A Spark Based Big Data Analytics Framework for Competitive Intelligence

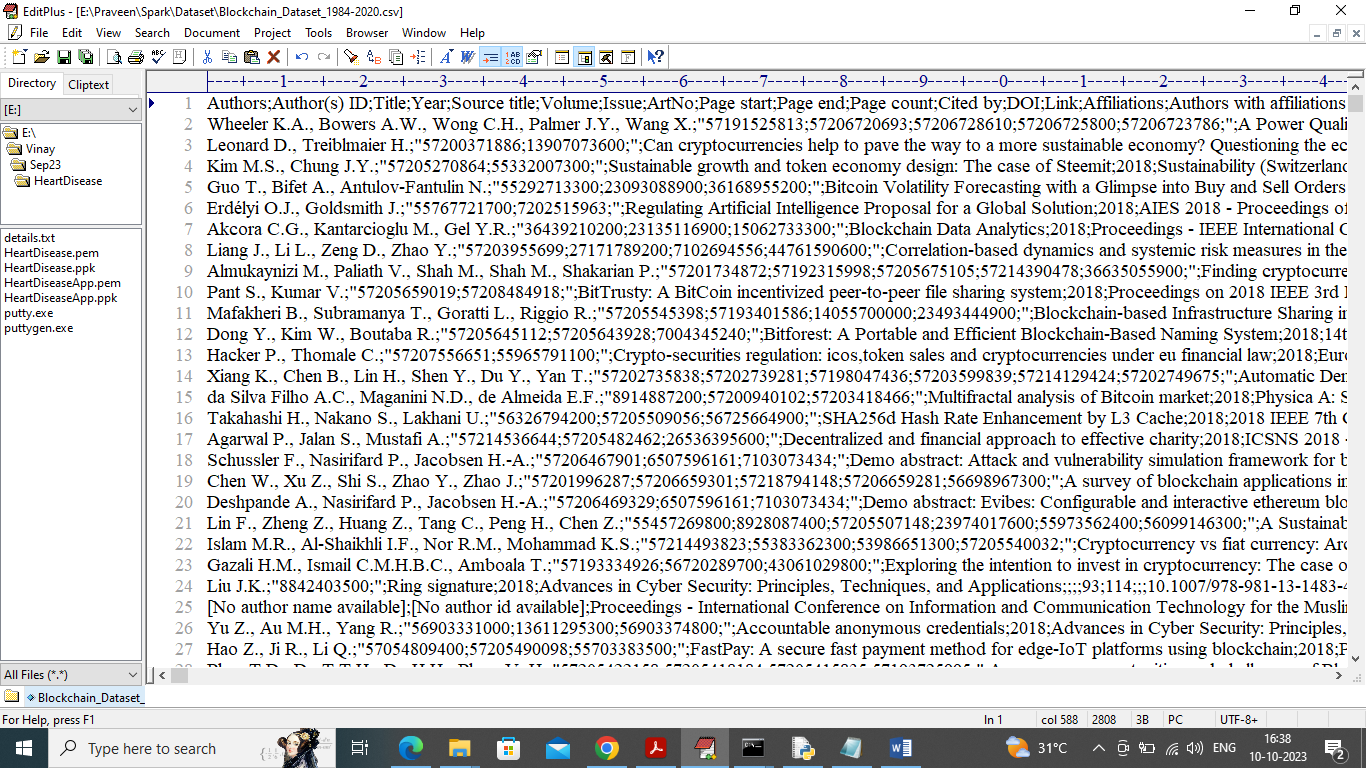
Due to internet all companies are migrating their business to online sale and this sales contains useful data information such as products with highest sales, customers are satisfying with product or not (they will write as reviews), frequent visiting of customers for product recommendation and may contains lots of other data. This data will get accumulated from internet in huge size and analysing such huge data with traditional algorithms may consume lots of time and this data often called as Bigdata.

To handle such Bigdata author of this paper employing Distributed processing frameworks like HADOOP for data storage, PYSPARK for data analysis and data modelling and this modelling help us in classification, topic identification and many more. GraphX for Bigdata visualization.

Bigdata can be stored either in HADOOP or MONGODB and Bigdata can be obtained by crawling websites of downloading required dataset from repository. Collected data will get stored in HADOOP and then SPARK will read data from HADOOP and perform various data analysis, visualization and modelling.

In Propose paper author downloading scientific papers on Blockchain technology and then training all those papers using LDA topic modelling to identify topics of given document or we can say document is describing about what topics. LSA is latent semantic analysis which identify most semantic similar words.

To implement above LDA and LSA algorithm we have downloaded IEEE papers dataset from KAGGLE and then used this dataset for modelling and in below screen we are showing dataset details



In above dataset screen first row contains dataset column names and remaining rows contains dataset values like AUTHOR, publisher, abstract and many other details.

Above dataset we have stored in HADOOP so SPARK can read from HADOOP and perform modelling. In below screen we are describing points to store data in HADOOP

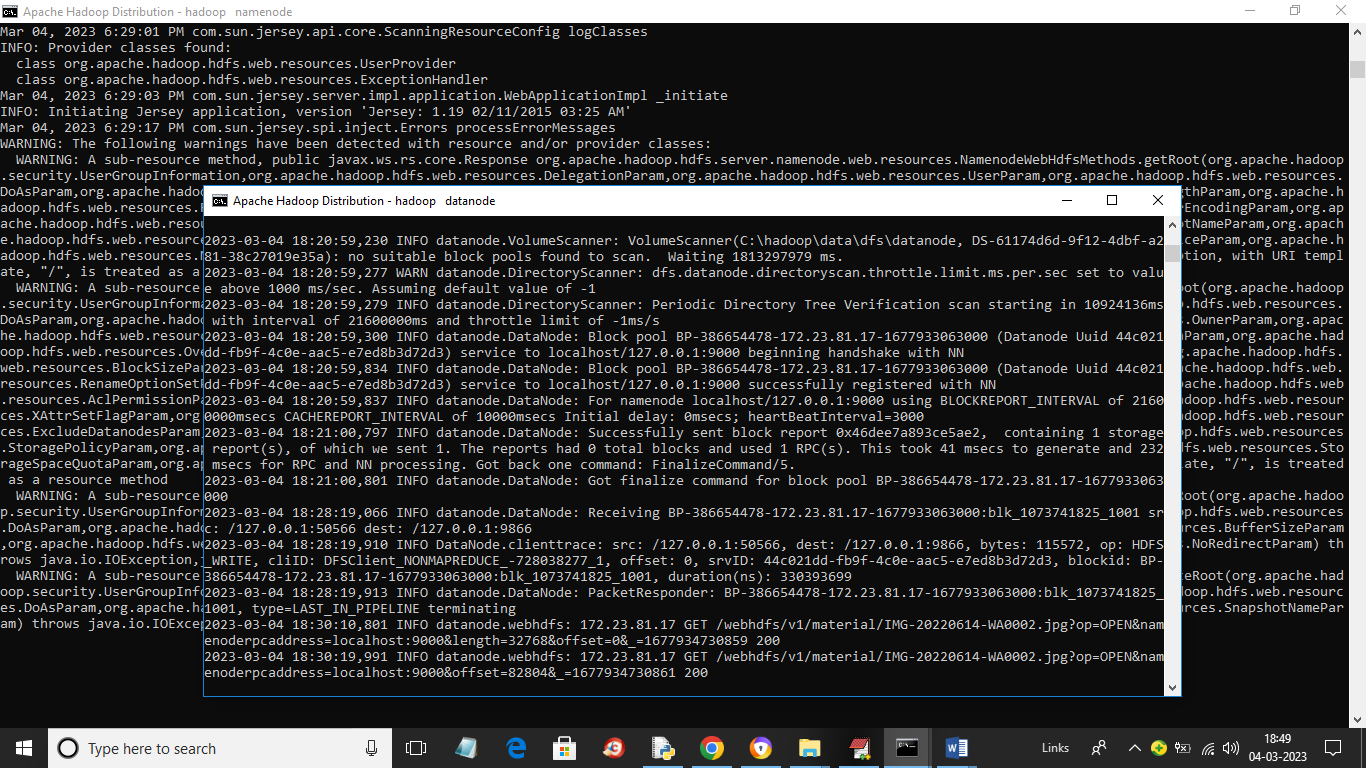
To install Hadoop first copy ‘hadoop’ folder and put in your system C directory

Next install java in C directory and give java install directory name as ‘Java’ and set JAVA\_HOME as C:/Java in environment system variable and set path as C:/Java/bin

Next set Hadoop path as ‘HADOOP\_HOME’ C:/hadoop and path as C:/hadoop/bin

Now open console and set location to C:/hadoop/bin and then type command as ‘hdfs namenode –format’ and then press enter key to create Hadoop data folder

