**MBAN Project 1: (Based on a True Story):**

Your team of 1-4 group members is working for the fictional “Massive Data Consultants”.

You have a client who has a nascent social networking site focused on users being able to follow other users based on common declared interests. Users can follow each other (similar to twitter) however, users are able to sort and search other users based on shared social connections as well as variety of declared interest categories. Currently connections happen due to customer pull, or agency. Also, similar to twitter, users can broadcast messages to all their followers (think of tweets or posts). Thanks to generous support programmes this client has somehow gotten this far without performing any data analytics and has no internal capacity to do so!

To take things to the next level, VC’s have told the client they need to show that they can increase the connectivity and potential information flow within their network. To this end, the client has hired your company, Massive Data Consultants to help them understand their data as well as leverage it for competitive advantage.

You have received an initial “data dump” performed by the client’s engineers so you can perform your analysis. There will not be another opportunity to pull more data (it is what it is) until after you deliver your report. The social network of “followers” and “followees” is available as well as each user’s declared interest categories (each user can have zero or more declared interests). User’s broadcasted messages are not available in this data dump.

**Part A:**

Help the client gain some basic understanding of their data. You should answer basic questions such as:

* How many users are there?
* How many interest categories were created by users?
* What is the distribution of interests in the entire population of individuals?
* How many declared interests do individual users have (histogram)?
* Determine and plot appropriate social network measures such as connection degree distribution.
* Can you make any comparison to well-studied social networks such as Twitter?

**Part B:**

Train a classifier capable of predicting which user will follow another user. The client hopes this will enable the app to make recommendations on who a user should follow and thereby increase the number of user follows, ultimately leading to increased revenue. In other words, they hope by leveraging technology they will be able to move towards a "push" model, where the agency in initiating connections moves to the application, by suggesting appropriate new connections.

Validate, and assess the performance characteristics of your classifier.

State any assumptions you are making.

**Part C:**

Produce a report including your results for A) and B)

Also answer the questions:

* What recommendations do you have for the client regarding additional analysis?
* What recommendations do you have for future “Data dives” or dumps of their data so that:
  1. performance of your classifier/recommender might be better assessed
  2. your analysis and performance of the classifier/recommender could be improved

At a high level, describe a strategy the client could use (next steps) to evaluate your classifier’s operation with respect to the client’s business?

*As “Appendix A” answer the following:*

* What kind of hardware did you run the analysis on?
* How long did this take?
* What are your recommendations regarding the volume of data you could manage for further analysis?
* The “%timeit magic command” included in ipython may be handy here!

*As “Appendix B” attach all your source code:*

* You must use a combination of unix command line tools, and/or python/matplotlib/scikit to perform all analysis.

***Hints/Notes/Asides:***

Notice: Projects that fulfil the requirements will receive an A grade. To earn an A+ I want your team to demonstrate some initiative or thought leadership

Hint: often it’s a good idea to pre-process your data, and even produce intermediate results which could be stored in files. This makes later analysis steps easier because data is already in a form amenable to learning. An example would be transformation or reduction of dimensionality (see: <https://en.wikipedia.org/wiki/Jaccard_index> )

Hint: when handling text, it’s often a good idea to strip any leading or trailing whitespace.

Aside: the data was encoded in a way such that id’s and interest categories have been substituted to protect privacy. The network structure and other relationships have been preserved, however. We are not interested in insights such as “users interested in government are also interested in cabbage” here, anyway.

Hint: The following code snippets might be helpful to save/load a trained model to/from a file:

from sklearn.externals import joblib

joblib.dump(clf, 'filename.pkl')

#filename.pkl can be replaced with a filename of your choice

#clf is the variable name of the classifier.

#elsewhere in the code I can load the same trained classifier:

clf = joblib.load('filename.pkl')

#make sure you use the same file name as before!

***Submission***: Hand in a printed paper report to the OMIS offices before it closes on March 1. Make sure to include cover page with the course number, and the student names and id’s of your group.

Professor Marek says: You may ask me questions. I can pretend to be the client, or your professor depending on what kind of advice you are seeking :-)

Most importantly: have fun!