EasyChair Preprint

№ 13917

Text Clustering for Topic Identification: a TF-IDF and K-Means Approach Applied to the 20 Newsgroups Dataset

Marppan Sampath and S Vignesh

July 10, 2024

Text Clustering for Topic Identification: A TF-IDF and K-Means Approach Applied to the 20 Newsgroups Dataset

[1]Marappan Sampath, [2] Vignesh S

[1]MS Student, [2]MTech Student

Department of Artificial Intelligence,
Reva university, Bangalore, India

Abstract : In this paper, we present an efficient approach for topic modeling using Term Frequency-Inverse Document Frequency (TF-IDF) and K-means clustering, applied to the 20 Newsgroups dataset. The 20 Newsgroups dataset is a well-known collection of approximately 20,000 newsgroup documents, partitioned across 20 different newsgroups. Our method involves preprocessing the text data to remove noise, calculating the TF-IDF matrix to represent the documents in a high-dimensional space, and employing K-means clustering to group the documents into distinct topics. The effectiveness of the approach is demonstrated through the identification of coherent topic clusters, highlighting the key terms associated with each cluster. This straightforward yet powerful combination of TF-IDF and K-means clustering offers a robust solution for text clustering and topic identification tasks, making it suitable for various natural language processing applications. The results show that our method can effectively uncover the underlying topics within a large text corpus, providing valuable insights for further text analysis and information retrieval.

I. INTRODUCTION

Text clustering and topic modeling are critical techniques in natural language processing (NLP) that facilitate the organization, summarization, and interpretation of large text corpora. These techniques are widely used in various applications, including information retrieval, document classification, and content recommendation. One of the most well-known datasets for benchmarking these methods is the 20 Newsgroups dataset, which comprises approximately 20,000 documents categorized into 20 different newsgroups. This dataset provides a robust foundation for exploring and comparing text clustering algorithms.

In this paper, we explore an efficient approach for topic modeling using Term Frequency-Inverse Document Frequency (TF-IDF) and K-means clustering. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents, effectively transforming textual data into numerical feature vectors. K-means clustering is a popular unsupervised learning algorithm that partitions data into k distinct clusters based on feature similarity.

Our methodology involves preprocessing the text data to remove noise and standardize the content, calculating the TF-IDF matrix to represent each document as a high-dimensional vector, and applying K-means clustering to group the documents into meaningful clusters. Each cluster is then analyzed to identify and interpret the underlying topics, represented by the most significant terms in each group.

1.1 Population and Sample

The dataset employed in this study is the 20 Newsgroups dataset, a widely used collection of newsgroup documents originally compiled by Ken Lang. This dataset is frequently utilized for evaluating and comparing text classification and clustering algorithms due to its diverse topics and substantial size. The dataset consists of approximately 20,000 documents, partitioned across 20 different newsgroups, covering a broad range of topics including politics, sports, technology, and religion.

Each document within the 20 Newsgroups dataset represents an article or post from a Usenet newsgroup, a popular discussion system widely used before the advent of the World Wide Web. The dataset is structured such that each document belongs to one of the 20 predefined categories or newsgroups, providing a labeled corpus for supervised and unsupervised learning tasks in natural language processing.

Preprocessing of the dataset involves several steps to standardize and clean the text data. These steps typically include tokenization, removal of stop words, punctuation, and numerical digits, as well as stemming or lemmatization to reduce words to their base form. The goal of preprocessing is to transform raw textual data into a format suitable for further analysis and modeling.

1.2 Theoretical framework

The implementation employs Term Frequency-Inverse Document Frequency (TF-IDF) to quantify the importance of terms in documents relative to a corpus, combining Term Frequency (TF) to measure term occurrences within documents and Inverse Document Frequency (IDF) to assess term rarity across the corpus. This statistical measure transforms textual data into numerical vectors, facilitating analysis. K-means clustering, an iterative algorithm, partitions these TF-IDF vectors into k clusters based on similarity, minimizing intra-cluster variance by iteratively adjusting centroids to represent the mean of data points assigned to each cluster. This framework enables systematic exploration and organization of text data, supporting efficient topic modeling and clustering in natural language processing applications.

