id	الاسم
20221468416	ابراهيم محمد ابراهيم النقيب
20221441990	حازم محمد احمد بكر
20221450304	أحمد سمير عبدالفتاح أمين
20221041599	خالد عبدالحميد حمدي محمد
20221373780	احمد سمير عبدالعظيم
20221375760	احمد السيد عجمي احمد
20221150099	نور الدين انور محمد
20221449224	أحمد ناصر عبدالفضيل
20221442040	مروان اشرف محمد عبدالباقي
20221453750	مصطفى صالح مصطفى

First The Reuters Dataset:

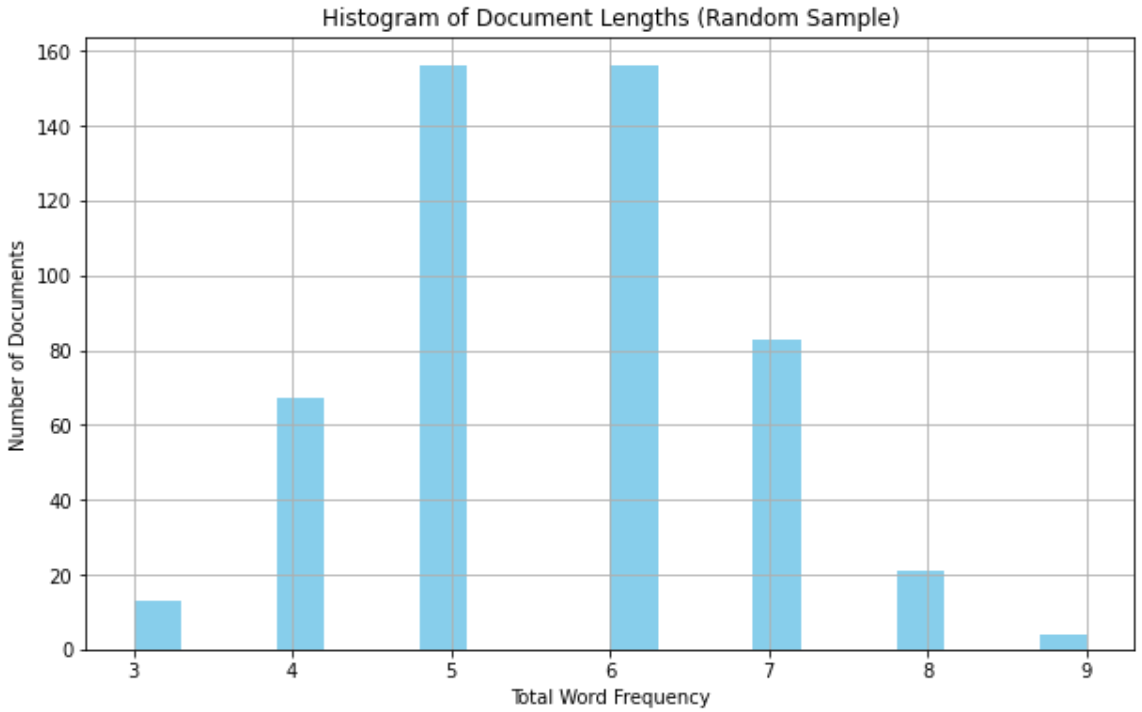
First the dataset documentation:

- The Reuters dataset is a text categorization text collection containing articles' titles, main article, topic, and other information about the article.
- The dataset contains about 19,000 articles with around 114,671 unique words.