Now change console location as ‘C:/hadoop/sbin’ and then type command as ‘start-dfs’ and then press enter key to start Hadoop and get below page



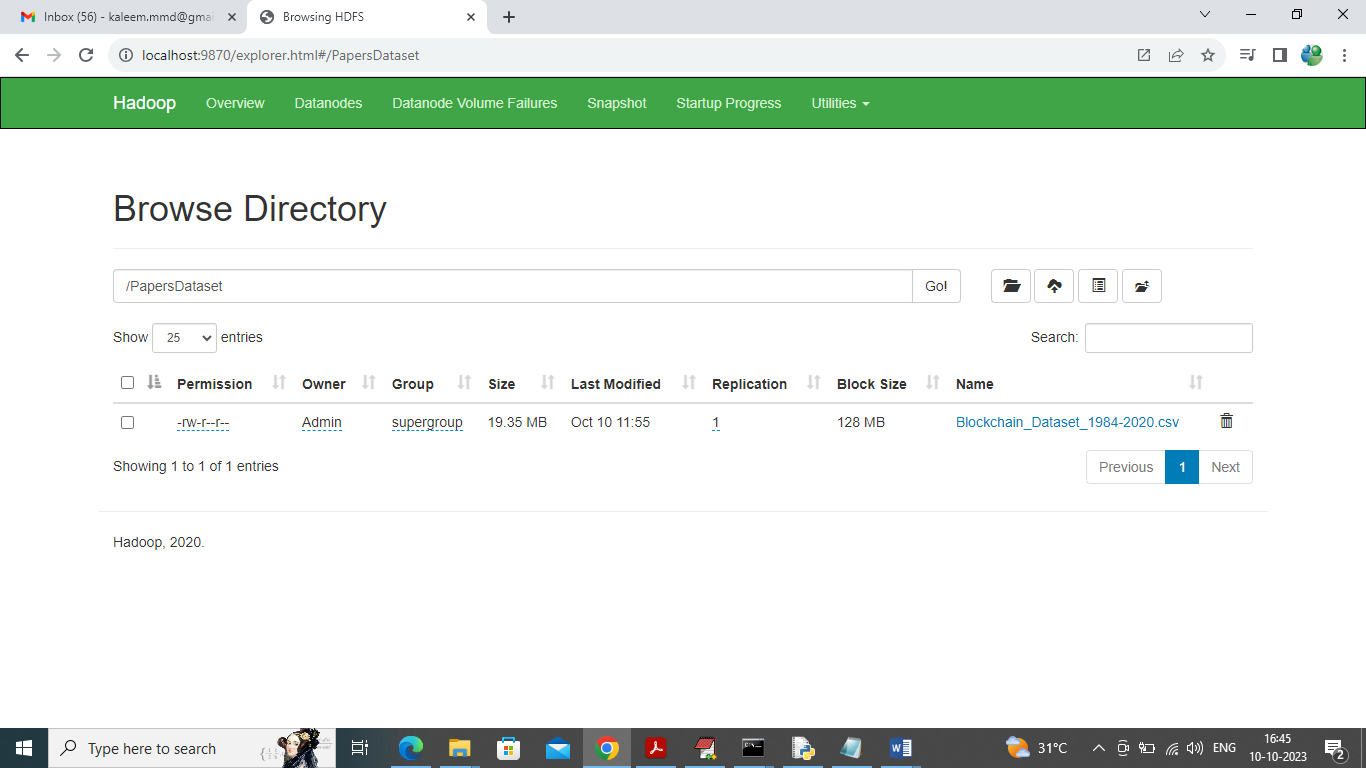
In above two screens Hadoop server started and now in console run below command to create ‘PapersDataset’ folder in Hadoop

hdfs dfs -mkdir -p / PapersDataset

now execute below command to store dataset in Hadoop

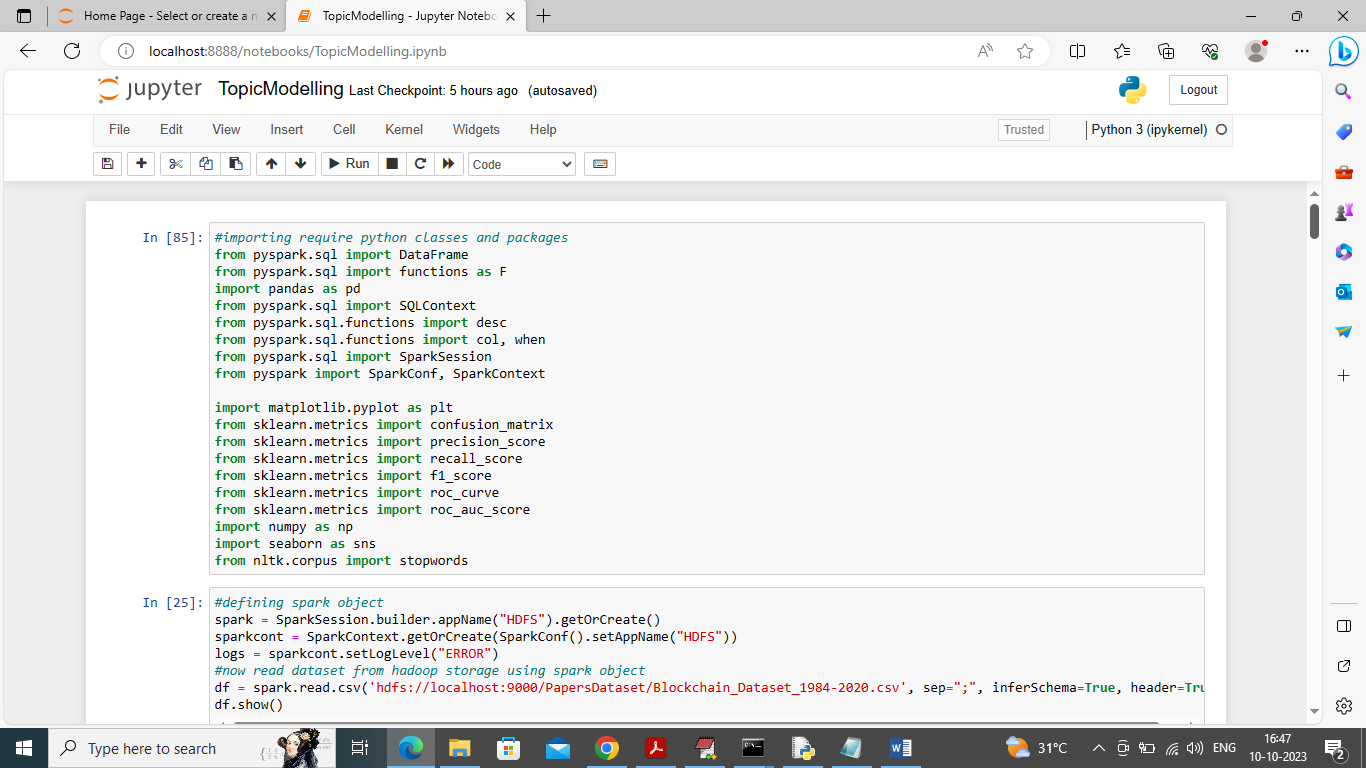
hdfs dfs -put Dataset/Blockchain\_Dataset\_1984-2020.csv hdfs://localhost:9000/PapersDataset

now for above command press enter key and then open browser and enter URL as ‘http://localhost:9870’ and press enter key to get below page

