Utilizing Clustering Techniques to Generate Reddit Recommendations

Jonathan Wiseman

Elizabethtown College

(717)857-4069

wisemanj@etown.edu

**ABSTRACT**

Document clustering is an application of cluster analysis that operates on collections of text. Its aim is to produce a clustering such that documents in the same cluster are more similar to each other than to documents in other groups. Reddit is a popular content aggregation site that allows users to make and comment on “posts” in dedicated communities known as “subreddits.” Treating each subreddit as an aggregation of comment data allows the creation of documents that can be used in clustering. Document clusterings can be used to make personalized recommendations for active users that have defined interests. Additionally, subreddits can be visually represented based on their textual content, enabling a holistic view of Reddit’s communities.

# INTRODUCTION

Reddit is a popular news aggregation and content hosting website that allows users to form directed communities, called subreddits, for discussion of relevant topics. The fifth most visited site in the United States, Reddit boasts upwards of 300 million unique users. Users can share content in a variety of ways: uploading of original content (including images and videos), links to content hosted elsewhere (such as news articles), and by making text-only “self-posts;” users are also able to comment on content. A positive-negative voting system, collectively referred to as karma and accumulated through “upvoting” and “downvoting” content, gives explicit feedback on comments and posts.

Cluster analysis is an unsupervised learning technique that seeks to group a set of objects together such that objects in the same group are more similar to each other than to objects in other groups. Document clustering is a problem in cluster analysis that seeks to cluster together “documents” based on their natural language content. In this context, documents may be web pages, books, news articles, etc. Document clustering is being used for a variety of tasks in information retrieval, including search engine optimization, topic extraction, fast information retrieval, and resolving lexical ambiguity [1].

Document clustering often relies on the representation of documents as word vectors; this is known as the bag-of-words approach. Bag-of-words entails representing each document as a sparse matrix consisting of binary attributes (with features being words: a 0 denotes the absence of a word, a 1 indicates the presence of a word) for each word in the collection’s vocabulary. To reduce the size of the vocabulary, stop words (common words like: a, the, is, etc.) are removed and some textual preprocessing (like lemmatization or stemming, both of which seek to get the root form of words) are performed [2]. Next, a measure of similarity or distance (called the metric) between word vectors must be chosen. Examples include Euclidean distance, cosine similarity, the Jaccard coefficient, Pearson’s correlation coefficient, and averaged Kullback-Leibler Divergence [2]. With document vectors and a sufficient similarity metric, a clustering algorithm must be chosen. Examples include k-means (in which documents are separated into k groups based on distance from a group’s centroid) and hierarchical clustering methods. Additionally, other preprocessing can be performed to increase model performance, such as feature selection (further selection of important words), word weighting (ex. using tf-idf weights), and standardizing of vector lengths to 1 to avoid long documents dominating the analysis [2].

# Design

Figure 1 displays the overall project outline:

A picture containing map

Description automatically generated

**Figure 1: Project Steps Overview**

## Approach

## Datasets: Cleaning, Calculating, Merging

## Machine Learning Algorithms

## Exploratory Data Analysis

# Experiments

# Conclusion

# REFERENCES

1. A. Di Marco and R. Navigli, “Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction,” *Computational Linguistics*, vol 39, no. 3, August 6, 2013. [Online]. Available: <https://www.mitpressjournals.org/doi/10.1162/COLI_a_00148>. [Accessed September 30, 2019]
2. A. Huang, “Similarity Measures for Text Document Clustering,” *Proceedings of the 6th New Zealand Computer Science Research Student Conference*, January 1, 2008. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.4633&rep=rep1&type=pdf>. [Accessed September 26, 2019]