**VIETNAM NATIONAL UNIVERSITY- HO CHI MINH CITY**

**INTERNATIONAL UNIVERSITY**

**DEPARTMENT OF MATHEMATICS**



**REPORT**

**K-MEANS CLUSTERING AND LATENT SEMANTIC ANALYSIS**

**Lecturer:** Assoc. Prof. Dr. Tran Manh Ha

**TOPIC:**

1. K-MEANS CLUSTERING (Clickstream)
2. LATENT SEMANTIC ANALYSIS and APPLICATIONS

Group members:

| 1. Tran Hoang Gia An 2. Vo Thi Khanh Linh 3. Phan Thanh My | MAMAIU17037  MAMAIU20074  MAMAIU20076 |
| --- | --- |

Table of contents

[1. THE K-MEANS CLUSTERINGS 3](#_heading=h.9b6js11ks9y3)

[1.1. Introduction 3](#_heading=h.ijhws6bsr59u)

[1.2. Applications 3](#_heading=h.uu744zmnueyp)

[1.3. Project applications 4](#_heading=h.uk1abcno3t3w)

[Data description 4](#_heading=h.ep6qcj97agjb)

[Pre-processing Data 5](#_heading=h.9xnl633p8c2y)

[Elbow Method 10](#_heading=h.65hu7beliou0)

[Result 11](#_heading=h.f33oim32ff0b)

[2. LATENT SEMANTIC ANALYSIS 12](#_heading=h.gmamodc8r736)

[2.1. Introduction 12](#_heading=h.10fjbidphpj)

[2.2. Project applications 13](#_heading=h.17tad0xf63ak)

[Data description 13](#_heading=h.wr1acy1mnmka)

[Pre-processing Data 13](#_heading=h.dhfwn4xwsdmc)

[LSA method 13](#_heading=h.nq0s5gvp3b54)

[Result 13](#_heading=h.vzdgeruvwx0i)

[2.3. References 13](#_heading=h.36qe8mq9xquz)

[REFERENCES 14](#_heading=h.gjdgxs)

# 1. THE K-MEANS CLUSTERINGS

## 1.1. Introduction

K-means clustering is a type of [unsupervised learning](https://blogs.oracle.com/ai-and-datascience/post/supervised-vs-unsupervised-machine-learning), which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

* The centroids of the K clusters, which can be used to label new data
* Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

## 1.2. Applications

Common usages of **K-Means Clustering**  include the following:

* **Customer Segmentation:** Clustering helps marketers improve their customer base, work on target areas, and segment customers based on purchase history, interests, or activity monitoring. This segmentation helps companies target specific clusters/groups of customers for specific campaigns.
* **Document Classification:** Cluster documents in multiple categories based on tags, topics, and the content of the document, this is a very standard classification problem and k-means is a suitable algorithm for this purpose. The initial processing of the documents is needed to represent each document as a vector and uses term frequency for identifying commonly used terms that help classify the document. The document vectors are then clustered to help identify similarity in document groups.
* **Identifying crime localities:** The type of crime, the location of the crime, and the relationship between the two can provide valuable insight into crime-prone areas within a city or locality when data on crimes is available for specific places in a city.
* **Call record detail analysis:** The data that telecom firms record about a customer's calls, messages, and online behavior is called a call detail record. When combined with demographic data about the customer, this information offers deeper insights into their demands. The K-means clustering technique can be used to understand client groupings based on their usage patterns over time.
* **Automatic clustering of IT alerts:** Large company IT infrastructure technology components like network, storage, or databases produce a lot of alert signals; these messages need to be manually assessed for priority for subsequent operations because they could indicate operational concerns. Data clustering can help in failure prediction and can shed light on different alert types and mean times to repair.
* **Rideshare data analysis:** The publicly accessible Uber ride information dataset offers a wealth of useful information on traffic, transit times, peak pickup locations, and more. Planning for future cities can be aided by understanding the traffic patterns in cities through the analysis of this data.
* **K-Means Clustering** can also be used for performing **image segmentation** by trying to group similar pixels in the image together and creating clusters. The different clusters formed are different objects in an image. K-means clustering can be used in recommendation engines as well. For instance, in a music streaming application, comparable genres or song types are clustered together for a user based on their listening habits, and the application can then suggest the songs that are the most similar to those groups of songs.

## 1.3. Project applications

### Data description

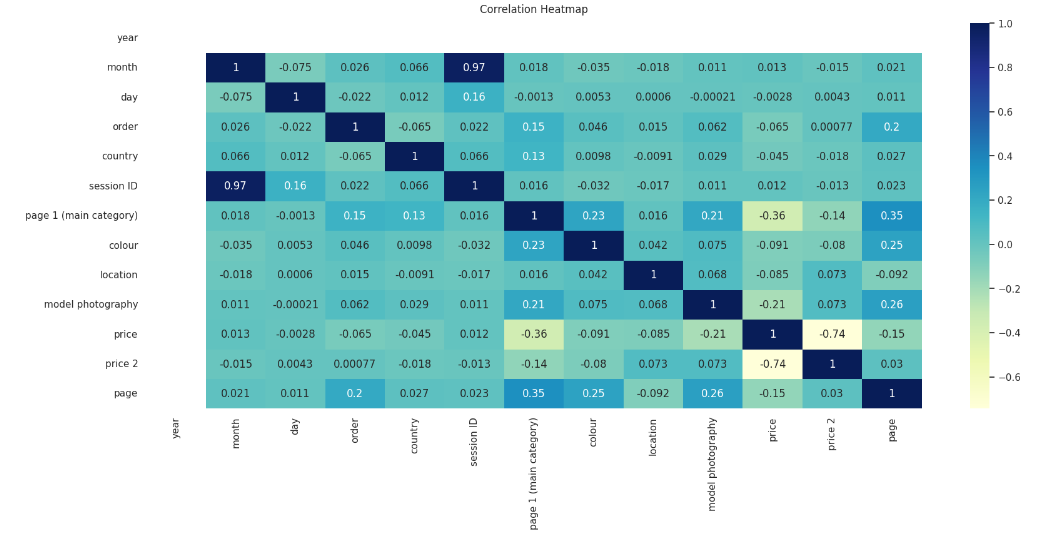
The collection includes clickstream data from an online store that offers maternity clothing. The information comes from the first five months of 2008.

