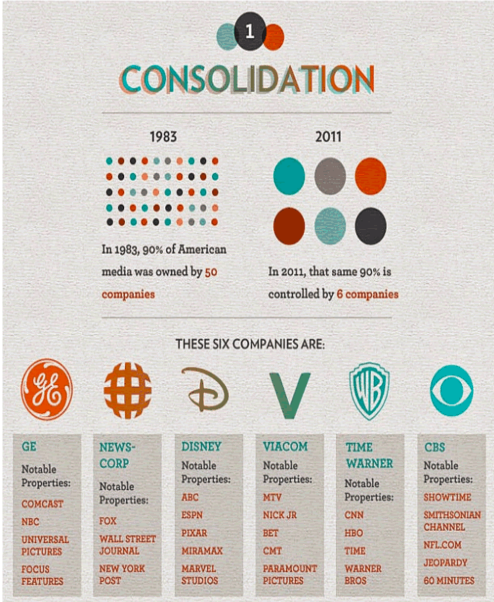
Ali Raza

Gary McKenzie

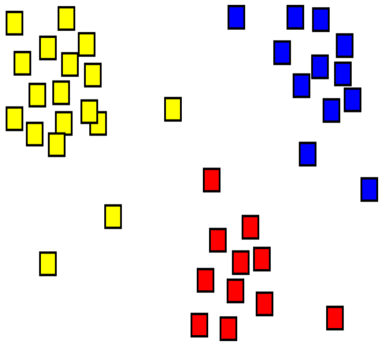
Capstone I

20 April 2017

Classification and Analysis of News Articles using Machine Learning



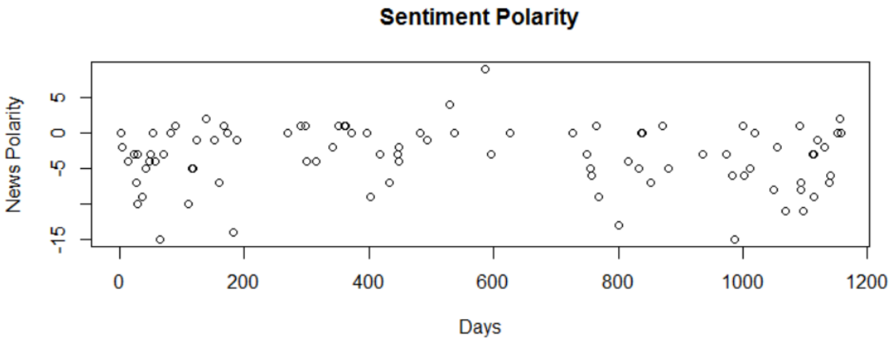
The Illusion of Choice is a phrase that refers to the monopolization of major services in a capitalist society. This phenomenon can be observed in the production of media in the United States. In 2012, Business Insider reported that 90% of media in the US is controlled by 5 companies (Lutz). Following Kellyanne Conway's use of the phrase "alternative facts" to defend misinformation disseminated by the White House, there exists a legitimate concern regarding the integrity of available news. Irrespective of political ideologies, it is evident that there is a pertinent need to analyze news content in order to make meaningful inferences and discover hidden trends and patterns. There have been many impediments to analyzing news in the past. Primarily, we have to analyze a massive amount of data before we can make any meaningful inferences. Furthermore, human analysis of news will always be conducted with bias, whether intentional or not. Advancements in distributed computation have made it possible for Computer Scientist to analyze massive data sets. Furthermore, various machine learning algorithms have made it possible for us to study media content while limiting bias. In this paper, we will focus on techniques for classifying news articles, an essential pre-processing step that will allow us to discover similar articles. We will then discuss whether it is possible to perform further analysis on articles that fall within a particular classification to discover meaningful trends and patterns pertinent to a specific classification.

