University of Southampton

Foundation of Data Science
Group 4

Yelp Challenge

Nicola Vitale
Antonios Andronis
Adisorn J.
Hujiang Lin
Nikolaos Perrakis

Supervisor:
prof. Elena Simperl

The Challenge

- What is Yelp?
- Why Yelp?

Site: http://www.yelp.com/dataset_challenge

The Yelp dataset

Yelp provides 5 datasets. We used 3 of them.

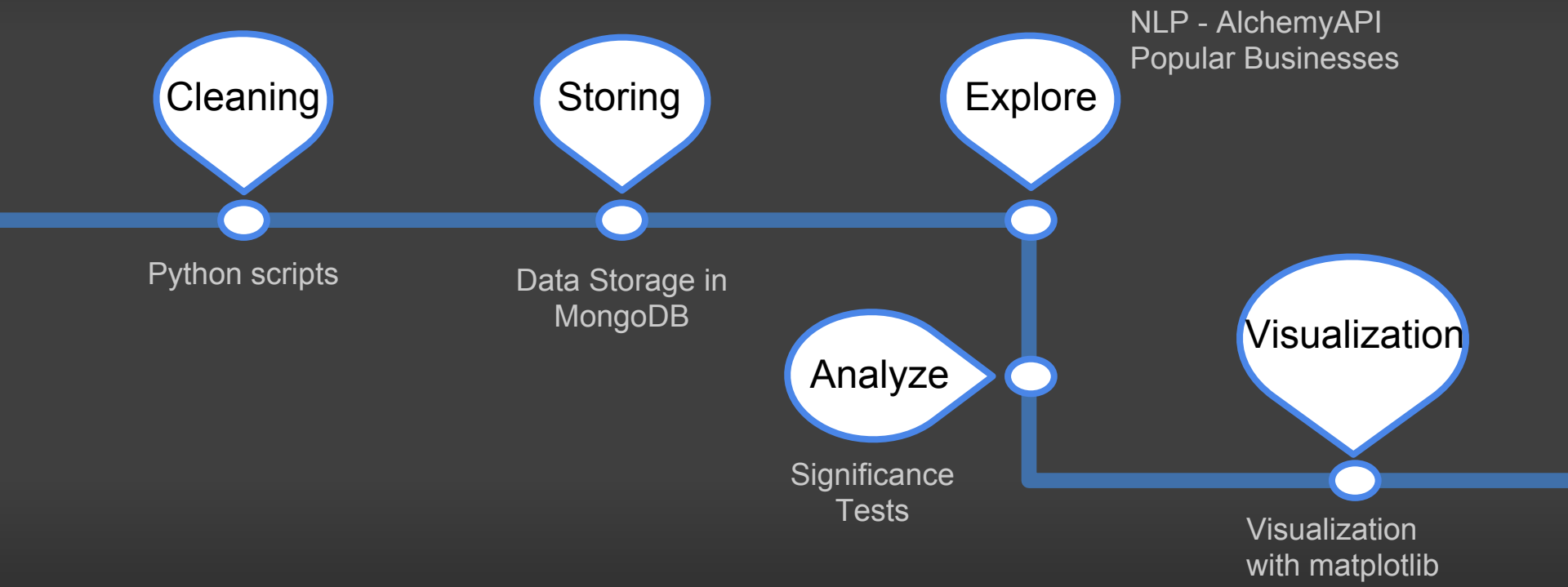
- User 366,715 records
- Review 1,569,264 records
- Business 61,184 records