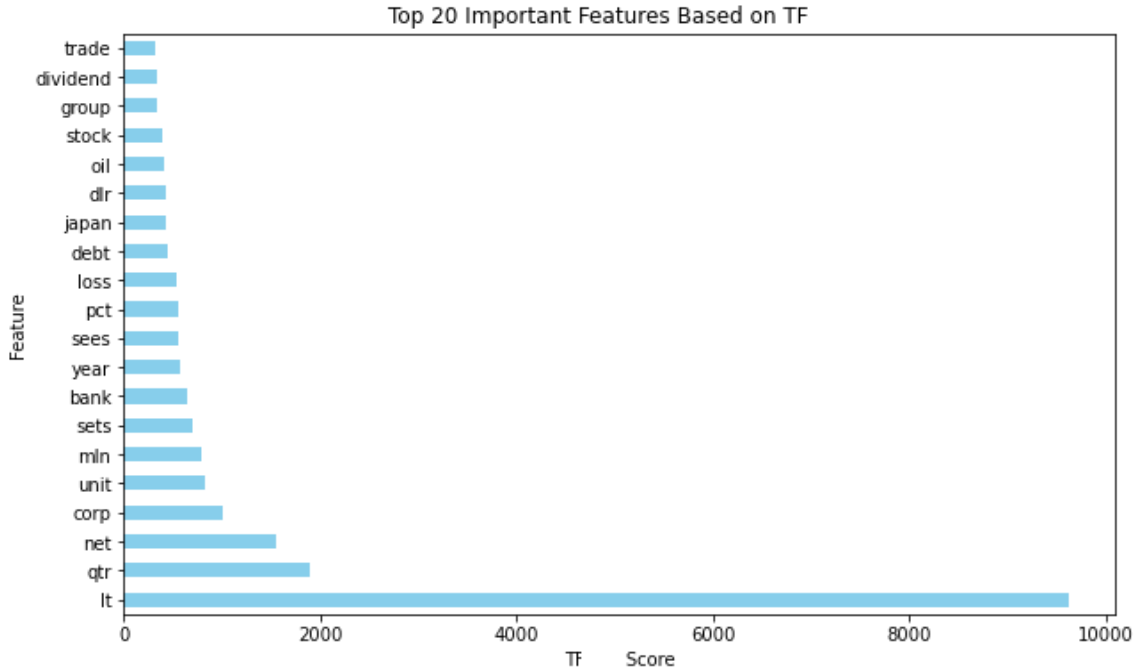
Second, project workflow:

- In the beginning of the coding process, we started by importing needed libraries and packages for the project.
- Then we started gathering the articles using regular expressions in a dictionary containing the article's title and its content.
- Now we gathered all the words of the articles' content in a list its name is "all_words" the length of this list is about 2,000,000 words.
- Then we used the unique property of sets to acquire unique words from the "all_words" list, the length of the "unique_words_list" is about 114,000 words.
- Now comes the filtration stage we removed all the words containing '$' or a number and we removed any of those characters from any word containing it '"()<>,:.?' for example the word "(ahmed" is converted to "ahmed" and so on the length of the filtered words list came to about 84,000 words.
- Then we started removing all the stop words like "is", "the", and so on from the "unique_words_list", the new filtered list is now called "unique_words_list_filtered".
- The "unique_words_list_filtered" contains about 49,000 words.
- Then we used the CountVectorizer to vectorize the words list and we fitted it to the titles we obtained.
- Then we created the term frequency matrix.

Third, the visualizations:



In this Histogram we notice that the word frequencies are normally distributed with mean 5.5