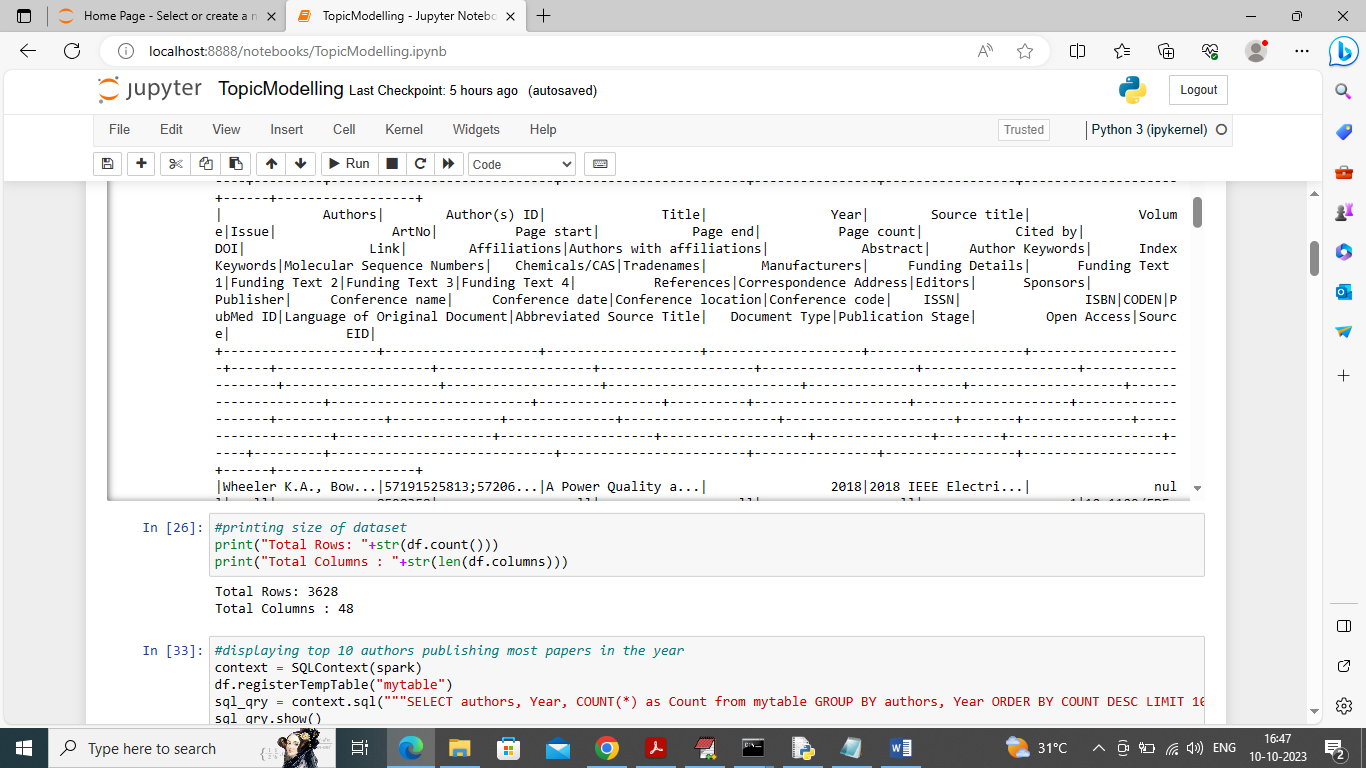


In above screen dataset is loaded in HADOOP and now run application by double click on ‘run.bat’ file

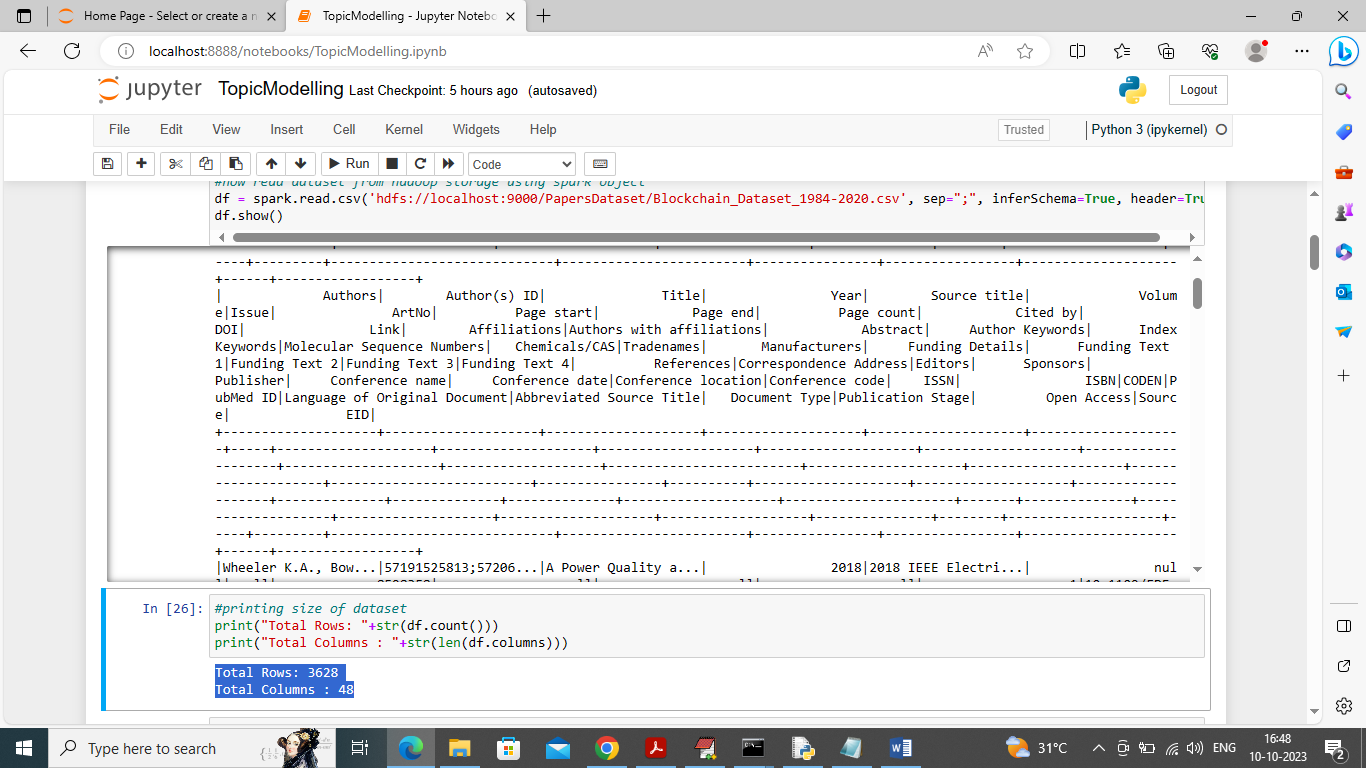
We have coded this project using JUPYTER notebook so you will get below screen with code, output and blue colour comments



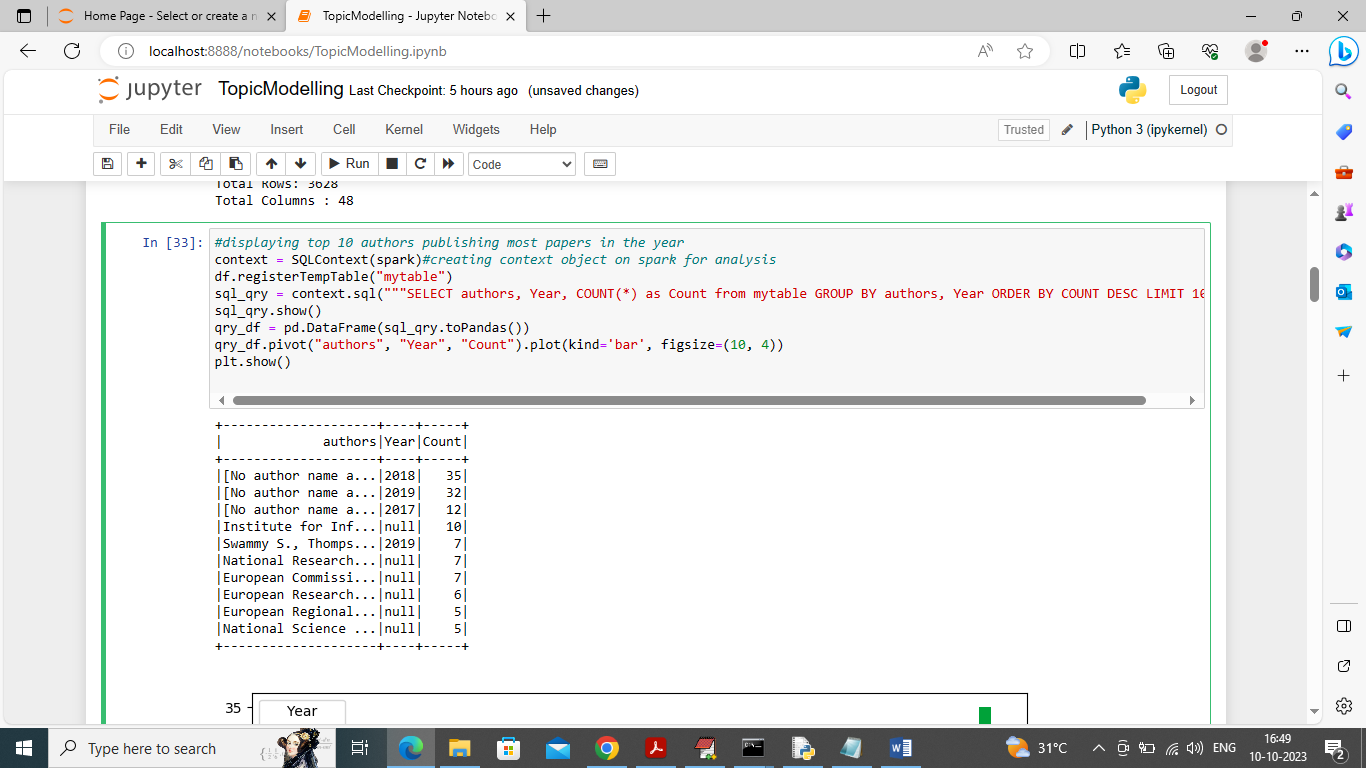
In above screen we are importing required SPARK packages and then in next block we are initializing spark session which will read dataset from HADOOP and then display below output