Our Objectives

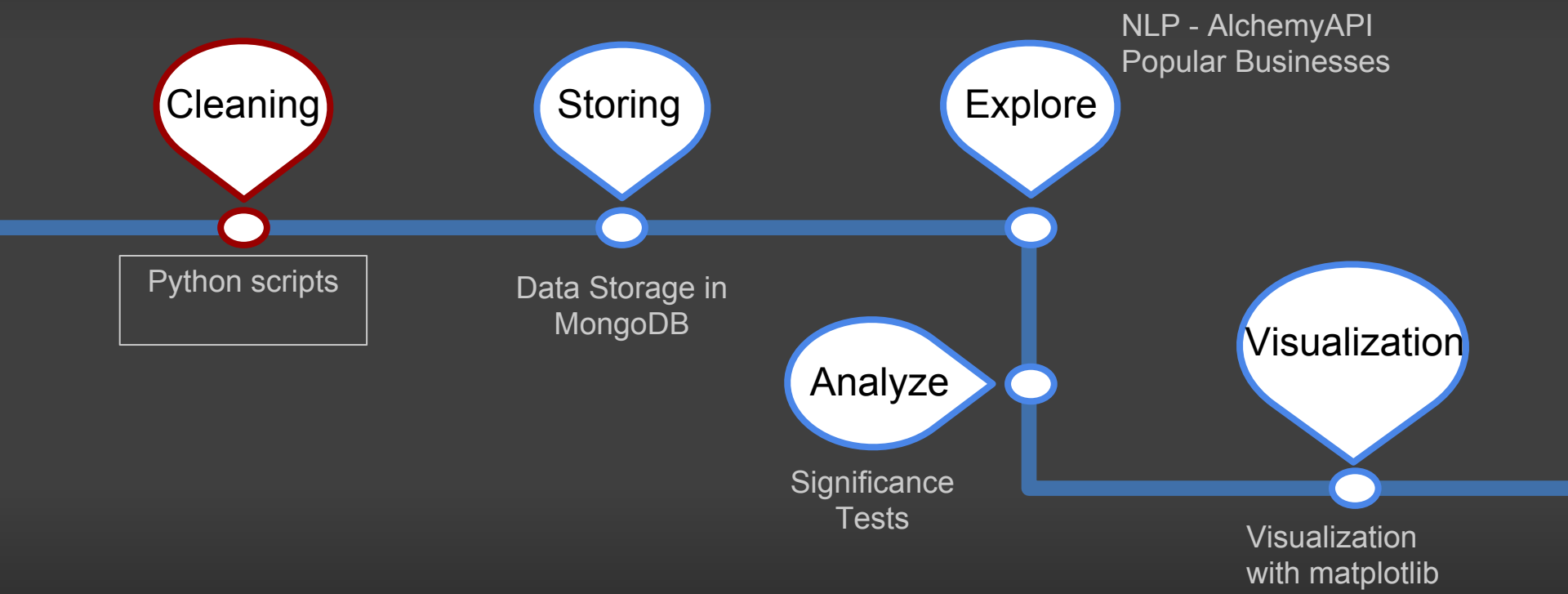
Use Yelp reviews to identify and predict business issues

- Identify strong negative reviews based on significance of the sentiment
- Quantify user contribution
- Identify negative review clusters around strong negative reviews
- Identify issues that generate increased negative reviews

