# Spark Practical Work – Report

Institution: UPM   
Lecture: Big Data  
Due Date: 21st January 2018

Authors: ECHAVARRIA CABALLERO, LADY CAROLINA  
 HUSSNAETTER, MICHAEL

Group: 22

# Workflow implemented

1. Dataset Analysis
2. Problem Characterization
3. Data Task and abstraction
4. Interaction and visual encoding
5. Algorithmic implementation

Dataset Analysis

The selected dataset corresponds to the Yelp Dataset, which is a service that allows users to review different business. The Dataset itself is divided into different datasets:

* Business
* Review
* User
* Checkin
* Tip
* photos

Problem Characterization

**Issues of the application domain and end users involved.**

**The main use of the platform is to consume information, but we are facing with two different users:**

* **The first users want to know which one is the best restaurant to go in a certain location? When should users go to that selected restaurant regarding to the amount of checkins it has? Has the place improve according to the number and sentiment of the reviews?**

Those End users want to know which restaurant to select according to the location and the rates. Then, see the amount of checkins per hour in order to select the best time to visit the place, and see the amount of reviews in a period of time for that specific place to ensure or reject the selection according to the behaviour of the reviews, which are directly related with the improvement or not of the restaurant.

* **The next users are those concerned by human behaviour or interested in marketing strategies, and would be interested about knowing the relationship between the number of stars and the amount of reviews. The initial hypothesis is that users post reviews when are extremely satisfied or** extremely disappointed. Results show that the amount of reviews increases with the amount of stars, being the stars 4 and 5 the ones with more reviews. This unexpected behaviour can be explained through this second hypothesis: The older the people are, the more exigencies they have and the less access to social networks and platforms like Yelp they have.

Given the second hypothesis, it would be interesting to have the age of users posting reviews, in order to know the relationship between the age of users and the amount of reviews.

## Data and Task abstraction

Basically is why the visual analytic tools are used for?

Given that we are facing two different kinds of users, as mentioned previously:

In the first case, the visualization tool is used to consume information about restaurants, in order to select the best place to go according to the users preferences.

**Task – Identify the tasks required by end users in their workflow**

* Explore/search the restaurants according to the location in the map and number of stars.
* Filter restaurants in: Free WIFI availability, if takes reservations, take out and caters.
* Select a place according which restaurant he/she finds more interesting.
* Explore the time to visit the restaurant according to the amount of checkins in a day of the week and the hour.
* Explore the amount of reviews in a period of time according to the number of stars (rates), in order to see the behaviour of the user’s comments and conclude about the improvements of service on the place in time.

**Data – Determine the representation that best fits user’s needs.**

* Location of restaurants in a map, using coordinates and a map server
* Filter restaurants using check boxes according to WIFI, etc.
* Show plots with amount of checkins per day of the week and hour
* Show plots with amount of reviews per year and filter per number of stars

In the second case, the visualization tool is used also to consume information about restaurants, but focused on the behaviour of the reviews respect to the rate of the place (or number of stars). This relationship is updated according with location and amount of business.

**Task**

* Explore/search a group of restaurants according to the location in the map and number of stars.
* Visualize the behaviour of the reviews according to the amount of stars (rate) of the selected group of restaurants in the map.

**Data**

* Positions of restaurants in a map, using coordinates and a map server, grouped by location.
* Plot the amount of reviews versus the rates.

Interaction and visual encoding

Determine the specific design choice for creating and manipulating the visual representation of the abstract data.

* As a first view, users will see a cartographic arrangement of the restaurants grouped by location and using the leaflet as a Mapping Library.
* On the first view, the users will be able to change the restaurants they are seeing on the map by filtering of the data according to the rates, free WIFI availability, if takes reservations, take out and caters.
* Also, in the current view, the second kind of users mentioned in the previous sections, will be able to see a scatter plot with the number of reviews respect to the number of stars.
* Once the users have selected a restaurant that fits their interests, they will obtain detailed information about the place in a second tab. By clicking on the selected place, two plots are displayed:
  + A histogram representing the frequency of checkins per hour, selecting the day of the week.
  + A histogram showing the frequency of reviews per year.

Algorithmic implementation

**Efficient implementation to achieve what was designed in the previous steps.**

# Final comments and remarks