Cairo University
Faculty of Engineering
Computer Engineering Department

CMP4011
Spring 2025

Big Data and Cloud Computing

Project Document

Project Description:

You are required to find an innovative *idea* and a dataset that supports your idea and on which you will apply the analytical techniques we use throughout the course.

Your idea must address a business problem, bring a business solution or provide values, insights or recommendations for business users; those who will benefit from the results of your work.

You are required to use one of the big data processing frameworks (e.g. Hadoop, Spark, Flink, etc.) integrated with your preferred language. The system should be implemented in pseudo-distributed mode (single-machine multiple processes). The implementation in fully distributed mode (multiple-machines multiple-processes) will be a bonus. Using cloud computing providers (Azure or AWS) will be a bonus.

FAQ:

1. **What do you mean by innovative idea?**

   Innovative idea is a new business problem you are trying to solve. For example:

   - A business problem for a bank may be "*Should we give this customer a loan or not?*" or "*Based on what factors/conditions should we give our customers a loan?*"
   - A business question for a retail store concerning shelf management may be "*What are the items commonly bought together by a sufficiently large number of customers?*" This is commonly referred to as market basket analysis.
   - A business problem for a magazine/newspaper may be "*What determines a customer's decision to subscribe or not?*"

   Your idea is not limited to business only but can be extended to governmental bodies, environmental institutions, and societal organizations. For example:

   - *"Which people are more likely to vote for/against a law?"*
   - *"How will the climate on Earth change for the next ten years?"*

2. **Is there a restriction on the dataset?**

   Not really. You should find a dataset that's large enough. We are talking here about datasets in order of hundreds of megabytes up to gigabytes. It needs to contain several features (20 or more columns). It needs to have 1 million rows or more.

3. **What programming language(s) can I use?**
   This is an analytics project, so it is recommended to use either *R* or *Python* for this purpose. Both languages are very popular for data analytics. However, there is no real restriction on which programming languages to use.

4. **What analytical techniques can I use?**
   You can use (but not limited to) all the analytical techniques studied in this course. Example:
   - For a ***classification*** problem, you can build a classifier using logistic regression or K-NN, train a neural network, build an SVM, among many other techniques.
   - For a **prediction** problem, you can go with MLR (Multilinear Regression) or PCR (Principal Component Regression).
   - For a **segmentation** problem, you can try different clustering techniques (K-means, hierarchical clustering, etc.)

5. **We are seniors and about to graduate, will the project consume too much time?**
   No.

6. **I have no idea in my mind.**
   We recommend having a look on Kaggle and Analytics Vidhya where you will find lots of ideas and datasets. You will also find many active competitions which you can actually contribute and participate in.

7. **Is this project considered a machine learning project?**
   No. This is a data analysis project. Proposals that focus on collecting a dataset and just training machine learning models and reporting accuracies will be **rejected**. You need to think to provide some technical and business questions related to your problem and use the studied concepts find answers to it.

8. **How should we use Hadoop or Spark in this project?**
   The target is to pick one of the algorithms you will apply (e.g. KNN) and implement it yourself in a distributed way with MapReduce.

9. **How to avoid my idea being rejected?**
   Your proposal needs to contain your idea, dataset and planned approach. Here are some guidelines for picking the dataset and planning your approach:
   a. The dataset needs to contain many features for broad analysis (minimum 20 columns).
   b. Mention how you imagine the EDA phase in your approach (just in one or two lines).
   c. Mention what algorithm will be implemented with MapReduce.

d. You need to use at least one descriptive analysis method (Association rules or clustering).
e. You need to try several predictive analysis methods (classifiers, regressors, etc)
f. Avoid datasets with encrypted or anonymized features.

Deliverables:

1. **Project Proposal**:

The proposal should not exceed one page specifying the following points clearly:

a. Idea:

The problem statement should be described clearly.

b. Dataset(s):

Links to the dataset(s) that will be used.

Clear descriptions of the dataset and its features that **you write by yourself**

c. Planned approach or Proposed solution.

A very brief plan of your proposed approach (you can change it later during the implementation phase).

2. **Final Delivery**:

a. Final Document containing:

i. Brief problem description.

ii. Project pipeline.

iii. Analysis and solution of the problem:

- Data preprocessing.
- Data visualization.
- Extracting insights from data.
- Model/Classifier training.

iv. Results and Evaluation.

- Model accuracy on train, test, and validation data.

v. Unsuccessful trials that were not included in the final solution.

vi. Any Enhancements and future work.

c. Codes

d. Presentation:

i. Business part.

ii. Technical part.

Project Schedule:

Phase	Due date
Team Formation & Project proposal.	TBD
Final Delivery.	TBD

Notes:

- There is a penalty for late submissions in any of the three mentioned phases.
- **Any sign of cheating or plagiarism will not be tolerated and will be graded ZERO in the project.**

Suggested Ideas:

1. Kaggle competitions and specially active ones (https://www.kaggle.com/competitions)
2. Kaggle datasets https://www.kaggle.com/datasets
3. Analytics Vidhya has some datasets and sometimes competitions https://datahack.analyticsvidhya.com/contest/all/
4. 17 places to find datasets for data science projects: https://www.dataquest.io/blog/free-datasets-for- projects/
5. Data Science for Social Good: It has ideas but no data. http://www.dssgfellowship.org/projects/