The Pipeline



The Pipeline



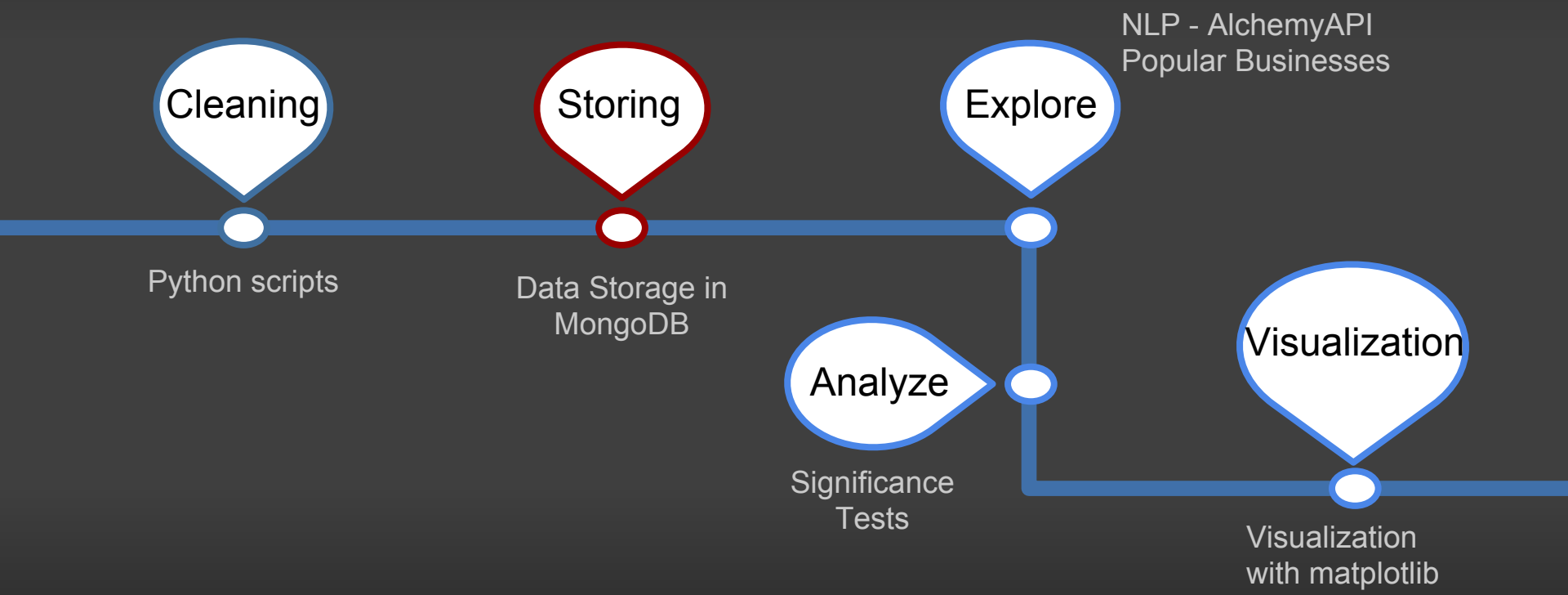
Data Cleaning

The dataset from Yelp is quite clean. However, we found some problematic values that affect NLP analysis.

- Having new line character(‘\n’) in some review texts:
  These strings are hard to handle and can not work well with NLP analysis, so we remove these.

- Having dashes in some review texts or even worse that the text contains only dashes like “----------”:
  It causes unsupported-text-language error when sending it to NLP.
  When the text has some dashes, we remove them out of the text.
  If the text itself is only dashes, we ignore the text.