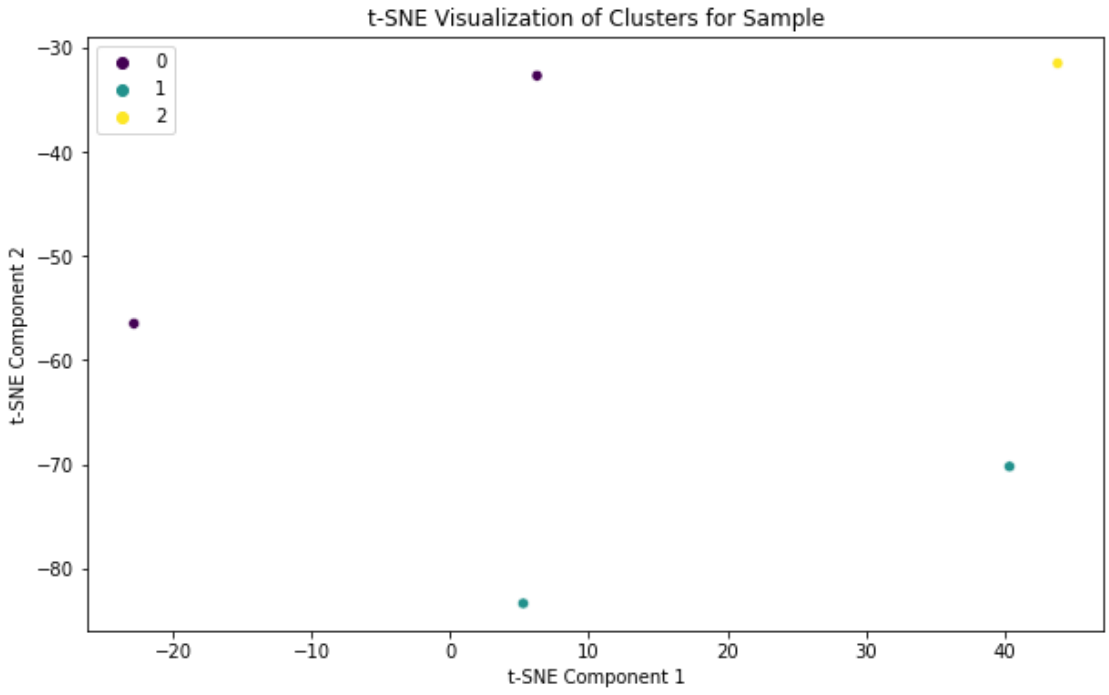


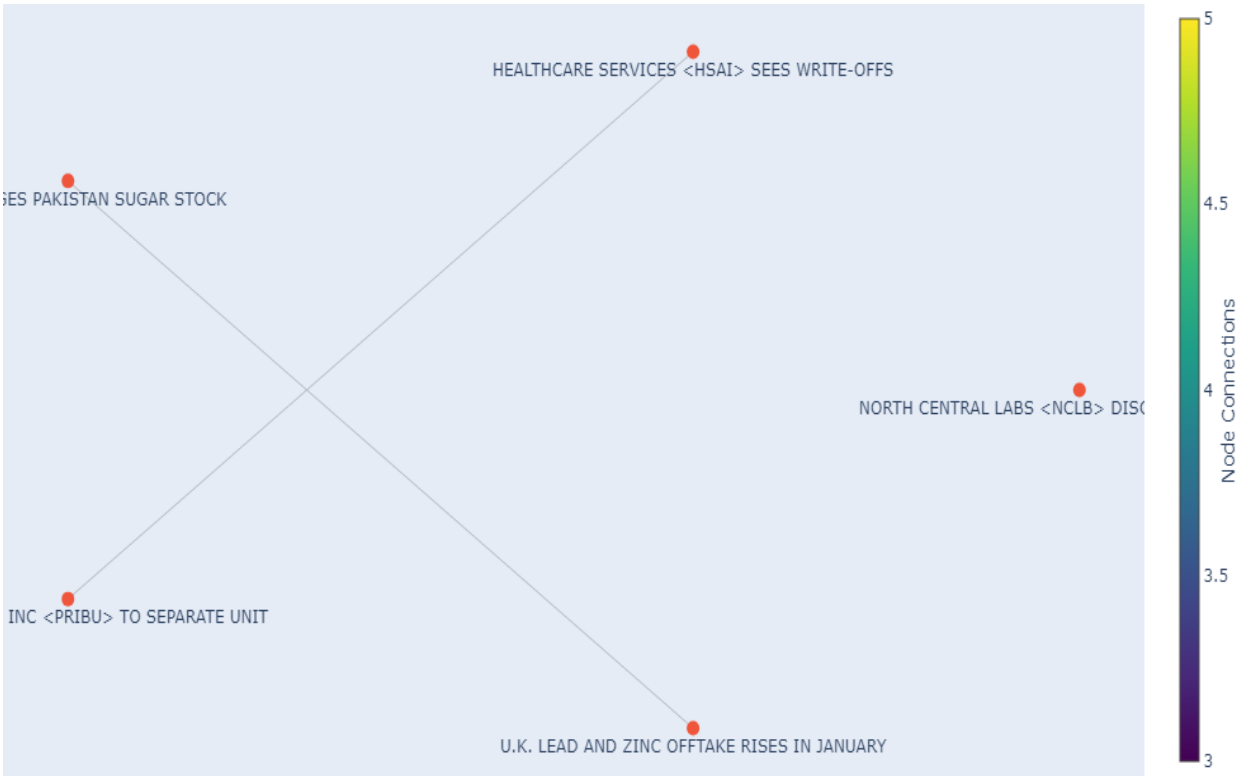
we plot the most popular 20 words in all the documents.

Clustering Machine learning model¶

We build a clustering model to make clustering for a random sample of documents based on KMeans algorithm and we cluster them into 3 groups.

