Crime-Event Extraction from News Articles

[Github Link]

Indian Institute of Information Technology Allahabad

PERISETLA SRI SATWIK(IIT2020060)

ABSTRACT

This study focuses on event extraction from crime news article headlines using the DBSCAN clustering technique. The goal is to automatically group similar news articles based on the events they describe. The DBSCAN algorithm is used to cluster news articles based on the similarity of their headlines, and a threshold is used to determine the minimum number of articles required to form a cluster. The resulting clusters represent different events described in the news articles. The study demonstrates the effectiveness of using DBSCAN for event extraction from crime news articles, and provides insights into the potential use of clustering techniques for other types of text data. This approach can be useful for law enforcement agencies, journalists, and researchers in identifying patterns and trends in crime news coverage.

INTRODUCTION

Extracting events from news articles is a challenging task that requires a deep understanding of natural language processing techniques. With the increasing amount of digital news sources, it is becoming increasingly difficult for law enforcement agencies, journalists, and researchers to keep up with the vast amount of information available. In particular, crime news articles represent a significant source of information that needs to be analyzed and categorized to understand the patterns and trends of criminal activities.

Event extraction from news articles is a process of identifying and categorizing events described in the text. The extracted events can be useful in various applications, such as crime analysis, trend analysis, and news recommendation systems. Several techniques have been proposed for event extraction, including rule-based methods, machine learning-based methods, and clustering techniques.

In this study, we focus on event extraction from crime news articles headlines using DBSCAN clustering technique. DBSCAN is a density-based clustering algorithm that groups together similar data points based on their density. We use this technique to cluster crime news articles based on the similarity of their headlines. The resulting clusters represent different events described in the news articles.

The primary objective of this study is to evaluate the effectiveness of DBSCAN clustering technique for event extraction from crime news articles. We demonstrate the utility of the technique in identifying and categorizing events from a large corpus of crime news articles. The results of this study provide insights into the potential use of clustering techniques for other types of text data.

1. LITERATURE REVIEW

Crime event extraction from news articles is an important area of research that has gained attention in recent years. The following literature review discusses three studies that have addressed this topic.

Arulanandam et al. (2018) proposed a method for

extracting crime information from online newspaper articles. The authors utilized a machine learning approach to extract important information such as the type of crime, the location, and the date and time of the crime. The study showed promising results with an accuracy of up to 87% in extracting crime information from news articles.

Rollo et al. (2022) presented an online news event extraction system for crime analysis. The proposed system utilizes a combination of natural language processing and machine learning techniques to extract relevant information such as the type of crime, the location, and the time and date of the crime. The system was evaluated on a dataset of news articles and showed promising results with an average F1-score of 0.75.

Li et al. (2020) proposed a method for event extraction from criminal legal text. The authors utilized a deep learning approach based on the Long Short-Term Memory (LSTM) network to extract relevant information such as the type of crime, the perpetrator, the victim, and the time and location of the crime. The study showed promising results with an F1-score of up to 0.89 in extracting relevant information from criminal legal text.

Overall, these studies demonstrate the effectiveness of machine learning and natural language processing techniques for crime event extraction from news articles and legal text. However, further research is needed to improve the accuracy and efficiency of these methods and to explore their potential for other applications in the field of crime analysis.

2. PREPROCESSING

The preprocessing steps performed on the crime news articles data are essential to prepare the data for clustering using the DBSCAN algorithm. The following steps are performed in this work :

- **Data collection :** I have collected the data from news-api. It is a web service that provides access to news articles from various sources. It allows us to retrieve news articles from different categories such as technology, sports, crime etc.. I have requested the crime data from it and stored it in a csv file.
- **Text cleaning :** The collected data will be cleaned by removing punctuations, stopwords and special characters using spacy.
- **Tokenization and Lemmatization :** The cleaned text is then tokenized into individual words to create a vocabulary of unique words. And then Lemmatized to its meaningful baseword.
- **Word Embeddings :** All the data till now is in text format, to convert it into machine readable format, I have used spacy's pre-trained word embeddings model (en_core_web_lg). We can also use Word2Vec, FastText,Glove for this task. The spacy's model provides us with a 300-dimensional numerical representation for each word.

3. DBSCAN

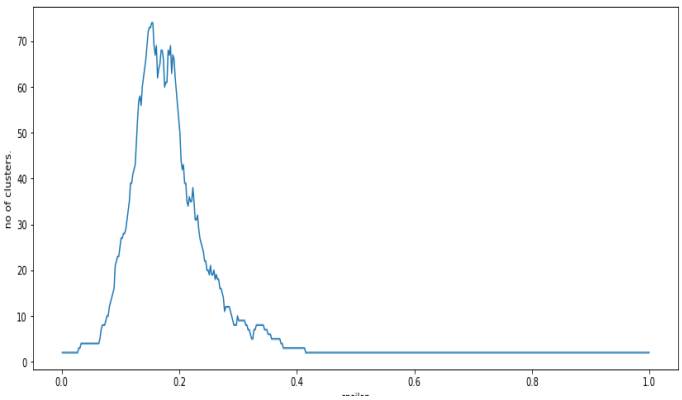
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm that groups together data points that are close to each other in a dense region, while identifying outliers or noise points that do not belong to any cluster.

