«Системні технології» 1 (138) 2022 «System technologies»
DOI 10.34185/1562-9945-1-138-2022-10
UDC 004.421

I.V. Ponomarev

DEVELOPMENT OF AN AUTOMATED SYSTEM FOR CLUSTERING TEXT DOCUMENTS

Abstract. Grouping texts into groups similar in content is a common task in various fields of human activity. Text document clustering is used to automatically categorize text documents, filter emails, group web pages in search engines, and so on. Automation of this process can significantly reduce the time spent on this task.
Keywords: clustering, text mining, TF IDF, HDBSCAN, tokenization, lemmatization, stop words, PYTHON.

Formulation of the problem. Data mining is the process of identifying interesting and useful patterns and relationships in large amounts of data [1]. This area combines statistics and artificial intelligence tools with database management to analyze large digital collections known as datasets.

Data mining is widely used in business, research, public security and many other areas. For example, in the field of modern Internet marketing, various parameters of consumers are analyzed (age, interests, etc.) and personal recommendations are formed for certain groups of people and for each person individually. Mobile operators process a large amount of data in order to identify people who can switch to another operator and offer them certain advantageous offers.

Text Mining is engaged in obtaining information from text collections. It covers a wide range of related topics and text analysis algorithms from various fields, including information retrieval, natural language processing, machine learning, etc. Text mining does not deal with databases, but with electronic libraries and corpora of texts (sets of texts, metadata, as well as grammatical and other rules used for language research).

Purpose of the research. It is necessary to analyze the methods and algorithms of text clustering and to develop an automated text clustering system for grouping text files by their content.

Main part. Clustering is the process of dividing a set of data objects (or just data or observations) into subsets. Each subset is a cluster such that the objects in

the cluster are similar to each other but different from the objects in the other clusters.

There are methods that are specific to the clustering of text documents: clustering of relative frequencies of words in documents (TF-IDF), clustering with suffix trees, thematic modeling, Scatter-Gather approach.

The analysis of text document clustering methods allowed to choose the clustering of relative word frequencies in documents (TF-IDF approach), which has the following key advantages:

- does not require the use of complex metrics to calculate the distance between signs;

- flexibility of the approach, namely the possibility of using different algorithms of cluster analysis to achieve the desired results, including algorithms that work with an unknown initial number of clusters.

The central idea of this approach is to group documents by topic, which are determined by the most commonly used words in a particular document for the entire collection of documents [2].

Modern clustering algorithms, such as k-means algorithm, BIRCH, DBSCAN, HDBSCAN, STING, are analyzed.

The HDBSCAN algorithm was chosen as the clustering algorithm [3]. This choice is justified by the fact that:

- the algorithm performs "hard" clustering. In the case of publications that should belong to one group, this is a logical choice.

- the algorithm works with an unknown number of clusters;

- the algorithm requires a minimum number of parameters: one really influential parameter is the minimum number of elements in the cluster;

- the ability of the algorithm to find "noise" - points that do not belong to any of the clusters;

- hierarchy of clusters is not required for the task.

Python programming language was chosen for the development, for which many effective libraries for data mining and natural language processing have been created.

The process of clustering text documents takes place in the following sequence: pre-processing of texts, formation of the TF-IDF matrix, application of the clustering algorithm.

Tokenization and lemmatization were performed using spaCy library tools. The TfidfVectorizer class of the scikit-learn library was used to create the TF-IDF matrix.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

```
        #TF-IDF
    tfidf = TfidfVectorizer(lowercase=False, preprocessor=PrePr
oc,
strip_accents=True, tokenizer=TokenizerLemmatizer, max_df=0.85
, min_df=3) vectors = tfidf.fit_transform(list_of_texts)
```

The HDBSCAN class from the hdbscan library was used to implement the clustering.

```
        #Clustering
    res = hdbscan.HDBSCAN(min_cluster_size = minfiles.get(),
    min_samples=2).fit(vectors)
    labels = res.labels_
    clusters = len(set(labels)) - (1 if -1 in labels else 0)
    noise = list(labels).count(-1)
```

Figure 1 shows the result of clustering one of the sets of texts.