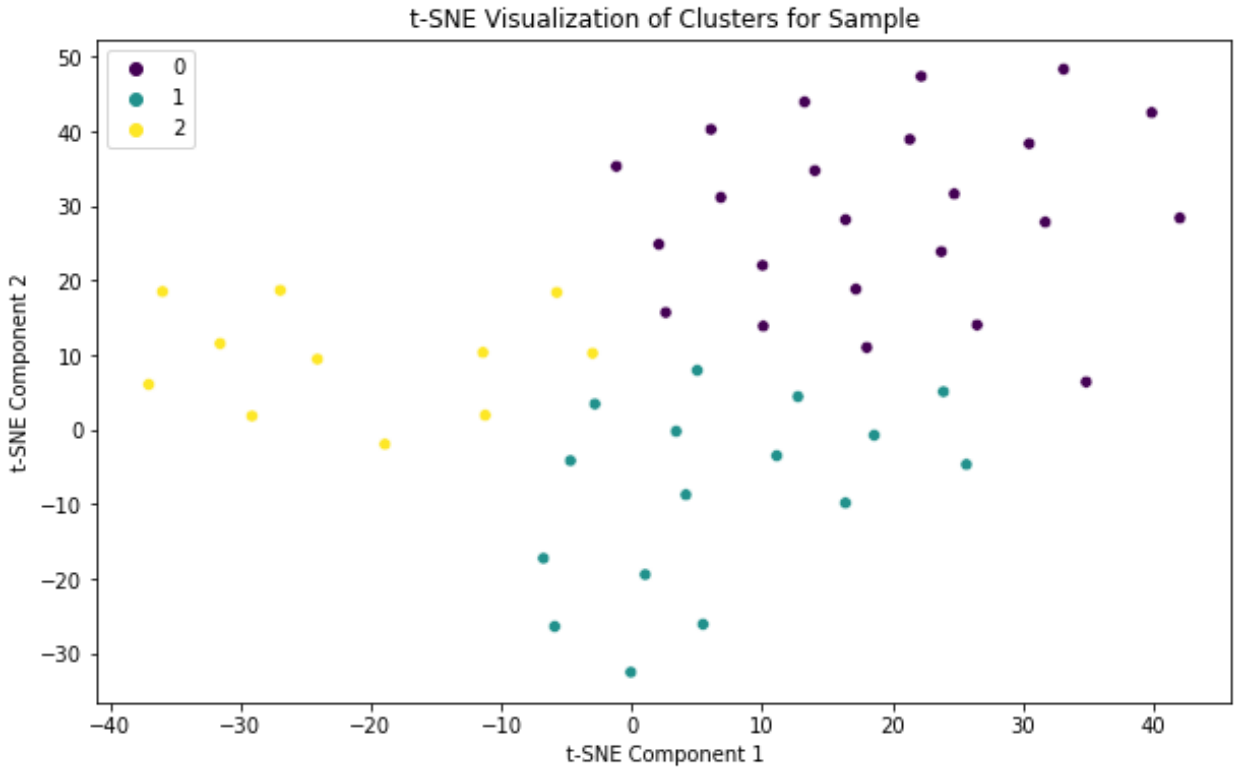
This is a simple example for only 5 documents.

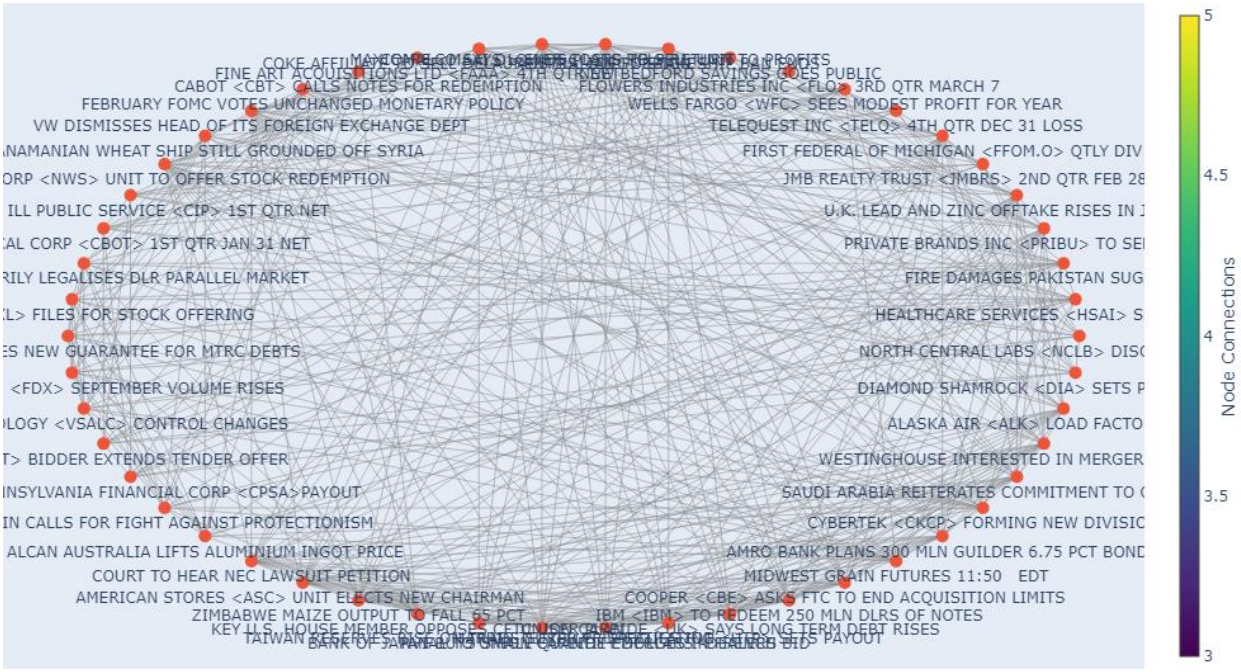




This network for the relation between documents that share the same class.

