M1 BI-ISCE 2020/2021 Dr. Hanen OCHI

Big Data - Project

Instructions

The last few weeks of ST2IBD will be dedicated to projects. This project may be done independently or in pairs. We will have higher standards for those working in pairs, but either way we expect it to be a substantial project on which you devote significant effort. It's difficult to quantify "significant effort" and there's no detailed grading rubric. Part of the purpose of the proposal is for the lecturer to provide feedback on whether the project appears to be of the appropriate scope, with higher expectations for pairs than for individuals.

Report

The report should briefly cover the following topics :

— Problem Definition : What is the problem that you are trying to solve ? What are the challenges of this problem ?

— Methodology : What is your methodology to attack the problem and the associated challenges ? What is the computational and space complexity of your solution in terms of input size ?

— Results and Discussion : What are the outcomes of the project ?

— Guideline : Briefly explain which code was used for which task.

— The code and final report of your project should be submitted before April 10, 23 :59 GMT.

Note that your report should not exceed 8 pages. Y

The rest of this document presents example project subjects to help you get started. While you can just pick a project from the list or get your inspiration from one of them, you are more than welcome to come up with your own ideas. Be creative !

Project proposal n°1 - Real Time Sentiment Analysis of Twitter Data Using Hadoop

Social media for many people has become integral part of their daily life. Social media metrics are now considered parts of altmetrics, which are non-traditional metrics proposed as an alternative to more traditional metrics.

Twitter is an online social networking service that enables users to send and read short 140-character messages called “tweets”. Registered users can post and read tweets, but general public can also read them. This is unlike Facebook, where social interactions are often private. Users access Twitter through the website interface, SMS, or mobile device app.

You can develop the application based on Apache Storm, a distributed computation framework, which adds reliable real-time data processing capabilities to Apache Hadoop. It is fast, scalable, reliable and can be programmed using a variety of programming languages (Python, Java, Scala).

Algorithm Sentiment analysis or opinion mining refers to the use of natural language processing and text analysis to identify and extract subjective information in source materials. Normally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual of a document(s).

Possible tools

- Big Data : Apache Strom, Apache HBase, Twitter’s Search and Streaming APIs,
- Visualization tools : D3 Visualization, Tableau visualization.
- Development tools : Python, R, Java, and Scala.
- Natural Language Processing Algorithms : Python Natural Language Toolkit (NLTK) and AlchemyAPI Service

Project proposal n°2 - Predicting Airline Delays with Hadoop

One of the main goals is using machine learning algorithms to build predictive models with Python packages and data analysis programs. Training the original datasets is important to build models with its performance. Finding a good combination of technologies and programming languages would be cruicial to make a successful project.

Dataset The data can be downloaded from Bureau of Transportation Statistics where it is described in detail. An other link to more detailed data can be found here.

Possible tools

- Apache Pig
- Hadoop
- Python
- scikit-learn

Project proposal n°3 - Vocabulary size of singers and writers

Matt Daniels did a very interesting analysis of the vocabulary size of rappers. He used token analysis to determine each artist vocabulary, and produced various graphs. The complete analysis can be found here :

http://rappers.mdaniels.com.s3-website-us-east-1.amazonaws.com/

Work on a similar analysis, either for a different musical genre, or for writers, for instance. You will first have to build a dataset, and to then write programs that extract various statistics from that dataset. Feel free to go beyond Matt Daniels' work and to produce graphs that show other interesting trends.

Projects from the Big Data for Social Good Challenge

If you need more project ideas, you can check IBM's Big Data for Social Good Challenge. Check the submission page to find Big Data projects that you can take inspiration from, or re-implement your own way : **http://ibmhadoop.challengepost.com/submissions** This page includes, for instance :

— A website with various graphs and statistics about 311 calls in New York City.
— A tool that shows the progression of influenza epidemics in the US and forecasts their evolution.

— A tool that makes it possible to correlate housing price in different areas of London with various demographic variables (crime in particular).

IBM provides many datasets to work with that you can use in your projects :
http://ibmhadoop.challengepost.com/details/data