The data has 165474 rows and 14 attributes columns.

| year: | 2008. |
| --- | --- |
| month: | 4 (April) to 8 (August). |
| day: | 1 to 31. |
| order: | Sequence of clicks during one session. |
| country: | The IP address of the country. |
| session ID: | Variable indicating session-id (short record). |
| page 1(main category): | Main product category. |
| page 2(clothing model): | The code of each product. |
| colour: | Colour of the product. |
| location: | Photo location on the page. |
| model photography: | Variable with two categories. |
| price: | Price of product (unit: US dollars). |
| price 2: | Indicates whether a product's price is higher than the average price for its entire category. |
| page: | Page number in the e-store website (from 1 to 5). |

### Pre-processing Data

* Correlation heatmap



This correlation heatmap shows all about the dataset's feature variables interdependence.

* Month and session ID are strongly correlated.
* Since the year feature only has one value, it is useless.
* Model photography and price, page 1 and price are negatively correlated.
* IQR Method

Interquartile Range Method is a way to measure the spread of the middle 50% of a dataset. It is calculated as the difference between the first quartile\* (the 25th percentile) and the third quartile (the 75th percentile) of a dataset.

IQR = Q3 - Q1

Lower fence = Q1 - 1.5IQR

Upper fence = Q3 + 1.5IQR

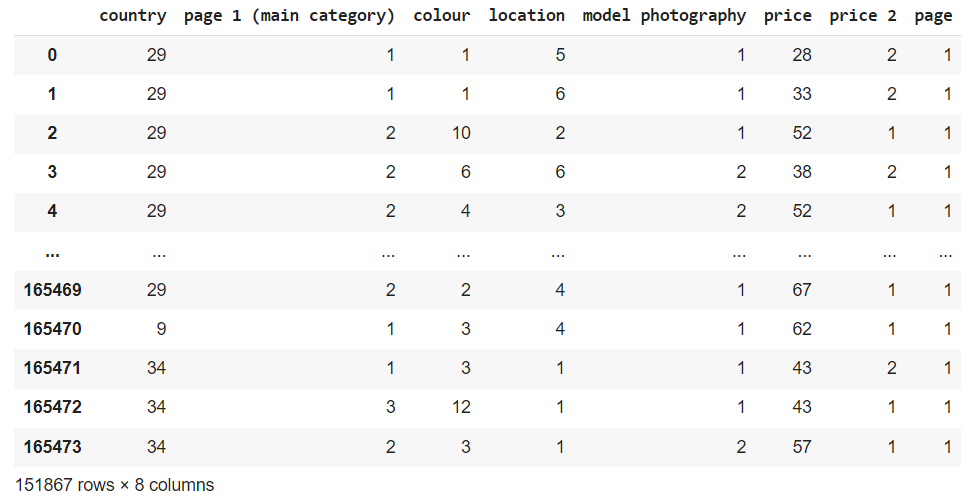
Using IQR Method to identify outliers of the price and the page. We get the distribution after IQR:



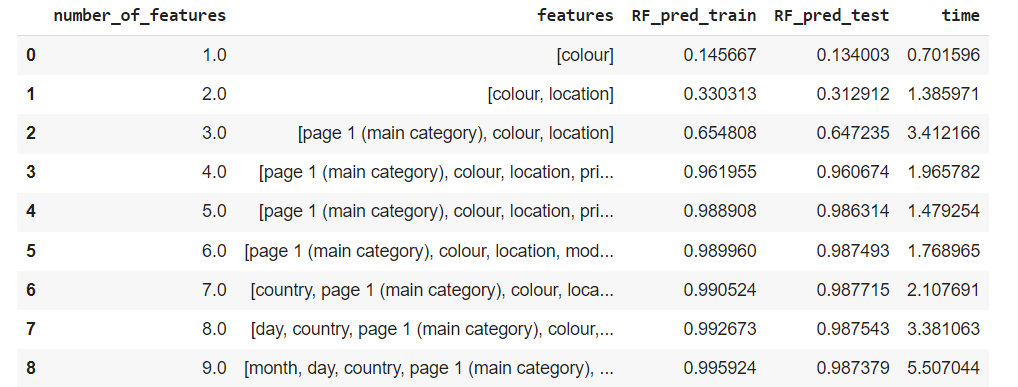


* Session ID, order and the page 2 (clothing model) field should be removed because they are unrelated to modeling.

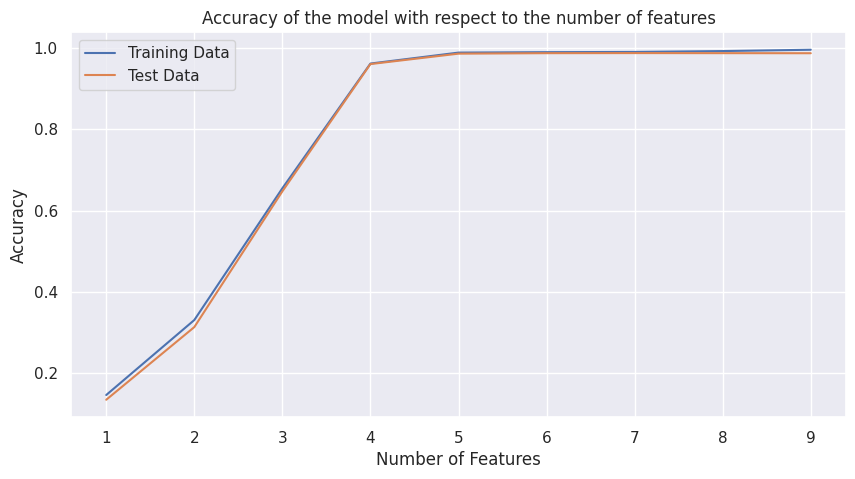
Then, the screenshot of the DataFrame after using IQR method and dropping some irrelevant columns is given below.