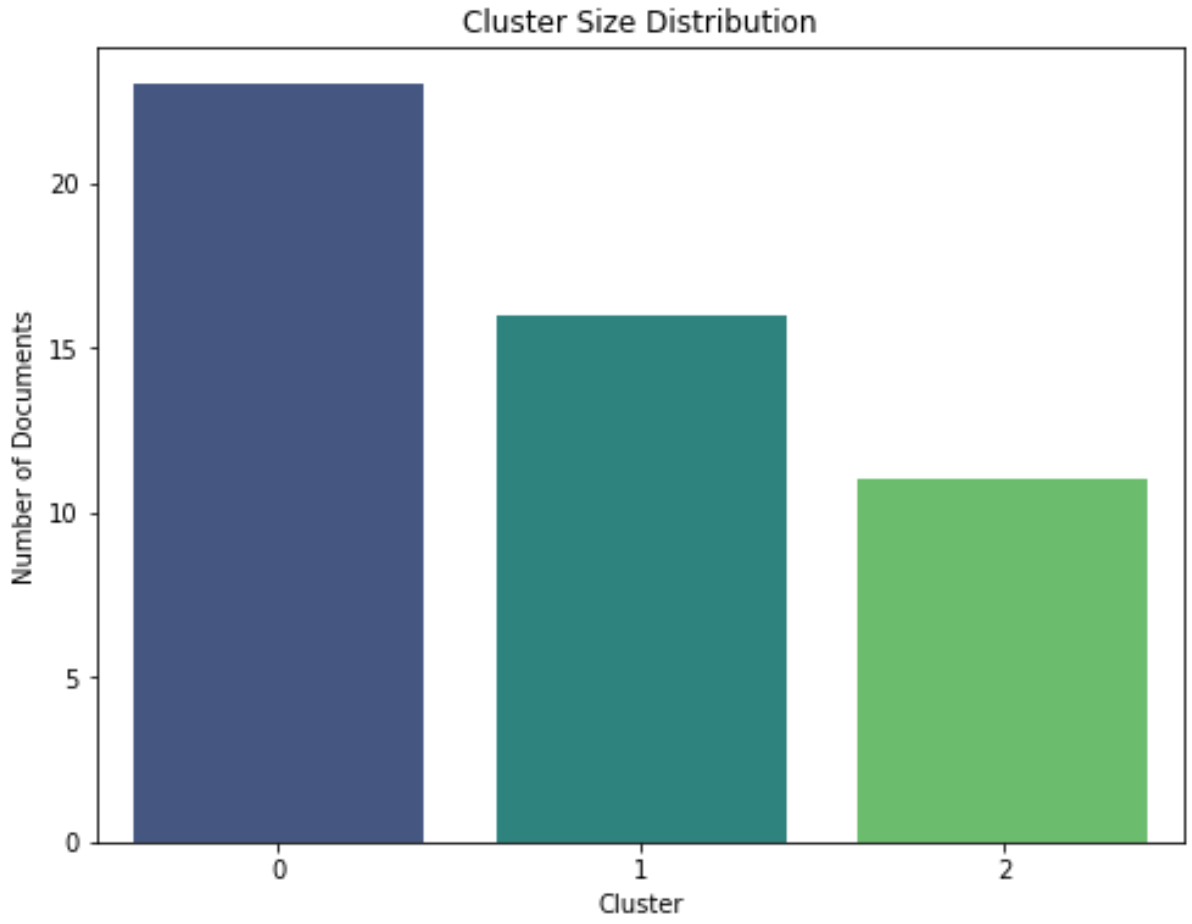
This is with 50 documents as a sample.