The Pipeline



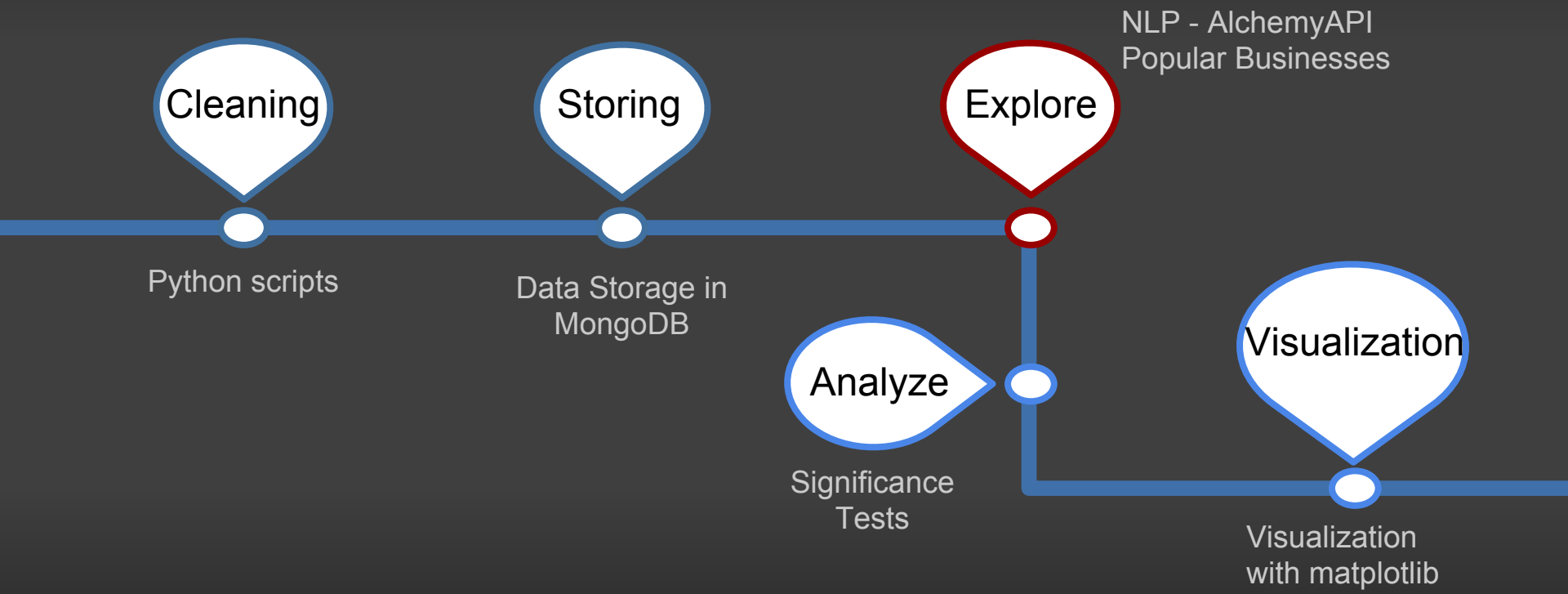
Data Storage

- **Dataset's size ~ 2.5GB, 5 datasets in JSON format**

  mongoDB version 3.0.x

  - Schema-less database -> data structure is quite flexible
  - Can contain large volume dataset BUT still has good performance
    -> NO significant variation of execution time
  - Easy to import and export, which is good for team's collaboration
  - Get started using it with just a few settings

The Pipeline



Alchemy API

AchelmyAPI is a part of IBM Watson.

IBM Watson is a technology platform that uses Natural Language Processing and Machine Laerning to reveal insights from large amounts of unstructured data.

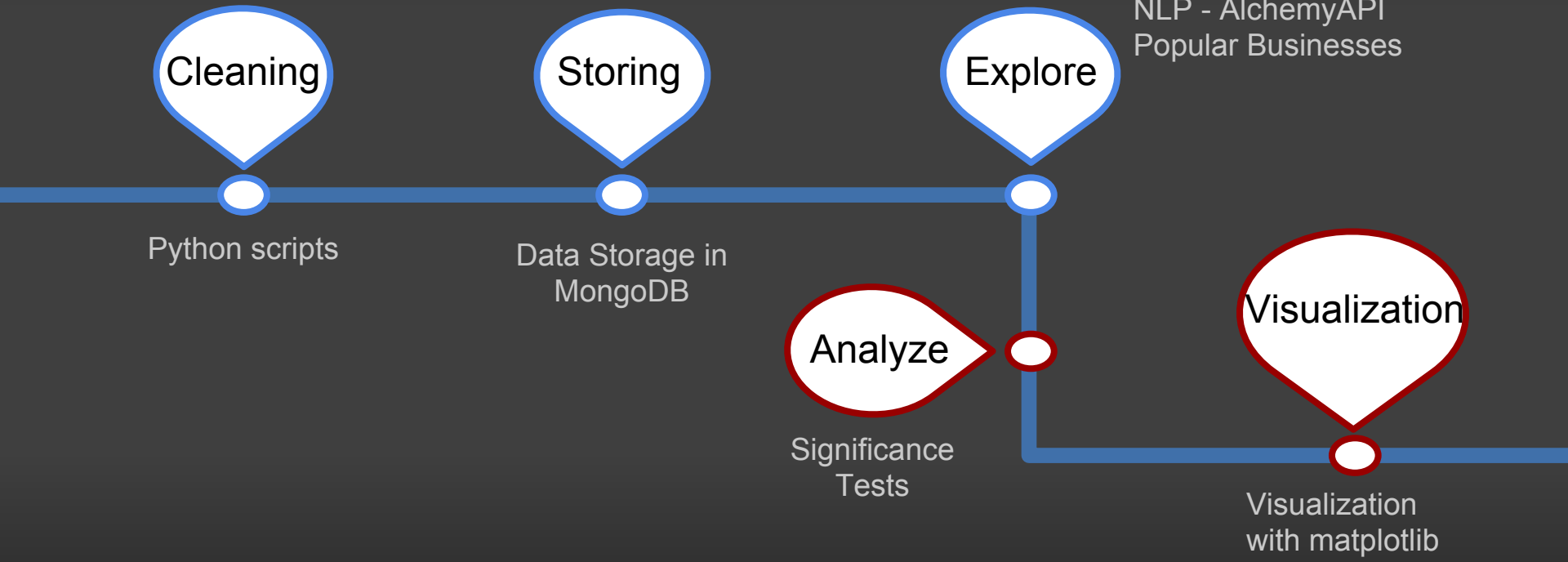
- Why choosing it?

