CS221
Project Proposal

Matthew Radovan (mradovan 06250254)
Christopher Cross (chrisglc 06281089)
Sasankh Munukutla (sasankh 06124893)

Labeling News Headline Topics through Unsupervised Learning

Motivation

News is constantly being updated, from the local scale to the global scale. With this huge range in scales comes a huge range of topics, and it can be hard for an individual to understand the whole landscape of major news topics and see where a given article may lie. Moreover, with the rise of fake news, it is critical to be able to discern the overall meaning easily. As a result, we decided to base our project around identifying the major topic in a given news article.

Project

Our project's core goal is to be able to take a given news article as an input and receive the article's primary topic, as a set of key terms. The set of possible topics will not be predetermined; instead, they will have been automatically found through clustering the articles and identifying the topics of the articles in those clusters. As a result, passing a new news article to the system should cause the system to assign the article to a cluster and return the topic associated with that cluster. Although this system has the potential to generalize to any news article, we will limit our scope to political news articles. This is in part because political news articles reflect major societal issues and concerns.

Evaluation

In order to evaluate the success of our clustering and topic extraction, we can evaluate it based on two metrics. The first metric would be to define a logistic regression classifier that predicts the cluster label given the article headline, which would help us understand how well the clustering went and give a proxy to the K-means model. The second metric is similar, instead using the same classifier on only the set of cluster-based key terms in the article title. By comparing these metrics, we can determine how well how well the cluster keywords actually describe the cluster (as compared to describing other clusters).

As a baseline, we have used a Latent Dirichlet allocation on 10 topics, as it is a commonly used topic modeling algorithm [1]. The topics of this give ~60% accuracy on the first metric and 15-30% accuracy on the second, suggesting that the LDA keywords are not enough to classify the headline into a category. For an oracle, we have manually labeled a subset of our articles by hand, as well as run the articles through Google Cloud NLP to have an advanced NLP model extract the keywords/topics of the article as a frame of reference. With this definition of an oracle, the performance is 100% accuracy on each metric as the actual topic is defined to be whatever the human or Google NLP label is. The gap here is major - the LDA uses only the corpus of words we provide, has highly overlapping topic keywords, and is purely probabilistic, while Google uses a huge corpus of knowledge and contextual data to infer what the key topic in a sentence is.

Proposed Approach

Our data comes from the *newsapi.org* API that returns news articles related to a given query. At the moment, we have approximately 4000 articles. Because our raw data is all text, we will first create feature vectors by extracting keywords from the text. Then, to find the set of major topics automatically, we will apply a K-means algorithm repeatedly to find the ideal K that minimizes loss. This will establish a set of clusters, so we can then take the articles assigned to a given cluster and calculate a sparse keyword vector for the cluster, such as through averaging the keyword vectors of all articles in the given cluster. And, any new article introduced to the project can be assigned to a cluster and return the cluster's topic. For instance, the article *"2020 Democrats Will Face Off on One Stage for the First Time This Week. Here's Everything to Know About the September Debate"* could return the topic vector ["democrat", "debate"] as an output. Another concrete example of input is *"6 in 10 fear a mass shooting; most think gun laws can help: POLL"*, which could return the topic vector ["shooting", "laws"] as an output. From these clustered articles, with corresponding key terms, we can run our evaluation metrics to determine how well those key terms define the clusters.

Related Work

The problem space of topic modelling and article labelling has been explored extensively by recent NLP literature over the last few years. Latent Dirichlet Allocation (LDA) is a generative Bayesian probabilistic models designed to assign to an article a mixture of topics [2]. Dirichlet distribution draws out a series of topics that are most salient, and assigns a relevancy score from each topic for a given article. Some works also use word embeddings to automatically label the topic of articles using vector spaces and distance metrics to evaluate topic saliency [3]. Lau et al. used candidate labelling and a ranking algorithm to decide between predefined topics when interpreting a given article [4]. More general models, especially Google Cloud NLP, are significantly more powerful; however, our proposed system can be run locally and without the same training data that backs up those larger models.

Social Impact

By creating a system to identify the set of current major political news topics, as well as to rapidly identify the primary topic of a given article, our system should give the ability to easily find out the trending topics in the news. This allows, for example, understanding how much the public values critical issues like climate change as well as identifying issues that people may attach a lot of importance to, which can be a way for organizations to identify and tackle pressing societal issues.