****

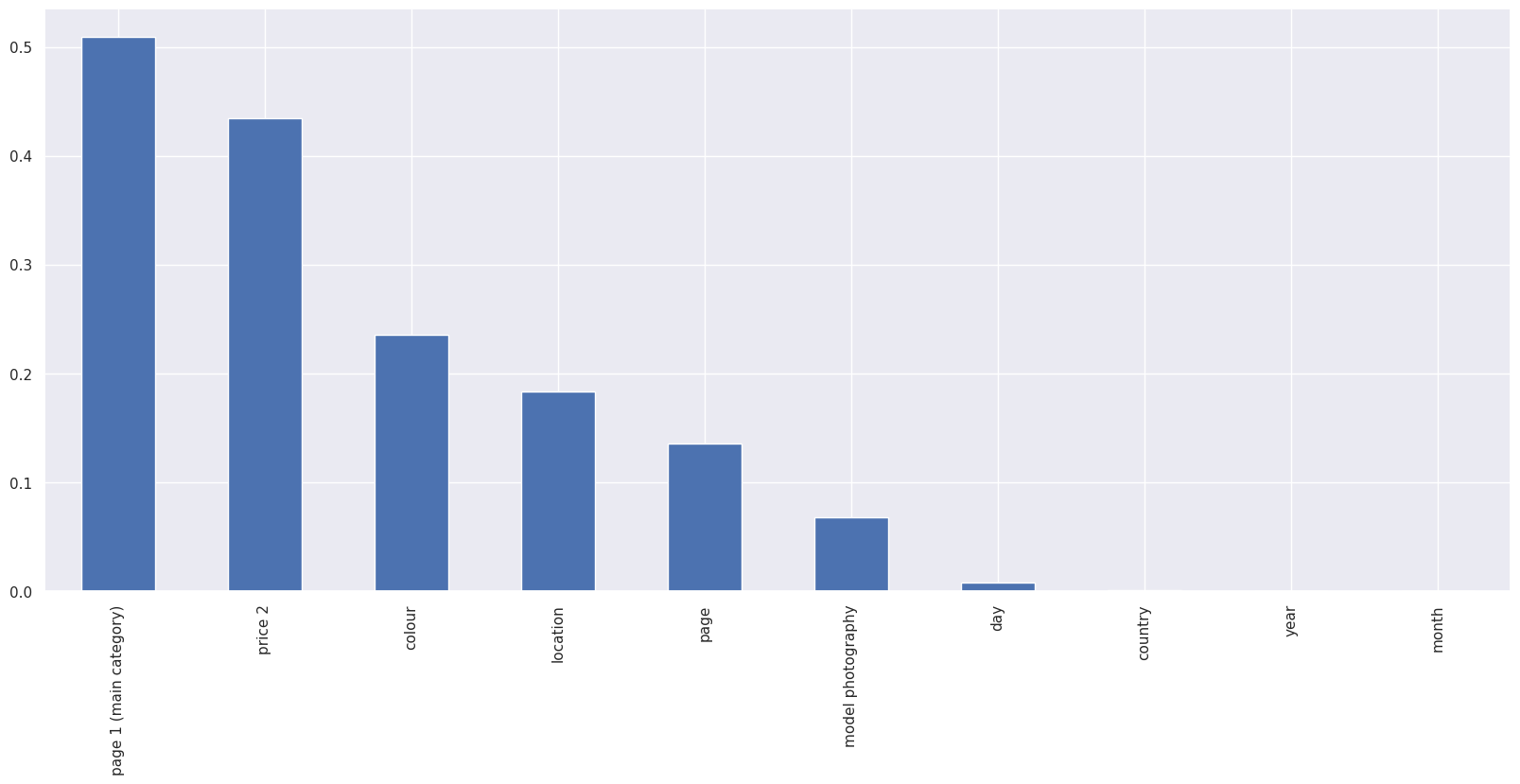
* Find features



* Accuracy



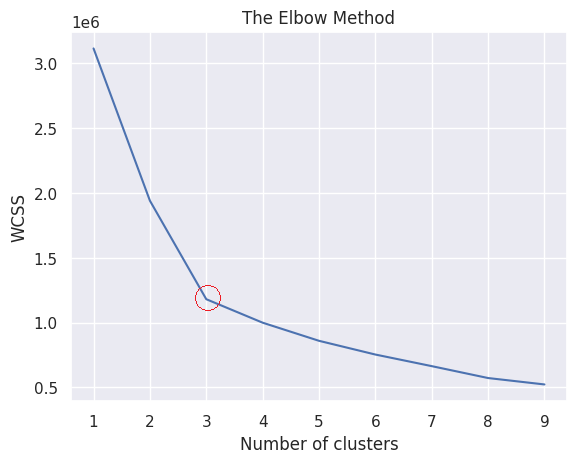
* From the figure below:



We can choose 6 features: page 1 (main category), color, price 2, location, model photography and page

### Elbow Method

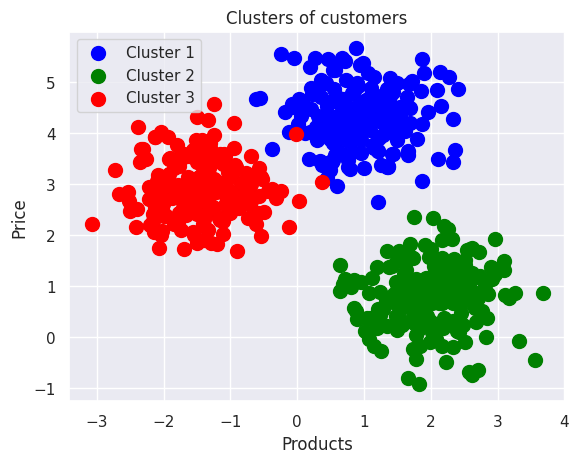
Taking the variable page1 and apply the algorithm to find out clusters

****

The point of bend or a point of the plot appears to be like an arm, then that point is considered the best value of K. Hence, in this case, k=3 is the optimal number.

