Using K-Means Clustering to Group Articles Together in a Large Wikipedia Corpus

John Speaks
jspeaks2@illinois.edu

Abstract

This paper analyzes of a substantial Wikipedia corpus consisting of over six million articles, exploring the K-Means clustering technique to identify thematic patterns and semantic similarities within the text. Using K-Means clustering, the paper compares the clustering results with human-annotated categories included in the corpus, aiming to assess the efficacy of automated clustering methods in capturing semantic cohesion. Through careful preprocessing and thorough evaluation, the research demonstrates K-Means' ability to organize articles into cohesive clusters, providing insights into the strengths and limitations of automated clustering compared to manual categorization processes. The findings verify the scalability and efficiency of K-Means clustering in handling large datasets while highlighting its potential to uncover broad associations among articles for further corpus research.

1 Introduction

Understanding the organization and thematic patterns within large text corpora is essential for various applications in natural language processing (NLP) and computational linguistics. With the the digital age and rise of digital content, the need for effective methods to analyze and categorize vast text data has become increasingly vital. One example of a corpus widely used for research purposes is Wikipedia, an online repository of articles covering millions of topics across numerous domains.

This paper explores the analysis of a large corpus derived from Wikipedia articles, encompassing a diverse range of subjects and domains. The corpus, comprising over six million articles, is the basis for the investigation of clustering techniques to identify thematic patterns and semantic similarities within the text. Due to the massive volume of data, a careful approach was adopted, which included the sampling of articles to ensure computational feasibility while maintaining a good representation of the diversity of Wikipedia across various topics.

The objective of this paper is to explore the efficacy of K-Means clustering in organizing and grouping articles based on their topical content. Additionally, this paper aims to compare the results obtained from K-Means clustering with human-annotated categories available within the Wikipedia corpus. By evaluating the clustering performance against a baseline and human annotations, insights into the strengths and limitations of this automated clustering method in capturing semantic cohesion in textual data are gained.

The paper follows an organized approach. Firstly, it provides background information on the corpus utilized for analysis, detailing the sampling procedure and preprocessing steps taken to prepare the data for clustering. Next, the methodology section details the application of K-Means clustering, highlighting the process of feature extraction and cluster forming. After this, the results of the clustering analysis are presented and evaluated against human-annotated categories and a baseline, providing insights into the effectiveness of K-Means clustering in identifying thematic patterns.

With this research and evaluation, further research is possible using broad clustered categories of Wikipedia articles, allowing linguistic research to use sub-corpora of data separated by topics that are not as limited in scope as those categories already provided in the corpus by Wikipedia. In addition, this research provides insights into the different uses that are best for human annotated groupings when compared to clustered groupings in the future.

2 Background on the Corpus

The corpus being used for analysis in this paper is a large corpus of all Wikipedia Articles that was created July 1st, 2023. It contains 6,286,775 articles

and contains their titles, raw text, and categories assigned by Wikipedia. The corpus was accessed from an entry on Kaggle[1]. Wikipedia is an important avenue for linguistics research because it is a vast online encyclopedia which encompasses knowledge about topics across time and the world, being updated often. This opens avenues for research into language as a whole with the overall corpus describing millions of topics, but also research delving into the language used in specific sub-fields which Wikipedia extensively details. Currently, the available method for researching these sub-fields is by using the given human annotated labels in the corpus. These sub-fields however are often too limited in scope and will only encompass very specific categories with under 100 articles. This can be limiting for linguistic analysis. The approach taken by this paper will allow broader topics to be created from the corpus with potential for specified linguistics research with large enough corpora of Wikipedia data.

The data was separated into parquet files A-Z for the starting letter of each article with two additional files for articles starting with a number and articles starting with a symbol. Due to limited computing resources, all of the data could not be loaded at once and two percent of each file was sampled to run the following experiments. The sampled corpus was made up of 125,736 articles. The data was randomly sampled from across the dataset to ensure an even representation of topics even with a small sample taken of the larger corpus. This was critical because the value of Wikipedia as a corpus comes from its diversity of topics and that had to be maintained even in a limited sample. The implementation of the sampling can be found on this paper's GitHub repository[2].

3 Preparing the Data

Before the data was going to be clustered, it was vital to identify multi-word phrases and group them as one item and identify stop words.