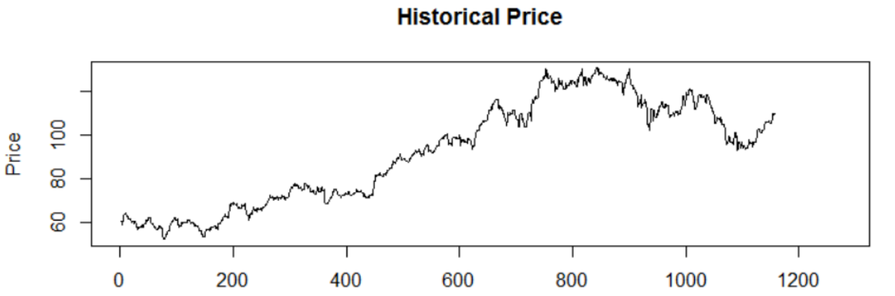


Designing efficient algorithms that group objects into a set is known as clustering, a focal point of machine learning research. The primary goal is to create sets such that objects within a set are similar to each other, based upon some metric. In exploratory data mining, clustering is used as a pre-processing tool. To understand its necessity, consider the image above. Suppose each square represents a news article. The clusters (distinguished by the color of the square) group articles that are similar based on a criterion. For example, if the goal of our analysis is to identify how prevalent buzzwords are in articles based on their publishing source, the criterion would be the publishing source. Why would this step be necessary? Since each article (regardless of publishing source) will have a different percentage of buzz words, we would need to aggregate the percentage of buzzwords for all articles. However, it does not make sense to collectively analyze articles from all sources because our goal is to find how prevalent buzz words are for each source. Here, we can use a clustering algorithm to separate news articles by source, and then subsequently aggregate and measure the average percentage of buzzwords for all articles within a cluster. Clustering is essential in data analysis because data in real life forms “natural categories”. For example, news articles can be clustered by author, source, topic, emotion conveyed, tone, etc. However, when data is collected, a computer cannot distinguish whether two articles are the same topic. This is where clustering algorithms come into play: they identify and form the aforementioned “natural groups” so our analysis can be focused upon the topic we are interested in.

Clustering is a form of unsupervised machine learning because clustering algorithms work without a predefined knowledge of the dataset. In contrast, there is another methodology: supervised learning, which generates inferences from trained data sets. In 2009, MIT students Dennis Ramdass and Shreyes Seshasai attempted to classify news articles using Naïve Bayes Classification. A Bayesian Classifier is an approach that makes assumptions about the generation of data, and then subsequently conceives a probabilistic model that embodies these assumptions. Bayesian classifiers use supervised learning on training sets to approximate the parameters of the generated model. Classification is performed with Bayes’ rule by selecting the category that is most likely to have generated the example. The naïve Bayes is the simplest classifier: it assumes that all features are independent of each other within the category. This is consider naïve because it is false in real-world applications. Despite this assumption, the naïve Bayes performs well in classification and is simpler to implement because the independence assumption allows for features to be learned separately. Ramdass and Sesahsai found that “with the right feature selection and a large enough training size, [it is possible to] create a classifier to classify documents with an accuracy of 77% into their respective sections” (Ramdass). Ramdass and Seshasai’s work is one of many examples of using machine learning to categorize news articles. Their concluding results support that it is possible to classify text documents into categories at an acceptable level of accuracy.

Having discussed the essential pre-processing step of clustering data in order to categorize articles and having provided evidence that machine learning is capable of generating categories at a respectable accuracy, we transition to the analysis of articles once categories have been formed. In 2016 researcher from KJSCE, a university in Mumbai, performed sentiment analysis on news articles to predict stock trends. The methodology was as follows: The documents were tokenized into word vectors which were subsequently compared to pre-built dictionaries of positive and negative words. The difference between the frequency of positive words and the frequency of negative words was used as a polarity measure for the news articles. They concluded that the sentiment of an article was able to accurately predict stock trends with up to 92% accuracy (Joshi).





In conclusion, machine learning has been a promising tool in analyzing news articles. Both supervised and unsupervised machine learning algorithms can be used to categorize news articles. Furthermore, once categorization has been achieved we can use various techniques, such as sentiment analysis, to predict trends and patterns based upon data gathered from news articles.

Work Cited

Kalyani, Joshi, Prof Bharathi, and Prof Jyothi. "Stock trend prediction using news sentiment analysis." arXiv preprint arXiv:1607.01958 (2016).

Lutz, Ashley. "These 6 Corporations Control 90% Of The Media In America." Business Insider. Business Insider, 14 June 2012. Web. 20 Apr. 2017.

Ramdass, Dennis, and Shreyes Seshasai. Document Classification for Newspaper Articles. MIT, 18 May 2009. Web. 23 Apr. 2017.