

A top-down view of a wooden desk. In the top left is a small potted plant with green grass-like leaves. To its right is a white computer keyboard. In the bottom right is a white coffee cup on a saucer. Below the keyboard is a black spiral-bound notebook with a pen resting on it. The background is a light brown wooden surface with some faint, wavy white lines. The title 'Text Mining and Search' is written in large, bold, white serif font, centered on the desk.

# Text Mining and Search

Topic Modeling and Text Classification

+

Lorenzo Lecce 830881

Andrea Lucini Paioni 826578

Academic Year 2023/24

# The Yelp Dataset



## Yelp Open Dataset

An all-purpose dataset for learning

```
df_review.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 6990280 entries, 0 to 6990279  
Data columns (total 9 columns):  
#   Column      Dtype  
---  ---  
0   review_id   object  
1   user_id     object  
2   business_id object  
3   stars       int64  
4   useful      int64  
5   funny       int64  
6   cool        int64  
7   text        object  
8   date        datetime64[ns]
```

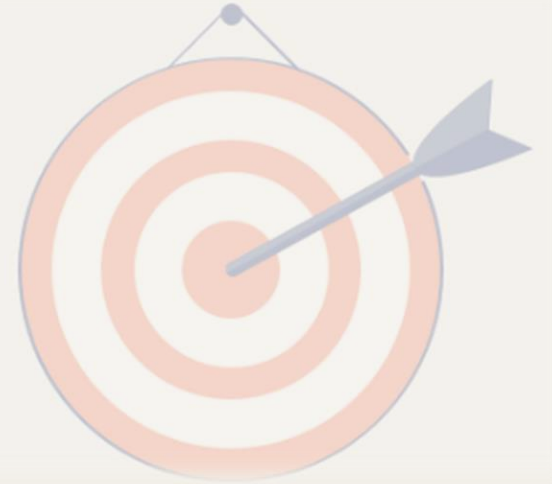
```
df_business.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 150346 entries, 0 to 150345  
Data columns (total 14 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   business_id  150346 non-null object  
1   name         150346 non-null object  
2   address      150346 non-null object  
3   city         150346 non-null object  
4   state        150346 non-null object  
5   postal_code  150346 non-null object  
6   latitude     150346 non-null float64  
7   longitude    150346 non-null float64  
8   stars        150346 non-null float64  
9   review_count 150346 non-null int64  
10  is_open      150346 non-null int64  
11  attributes   136602 non-null object  
12  categories   150243 non-null object  
13  hours        127123 non-null object  
dtypes: float64(3), int64(2), object(9)
```



