
ISyE 6740 – Spring 2021

Project Proposal

Team Member Names: Anthea Mitchell (1 person)

Project Title: Picturing the News: Kmeans Analysis of Image Choice Patterns In Fox, NYT, and BBC

Problem Statement

Image analysis using clustering has a great deal of background in the scientific context; x-ray analysis, plant or animal identification, grouping cell images, etc. There are also a number of unsupervised learning applications around manufacturing and industry; recognizing when a problem arises on a conveyor belt based on image analysis, or even facial recognition in the context of photo apps. The goal of this project is to reveal unique patterns behind image choices within and between online news outlets. It's possible that crime sections across publications will tend to have high-contrast visuals, or perhaps NPR uses more photos centering human faces as compared to Fox. The project is intended as a starting point that could be honed in follow-up studies to look at bias both in publication and readership. Often engagement metrics and page views influence either directly or indirectly the type of content editors chose to place alongside links, including images. It could also be used as a means to inform photo banks like Getty on the type of pictures most in demand for different publications.

Methodology

There is no specific theory being posited on what patterns will be found between outlets or within outlets and between news topics; rather any differentiation or pattern would be of interest, making unsupervised Kmeans an ideal starting point. After initial exploration, Kmeans will be used in combination with PCA, and the data (after being collected by web scrape) for each publication will be split into training and testing sets. These will be used to see how well the model performs in classifying between sections of each respective publication.

There is precedence for this application of Kmeans as seen in the work of Mingjie Qian and Chengxiang Zhai, whose interest was in applications of feature selection on web-news data using Kmeans on both text and images (Qian & Zhai, 2014). Their work emphasized that “unlike traditional document clustering, images play an important role in web news articles as is evident from the fact that almost all news articles have one picture associated,” which supports the importance being placed in this project on isolated images. Rather than pairing images with significant portions of text, it will consider them with the bare minimum context in order to see how clustering results change with a paired-down data set.

A similar work from Zhan et al. was interested in clustering news based on the topic (Zhan et al., 2019). The application was for improving multiview learning to cover the same topic but with different viewpoints. This approach did consider images, however not any pulled from selected news publications; rather it more broadly considered applicable algorithms as applied to different libraries (specifically the Columbia object image library).


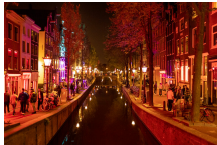
The addition of the title metadata allows for the option to cluster on both title and image or create separate text or image-based clusters. Previous work has also been done focusing only on text elements of news data, specifically as it applies to news publication work in the language of Bangla, using both K-Means and MiniBatch K-Means algorithms (Hasan et al., 2022). Despite not focusing on image processing, this work found machine learning clustering as applied to both text and image processing in the news context promising. Given time, comparing to sklearn's MiniBatch K-Means might be an additional approach worth experimenting with in this project.

Unlike the work of Qian and Zhai, which looked at two U.S. publications (CNN and Fox News), and Hasan, Ruiqin, and Hussain, which focused on the language of publication rather than publications themselves, this project will consider 3 publications, two within the United States from differing political groups, and one major publication from outside of the United States.

Data Source:

The data set is comprised of scraped images and metadata using a combination of the Selenium and BeautifulSoup python packages and a web scraping software called Octoparse. Three news sources (Fox, NYT, and BBC) are included in the set, with images pulled from across their respective US, World, Science, and Health sections. Data will be scraped on the same date, and delimited by whichever publication has the smallest number of images (i.e. if NYT only has 100 images on their front page, and BBC has 400, only 100 will be used across all publications). Metadata includes the article title, publication section, and the photos associated, represented as pixels in a data frame.

Sample Row of Data for Each publication:

 <p>Publication: BBC</p>	<p>Title: Bolsonaro touches down in Brazil after self-imposed exile</p> <p>Section: World News</p>	 <p>Publication: Fox</p>	<p>Title: Sheikh Mohammed, UAE premier, appoints eldest son as successor</p> <p>Section: World News</p>	 <p>Publication: New York Times</p>	<p>Title: Amsterdam Has a Message for Male Tourists From the UK: 'Stay Away'</p> <p>Section: World News</p>
------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------

Evaluation and Final Results

Models will be evaluated by a couple of metrics. First, a qualitative assessment of the clustering will be done in order to determine any unique insights offered by unsupervised learning. Secondly, a quantitative approach will be taken to determine the percent of images correctly clustered both by section and by publication, both within and between publications.

Works Cited:

Hasan, S. A., Ruiqin, W., & Hussain, M. G. (2022). Clustering Analysis of Bangla news articles with TF-IDF & CV using mini-batch K-means and K-means. *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*.
<https://doi.org/10.1109/cyberneticscom55287.2022.9865339>

Qian, M., & Zhai, C. (2014). Unsupervised feature selection for Multi-view clustering on text-image web news data. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*.
<https://doi.org/10.1145/2661829.2661993>

Zhan, K., Nie, F., Wang, J., & Yang, Y. (2019). Multiview consensus graph clustering. *IEEE Transactions on Image Processing*, 28(3), 1261–1270. <https://doi.org/10.1109/tip.2018.2877335>