BIG DATA ANALYTICS TECHNICAL REPORT

GROUP 2

Abstract

Perform data extraction from different sources and perform cleansing on the gathered data using several techniques according to the requirements. Apply topic modelling techniques [text-mining] to identify patterns in corpus of files. Summarizing and describing the collected data using Descriptive Statistics and reporting the same via Visualization tools for better graphical analysis.

15-March-2016

Problem Statement:-

Gathering the data from various sources and extract meaningful insights using several mining and statistical techniques.

Task and Tools used:-

- **Fortune 500 Excel:-**
  - » Excel files contacting List of Fortune 500 companies, Revenue and KLD report. Collected and perform elimination of irrelevant columns from the goal.
  - » Gathered data further uploaded on Hive for processing & creating consolidated sheet.
  - » Handling of missing and null values handled in Hive and R programming.
  - » Summarizing and descriptive statistics [min, max, median] implemented on refined data in R ddply package for every column w.r.t company names.
  - » IBM Java Cloud displaying states name with most number of companies.

- **Patent Data:-**
  - » Unzip the patent zips in Hadoop file system.
  - » Fetched the extracted files in the local folder using Hadoop fs –get command.
  - » Downloaded the files into local system for further analysis.
  - » Java code to extract .xml file names from a downloaded directory to text for next step.
  - » Understand the xml pattern and implemented Java code to pull the strings iteratively from all the .xml files.
  - » Run the program over all the files and stored it on Hive – 4.5+ million records fetched.
  - » Tableau Data visualization performed based on year and companies with most patents.

- **Annual Revenue Report:-**
  - » Links for Fortune 500 companies Annual Revenue Report has been extracted using Google Search on www.sec.gov/Archives/edgar/data sites via Java Program – [Jsoup.jar Java HTML Parser].
  - » For each firm**,** Annual Report for past 3 years has been downloaded using Linux wget command over web via extracted links.

- **CSR Coding: -** Understand the parameters looked-for or impacting the ranking of the firms and attached it in the consolidated sheet.

- **Company background and competitiveness information**
  - » Can be extracted from Hive SQL and plotted visualization graph between few parameters for analysis in Tableau.

- **Topics Analysis:-**
  - » Performed techniques on each firm Annual Report of past 3 years.
  - » Attempted using Mallet but it did not provide effective results.
  - » Applied Topic Modelling using R & Python Programming on the downloaded files.
  - » Uploaded the same set of data on Tableau for better analytics and visualization.

- **USA Presidential Election: -** Accumulated data with Year and Elected political party for future analysis.

Appendix:

Sr. No	Task	Object File	Description
1	Fortune 500 Consolidation	Hadoop_Hive_Commands.docx	Hive commands to create table and consolidate 3 sheets into 1.
2	Fortune 500 Descriptive Statistics	R scripts.txt	Min, Max, Median, Standard-deviation w.r.t company name based on year, specialty.
3	City - IBM Java Cloud	cities.png	Cloud showing cities having most number of companies.
4	Patent Data	Step 1	Extract the zip files and download in local system for processing.
5	Patent Data – Extract xml file names	ListFiles.java	Extracting all the .xml file names in directory to a single text file for iterations.
6	Patent Data – XML pull parser	XMLPullParser.java	Reading strings between XML start and end tags – storing it in excel file. 233 MB Excel – 4.7 million records.
7	Patents Data – Hive	Step 1	Uploading Big-data in Hive for querying purpose.
8	Annual Revenue Report	GoogleRes.java	Searching links of Annual Revenue Report for each firm using Google Search.