Document name	cluster
NORTH CENTRAL LABS <NCLB> DISCUSSES AIDS TESTS	2
HEALTHCARE SERVICES <HSAI> SEES WRITE-OFFS	1
FIRE DAMAGES PAKISTAN SUGAR STOCK	1
PRIVATE BRANDS INC <PRIBU> TO SEPARATE UNIT	0
U.K. LEAD AND ZINC OFFTAKE RISES IN JANUARY	0
JMB REALTY TRUST <JMBRS> 2ND QTR FEB 28 NET	2
FIRST FEDERAL OF MICHIGAN <FFOM.O> QTLY DIV	0
TELEQUEST INC <TELQ> 4TH QTR DEC 31 LOSS	2
WELLS FARGO <WFC> SEES MODEST PROFIT FOR YEAR	1
FLOWERS INDUSTRIES INC <FLO> 3RD QTR MARCH 7	2
NEW BEDFORD SAVINGS GOES PUBLIC	2
AUSTRALIAN FOREIGN SHIP BAN ENDS	0
COMALCO SAYS LOWER COSTS HELP RETURN TO PROFITS	0
MAXICARE <MAXI.O> ENDS PLANS TO SELL UNIT	0
COKE AFFILIATE TO SELL DELAURENTIIS <DEG> STAKE	0
FINE ART ACQUISITIONS LTD <FAAA> 4TH QTR NET	2
CABOT <CBT> CALLS NOTES FOR REDEMPTION	1
FEBRUARY FOMC VOTES UNCHANGED MONETARY POLICY	1
VW DISMISSES HEAD OF ITS FOREIGN EXCHANGE DEPT	0
PANAMANIAN WHEAT SHIP STILL GROUNDED OFF SYRIA	1
NEWS CORP <NWS> UNIT TO OFFER STOCK REDEMPTION	0
CENTRAL ILL PUBLIC SERVICE <CIP> 1ST QTR NET	2
CABOT MEDICAL CORP <CBOT> 1ST QTR JAN 31 NET	2
SARNEY TEMPORARILY LEGALISES DLR PARALLEL MARKET	1
GRANGES <GXL> FILES FOR STOCK OFFERING	1
HONG KONG PROVIDES NEW GUARANTEE FOR MTRC DEBTS	1
FEDERAL EXPRESS <FDX> SEPTEMBER VOLUME RISES	1
VISUAL TECHNOLOGY <VSALC> CONTROL CHANGES	1
NECO <NPT> BIDDER EXTENDS TENDER OFFER	2
CENTRAL PENNSYLVANIA FINANCIAL CORP <CPSA>PA...	2
BRITAIN CALLS FOR FIGHT AGAINST PROTECTIONISM	1
ALCAN AUSTRALIA LIFTS ALUMINIUM INGOT PRICE	1
COURT TO HEAR NEC LAWSUIT PETITION	0
AMERICAN STORES <ASC> UNIT ELECTS NEW CHAIRMAN	0
ZIMBABWE MAIZE OUTPUT TO FALL 65 PCT	0
KEY U.S. HOUSE MEMBER OPPOSES CFTC USER PLAN	1
TAIWAN RESERVES RISE ON TRADE SURPLUS, SPECULATION	0
BANK OF JAPAN BUYS SMALL QUANTITY DOLLARS -DEALERS	0
RVIAL TO UNION CARBIDE EMERGES IN FRENCH BID	0
HARRIS-TEETER PROPERTIES INC <HTP> SETS PAYOUT	0
UNION CARBIDE <UK> SAYS LONG TERM DEBT RISES	0
IBM <IBM> TO REDEEM 250 MLN DLRS OF NOTES	0
COOPER <CBE> ASKS FTC TO END ACQUISITION LIMITS	0
MIDWEST GRAIN FUTURES 11:50 EDT	1
AMRO BANK PLANS 300 MLN GUILDER 6.75 PCT BONDS	0
CYBERTEK <CKCP> FORMING NEW DIVISION	1
SAUDI ARABIA REITERATES COMMITMENT TO OPEC ACCORD	0
WESTINGHOUSE INTERESTED IN MERGER OF RADIO UNIT	0

```
ALASKA AIR &lt;ALK> LOAD FACTOR UP IN MAY                0
DIAMOND SHAMROCK &lt;DIA> SETS PRORATION FACTOR          2
```



We plot cluster size distribution and we note that most of our documents are in cluster number 0.

Second IRIS Dataset:

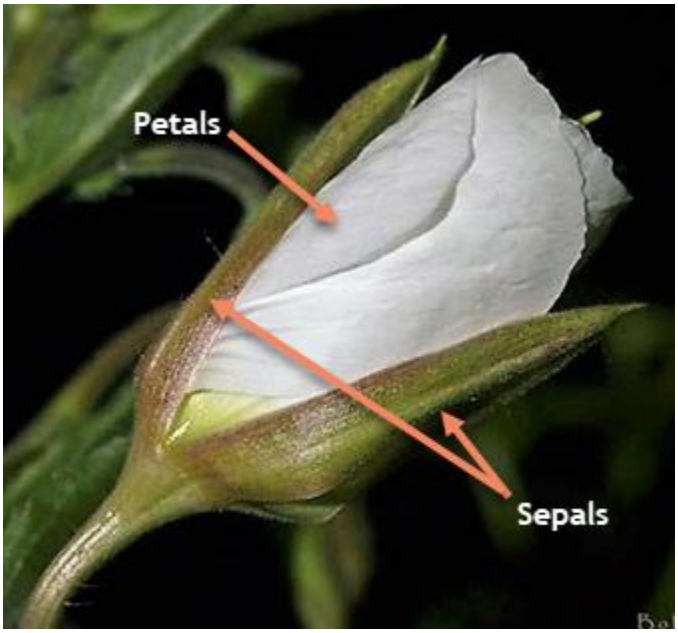
Dataset attributes:

Before we can explain the meaning behind the attributes of the iris dataset, we need to be familiar with what is petals and sepals.