II. MODEL IMPLMENTATION

2.1 Objective:

The objective of this paper is to implement and evaluate a methodology for efficient topic modeling using Term Frequency-Inverse Document Frequency (TF-IDF) and K-means clustering on the 20 Newsgroups dataset. Specifically, we aim to:

- Utilize TF-IDF to transform textual data into numerical representations that capture the importance of terms within documents relative to the entire corpus.
- Apply K-means clustering to partition these TF-IDF vectors into clusters, enabling the identification of coherent topics across the dataset.
- Demonstrate the effectiveness of our approach through systematic analysis and interpretation of the resulting clusters, evaluating their relevance and coherence.
- Provide insights into the practical application of text clustering techniques for topic identification in natural language processing, contributing to the broader field of text analysis and information retrieval.

2.2 Procedure:

Data Loading and Preprocessing:

Download the 20 Newsgroups dataset, which consists of approximately 20,000 documents across 20 different newsgroups. Preprocess the text data by tokenizing, removing stopwords, punctuation, and performing stemming or lemmatization to standardize the text.

TF-IDF Calculation:

- Compute Term Frequency (TF) for each term in every document:

$$\mathrm{TF}(t,d) = \frac{\mathrm{count}(t,d)}{\sum_{t' \in d} \mathrm{count}(t',d)}$$

  Where count(t,d)\text{count}(t, d)count(t,d) is the number of times term ttt appears in document d.

- Calculate Inverse Document Frequency (IDF) for each term:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{|\{d : t \in d\}|}\right)$$

- Compute TF-IDF scores for each term-document pair:

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

- Construct a TF-IDF matrix where each row represents a document and each column represents a term, with values being the TF-IDF scores.

K-means Clustering:

- Choose the number of clusters *k* based on the desired number of topics or through evaluation metrics.
- Initialize *k* centroids randomly or using a heuristic method.
- Assign each document to the nearest centroid based on Euclidean distance in the TF-IDF space.
- Update centroids by computing the mean of all documents assigned to each cluster.
- Repeat the assignment and update steps until convergence (centroid positions stabilize or a maximum number of iterations is reached).

2.3 Selection Criteria:

TF-IDF and K-means clustering are suitable for text analysis tasks where the dataset consists of textual documents varying in length and containing diverse vocabulary. They excel in identifying topics or themes within a corpus, making them ideal for exploratory analysis and topic modeling. These algorithms handle sparse data effectively and are scalable to large datasets, provided adequate computational resources. They offer interpretable results through cluster analysis, where TF-IDF facilitates the identification of key terms driving each cluster. Evaluating their performance using metrics like silhouette score ensures the quality of clustering results. Consideration of preprocessing needs and domain-specific requirements further guides their applicability in natural language processing tasks.

2.4 Benefits:

This paper on TF-IDF and K-means clustering for topic modeling on the 20 Newsgroups dataset provides a streamlined method to extract and organize themes from textual data. Using TF-IDF, it accurately assesses term importance, while K-means clustering groups documents based on their similarities in TF-IDF space, enabling the systematic discovery of coherent topics. This approach enhances interpretability, supports efficient data analysis, and is scalable for large datasets, making it applicable across various domains where understanding textual content is crucial. Moreover, the study serves as a benchmark for evaluating clustering algorithms in text mining, contributing valuable insights into effective topic modeling techniques.