First, the Gensim Phrases model is used to identify multi-word expressions or phrases within the text data. The model was provided the set of text data, a min_count parameter, and a threshold parameter. The min_count was set to two, requiring that a phrase must appear at least twice in the dataset to be included. The threshold was set to 0.7 and controls how strong of a correlation the words in a phrase must have in where they appear to be considered a phrase. In order to score the phrases, the model uses Normalized Pointwise Mutual Information (NPMI). Once the phrases are learned, the individual words making up the phrases are replaced in the text with the single string representing the multi-word expression. This process helps in recognizing and consolidating multiword expressions, which can often carry more meaning collectively than individually, enhancing the semantic understanding and analysis of the text.

Next, stop words needed to be removed from the data. Using the CountVectorizer from scikit-learn, the data is tokenized into words with there corresponding counts in the data. Words were chosen as the tokenization method for this experiment because they are more content based, and this approach is better for identifying topics. After the CountVectorizer is fit to the data, the top 500 most common words are removed from the dataset as stop words. This approach is commonly used to identify and exclude words which are common and may not carry significant meaning in the context of text analysis.

4 Performing K-Means Clustering

Once the data is prepared, it is ready to be clustered. A vectorizer and is fit to the data, keeping into account the stop words found in preprocessing. The data is transformed into a feature matrix which is then clustered using the K-Means approach from sklearn. K-means clustering is an algorithm used to partition a dataset into K clusters, in this case we partition into ten clusters in each iteration. It adjusts the clusters repeatedly until it converges to a point where the clusters to minimize the within-cluster sum of squares distance of all article vectors from each other. This is repeated on topics that contain over 5% of the data to further divide the topics until no such large topics exists. The clustering process aims to group similar documents together based on their content features, ultimately facilitating topic extraction and analysis from the given text data. K-Means clustering is a good, efficient approach to data clustering, but is very dependent on the initial random start each cluster has. For this reason the clusters may converge to local minima instead of global ones. Other approaches could be

[1] https://www.kaggle.com/datasets/jjinho/wikipedia-20230701

[2] https://github.com/JTSIV1/Wikipedia-Article-KMeans-Clustering/blob/main/Project2_1.ipynb

used in further studies, but in this case it was a sufficient and computationally feasible choice. The implementation of this clustering and details on how to perform it independently are on the GitHub repository for this paper[3].

5 Results and Evaluation

Method of Grouping	K-Means	Manual-Human
Number of Groups	46	125736
Mean Size	2733	444
Median Size	1484	3
Standard Deviation	3070	2928

Figure 1: Simple statistics on distribution of clusters based on K-means and human annotated approaches

The results of the clustering can be seen in figure 1. The K-Means clustering formed 46 clusters with an mean size of 2733 articles per cluster and median of 1484. The standard deviation for cluster size is 3070, meaning that the size of each cluster varies highly but at the same time is mostly relatively small clusters with some very large outliers. A visualization of the clusters can be seen below in figure 2.

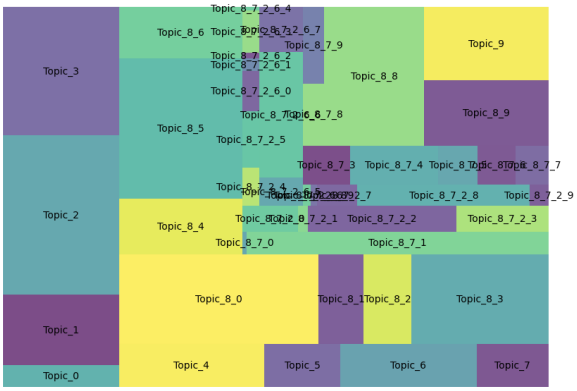

Figure 2: Topic and Subtopic Clusters Graphed by Share of Data

To compare this to a pseudo ground truth (what is used today), the category labels given in the original corpus were used to form groupings. Using this approach, there are over 125 thousand categories in the data as each article has several category labels. This was too computationally intense to work with, so instead a random sample of 200 categories were used in the comparison with the K-Means clusters. The mean size of these category clusters was 444 articles, and the median was 3. The standard deviation was very large with 2928. This variation is caused by most categories containing a very small number of samples (many just one sample) and a few containing many. For example, the category Steam locomotives of Sudan had only one article within the sample of articles while Living People had 20,984. Living People was the only category with over 500 articles but caused a very large skew in the mean.

