**CIS 8045 Term Project**

**Project Team and Due Date**

You can do this project in team. **A team should contain no more than 3 people.** The project is due before the final exam.

**Project Data**

The project data is a set of real customer review data collected from Skytrax ([www.airlinequality.com](http://www.airlinequality.com)), a major website for customers to evaluate various airline companies. This dataset contains 41,396 review entries, with each entry capturing the rating of a particular airline company by a particular author. In this context, **an author is also a rater as well as a customer**. The data is stored in a MongoDB database. A sample document (corresponding to a review entry) is as follows:

{

"authorcountry" : "Israel",

"rating\_valuemoney" : 5,

"recommended" : 1,

"airlinename" : "el-al-israel-airlines",

"travellertype" : "FamilyLeisure",

"cabin" : "Economy",

"aircraft" : "Boeing 747-400",

"rating\_overall" : 9,

"reviewcontent" : "Flight was half-full. I had the whole row to myself. ...",

"rating\_inflightEnt" : 4,

"reviewdate" : "7/4/2015",

"rating\_cabinstaff" : 5,

"route" : "TLV to JFK",

"rating\_seatcomfort" : 4,

"rating\_foodbeverage" : 5,

"authorname" : " Moam Ben-Shalom"

}

The overall rating is on a scale of 1-10, with a higher level indicating more satisfaction. The ratings of individual aspects (e.g., cabin staff, seat comfort, food & beverage, etc.) are on a scale of 1-5. The property **“recommended”** indicates whether or not the author would like to recommend this airline to others. Other property names are mostly self-explanatory.

Since the data is provided to you in a MongoDB Backup format, you need to first import it to MongoDB and then use PyMongo to access it in developing your subsequent problem solutions.

**Project Tasks**

In this projects, you need to develop analytics solutions for the following two sets of problems. Unless explicitly specified, your solutions need to be realized using **Python**.

First, you need to develop text-analytics solutions for the following problems:

* 1. You need to develop a solution to assess the sentiment of authors in their reviews. Develop your solution to assess review sentiment in two ways: a lexicon-based way and a machine-learning-based classification (you can use ***rating\_overall*** for learning);
  2. You need to develop a model to explain what factors influence authors’ likelihood to recommend an airline. Your model should include the information retrieved from review content. You can use your own judgement to decide which text-related factors, alongside with other available variables from the data, to include in the model, and you need to justify them in your project reports;
  3. **(Optional)** You may consider enriching your above analysis (in part b) by extracting the topics from review content and using some topic measures (e.g., the number of topics and topic loadings).

Second, you need to construct a rating network between author and the airline companies they rated. Such rating network captures who rated which airline companies and can be used to realize the network-based analyses:

1. Use Py2Neo to construct this network in Neo4j;
2. For any airline company, it is important compare itself with its direct competitors based on customer rating and sentiment. Two airlines are considered direct competitors **only when** there are some customers rating both of them. So you need to develop a solution to compare airlines with their competitors. In terms of what to compare, it is usually important see how direct competitors are rated and perceived (in terms of sentiment) differently by their common raters;
3. **(Optional)** Network analysis can also enhance the analysis on recommendation likelihood (i.e., the part d above). For example, one can consider how the ratings and review content features (sentiment or topics) for a focal airline’s most direct competitor (provided by the common raters) influence the likelihood of recommendation (by the common raters as well) for the focal airline. You need to develop a solution to enrich the recommendation analysis by incorporating some of these measures that you would like to use.
4. Regarding raters, it is often useful to find raters with common interests. A basic analysis can be to find raters who rated and commented common airlines. You can develop a solution of your own to perform this analysis and generate some insights. Then **(optionally)**, you can further analyze how similar these raters are with each other in terms of their ratings of the common airlines and their sentiment toward the common airlines.

**What to Turn In:**

First, you need to turn in a report to clearly explain your solutions. For each problem:

* + - Justify your proposed solution;
    - Briefly explain how you do in your solution;
    - Show and explain your main results and findings.

You can make your report as professional as you can (e.g., using tables, plots or other visualization approaches for results presentation), but you do not have to worry too much about “decorating” your report.

Second, you need to turn in your complete Python code. Use sufficient in-code comments to make your code readable. Make sure that your code is executable, so that others should be able to replicate your solution results using your code.

**Grading:**

Since this is a project, the solutions are fairly open. Our objective is to solve problems, and therefore there is no absolutely correct answers. The grading will be based on the extent to which your proposed and developed solutions demonstrate that you understand the course materials and be able to apply them.