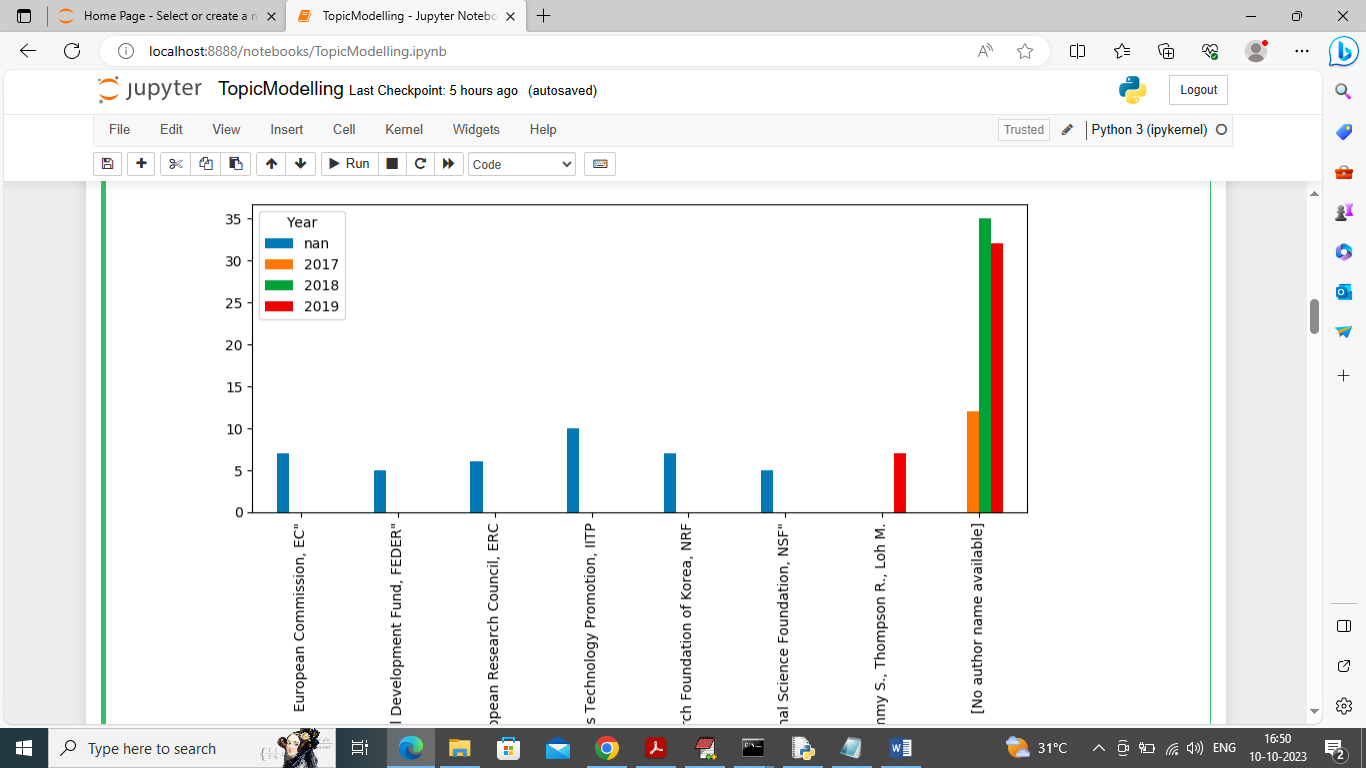
In above screen we can see dataset values loaded from HADOOP to spark and now spark will analyse this dataset for visualization and other analysis



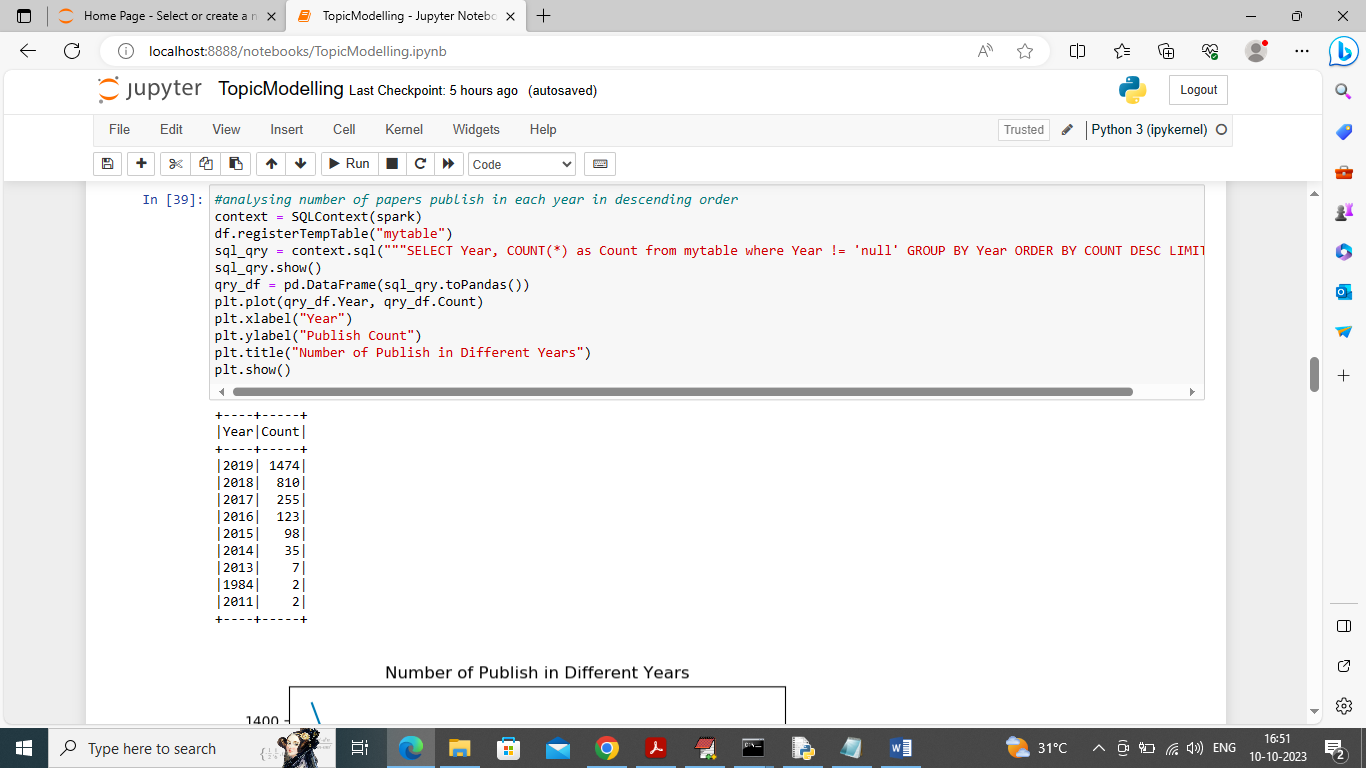
In above screen in blue colour text displaying total rows and column found in dataset



