Capstone Team Project
Sentiment Analysis and/or Recommendation Systems (DEFAULT)
From Business Questions to Business Decisions

Note:

- **All the other capstone projects should follow a similar format of 11 steps described below. The difference will be in the type of data used and the type of the business problem to be addressed. Those nuances/differences are highlighted in BLUE in this document.**
- **The Teams are self-formed, and should include up to 5 (five) students.**
- **The DE students will be presenting via WebEx during the on-campus class time unless they have on-campus students in their teams.**
- **The final Project presentations will be delivered on April 19 and April 22 during the regular class time (Tu/Th: 1:30 PM)**
- **If your Capstone project is BUSINESS-SENSITIVE, e.g., it is related to your work place, please, discuss with the instructor (Nagiza: nagiza.samatova@gmail.com) how we might approach evaluating your project (e.g., NDA agreement, etc.)**

Suppose that your team has just joined the Data-guided Business Intelligence (DBI) unit in your company. Your boss meets you on your first day and makes the following comment:

> *I keep hearing that Sentiment Analytics and Recommendation Systems are promising technologies that improved the efficacy of many businesses by more than 5%-7%. I feel that we are behind and should build the expertise within our unit and explore opportunities for using such technologies in our end-to-end data-guided business intelligence process. Before starting to collect the data specific to our business, let's first gain some experience using previously collected data by Yelp. I have heard that your team has proven already as being the team of infinite intelligence; hence, I am choosing your team to spear-head the effort!*

Your first assignment, as a team, within the DBI will include the following:

1. To examine and assess the business value of the Yelp data set (the CSV and/or JSON files will be provided to you):
   http://www.yelp.com/dataset_challenge

2. To create a list of four business questions that could utilize Sentiment Analysis and/or Recommendation System technologies to guide business intelligence.

3. To find the survey papers and tutorials (see Google Scholar, https://scholar.google.com/) that summarize the most recent R&D advances in Sentiment Analytics and Recommendation Systems.

4. To propose a **target** question your team will address using the Yelp Data Set with Sentiment Analysis and/or Recommendation Systems approaches. For example, *how to detect the FAKE reviews*? (Yelp has the ground truth data, i.e., manually curated data of fake and non-fake reviews for this question)

5. To justify the possible business value behind answering such a question.

6. To discuss how you may evaluate the quality of different answers: both from business and technology perspectives.

7. To recommend 3 research papers that could be relevant to solving your target question and write a paragraph summary for each paper:
    - When making your recommendations, pay attention to how many people cited such papers (*Cited By* field in Google Scholar): highly cited papers may indicate more impactful technology and/or earlier technology, and
    - Whether these papers have been published in high impact conferences such as SIGMOD KDD, ICDM, ICDE, NIPS, etc. (see computer science conference rankings: https://en.wikipedia.org/wiki/List_of_computer_science_conferences).

8. To describe which features/attributes from the Yelp Data Set have you used.

9. To prototype a relatively simple, 'bread-and-butter' solution that solves this question from end-to-end perspective (both the design and the implementation demo). Make sure to include at least two (2) components (in total); each from a different category/topic including but not limited to (see the roadmap of other technologies relevant to data science in general, https://datavizblog.files.wordpress.com/2013/10/image1.jpg):
    - Graph Embedding
    - Deep Learning
    - Time series forecasting
    - Apache Spark
    - Heterogeneous, dynamic, and multi-attribute graphs
    - Fraud/anomaly/outlier detection
    - Model inter-comparison, diagnostics, design of data science experiments
    - Generalized Linear Models
    - Sentiment Analysis and/or Recommendation Systems

While using Python/R is a desirable solution, you are not restricted to Python alone. Any other programming tools, packages, etc that are available in the open source domain is a fair game for this initial prototype implementation. ***We are not after the most optimized and accurate solution. We would like to test your ability to reason about the business intelligence problem from the end-to-end perspective!***

10. To submit your Project materials in Moodle by **October 22**
    - Clearly describe how each member of the team contributed to this assignment (in a separate file of your submission)
    - Include the identified survey/tutorial and research papers, and paper summaries
    - Include all the required codes & README on how to use your end-to-end solution to the target question as part of your GitHub portfolio. Provide us with the link to your GitHub.
    - Describe how you evaluated the quality of your solution to the target question
    - Include Power Point slides summarizing your project

11. To give a **10 minute presentation on April 19 and/or April 21** in front of the entire class (the schedule of which teams present when will be determined by the teaching staff).
    - Show the working **DEMO** of your program as part of your presentation