Proposal

Group members:

- **Cristina Giraldo**
- **Renzo Castagnino**
- **Siqi Jiang**

● What problem did you select and why did you select it?

When people want to plan a vacation or a trip, it is normal that we are interested in knowing the place, but it is also part of our interest to know where we are going to sleep. The best way to know about the lodging is knowing how good the reviews of the place are. However, commentaries sometimes mismatch ratings and it is the idea with this project to confirm if the commentaries and ratings are reliable.

• What database/dataset will you use? Does it need to be cleaned?

The dataset that corresponds to 10,000 reviews of 1667 hotels in the US for the period 2002 to 2018. The cleaning process will mainly consist in depurating the reviews for the text analysis. However, other columns also need cleaning process:

- reviews_text
- reviews_title
- reviews_usercity
- reviews_userprovince
- review_date
- reviews_dateseen

Source:

https://data.world/datafiniti/hotel-reviews

• What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

So far, we are considering to use algorithms related to *text classification/text mining*. The following are some of the algorithms that we could use in this process:

- Naive Bayes
- Supporting Vector Machine
- Grid search

However, at this early point we can not assure if we will have to customize the code or use the standard.

• What software will you use to implement the network? Why?

For this project we will use pycharm running over anacondas and possibly QT designer. The packages to be used are the following:

- Pandas
- Matplotlib
- seaborn
- PyQt5
- NLKT
- RE
- Scikit-learn

• What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

- https://monkeylearn.com/text-classification/
- Natural Process Language with Python, Steven Bird, Ewan Klein and Edward Loper
- Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data, Dipanjan Sarkar
- Shaik, Javed. (2017). Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK. Retrieved from https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a.

• How will you judge the performance of your results? What metrics will you use?

We can judge the performance of the algorithm according to the accuracy given for the metrics accuracy_score from sklearn.metrics

• Provide a rough schedule for completing the project.

N.	Activities	Week 04/01	Week 04/08	Week 04/15	Week 04/22
1	Search Dataset				
2	Verify if the dataset is clean or not				
3	Problem Specification				
4	Problem Understanding				
5	Proposal elaboration				
6	UI Planning				
7	**Code Planning**				
7.1	Data Cleaning				
7.2	Noise detection				
7.3	Noise Removal				
7.4	Feature Selection				
7.5	Algorithm Programming				
7.6	UI Programing				
8	Evaluation				
9	Result Exploitation				
10	Testing				
11	Final report				
12	Individual report				
13	Presentation				