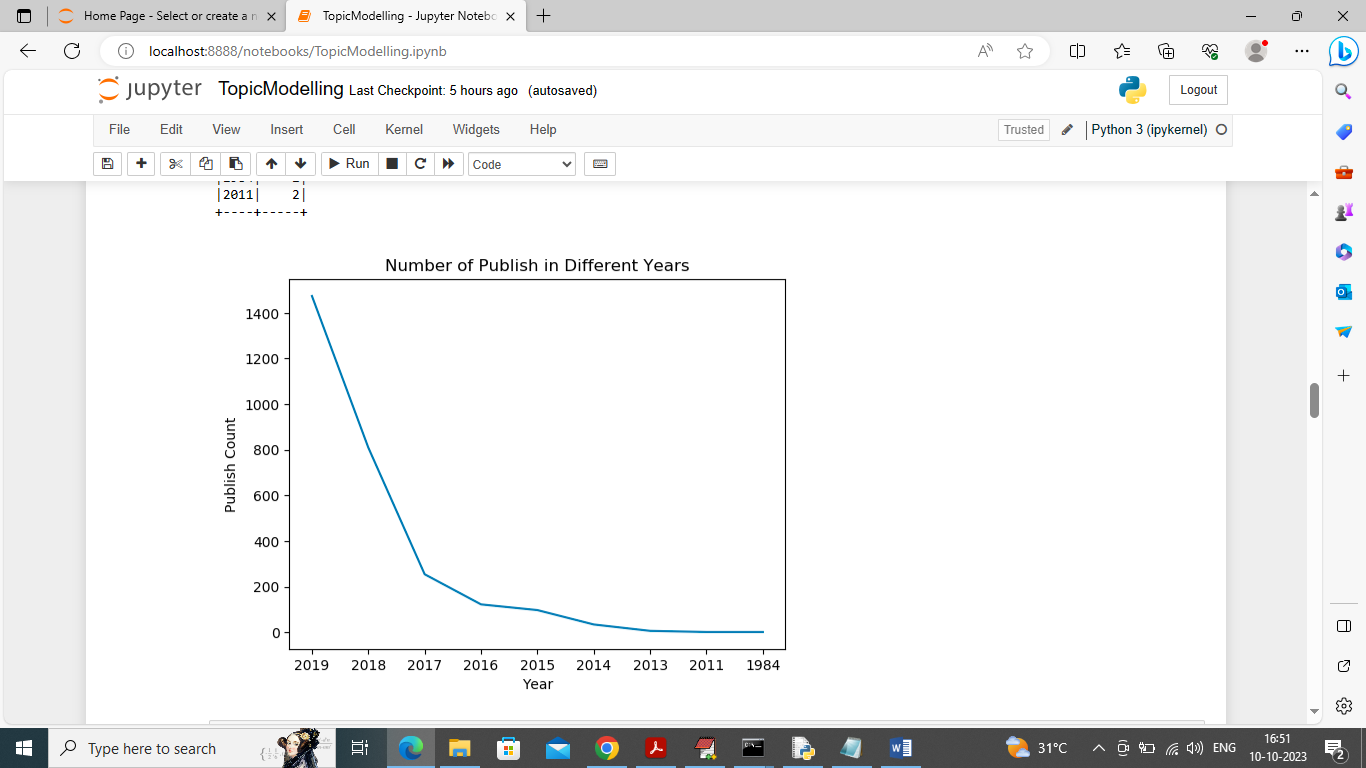
In above screen using spark we are executing query to find top 10 authors who publish more papers in particular year and in tabular format we got out as Author name, year and number of publish and in below screen we can see the graph



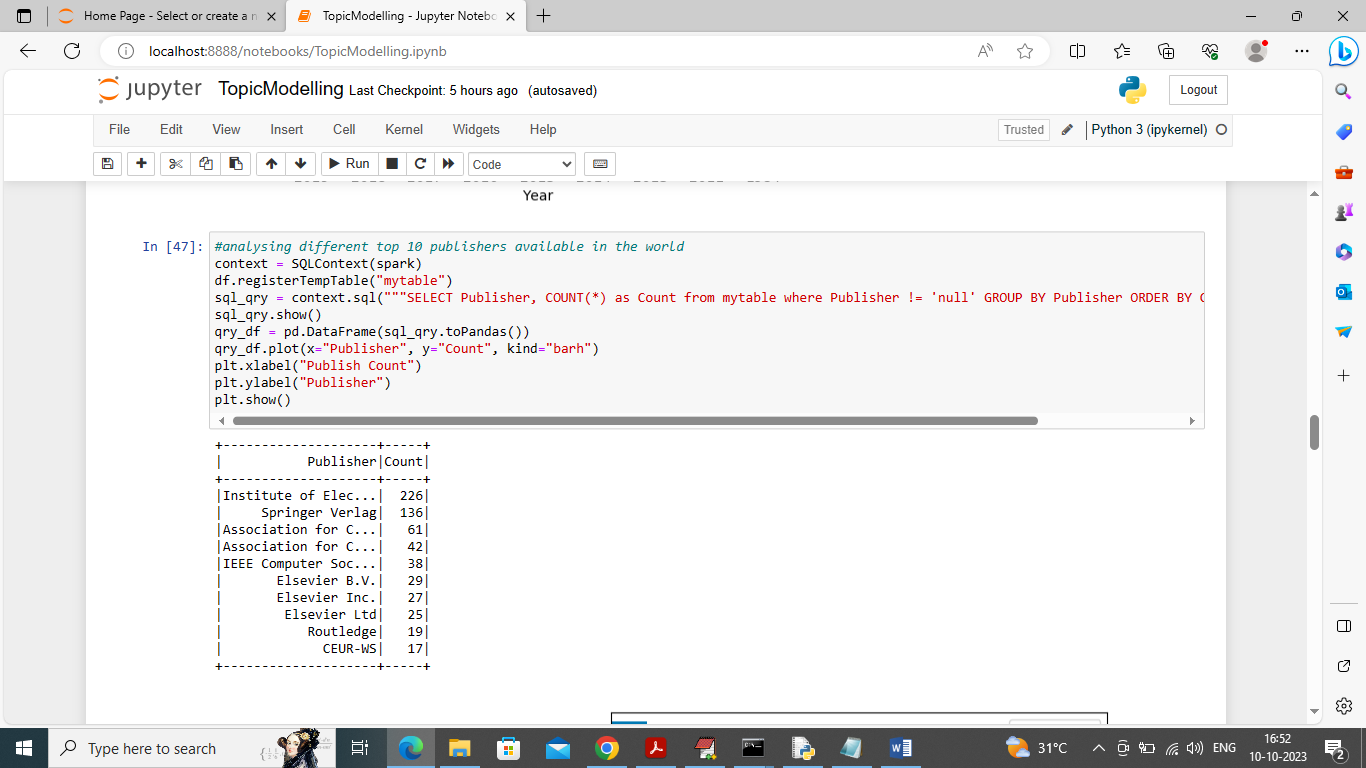
In above graph x-axis represents author names and y-axis represents number of publishing done by those authors



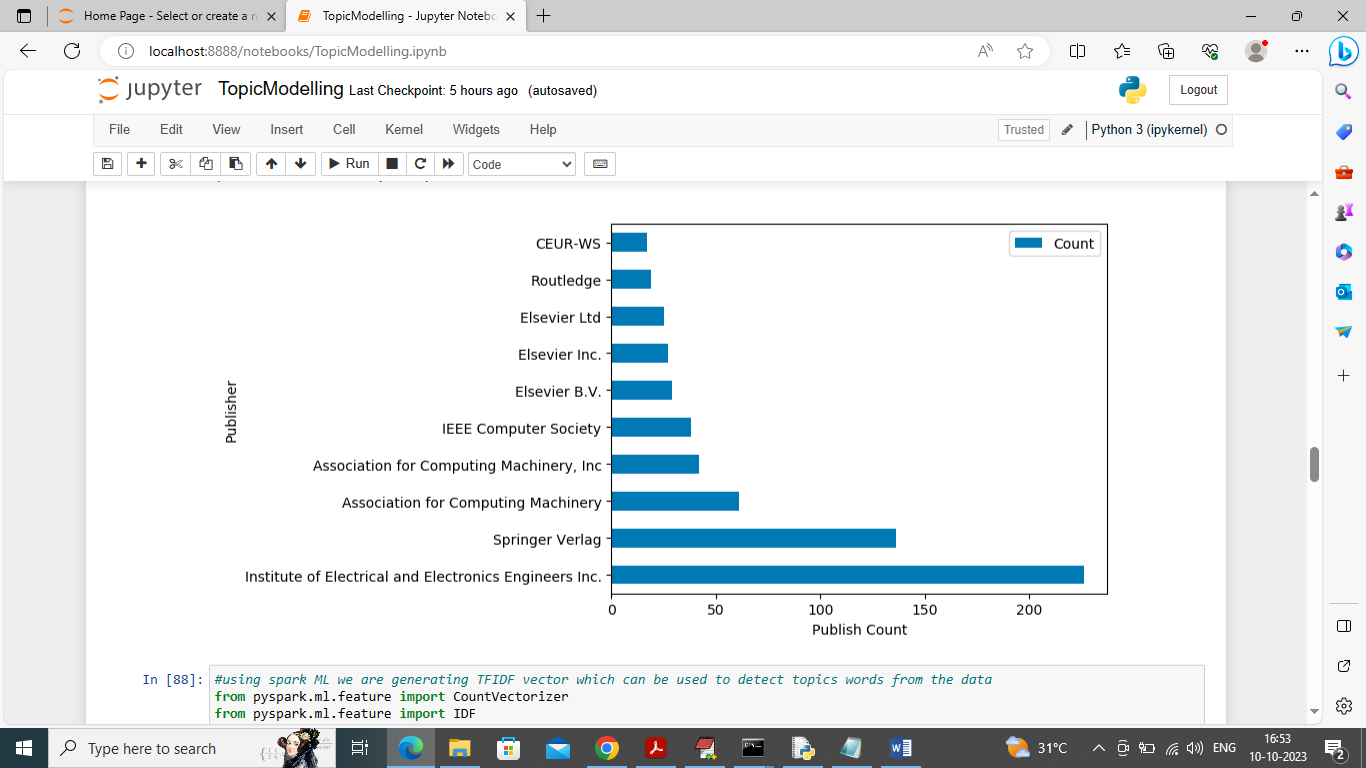
In above screen we are finding more number of paper publish in different years and in table format we can see the output with below graph