### Result

We have successfully identified the clusters based on the price of the products and the products. According to the Elbow Method Graph, there are 3 clusters formed on the feature page 1. These three clusters imply that there are 3 buying patterns-based on the products and the price of the products.

****

# 2. LATENT SEMANTIC ANALYSIS

## 2.1. Introduction

Latent Semantic Analysis (LSA) is a computational approach that allows for the identification of connections between different documents and the words they contain. This is achieved through the use of mathematical algorithms that scan the documents and uncover underlying associations between words and concepts. LSA has been in use since the 1960s, and it was patented in 1988. Its primary application is in the identification of concepts and the automated categorization of documents. However, LSA has also found utility in other domains, such as information retrieval, text summarization, search engine optimization, and software engineering.

TF-IDF is a tool used by computers to help determine which words in a document are the most important. It looks at how often a word appears in the document and how rare it is compared to other documents in a big group called a corpus. The more often a word appears in the document, the more important it is. But if the word appears in lots of other documents too, it's not as important. We use these numbers to calculate a final score that tells us how important each word is in the document. This tool is often used in fields like information retrieval (IR) and machine learning.

SVD is a mathematical technique that computers use to reduce the amount of information they need to understand how words are related to each other. In Latent Semantic Analysis (LSA), we typically assume that only the top 100 or so words are necessary to capture the meaning of a document. Words that frequently occur together will have similar scores, indicating that they are related. The remaining words are often considered to be irrelevant noise, and can be disregarded.

## 2.2. Project applications

### Data description

The dataset at hand comprises scientific papers from various conferences across the globe, in which contains:

* + The link to a given research paper can be obtained by combining the prefix "http://dx.doi.org/" with the paper's unique digital object identifier (DOI).
  + Other relevant information pertaining to the articles includes their type (ArticleType) and style (CitationStyle).
  + Lastly is a collection of HTML code obtained from the web. The code is structured such that the content of each tag is contained within the respective tags (citationStringAnnotated).

In this mini project, we were assigned to seek similar documents from this dataset to the provided queries, in which 7 1-billion-row csv files were chosen to be executed. However, the lack of resources prevented the execution, so that only smaller datasets were taken into this procedure. For the Preprocessing Data and LSA method, the algorithms will be applied on 2 sizes of datasets: around 150,000 rows (small datasets), and 3,000,000 rows (bigger data sets).

### Preprocessing Data

Here are steps that were implied and their pseudo-code to construct this progress.

1. Remove duplicate rows.   
   Pseudocode: dup = description.drop\_duplicates(subset=[‘column\_name’])
2. Lower case: to synchronize all texts into lower texts.  
   Pseudocode: lower = dup.str.lower()
3. Remove URl:  
   Pseudocode: url = lower.remove\_url()
4. Remove HTML tag  
   Pseudocode: htm = url.remove\_html()
5. Remove punctuation  
   Pseudocode: punctuation = htm.remove\_punctuation()
6. Remove numbers  
   Pseudocode: numbers = punctuation.remove\_numbers()
7. Remove word containing numbers  
   Pseudocode: word\_containing\_numbers = numbers.remove\_word\_containing\_numbers()
8. Remove short words: remove words that are less than 3 letters  
   Pseudocode: short\_words = word\_containing\_numbers.remove\_short\_words()
9. Remove long words: remove words that are less than 25 letters  
   Pseudocode: long\_words = short\_words.remove\_long\_words()

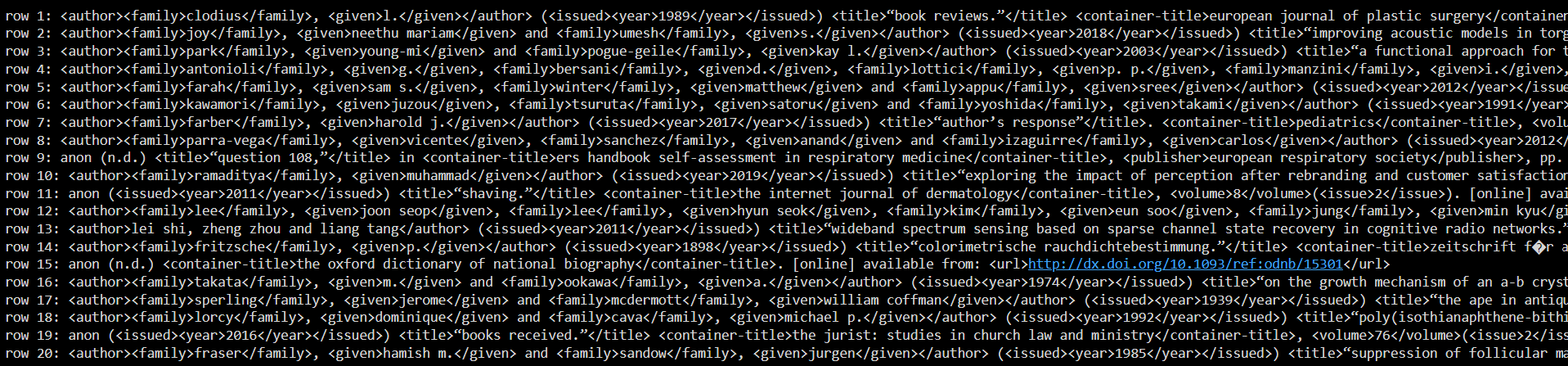
#### Small datasets (around 150,000 rows)



Observing the data set exposes several problems that could compromise the effectiveness of the LSA algorithm.



