Big Data Tools for Business: Real World Uses Cases Project

What is expected

- Organize yourself within groups of maximum 3 students
- Provide clear and meaningful results in your notebooks.
- Upload two notebooks (one for each part of the project) to the dedicated space for this project on Campus
- Deadline: 07th January 2021 at midnight

Part 1: Donald Trump Communication Analysis on Twitter

You are big data analyst for a communication agency who wants to analyze Donald Trump's communication on Twitter.

You have an history of all Donald trump tweets from 2009 to November 19th, 2020 in the form of a text file (***trump_tweets.txt***).

Each line of this text file is in the form: text_of_the_tweet**;**date_of_the_tweet

Each tweet can be an original tweet or a retweet, each retweet starts with the keyword 'RT'

You are tasked to explore this data by using Spark, because your analysis should also be able to apply on very large data sets distributed on a Hadoop cluster, for instance to analyze the communication of other public figures.

You should present your keys findings in form of lists, tables or visualizations.

You can for instance search for:

- Top positives and top negatives words used
- Overall sentiment scores (from positives and negatives words)
- Top contextual words or hot topics (e.g. covid-19, presidential elections) and associated opinions…
- Top hashtags (#) cited
- Top references (@) cited
- …

You can explore all these points over the time (per month or per year for instance), and make a differentiation between original tweets and retweets (RT).

Your final Jupyter notebook should contain at least 5 keys findings (lists, tables, visualizations, …) from these points. You can also comment all these key findings.

You also have additional resources files containing some positives words, negatives words and stop words that can help you for your analysis. Feel free to modify the content of these files if necessary, for removing or adding items during your analysis.

Part 2: Real Estate Market Data Exploration in major French cities

The objective is to perform real estate data exploration of major French cities.

Your company (Immo-Inv) is a real estate agency who wants to understand very well the real estate market in France.

You are the big data analyst of the company and you have access to a 5-years data history of real estate transactions in France (***real_estate_transactions.csv***). The dataset contains details for each transaction: sale date, localization (city, postal code), type of residence, type of sale, land area, living area, number of rooms, price, etc.).

You should use Spark for this analysis because you should be able to apply your analysis to a dataset with the entire real estate market for all cities in France for instance (big data file) distributed on a Hadoop cluster.

Challenges here are to explore all possible aspects of this real estate market (variables, relationships between variables, trends, patterns, outliers, etc.). But at the end you should focus on at least 5 keys findings (lists, tables or visualizations) in your final notebook. You can also comment these findings. You can explore for instance:

- The evolution of the prices (e.g. price per square meter)
- Identification of sales outliers
- Differences of prices per cities or postal codes
- Differences between apartment, houses or other type of properties

- Differences between types of sales
- Differences between cities
- High cost, low cost or emerging cities
- …