  It provides deeper information such as keyword entity, taxonomy, keyword sentiment. It has many versions of API language like Python, Java, etc.

- Alternative NLP candidates

  Data Calais, CoreNLP(Stanford), NLTK(Python)

The Pipeline



Exploratory Data Analysis

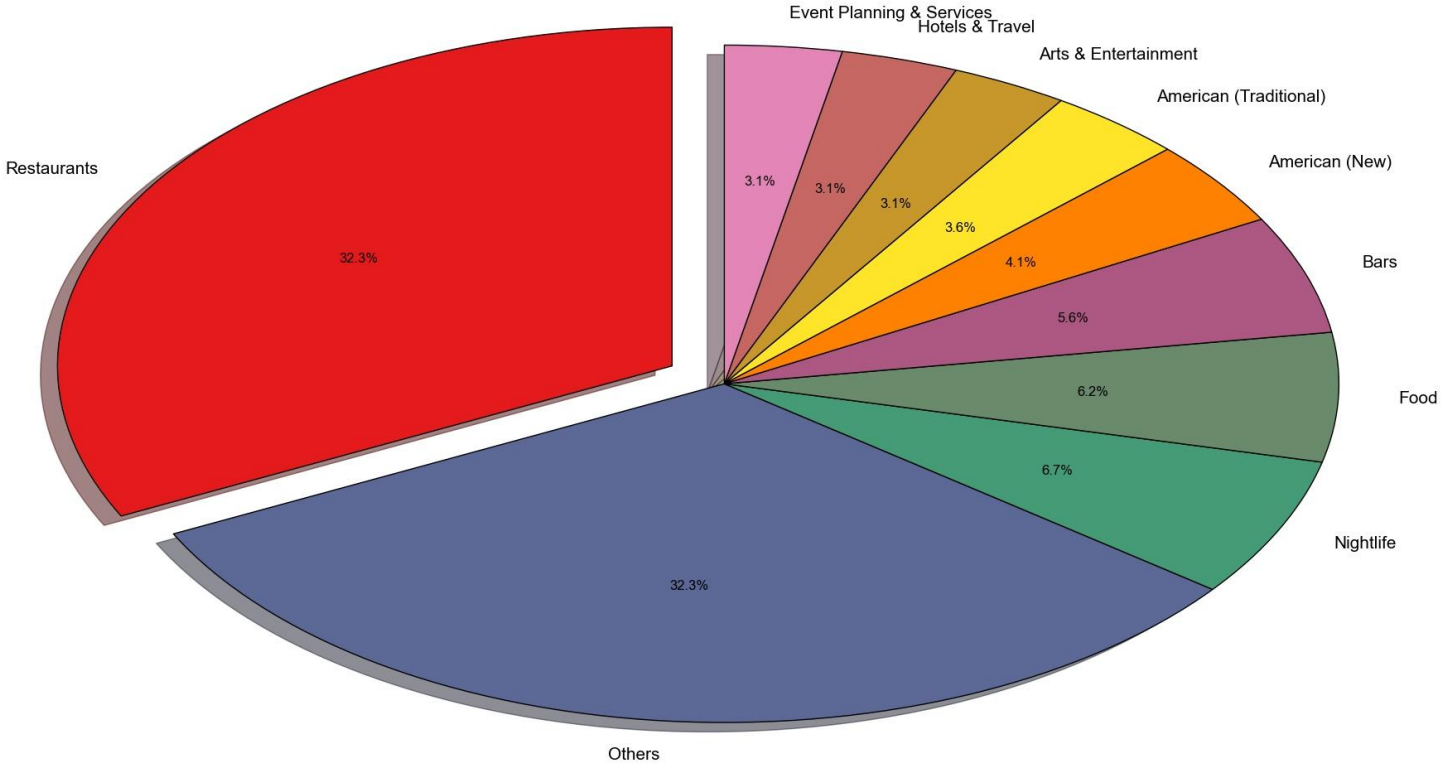
Tools

Python version 2.7 including build-in libraries: Pymongo, NumPy, Matplotlib.
Achelmy API as a NLP.