We drop duplicate rows.



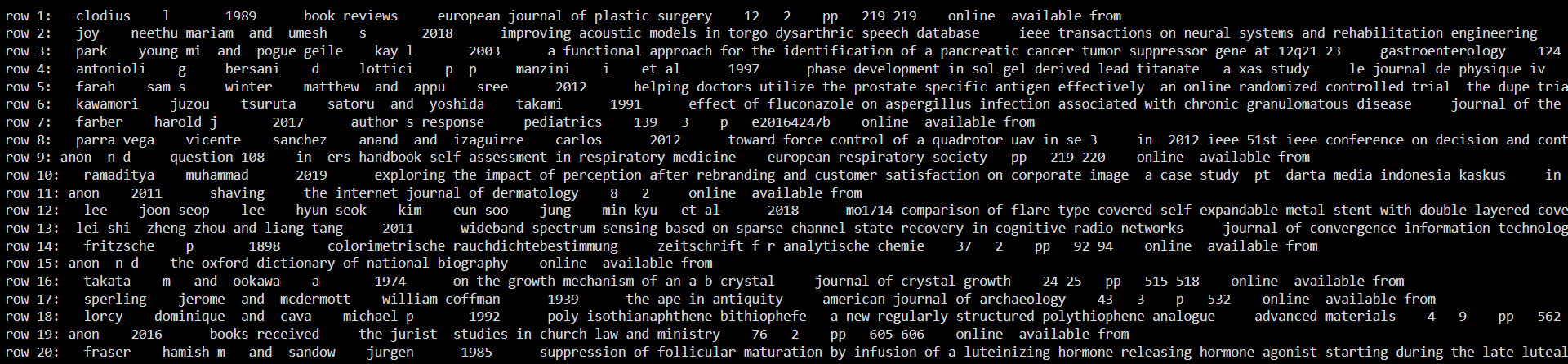
We change the strings to lowercase.



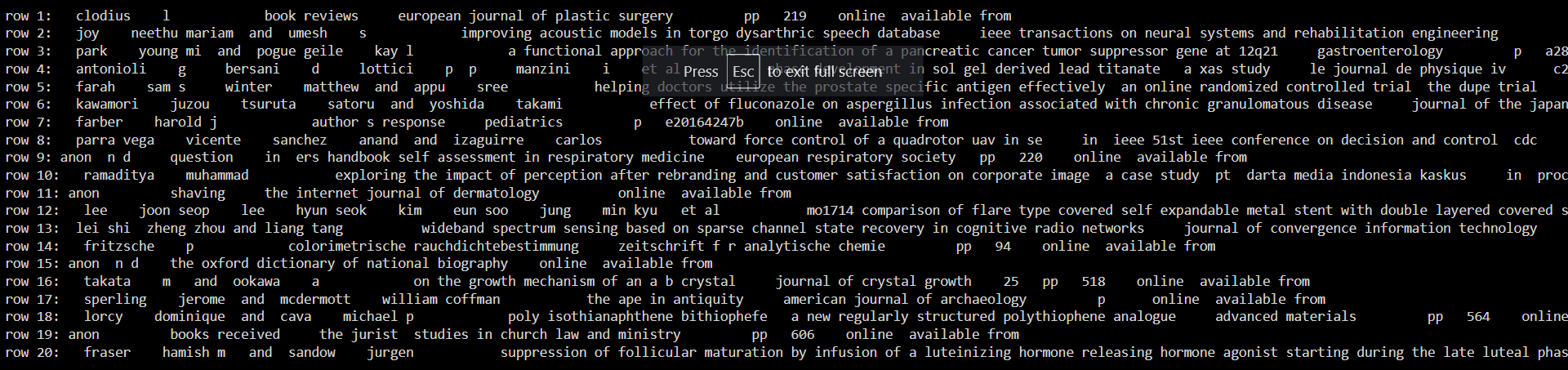
We eliminate URLs.



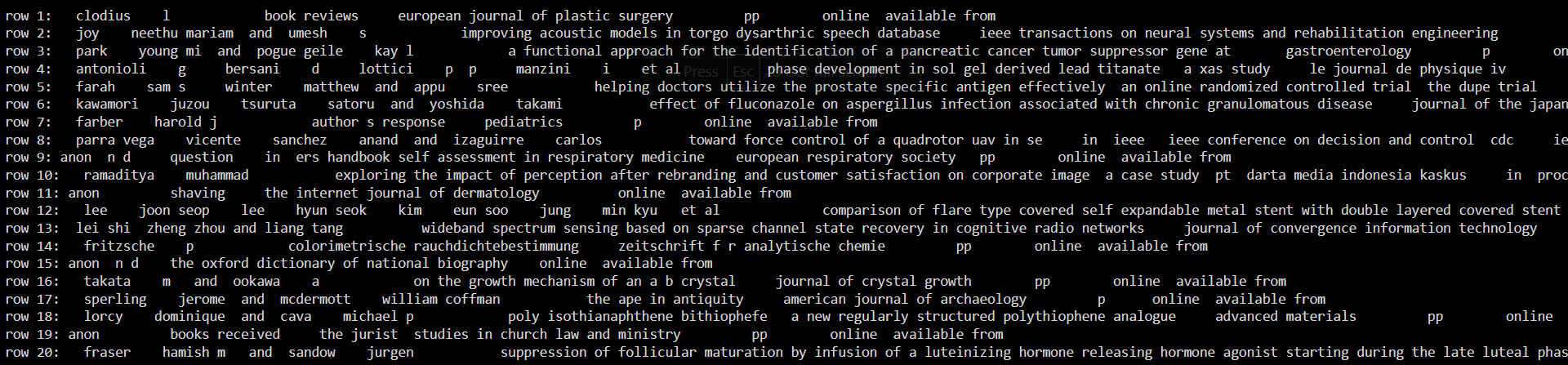
Next, we remove the html tag.



