Cairo University
Faculty of Engineering
Computer Engineering Department

CMP 461
Spring 2019

Data Mining, Big Data and Analytics

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.

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 megabytes up to gigabytes.

3. **What programming language(s) can I use?**
   This is an analytics project, so you can use either R or Python for this purpose. Both languages are very popular for data analytics.

4. **What about Hadoop and its ecosystem?**
   Again, we are focusing for this project on big data analytics. For big data processing and parallelization with MapReduce jobs (beyond the scope of this project), you can use Hadoop, Spark, Flink and other frameworks of course. This is a bonus point.

5. **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, and 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, mean shift clustering, etc.)*

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

7. **I have no idea in my mind.**
   Check the suggested ideas in the last part of the document. We recommend Kaggle where you will find lots of ideas and datasets. You will also find many active competitions which you can actually contribute and participate in.

Team Formation:

A team is formed of 3-4 members. Please fill in your names in this Google Sheet: ghttps://docs.google.com/spreadsheets/d/1yquGn7JM7hsPX4rCAMS14qgOgkBHZ51gadH_5oo4-qs/edit?usp=sharing

Deliverables:

1. **Project Proposal**:
   - **a.** ***Idea***:
     The problem statement should be described clearly.
   - **b.** ***Dataset(s)***:
     Links to the dataset(s) that will be used.
   - **c.** ***Planned approach or proposed solution***.
2. **Final Delivery**:
   - **a. Document containing:**
     - *i.* Brief problem description.
     - *ii.* Project pipeline.
     - *iii.* Analysis and solution of the problem:
       - Data visualization.
       - Data preprocessing.
       - Model building.
       - Model training.
     - *iv.* Results and Evaluation.
       - Model accuracy on test and cross validation data.
     - *v.* Unsuccessful trials that were not included in the final solution.
     - *vi.* Any Enhancements and future work.

b. **Codes**

c. **Presentation**:

   i. *Technical part.*

   ii. *Business part.*

d. **A readme text file** containing the names of the team members.

Project Schedule:

Phase	**Week**	**Due date**
Team formation.	Week 4	Saturday 2nd March, 11:59 pm.
Project proposal.	Week 7	Saturday 23rd March, 11:59 pm
Final Delivery.	Week 12	Sunday 28th April, 11:59 pm.

Delivery Details:

- Deliverables should be sent by e-mail to *submissions.bda@gmail.com* using the subject:
  - **[Semester][Proposal][Team #][Team Number]**: for the project proposal.
  - **[Semester][Final][Team #][Team Number]:** for the final delivery.
- Don't print any document or submit the project on a CD. All submissions are electronic.
- There will be a late 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)
   a. Quora Question Pairs competition on Kaggle: Can you identify question pairs that have the same intent? https://www.kaggle.com/c/quora-question-pairs
2. Kaggle datasets https://www.kaggle.com/datasets
3. The Data Incubator: Data Sources for Cool Data Science Projects
   a. http://blog.thedataincubator.com/2014/10/data-sources- for-cool- data-science-projects-part-1/
   b. http://blog.thedataincubator.com/2014/10/data-sources- for-cool- data-science-projects-part-2/
4. The SAPA Project: Explore your personality. The data is available at https://dataverse.harvard.edu/dataverse/SAPA-Project

5. Analytics Vidhya has some datasets and sometimes competitions https://datahack.analyticsvidhya.com/contest/all/
6. 17 places to find datasets for data science projects: https://www.dataquest.io/blog/free-datasets-for- projects/
7. Data Science for Social Good: It has ideas but no data. https://dssg.uchicago.edu/projects/