III. Results and Discussion:

Cluster	Keywords (Top 25 words)	Topic
Cluster 0	com gauges would gauge oil vcd hp basic set dave temp pressure coolant subject choice organization lines dmunroe must microware writes free reign design instrument	Automotive and Instrumentation
Cluster 1	muskingum jbuddenberg vax cns edu jimmy buddenberg vesa controller card college subject get organization lines dx mhz local bus would see much increase speed drives	Computer Hardware and Performance
Cluster 2	uk ac subject lines organization writes com edu co x article university would posting host nntp one like apr know reply c anyone mail please	Academic Discussions and University Topics
Cluster 3	sas com unx sasghm theseus institute inc extra gary merrill science methodology rational subject originator nntp posting host organization lines bath well schlotz kestrel xcutsel	Scientific Research and Methodology
Cluster 4	nz ac new zealand subject organization lines university edu nntp posting host canterbury writes article cri waikato otago comp apr gen grace like would kosmos	New Zealand Universities and Academia
Cluster 5	crchh brian antioc antioch edu bcash nosubdomain nodomain cash subject beleive either nntp posting host organization bnr inc lines article apr smauldin writes stopped believing	Personal Beliefs and Academia
Cluster 6	ti dseg com pyron skndiv dillon edu vax support subject lines organization writes article nntp posting host need apr glock reply unless opinions space one	Technical Support and Opinions
Cluster 7	nyx cs edu denver du access university organization subject lines public unix dept math u one x writes community fbi system responsibility disclaimer run know	University Systems and Public Access
Cluster 8	gear westminster wes fujii jkjec ac uk shazad barlas subject improvements automatic transmissions organization lines wheelspin auto keep n gas stick never tried sure works	Automotive Improvements and Performance
Cluster 9	alaska edu space nsmca aurora acad organization subject article lines nntp posting host university would fairbanks michael adams high jacked apr get writes make billion	Space Science and University Research
Cluster 10	wagon edu sumax nissan smorris seattleu v seattle subject organization lines morris station nntp posting host university com open letter wagons addiction studies eliot based	Automotive and University Studies
Cluster 11	com kubey sgi obp ken organization batting average edu root subject hbp bb	Baseball Statistics and Analysis

	big cat nntp posting host wpd distribution na lines article mjones writes	
Cluster 12	rochester edu cc uhura university subject organization lines nntp posting host new writes schnopia asb york com bi article cs get c gay andrew ny	University Communications and Discussions
Cluster 13	hp fc com rod cerkoney rodc hewlett packard fort collins co subject gqxf fekvh nntp posting host hpfcmrc organization x newsreader tin version pl lines	Hewlett-Packard and Technical Discussions
Cluster 14	loral des koontzd phobos lrmsc com david koontz subject triple keywords organization rolm computer systems lines please post news	Computer Systems and News Posts
Cluster 15	com sun convex subject lines organization edu east writes posting nntp host article microsystems like reply ed green corp distribution computer central know would world	Computer Corporations and Industry News
Cluster 16	ca maynard laurentian cs roger team ramsey writes best would subject player organization lines edu apr leafs say university game better article one many morris	Sports Discussions and University Topics
Cluster 17	ax edu com subject lines x organization would one writes article w q like people p c u university r posting know get host nntp	General University Discussions
Cluster 18	edu udel ravel cobra andrew cmu subject organization university delaware lines send take look two todd play think fred masterson nhlpa poll partial stats results	University Sports and Poll Results
Cluster 19	henry toronto edu zoo spencer u zoology writes article lines subject organization work utzoo svr one space would com high orbit like man earth kipling	Zoology and Space Research

1. Discussion:

1.1 Algorithm:
Combination of TFIDF and K-means provides a good classification of topics for 20 news group data set.

1.2 Model Evaluation:
Evaluating multiple feature sets provided insights into which combination of features yields better predictive performance.

1.3 Model Performance:
The words found in the cluster of topics are almost matches with the central theme of the discussion.

1.4 Future Work:
Future research will explore advanced machine learning algorithms with Bayesian network model like latent Dirichlet allocation (LDA).

IV. References:

1. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.
2. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513-523, 1988.
3. T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, Nashville, TN, USA, 1997, pp. 143-151.
4. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.
5. J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281-297.

6. D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New Orleans, LA, USA, 2007, pp. 1027-1035.
7. I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143-175, 2001.
8. J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.
9. A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010.
10. C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, Banff, AB, Canada, 2004, pp. 225-232.