PRINCIPLES OF BIG DATA MANAGEMENT

Warm Up Project

Objective:

To build a word cloud for NY Times articles, to build the world cloud for the top 5 news categories and to list the top 10 words that are present in news articles under the same category. Technologies used are python programming language and Apache Spark to implement the word cloud project.

Task 1:

- To install apache spark and related IDE
- Connecting spark to IDE
- Importing the required modules and installing the word cloud
- Excluding the stop words and plotting the top 100 words as a word cloud

Task 2:

- Importing the required modules
- Making the categories for the top 5 news categories
- Plotting the word cloud for the top 5 news categories

Task 3:

- Importing the required modules
- Dividing the article into categories
- Considering the top 10 words from each category published in most number of articles

Challenges faced:

Took a bit time to install and overcome the errors occurred, understanding the logic in the question became tricky.

Team Members (Team 9):

1. Sree Vaishnavi Veeragandham (svxfb@umsystem.edu)
2. Ravi Krishna Teja Dachepalli (rdbgm@umsystem.edu)
3. Madhuri Kondepu (mk9tm@umsystem.edu)
4. Rakesh Chilukuri (rccgt@umsystem.edu)
5. Chris Ojiaka (ccoqn5@umsystem.edu)

Github Repo URL : https://github.com/ravi4080/PBD_WarmUp.git