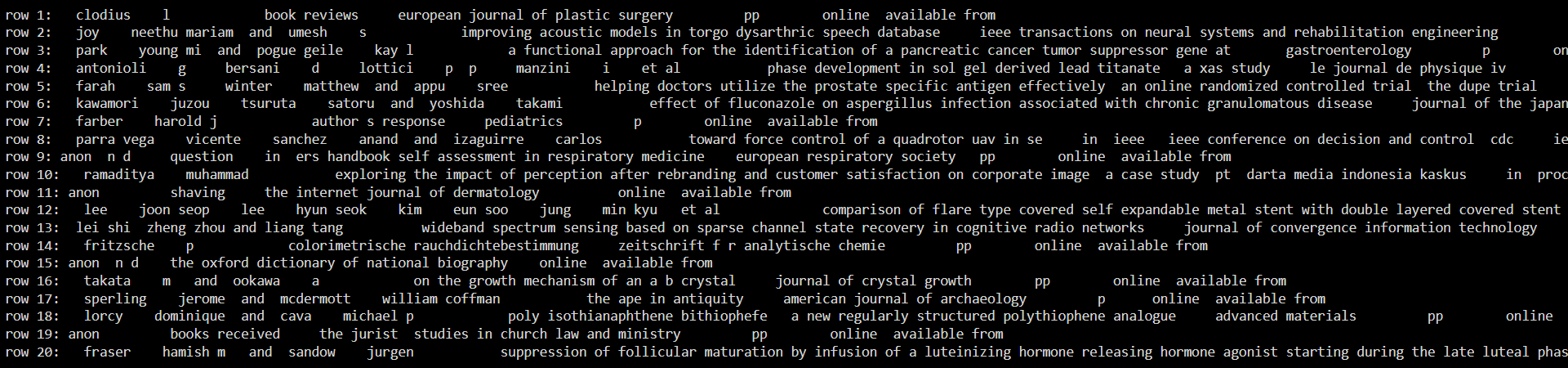
We remove punctuation marks from strings.

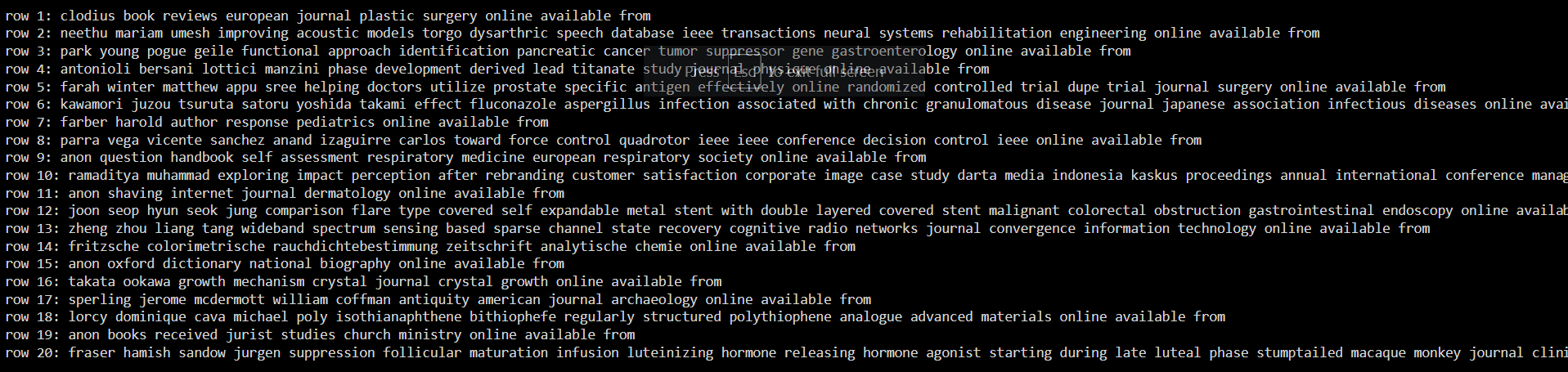


We remove numbers.



After that, we drop words containing numbers.

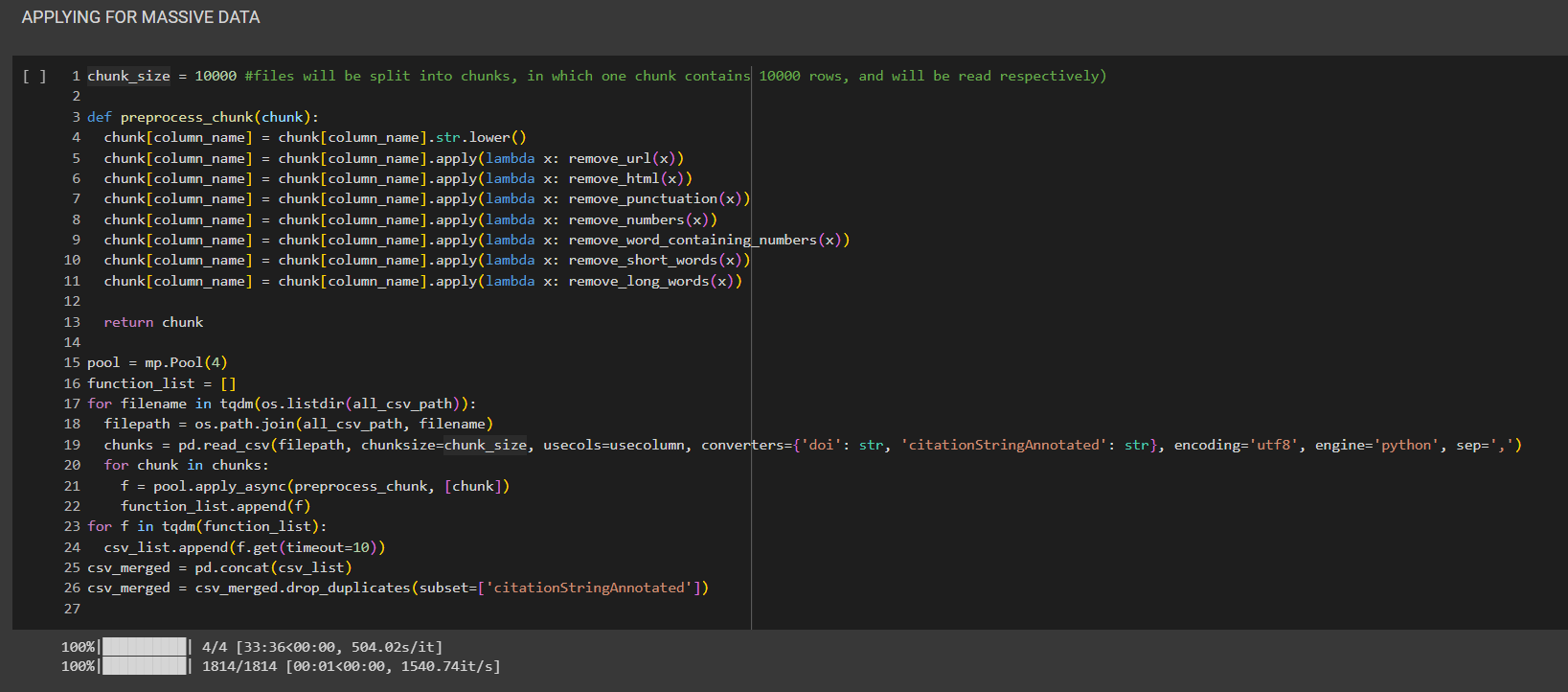
 We just eliminate words with fewer than 3 letters.