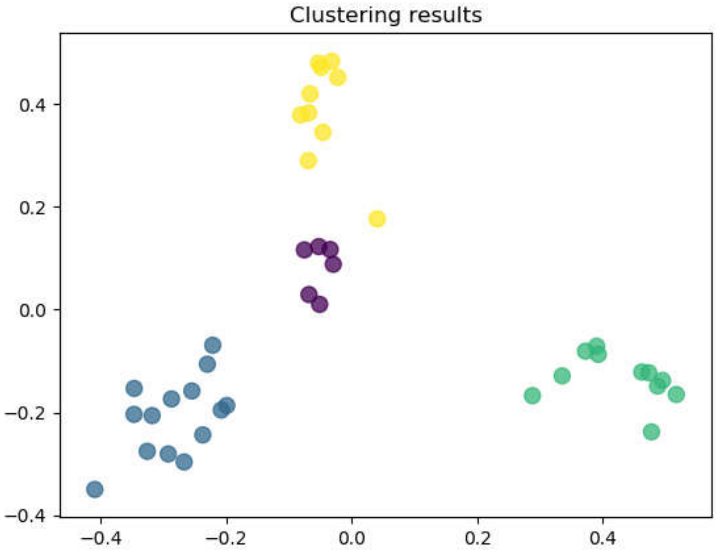

Figure 1 – The result of clustering a set of texts

Thus, the program copes well with the task of clustering text documents in most cases. The best results are observed for texts that differ greatly in content and are approximately the same in size. Problems arise with texts that are small in size or combined with some common theme, in which case a custom stopword mechanism can be used to improve clustering.

Conclusions. The issue of clustering text documents has been studied and on the basis of selected technologies a program has been developed for grouping text files into folders according to their content. To do this, we used tools for building the

TF-IDF matrix, implemented in the scikit-learn library, and natural language processing tools implemented in the spaCy library. The implementation of the HDBSCAN algorithm from the hdbscan library, which was previously tested on sets of numbers, was applied.

REFERENCES

1. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques 3rd Edition. Morgan Kaufmann, 2011, 744 pages.

2. Prafulla Bafna, Dhanya Pramod, Anagha Vaidya. Document clustering: TF-IDF approach. ICEEOT, 2016, p.61-66.

3. L. McInnes, J. Healy, S. Astels. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 2017, p.205-206.

Received 14.01.2022.
Accepted 18.01.2022.

Розробка автоматизованої системи кластеризації текстових документів

Групування текстів у групи схожих за змістом є частим завданням у різних областях людської діяльності. Кластеризація текстових документів використовується для автоматичної категоризації текстових документів, фільтрації листів на електронну пошту, групування веб сторінок у пошукових системах і так далі. Автоматизація даного процесу дозволяє істотно скоротити час, що відводиться на цю задачу.

Існують методи, що є специфічними для кластеризації саме текстових документів: кластеризація відносних частот слів у документах (TF IDF), кластеризація за допомогою суфіксних дерев, тематичне моделювання, підхід Scatter Gather. У цих методах застосовуються сучасні алгоритми кластеризації, такі як алгоритм k середніх, BIRCH, DBSCAN, HDBSCAN, STING.

Метою дослідження є розробка автоматизованої системи кластеризації текстових документів для групування текстових файлів за їх змістом.

Проведений аналіз методів кластеризації текстових документів дозволив обрати кластеризацію відносних частот слів у документах (підхід TF IDF), який має можливість застосування різних алгоритмів кластерного аналізу для досягнення бажаних результатів, у тому числі алгоритмів, що працюють при невідомій початковій кількості кластерів. Центральної ідеєю цього підходу є групування документів за темами, які визначаються найбільш уживаними словами у конкретному документі відносно усієї колекції документів. Процес кластеризації текстових документів відбувається у наступній послідовності: попередня обробка текстів, формування матриці TF IDF, застосування алгоритму кластеризації. У якості алгоритму кластеризації обрано алгоритм HDBSCAN, що працює при невідомій кількості кластерів.

Для розробки автоматизованої системи використовується мова програмування Python, для якої створено безліч ефективних бібліотек для інтелектуального аналізу даних та обробки природних мов.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

В роботі досліджено питання кластеризації текстових документів, розроблено програму для групування текстових файлів у папки за їх змістом на основі підходу TF IDF.

Пономарьов Ігор Володимирович, кандидат технічних наук, доцент кафедри електронних обчислювальних машин факультету фізики, електроніки та комп'ютерних систем Дніпровського національного університету ім. О. Гончара.

Ponomarev Igor Volodimirovich, candidate of technical sciences, associate professor of the department of electronic computers of the faculty of physics electronics and computer systems of the Oles Honchar Dnipro National University.