To evaluate the clustering, a similarity score was calculated for each of the clusters, each of the category groupings, and for a random sample of articles without any cluster or category grouping as a benchmark. A similarity score was created by converting two documents into vectors that store the importance of different words in the document depending on their frequency in the article and overall frequency in the corpus. These were then compared using cosine similarity to generate a similarity score between -1 and 1. The cosine similarity was used for comparison because it is able to compare multi-dimensional feature vector data and also provide a normalized result with a fixed range. It also is generally thought to capture the semantic similarity between documents by measuring the cosine of the angle between their vector representations. Semantically similar documents generally have similar orientations in the multi-dimensional vector space, leading to higher cosine similarity scores. This property aligns well with the goal of identifying thematic patterns.

For a whole topic or category, pair-wise similarity was calculated for every two articles in the set and then all the pair-wise similarities were averaged to produce the average similarity of the set.

For a fair comparison in the category-wise separated data, all of the categories with only one article were excluded because they would have a similarity of 1 by default, skewing the results and not providing information about the intra category similarity. The range of the topic cluster size was 36 to 11,176 so only the category similarities for categories withing that size range were included.

Method of Grouping	Average Cosine-Similarity Score
Random	0.048693847
K-Means Clustering	0.093855796
Human Annotated Categories	0.194613245

Figure 3: Average similarity scores by article grouping method

The similarity scores calculated can be seen above in figure 3. From the data, both the K-Means

[3] https://github.com/JTSIV1/Wikipedia-Article-KMeans-Clustering

clusters and the categories in the corpus provide a better method for grouping than the random baseline. K-Means performed 1.927 times better and the Human annotated categories performed 3.997 times better than the baseline. Additionally, human annotated categories performed 2.07 times better than K-Means clustering. These measures provide a way to evaluate the efficacy of the semantic cohesion with each method relative to each other.

It is clear that human annotations for article categorization was better at producing similar documents, but there are still clear advantages to the K-Means approach. Firstly, it is automatic and not labor intensive like the Wikipedia process. K-Means can cluster all of the articles at once, and with enough computing power, relatively quickly. By comparison, if human annotation needed to create a new category, this would require humans going through over six million articles and deciding its categorization manually which would be expensive and time consuming. K-Means clusters were also able to identify patterns between articles not identified in the manual categorization. For example, many Christian Clergy were clustered together in K-Means but share no common category label because they existed at different points in history. K-Means was able to create a broad category of Clergy, which is new information that is not provided by the human categories. This would allow further linguistics exploration and research into the language used to talk about Clergy, which would otherwise require crawling the data, since no broad labels existed. K-means was also able to provide groupings with much improved similarity in under 50 topics as opposed to the human labeled categories which divides the corpus into over 125 thousand categories. In 2,733 times more divisions, manual categorization was only able to provide 2.07 times more semantic cohesion, meaning that K-Means was much more effective with broad categories.

6 Future Explorations

Using knowledge gained from this paper's exploration, further work should be done to create clusters for the entirety of the Wikipedia data. Using a more computationally powerful machine, clustering should be performed to create a cohesive and large corpus for linguistics research. Using that quantity of data may require different thresholds for K-Means clustering, like requiring the topics to have even less than 5% because the dataset is simply so large that even 5% of six million would be 300,000 articles. That would likely be too much for a cohesive category.

After a larger corpus is created using a more powerful computer, exploration could be done into any of the topics. An example from this paper's exploration could be how modern sports are described when compared to historical sports issues. The clustering generated in this paper created the Topic 3 sub-corpus, which contains articles about athletes and coaches on sports teams from across time. Using the Wikipedia categories in the data to separate the articles into time periods, the language to describe athletes and sports now could be compared to that language for historical sports. This could provide insights into if humans use language differently when describing topics they are more temporally connected to.

7 Conclusions

This paper's Wikipedia corpus analysis unveils intriguing insights into the efficacy of clustering techniques in identifying thematic patterns. The findings highlight the efficacy of the K-Means clustering approach in identifying coherent clusters within vast text corpora, offering a computationally efficient alternative to the manual categorization processes that are used today. While human annotations are undoubtedly better at identifying semantic cohesion, the scalability and automation that come with K-Means clustering present advantages, particularly in handling large datasets.

Further, the evaluation metrics reveal a significant improvement over random baselines, proving the efficacy of both K-Means clustering and human-annotated categories in capturing semantic similarities within the corpus. Specifically, K-Means clustering was identified as a versatile tool capable of uncovering broad associations among articles, which helps to build our understanding of the corpus and identify groups of articles that describe similar concepts without being directly related. K-Means thus provides an effective way to find these categories for further corpus research.

Additionally, this paper was able to create new topic labels for the sampled Wikipedia data. Future studies could use these additional topic annotations on the corpus to explore how linguistic features vary across time or region while staying within a broad topic.