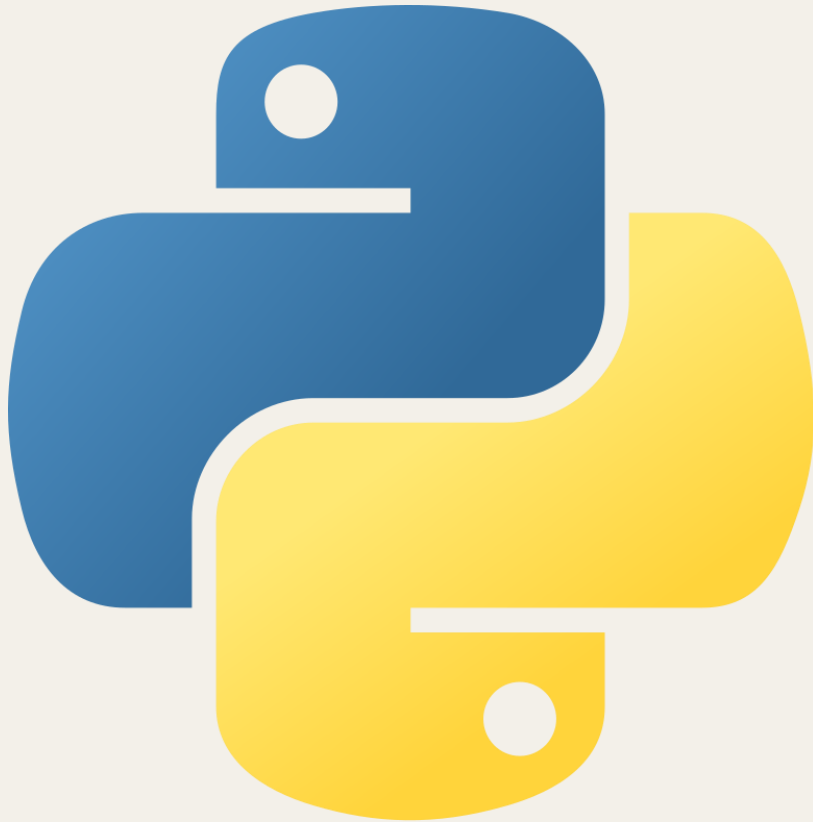
# Objectives

- + Evaluate the performances of different text representations considering the classification task.
- + Perform topic modelling techniques, to find some of the most discussed topics in the Yelp reviews.
- + Predicting the review stars considering the text of the reviews with classification.





# Data Preparation



```
df_ita.info()
```

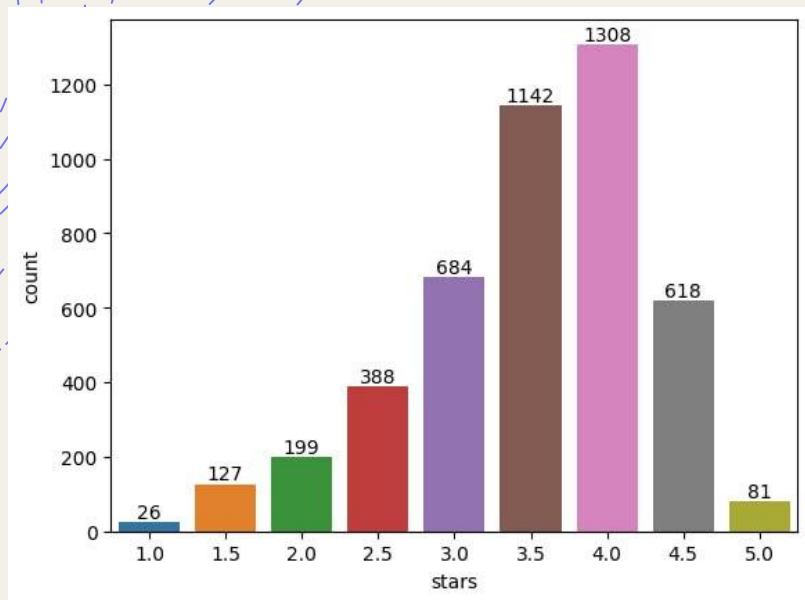
```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 439358 entries, 3306 to 6989409  
Data columns (total 15 columns):  
#   Column          Non-Null Count  Dtype  
---  ---  
0   review_id       439358 non-null object  
1   user_id         439358 non-null object  
2   business_id     439358 non-null object  
3   review_stars    439358 non-null int64  
4   useful          439358 non-null int64  
5   funny           439358 non-null int64  
6   cool            439358 non-null int64  
7   text            439358 non-null object  
8   date            439358 non-null datetime64[ns]  
9   name            439358 non-null object  
10  city            439358 non-null object  
11  stars           439358 non-null float64  
12  review_count    439358 non-null int64  
13  attributes      439063 non-null object  
14  categories      439358 non-null object
```

