Principles of Big Data Management

Computer Science 5540 0001

Team-7

Muni Kanaka Sri Shalini Chintam([mcnhy@umsystem.edu](mailto:mcnhy@umsystem.edu))

Manoj Desetty([mdkmy@umsystem.edu](mailto:mdkmy@umsystem.edu)

Anudeep Reddy Thippireddy([atf9d@umsystem.edu](mailto:atf9d@umsystem.edu))

Ghadah Almousa([gfap7c@umsystem.edu](mailto:gfap7c@umsystem.edu))

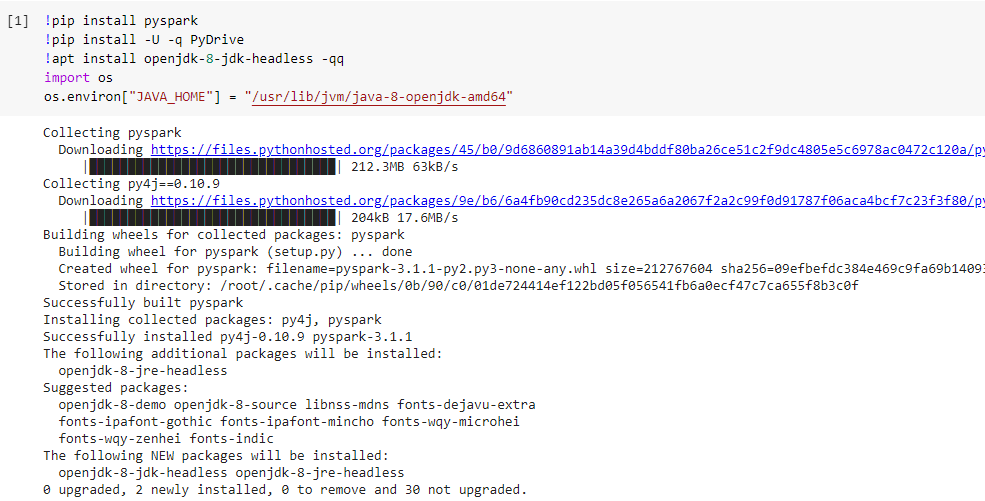
Santhosh reddy thada([stn60@umsystem.edu](mailto:stn60@umsystem.edu)

Github link: <https://github.com/MuniKanakaSriShaliniChintam/PB->

Warm Up Project

Objective : how many news articles mention ‘united states’ at least three times in all news catgories(group by category)?

Project Explanation:



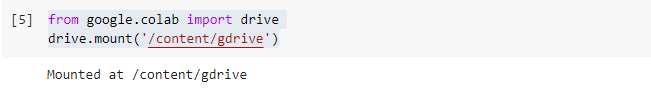
We have done our project using the google collab so we first have to connect spark to google collab we can do it by first installing the pyspark and installing the java and setting the path where it have to installed.



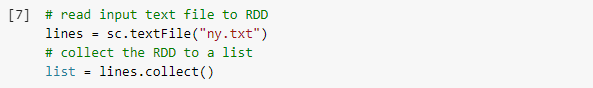
Creating the spark session and it is entry point of any spark application and builder method is used for it creation getorcreate() would either create the new session or get any running session.



Here spark content is created and the application is being run locally.

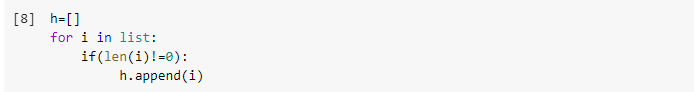


Here we are connecting the google collab with my drive so that we can save our data directly to our drive.



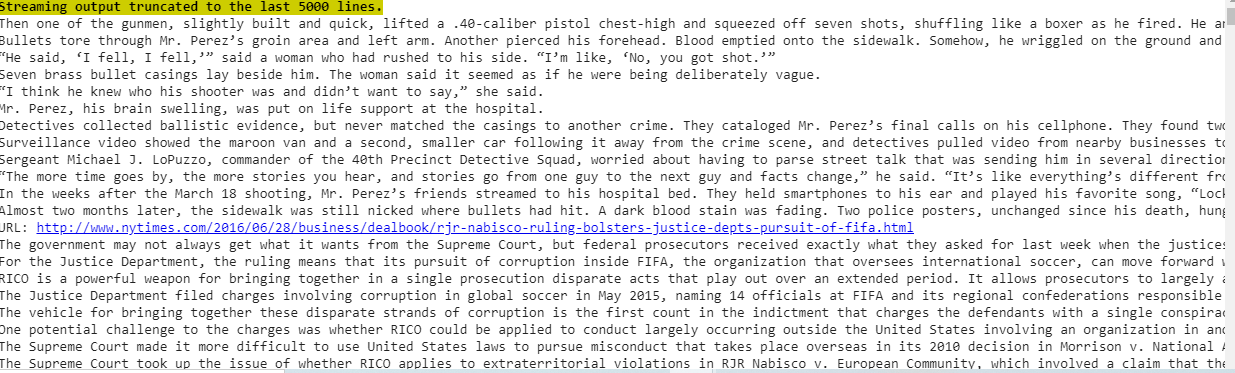
In the first line we are reading the text file named ny.txt

In the second line we are reading the rdd data to the list.

