**Candidate Name: Nailson Boaz Costa Leite**

**Date: 18/11/2017**

**Challenges selected:**

* Challenge 2 – Recommendation Engine
* Challenge 3 – Understanding user and product interaction

**Environment and tools:**

* Jupyter notebook version 5.0.0
* Python 3.6.1 | Anaconda 4.4 (64 bits) - pandas, itertools, collections
* R 3.4.2 – dplyr, ggplot2, arules

**Challenge 2 Recommendation Engine:**

Build a recommendation engine for company ABC’s video streaming product. Currently, the videos shown on their home page to new users are manually chosen and you need to implement a recommendation engine to increase the conversation rate.

You have been presented with the following business questions:

* Classify each video into the following buckets:
  + "Hot" - means trending up. These videos are candidates to be shown.
  + "Stable and Popular" - video view counts are flat, but very high. These videos are candidates to be shown too.
  + "Everything else" - these videos won't be shown.

Notebook: “Challenge 2.1 – Recommendation Engine.ipynb”

I divided the problem in two tasks:

1. Defining when a video is popular: Using the distribution of video counts I set the 3rd quartile as a threshold for popular videos. So I marked the videos above the threshold as popular.
2. Defining when a video is trending up: Using the angle degree of the linear regression line of video visualizations over days we capture the trendiness of each video. Visually I set videos with angle in the range of -30° to 30° as “stable” videos and “trend” videos above 30°.
3. Using the trends and the popularity information we classify videos as “Hot”, “Stable and Popular” and “Everything else”.

* What are the main characteristics of the "hot videos"?

Notebook: “Challenge 2.2 – Recommendation Engine.ipynb”

The videos classified as hot, are generally videos of **good quality** (1080-480p). Compared to other classes, it has the highest proportion of **videos in English** (~ 34%), on average **shorter videos**, comparable to the "stable and popular".

In comparison to age (publication day), the hot videos are in average a few days more recent than the others (28 days).

Some comments:

\* “Hot”, “Stable and Popular” videos are in general short videos with good quality.

\* Videos in English are the most of “Hot” videos.

\* Chinese videos are usually the most "Stable and Popular"

* After having identified the characteristics of the hot videos, how would you use this information from a product standpoint?

Notebook: “Challenge 2.2 – Recommendation Engine.ipynb”

I believe that with the information of the most important characteristics of some types of videos, the company could automate the selection of videos from the homepage for new users, highlighting videos that have similar characteristics with "HOT" and "Stable and popular" such as short videos with good video quality or prioritize videos in English (classified as hot) or Chinese (stable and popular). This approach could also help in the selection and ranking of new videos that are published. However we must always be careful about the impact and bias that this type of approach may entail.

**Challenge 3: Understanding user and product interaction**

VP of product in the company ABC has asked you to review how users interact with their online travel website.

They store their data in JSON files. Each row in these files lists all the different cities that have been searched for by a user within the same session (as well as some other info about the user).

Business questions:

* There was a bug in the code and one country didn't get logged. Can you guess which country was that? How?

Notebooks: “Challenge 3.1 - Understanding user and product interaction.ipynb” and “Challenge 3.2 - Understanding user and product interaction.ipynb”

It's probably Canada. There are many Canadian cities (such as "Toronto ON", "Montreal QC" and "Vancouver BC", with 457, 334 and 207 searches respectively) in user’s sessions with missing country field. In addition to this, Canada is one of the countries with many searched cities that does not appear in the list of distinct countries in the dataset.

* For each city, find the most likely city to be also searched for within the same session.

Notebooks: “Challenge 3.2 - Understanding user and product interaction.ipynb” and “Challenge 3.1 - Understanding user and product interaction.ipynb”

The safest way to solve this is to calculate the frequency and ratio by each cities relation and define the most likely searched city the one with the highest frequency. We could also use the ratio to filter cities that even having a higher frequency are not proportionately relevant.

Another way to analyze this kind of data is to transform the researched cities in a session to transactions, and then use association rules learning algorithms like Apriori to discover the frequent item sets (or cities sets) and find cities association.

* Travel sites are browsed by two kinds of users. Users who are actually planning a trip and users who just dream about a vacation. The first ones have obviously a much higher purchasing intent. Users planning a trip often search for cities close to each other, while users who search for cities far away from each other are often just dreaming about a vacation. Based on this idea, can you come up with an algorithm that clusters sessions into two groups: high intent and low intent.

Notebook: “Challenge 3.1 - Understanding user and product interaction.ipynb”

Using latitude/longitude, for each one of the cities researched by a user, calculate the pairwise the distance between them and extract some metrics from distances, such as the mean, standard deviation, minimum and maximum. Using a clustering algorithm, such as K-mean, we could segregate and visualize different user search behavior and cluster them in high intend and low intend groups.