# Text Preprocessing

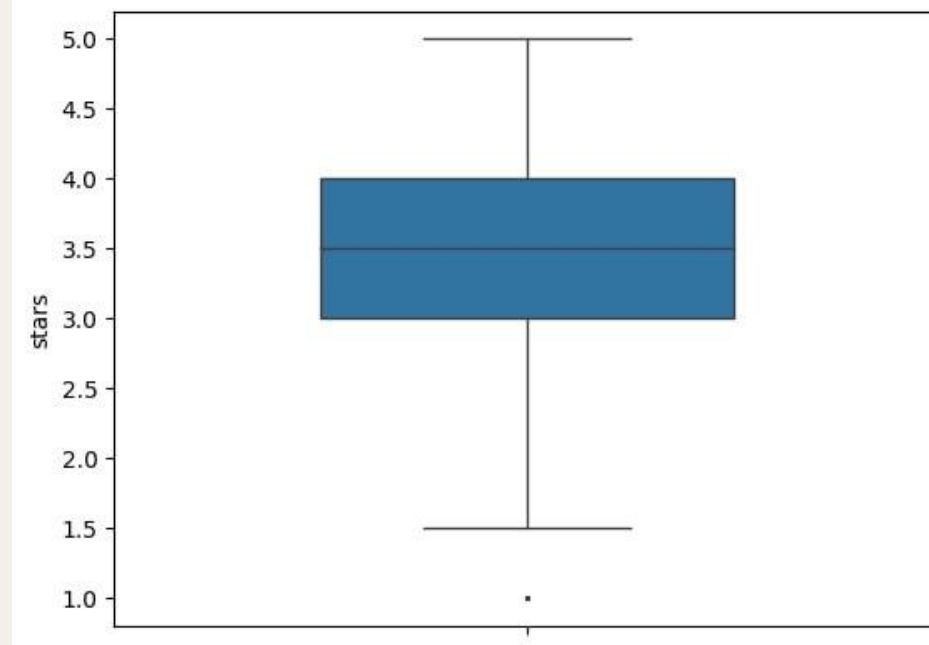
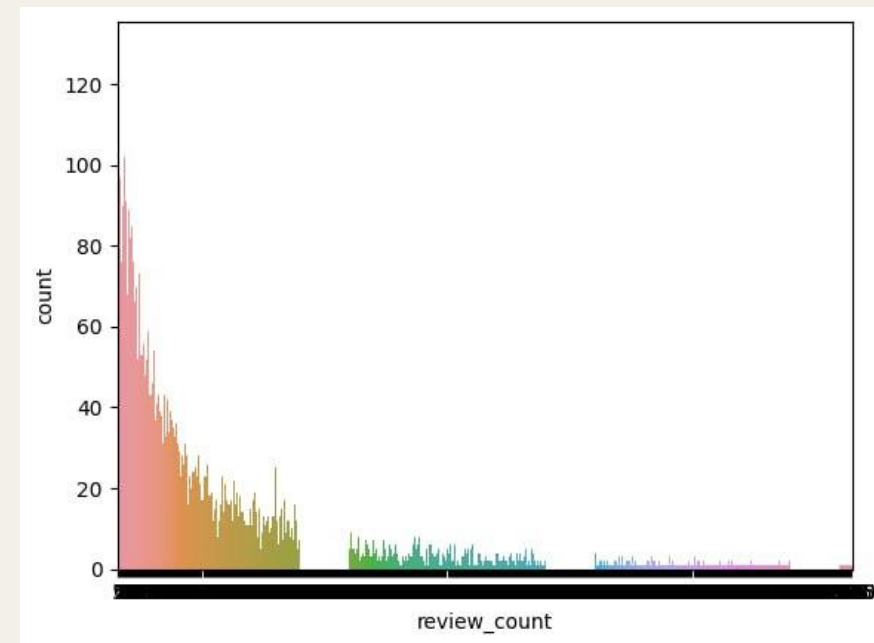
- + Case folding
- + No numbers
- + No empty lines
- + No URL/links
- + No whitespace
- + No emoji
- + No repeated characters
- + No punctuation
- + Tokenization and stopwords removing
- + Lemmetization