A petal is the colored part we see in a flower which gives the flower its attractive look.

A sepal is the green hard part of a flower which is used for protection.

Here's an image to explain it well:



The iris dataset contains four attributes:

1] The length of the petal in a certain flower measured in centimeters

2] The width of the petal in a certain flower measured in centimeters

3] The length of the sepal in a certain flower measured in centimeters

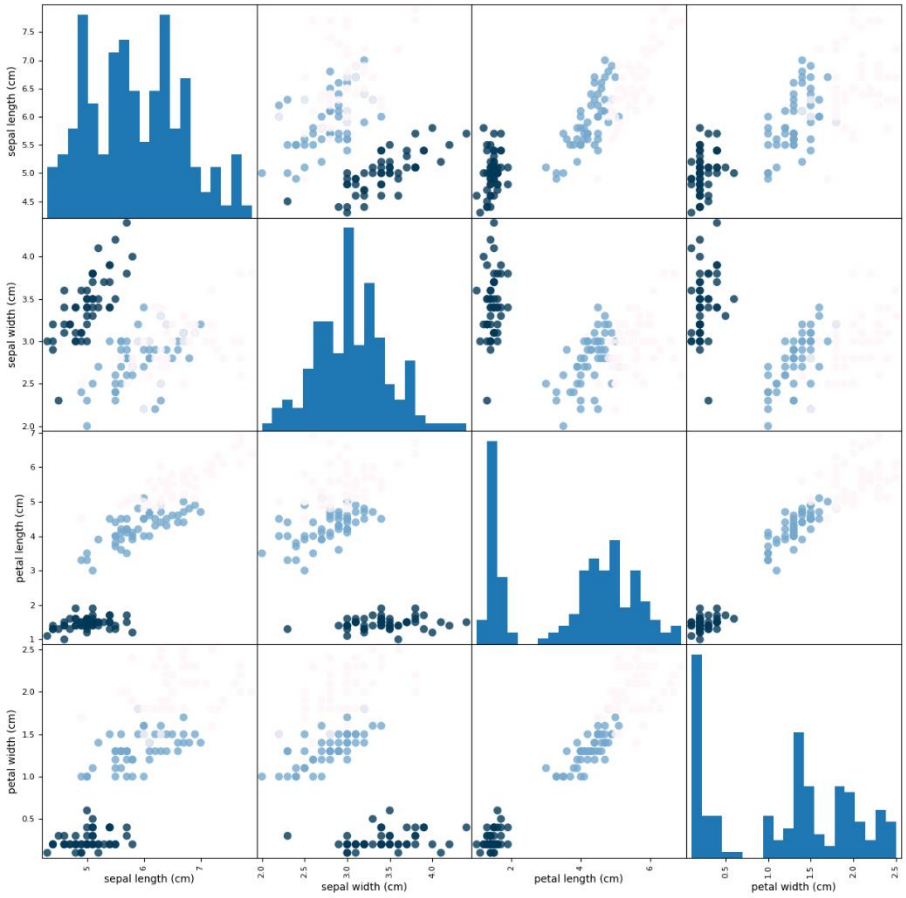
4] The width of the sepal in a certain flower measured in centimeters

5] The species of the flower (note that they're all iris flowers but this is a sub species from the iris)

All the attributes are numeric except of the species columns which is a categorical attribute and there is 150 rows or objects in the dataset.