Words longer than 25 letters should be eliminated.

#### Bigger datasets (around 3,000,000 rows)

We executed this using chunk\_size, which means the code will divide csv files into smaller parts (for this example, 10,000-row parts), then read them, respectively. We executed 4 1-billion-row csv files within approximately 35 minutes to get this result.

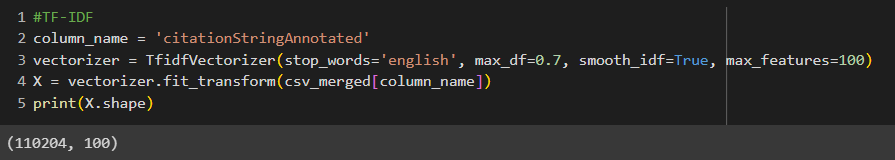


### LSA method

#### Small datasets (around 150,000 rows)

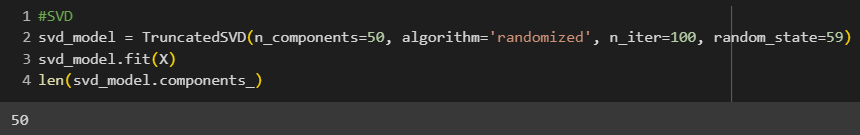
##### TF\_IDF (Term Frequency-Inverse Document Frequency)

We employ the TF-IDF to identify key terms with a high frequency of occurrence.



##### SVD (Singular Value Decomposition)

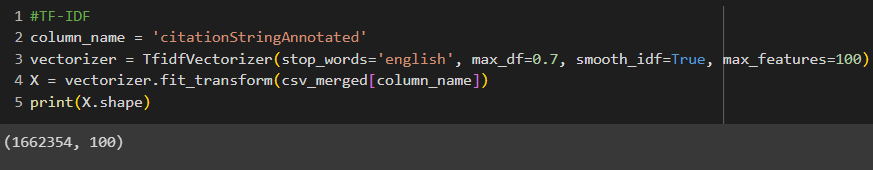
Next, we use SVD. Based on the Frobenius norm distance between the two matrices, the SVD approach will identify a class of matrices that most closely resemble a given matrix.



#### Bigger datasets (around 3,000,000 rows)

The step will be similarly from the small datasets’ parts

##### TF\_IDF (Term Frequency-Inverse Document Frequency)

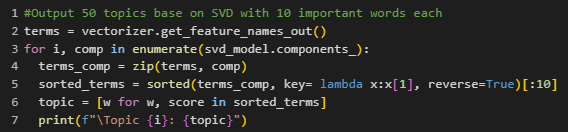


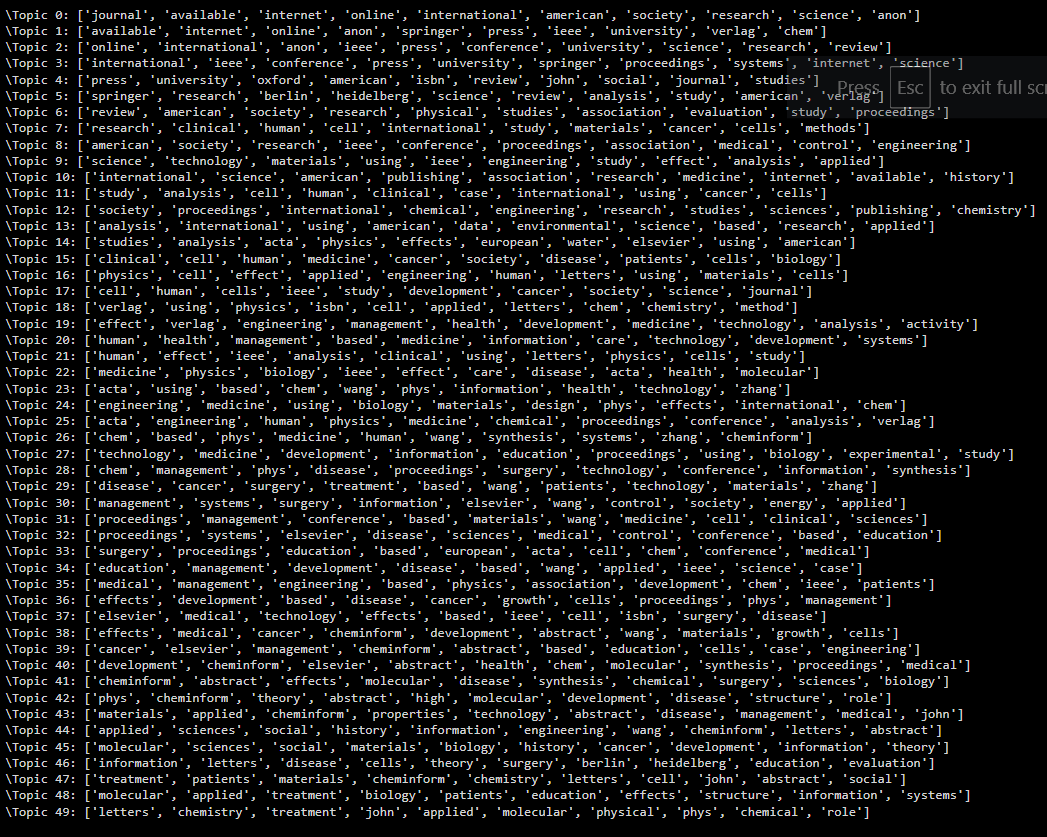
##### SVD (Singular Value Decomposition)