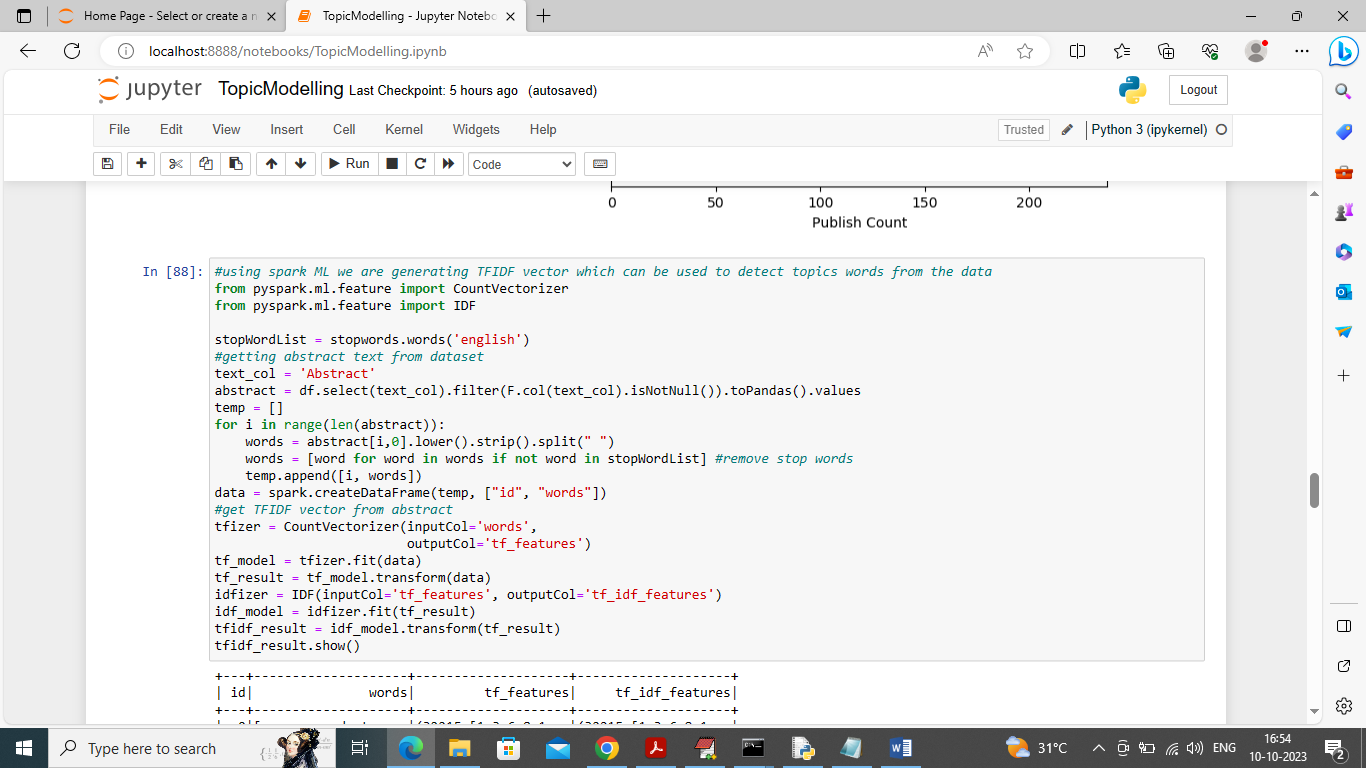
In above graph x-axis represents year of publish and y-axis represents total publish in that year



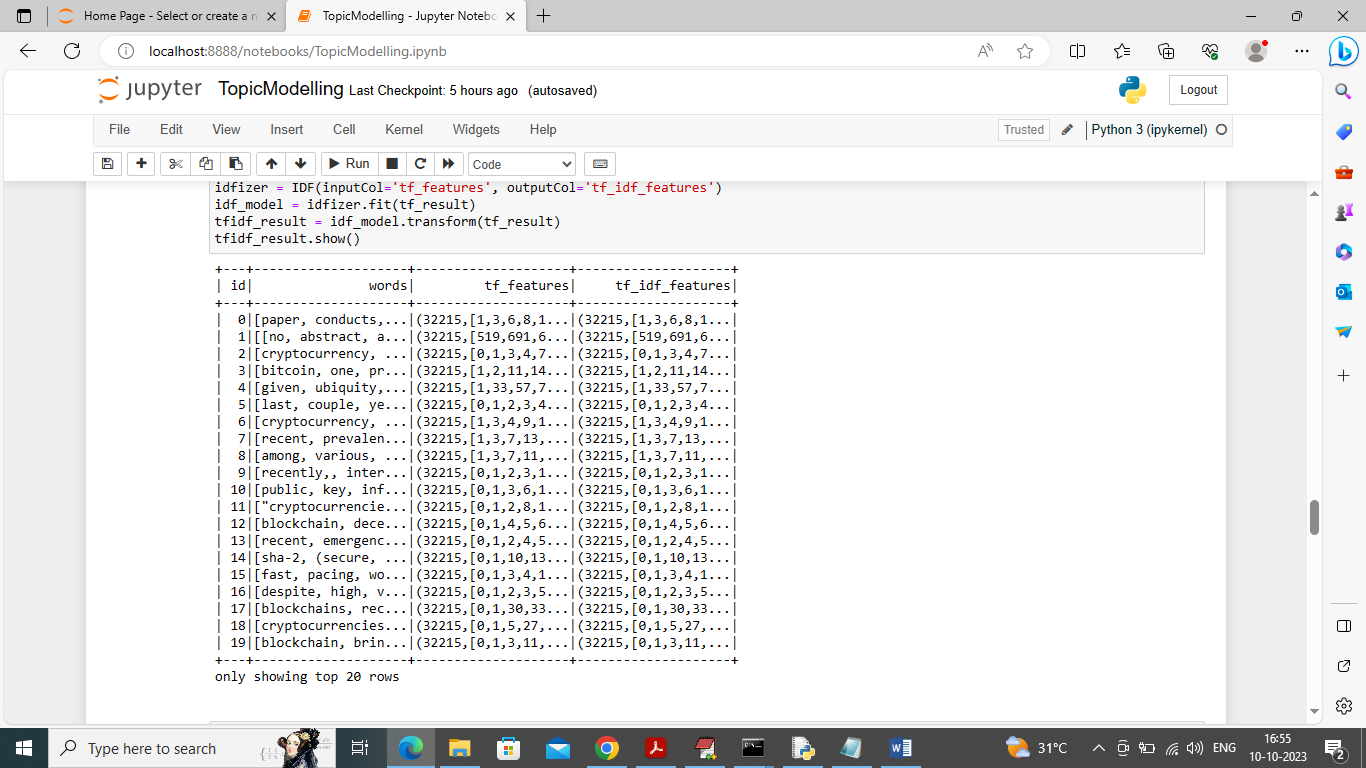
In above screen using spark we are finding top 10 publishers who publish more papers and in table format we can see publisher names and count of publish papers