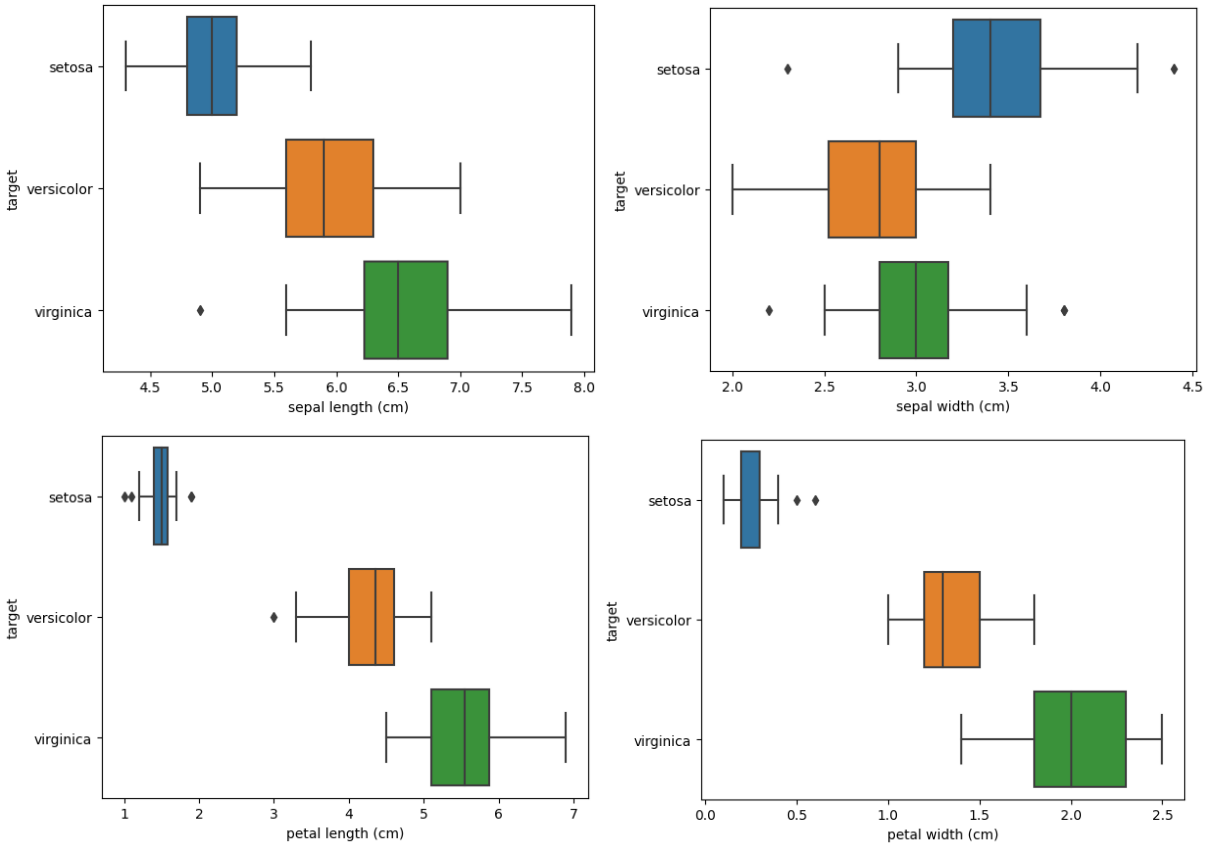
We can see that most data is normally distributed but the petal attributes are a little skewed.

Also, it's clear that there are multiple attributes that correlate together, which makes sense since the taller the petal the more likely it is for the sepal that surrounds it for protection will be.



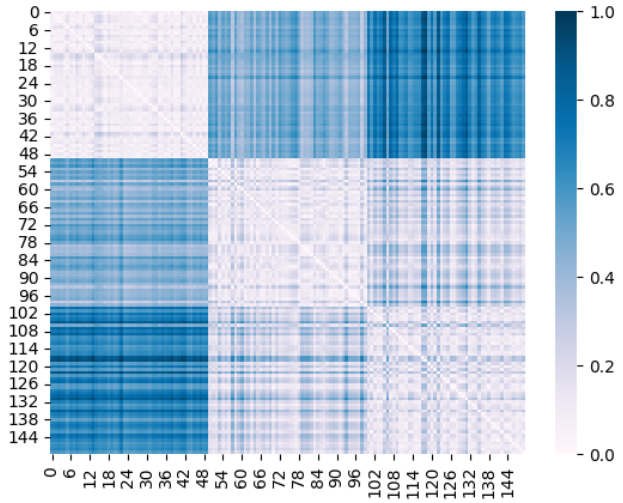
Box plots are used to observe the distribution of the data and to check the existence of outliers to be able to fix them.

From the plots we can see outliers is not a problem in our dataset except for a couple of values that will not cause problems.



The dissimilarity matrix which is the normalized Euclidean distance between each two objects in the dataset.

From our dissimilarity matrix we can see that versicolor and virginica flowers are closer to each other than they are to setosa as they are lighter in color which means closer from the color bar explaining the heatmap (lighter means the dissimilarity is lower which means they're more alike)



Workflow explained:

- At first, we had to load the data and build our data frame. The data was an object with different attributes which were as follows:

    Data -> the data itself

    Target -> the class column encoded

    Target names -> the 3 classes or species of the iris we have

    Descr -> description of the dataset

    Feature name -> the columns' names in the dataset

- Later we printed some information about the data and the plots we explained earlier

- Built the dissimilarity matrix and exported it

- We split the data into train and test, made a model using train data then test the model with our test data to see how well the model performed