Unfortunately, we ran out of resources, so we cannot compute any solutions for this part. Also, the final result will be the result of the small datasets.

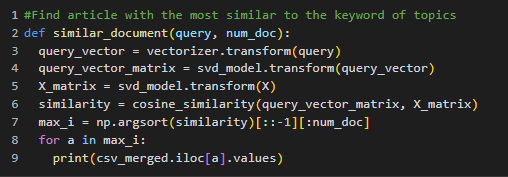
### Result

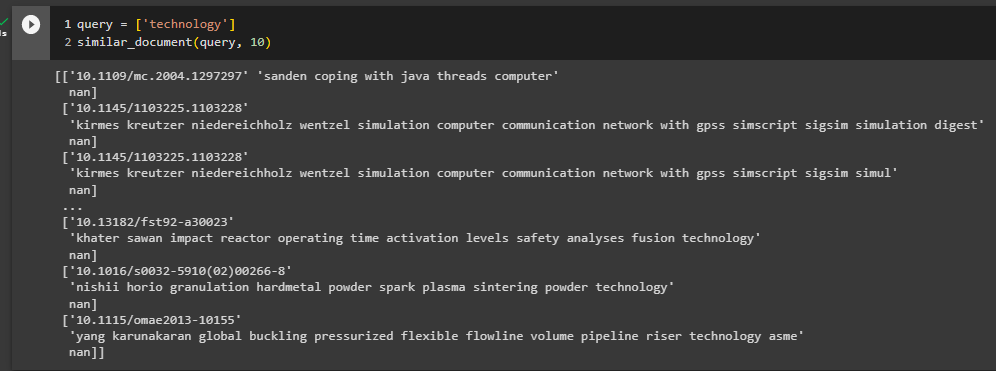
Our calculations have the capability to produce a comprehensive list of 50 distinct topics for the perusal. These topics are designed to encapsulate the most salient concepts and ideas, with each topic comprising a carefully curated selection of the 10 most significant words that are pertinent to the subject matter at hand. This functionality enables you to swiftly and efficiently identify the core themes and ideas of each topic, thereby obviating the need to sift through extraneous and irrelevant information.

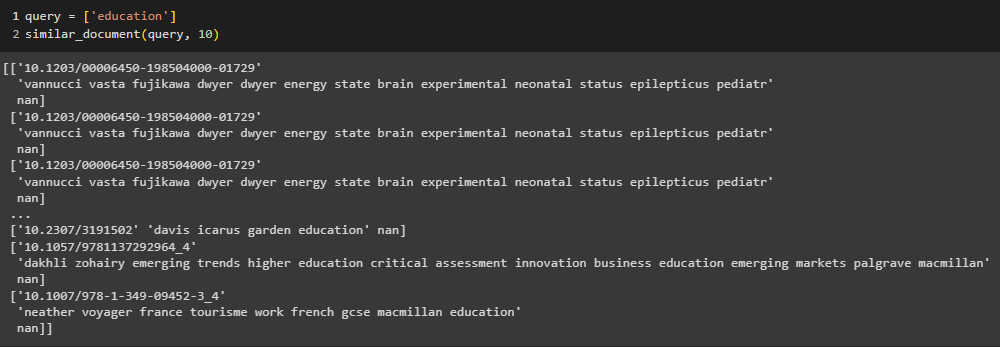




After that, we find the most similarity to the keyword in the ‘query’ section and print out 1o rows of our datasets with the highest similarity scores.







# REFERENCES

1. [Introduction to K-means Clustering:](https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering)
2. [Applications of K-means Clustering:](https://www.linkedin.com/pulse/k-means-clustering-its-applications-ritvik-ranjan)
3. [Dataset](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=ae466a9818ebce8dae3ccfeb17b0?persistentId=doi%3A10.7910%2FDVN%2FLXQXAO&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=name&fileSortOrder=desc&fbclid=IwAR3GhCOiUJpZ_nnDUhegvhfqyFl0DOp2LVKqJ2NIIvbdwoRx6ZdDbBPX1L4)
4. [Introduction to LSA](#_heading=h.30j0zll)
5. [TF-IDF](https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/)
6. [Dataset](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=ae466a9818ebce8dae3ccfeb17b0?persistentId=doi%3A10.7910%2FDVN%2FLXQXAO&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=name&fileSortOrder=desc&fbclid=IwAR3GhCOiUJpZ_nnDUhegvhfqyFl0DOp2LVKqJ2NIIvbdwoRx6ZdDbBPX1L4)
7. [IQR Method](https://online.stat.psu.edu/stat200/lesson/3/3.2#:~:text=We%20can%20use%20the%20IQR,add%20this%20value%20to%20Q3.)