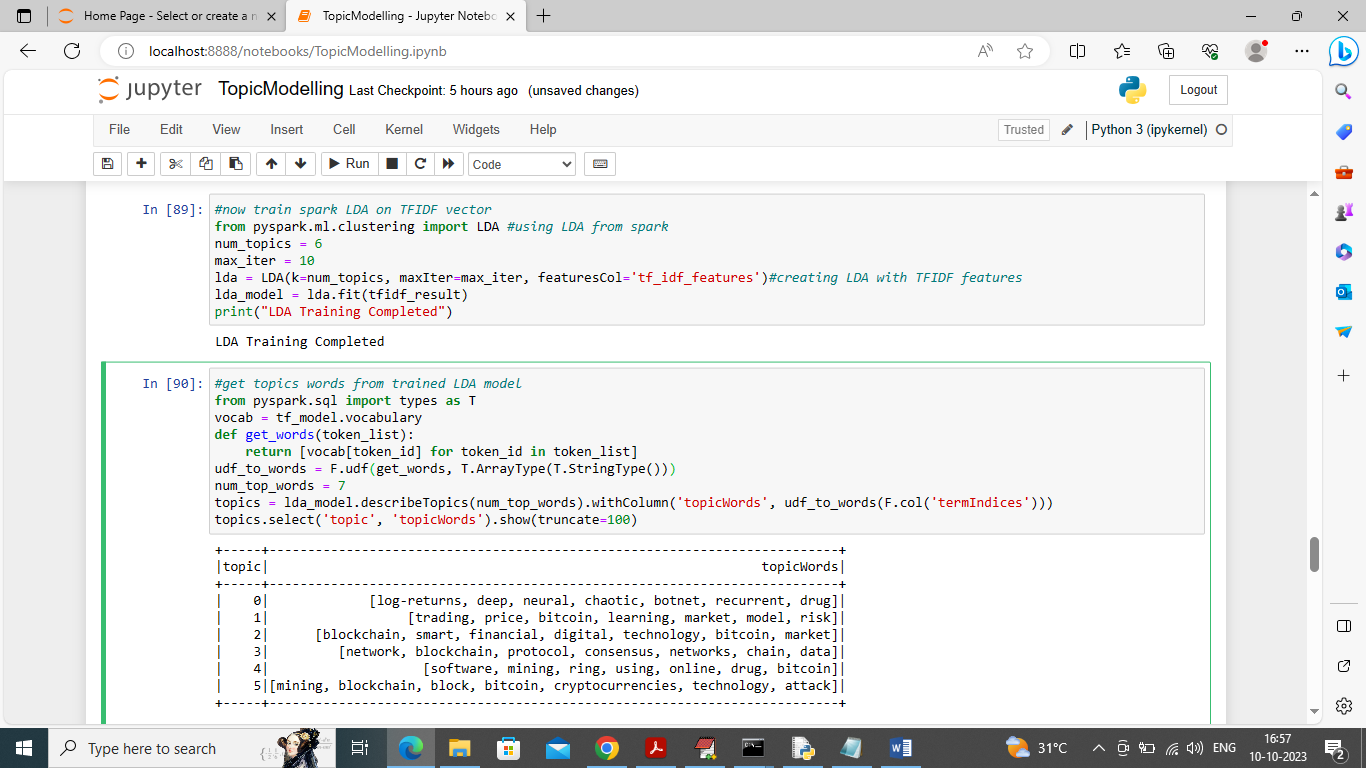
In above graph x-axis represents counts of publish papers and y-axis represents publisher names. So in above screens we did some analysis and visualizations using SPARK and now in below screen we are performing LDA modelling



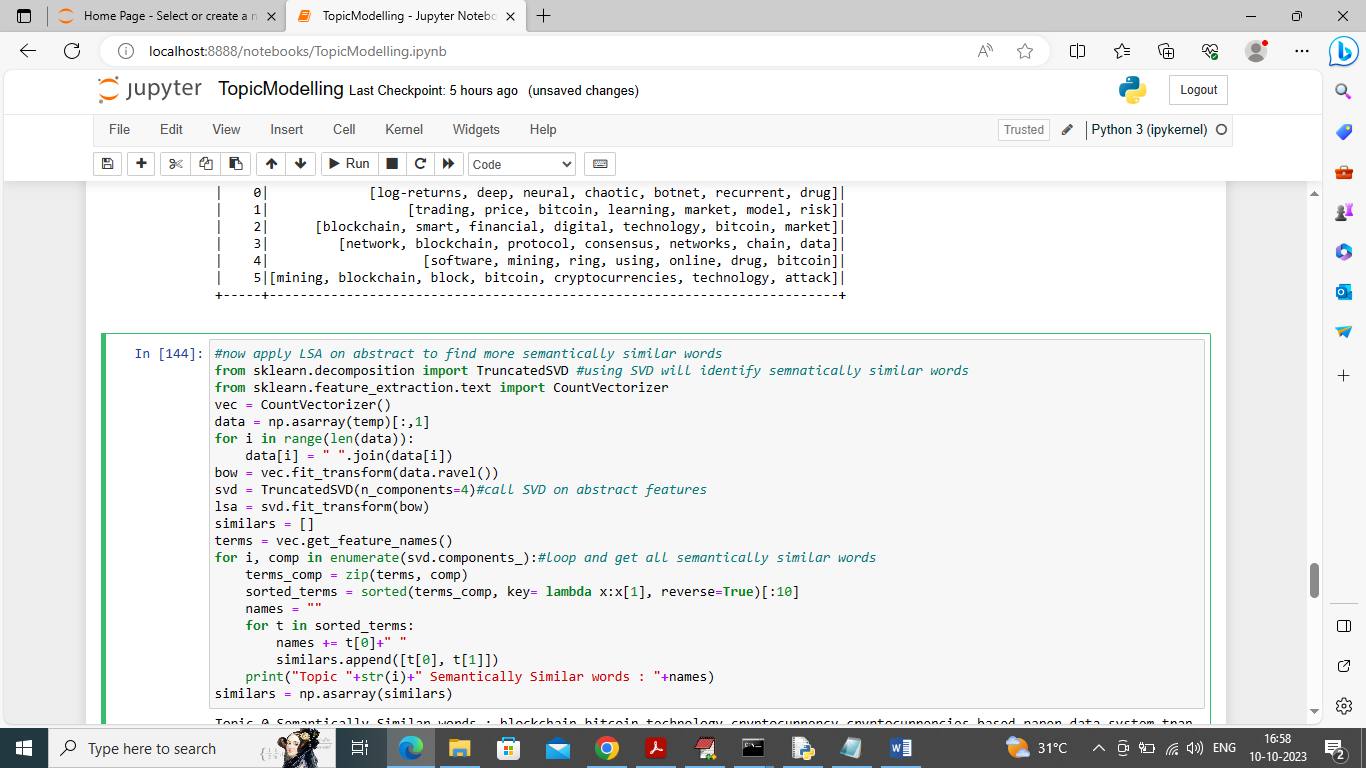
In above screen we are reading all abstracts from dataset and then converting them into TF-IDF (term frequency inverse document frequency) vector which contains average frequency of each words and after executing above block will get below output



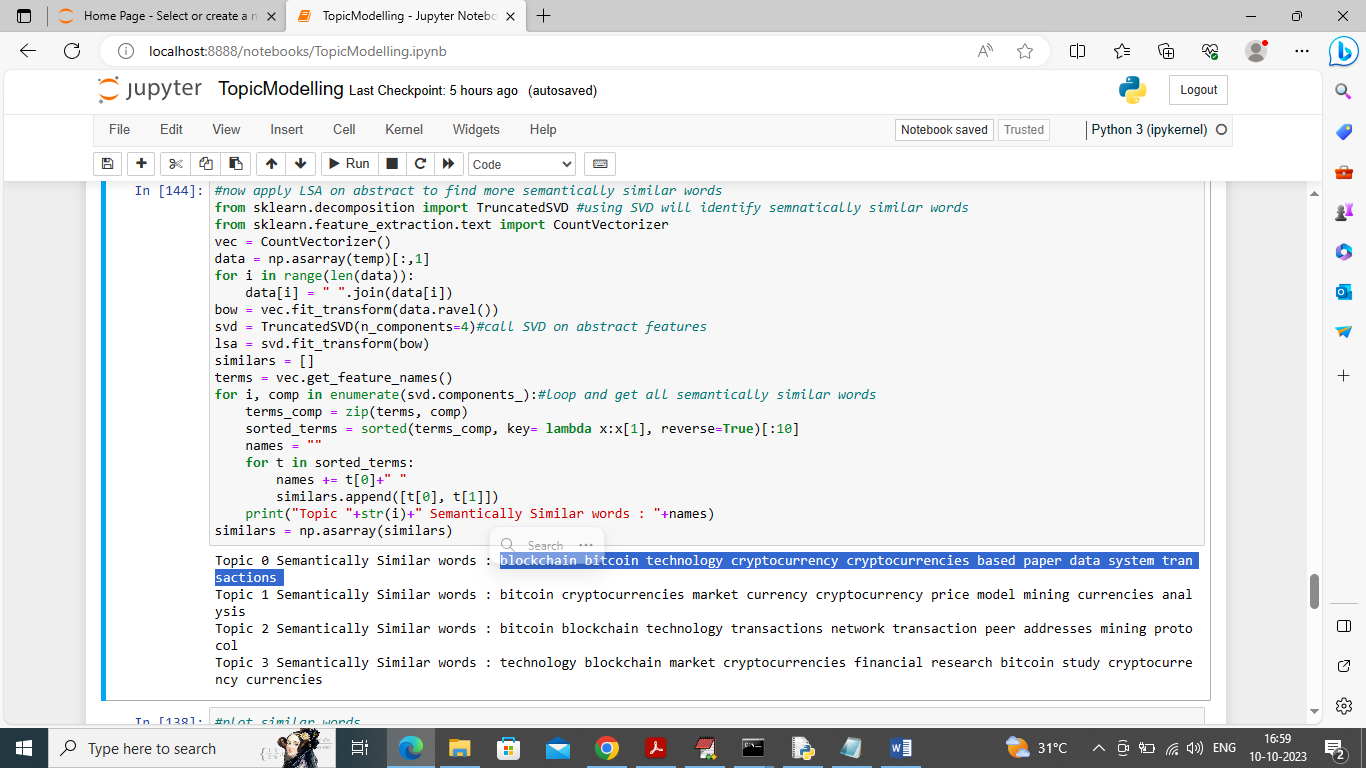
In above screen we can see words and their TFIDF frequencies and now this TFIDF values will be input to LDA algorithm to get topics words



