# 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 improved 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 check-ins 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.

Those users are also concerned not only about some general information of the amount of restaurants, the total reviews and the average review ratings by state, but also about the detailed review ratings by state, showing the percentage of each rate.

## Data and Task abstraction

Basically is why the visual analytic tools are used for?

Given that we are facing with 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 check-ins 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 over the 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 check-ins per day of the week and hour
* Show plots with amount of reviews per year and 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) and the average ratings per state.

**Task**

* Visualize the behaviour of the reviews according to the average rating scores.
* Explore the average review ratings by state and visualize more detailed proportion of review ratings by state

**Data**

* Plot the amount of reviews versus the rating scores.
* Distribution grid of average rating score per state

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 in the tab “Look for a restaurant”, where the identity channel is used due to the main objective is to provide information about identity and location, using a Heat map representing the amount of check-ins and the cartographic arrangement of the restaurants grouped by location, with leaflet as a Mapping Library to have the world map background:

* The Heat map shows the number of check-ins over the map, and gives to users an overview of the amount of check-ins distributed on the area, which is related with the most or less visited areas, and it will help the customers to select an area to go. Each number of check-ins is represented by two-dimensional marks to which has been assigned a colour scale that goes from red to yellow: the red represents the areas with more check-ins and yellow the areas with less check-ins.
* The clusters of restaurants are shown using area marks as circles with colours green, yellow and orange expressing how many restaurants are grouped: In the case of orange circles, clusters with more than 100 restaurants are shown; yellow circles represent clusters between 10 and 100 restaurants and green, cluster with less than 10 restaurants. The restaurants are grouped with the purpose of having a general idea of the location of restaurants without annoying the users with a big amount of points located in the map. Every time users zoom in on a certain area in the map, the circles are updated with more detailed clusters, obtaining at the end the single restaurants seen as point marks.

Given that the first kind of users (mentioned in the section Problem Characterization) want to select a restaurant that fits their interests, in the described first view, the customers are able to change the restaurants they are seeing on the map by filtering of the data according to: The number of stars from 1 to 5, due to some users want to select restaurants with an specific rating score; free WIFI availability, if takes reservations, take out and caters. Once the filter is done, the map is updated with the new amount of restaurants and the right panel will also show the number of restaurants in the current view from the total amount of restaurants.

Once the users have chosen a restaurant, by clicking on the selected location, they will be able to obtain detailed information about the place:

* + A bar char was chosen given that customers want to lookup on the time using the amount of check-ins as quantitative value attribute and the hour as categorical key attribute. Because of the users need to select the day and time to visit the restaurant according to the amount of visits, the number of check-ins is a good indicator since it has a close relationship with the amount of visits per hour.
  + To lookup the amount of reviews per year, a stacked bar chart was chosen, because it allows consulting values accessing with: two keys which are year and number of stars; and one quantitative attribute equivalent to number of reviews. The secondary key, in this case the number of stars, is used to build the one-dimensional vertical structure on the bars, that shows the part – of – whole relationship between the number of stars and the amount of reviews per year. The information offered by this plot, gives to users an idea about the behaviour of customer perception over the time, and helps the users to ensure or reject the selected restaurant.
  + A normalized stacked bar chart stretches each rating scores on the stack to the maximum possible length, in this case 5, representing percentages instead of absolute values as the stacked bar chart explained before. The current plot helps users to understand the contribution of each rating scores or number of stars to the whole numbers of reviews, over the time.

For those users concerned by human behaviour or interested in marketing strategies, mentioned in the section Problem Characterization, the second tab of the visual analytic tool, called “Inspect dataset” presents general information:

* In order to visualize the behaviour of the reviews respect to the average rating scores, a scatter plot was chosen because it is very effective to characterize the distributions encoding the variables using the horizontal and vertical spatial position channels to express the number of stars and number of reviews, respectively, with point marks. The result plot is showing that the most of the reviews are done for ratings equal to 4 stars, followed by 3.5 and 4.5 stars.
* A grid plot showing the detailed information about the amount of reviews for each state, according to the number of stars, through the magnitude channel to provide quantitative information using area marks and luminance colour: the proportion is shown with circles, where the size of the circle represents the percentage of reviews, and the colour gradient represents the average score. In this distribution grid of average rating score, each column indicates a proportion of each score by state while each row indicates a proportion at a specific review score in each state. From this grid, it is possible to see the pattern that most states have businesses with average rating scores between 3.0 and 4.5.

The information derived from the visual analytic tool becomes a tool for experts to analyse and make conclusions about the behaviour of reviews according to the average rating score and state.

Algorithmic implementation

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

# Final comments and remarks