurring class and predict it for all the inputs.

REFERENCES

[1] Classifying Reddit comments by subreddit (2016) , Jee Ian Tam
https://web.stanford.edu/class/cs224n/reports/2735436.pdf
[2] Text Classification of Reddit Posts (2015), Jacqueline Gutman Richard Nam
https://jgutman.github.io/assets/SNLP_writeup_gutman_nam.pdf
[3] Aggarwal, C.C. and Zhai, C., 2012. A survey of text classification algorithms. Mining text data, pp.163-222.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.7788&rep=rep1&type=pdf
[4] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
[5] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.