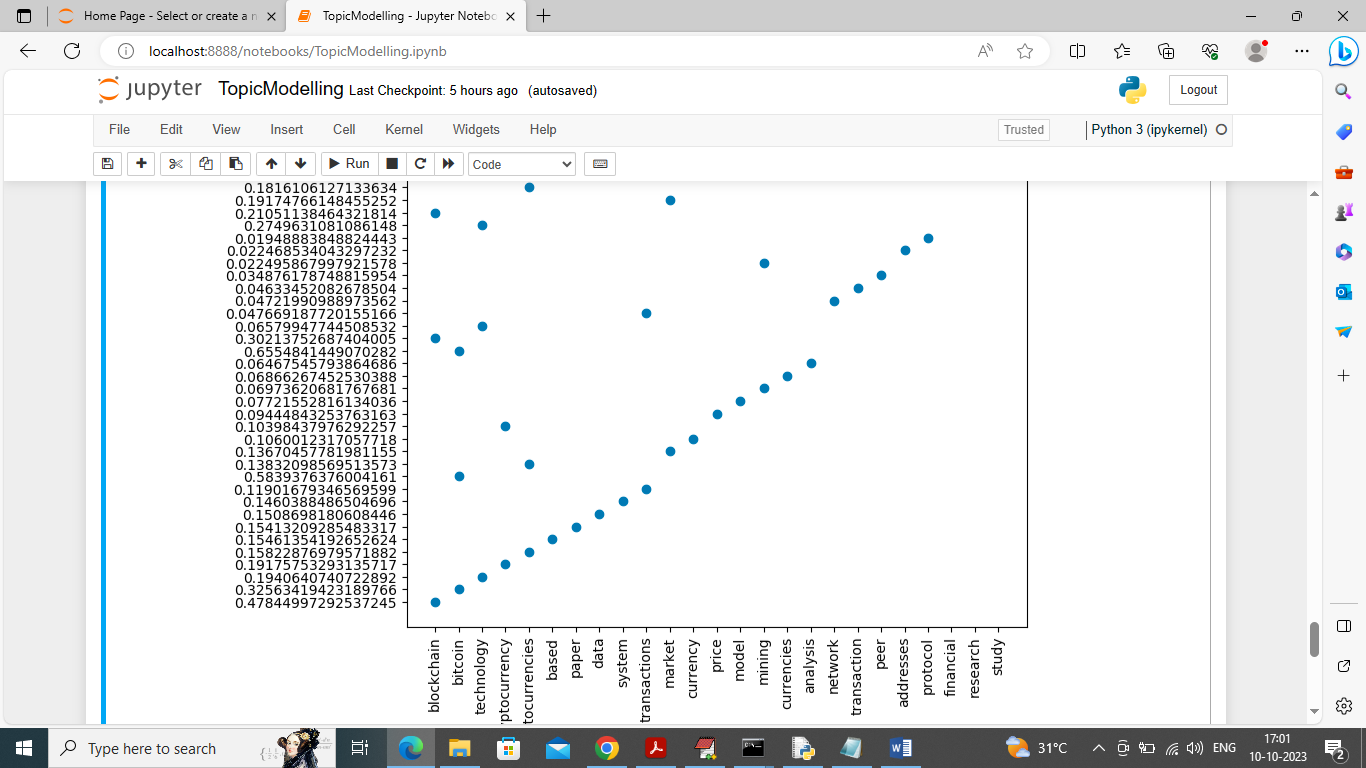
In above screen we are training all abstracts TFIDF features with spark LDA and then we are displaying LDA topics extracted for all abstract. All topics words you can see in tabular format



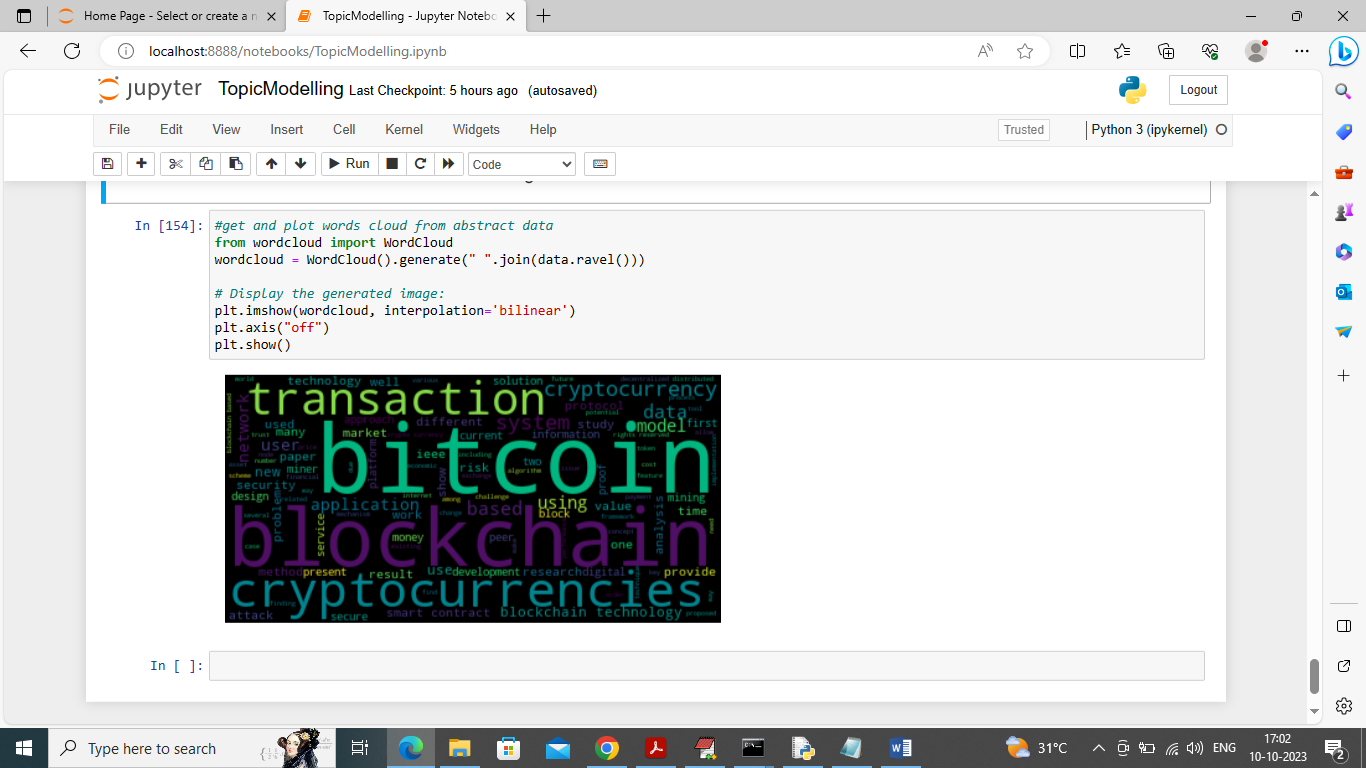
In above screen training LSA using SVD matrix on abstract features to identify more semantically similar words and after executing above block will get below output



In above screen from different topic NO we are extracting and displaying semantically more similar words and in above output we can see Bitcoin, Blockchain or crypto currency are similar words and in below graph we can see words which are similar will be closed to one and other



In above graph x-axis represents semantic similar words and y-axis represents similarity score and in graphs blue dots represents words and words which are similar will be closed to one and other



In above screen displaying cloud of words and can see which word appear most will have bold font.

So in above screen using spark framework we have analyse IEEE papers dataset and then train LDA model to identify topics of papers. We have selected all Blockchain papers so most topics are describing above Blockchain technology words only.