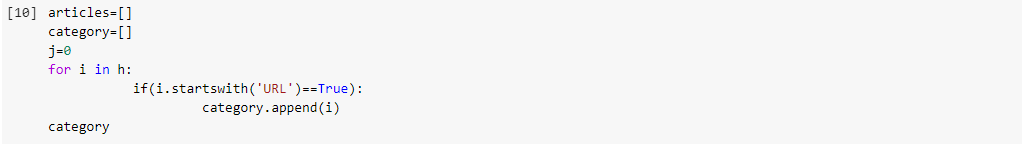


We are creating another list h and adding all the non empty lines to the list.



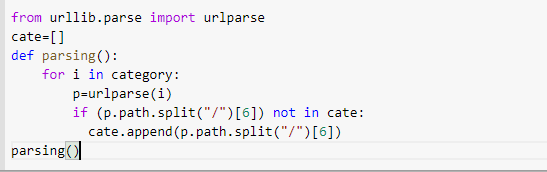


Printing the whole list using the for loop.

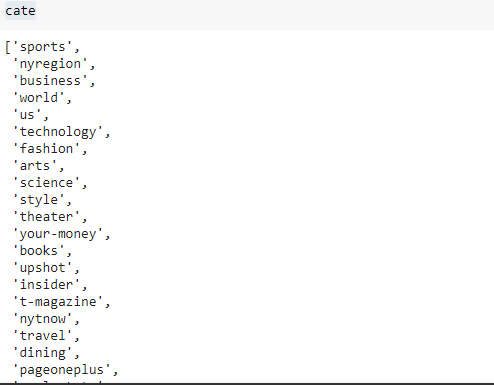


Here we are creating the category list containing all the url’s.

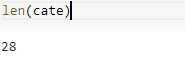




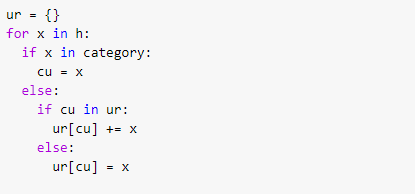
Here we are using the urlib.parse library for extracting the group category and we are new list to store all the group categories and list name is cate.



Here in the cate list we have all the group categories and they are displayed above.

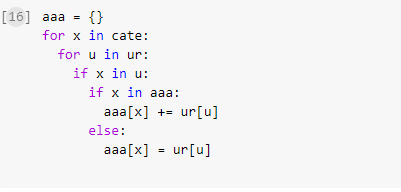


These the total number of group categories.

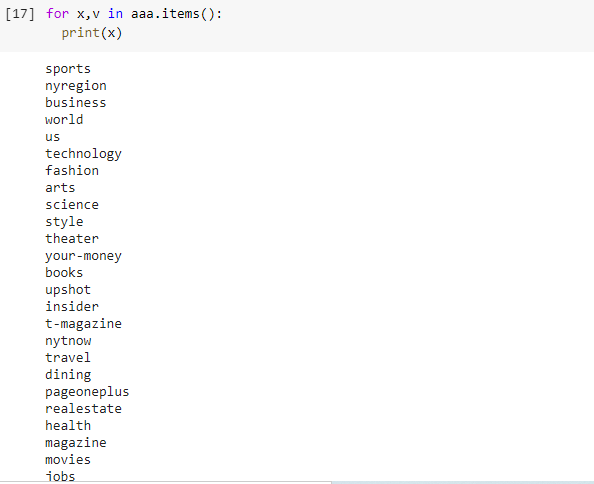


Here we are creating a new dictionary and storing the adding all the urls into the dictionary. And there associated data regarding to that url

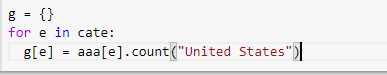




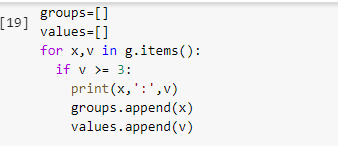
Here we are creating a new dictionary and storing the categories and there associated articles into it.



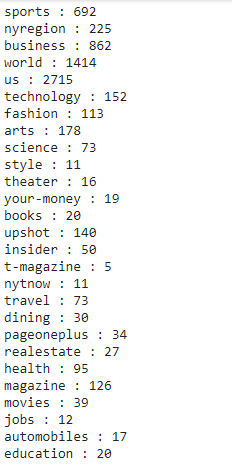
These are the categories who are articles are being stored in the dictionary.



Here in this we are finding the count for occurrence of the united states in them and storing that value into the



Here we are finding the categories who have united states occur in it more than 3 times. And the output would be like below and we creating two lists to store the categories and the count value.



Here by using the Tableau to visualize the categories and number of times the word united states has been occurred by using the bar graph and bubble chart.