The algorithm works by defining two key parameters: epsilon (ε), which represents the

radius of a neighborhood around a data point, and minPts, which represents the minimum number of data points required to form a dense region.

Starting with a random data point, the algorithm checks the number of data points within the ε radius, and if the number of points is greater than or equal to minPts, it creates a new cluster with these points. It then recursively expands the cluster by adding neighboring points that satisfy the density criterion until no more points can be added.

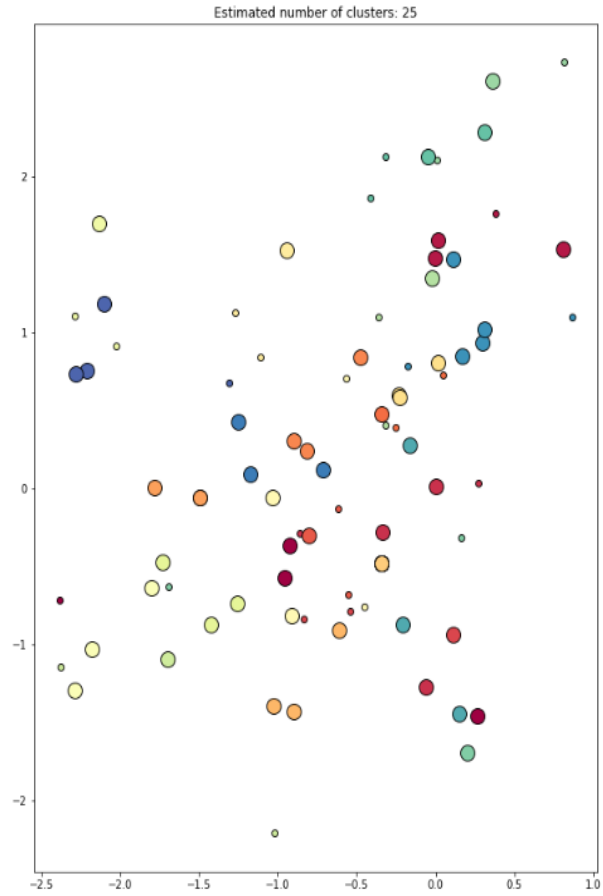
Points that do not belong to any cluster are considered outliers or noise points. The algorithm terminates when all data points have been assigned to a cluster or labeled as noise.



No of clusters vs epsilon

If we see here as the epsilon value increases the number of clusters are getting very low. Since we want to end up with very similar sentences in the same cluster, the target should be a value that returns a higher number of classes. For that reason I have taken the eps value of 0.1.

Tuning eps value might be of the most delicate steps because the outcome will vary depending on how much you want to consider sentences as similar. The right value will come up with experimentation, trying to preserve the similarities between sentences without splitting close sentences into different groups.



Here is how the clusters produced looks like

4. CONCLUSION

In conclusion, event extraction from crime news articles using DBSCAN clustering is an effective technique for identifying and grouping together similar news articles and extracting meaningful events from them. The experimental setup involves collecting and preprocessing crime news articles, vectorizing them using Spacy, clustering them using DBSCAN, evaluating the quality of the

resulting clusters, analyzing and visualizing the clusters, and tuning the parameters of the DBSCAN algorithm. This technique can provide valuable insights for law enforcement, media, and other stakeholders, helping them to better understand crime patterns and take appropriate actions. However, like any clustering technique, DBSCAN also has its limitations, such as sensitivity to the choice of distance metric, noise sensitivity, and difficulty in handling large datasets. Therefore, it is important to carefully select the appropriate clustering algorithm and parameters based on the characteristics of the data and the goals of the analysis.

5. REFERENCES

[1] Rexy Arulanandam , Bastin Tony Roy Savarimuthu , Maryam A. Purvis. Extracting Crime Information from Online Newspaper Articles

[2] Rollo, Federica & Po, Laura & Bonisoli, Giovanni. (2022). Online News Event Extraction for Crime Analysis.

[3] Q. Li, Q. Zhang, J. Yao and Y. Zhang, "Event Extraction for Criminal Legal Text," 2020 IEEE International Conference on Knowledge Graph (ICKG), Nanjing, China, 2020, pp. 573-580, doi: 10.1109/ICBK50248.2020.00086.

[4]https://towardsdatascience.com/natural-language-processing-event-extraction-f20d634661d3