# Exploratory Analysis

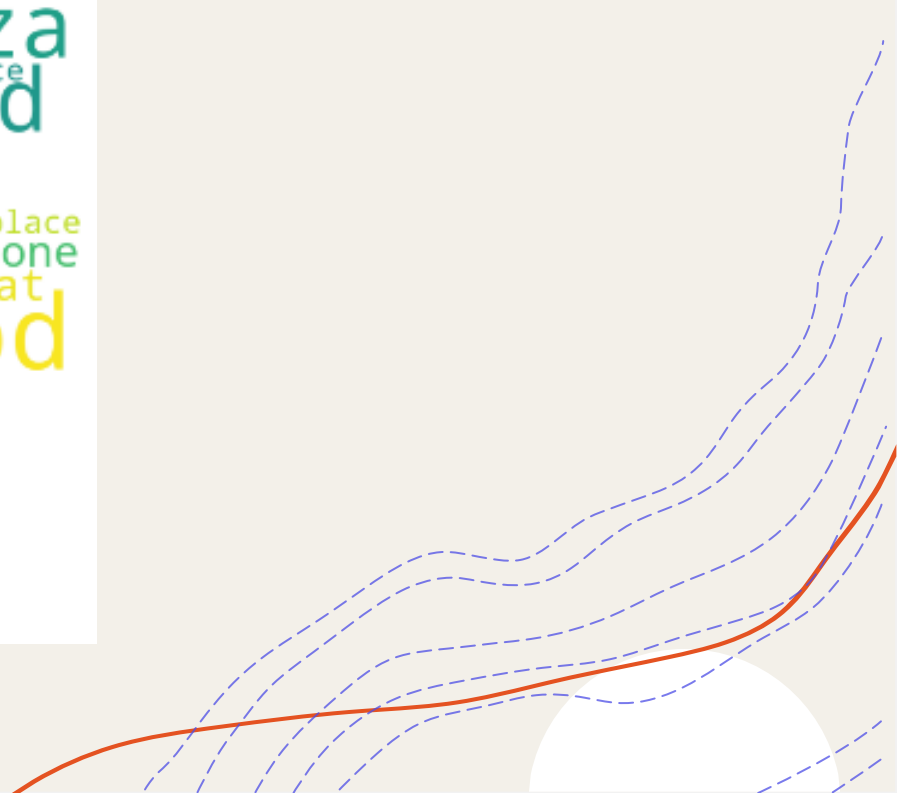


variables	mean	min	max	std	25%	50%	75%
Stars	3.51	1.0	5.0	0.78	3.0	3.5	4.0
Review count	92.39	5	4250	156.77	18	44	108



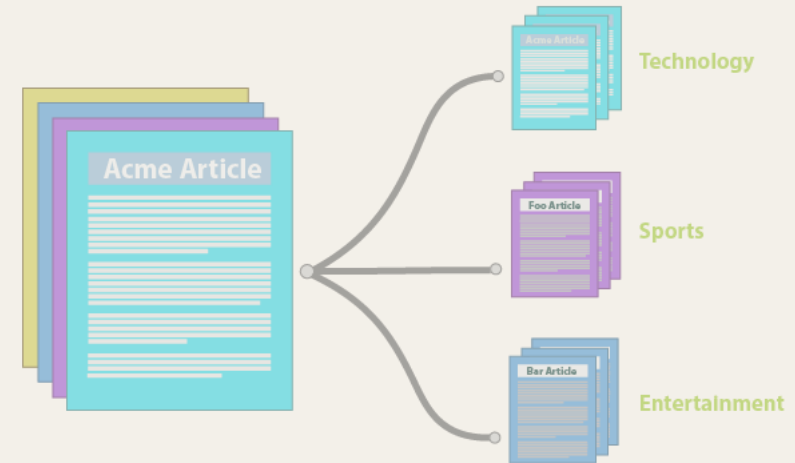


# Topic



# Text Classification

Class method countvec torizer()	Accuracy Score	F1 Score	Recall Score
DT	0.77	0.82	0.83
SVM	0.88	0.92	0.93
RF	0.86	0.90	0.96



Class method TFTDF	Accuracy Score	F1 Score	Recall Score
DT	0.77	0.84	0.84
SVM	0.90	0.92	0.95
RF	0.87	0.91	0.96



# Results and conclusion

- + Topic Modeling: big presence of words related to Italian restaurants and food ('pizza', 'pasta'...); in general, positive words and feedbacks.
- + Text Classification: among the two representation method and the three classification models, Support Vector Machine with TFIDF were the best; Decision Trees classifier was the worst (in particular with countvectorizer()).



The background is a light beige color. In the top-left corner, there is a white circle partially cut off by the edge, with several blue dashed wavy lines flowing downwards and to the right from it. In the bottom-right corner, there is another white circle partially cut off, with several blue dashed wavy lines flowing upwards and to the left from it. A solid orange line also flows from the bottom-left towards the bottom-right circle.

Thanks for your attention!