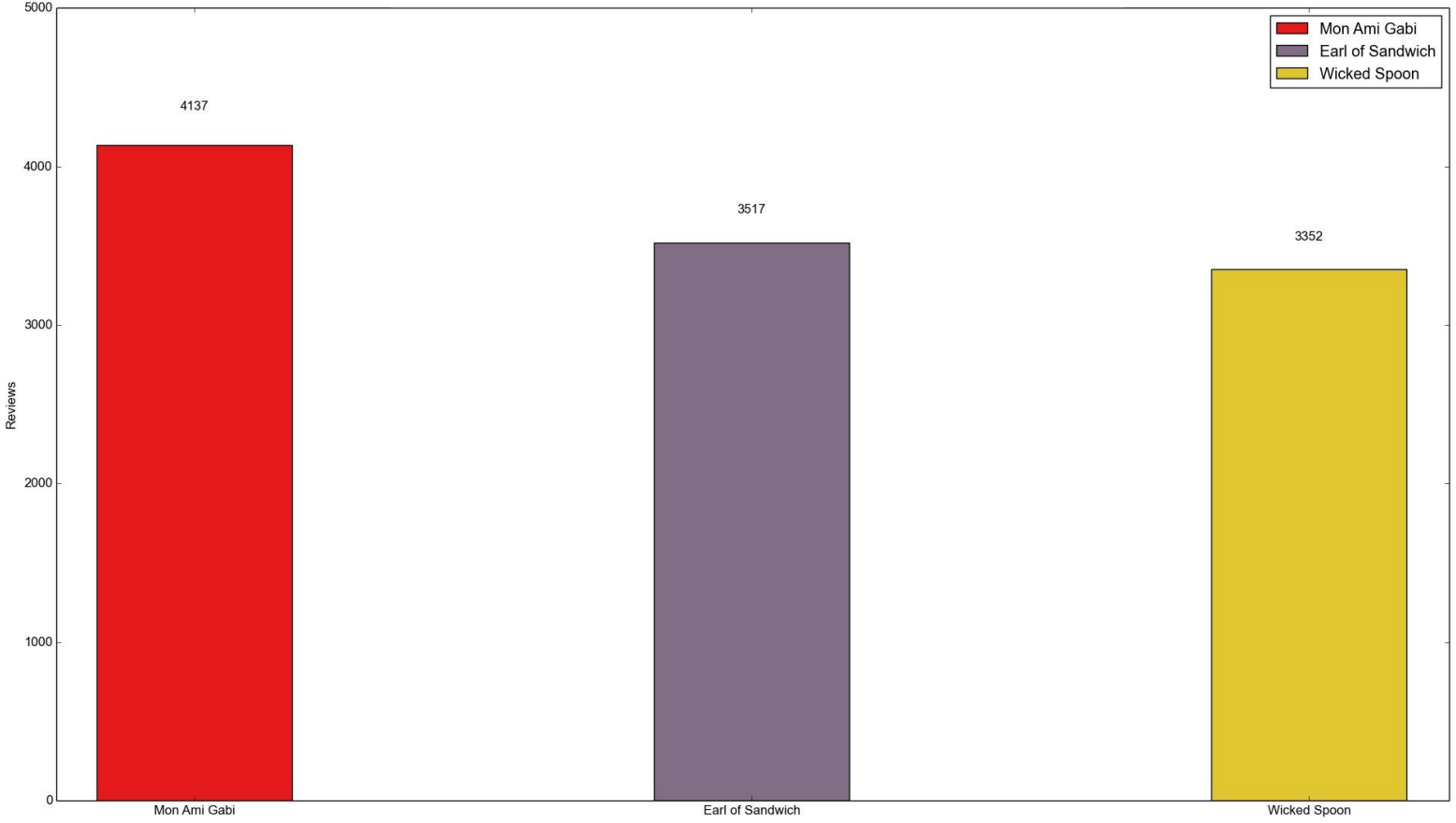
Techniques

Query the database to see the data strucure:
Time Range: December 2009 - January 2015
Business categories popularity.
Business Popularity
Reviews sentiment score Run Chart.

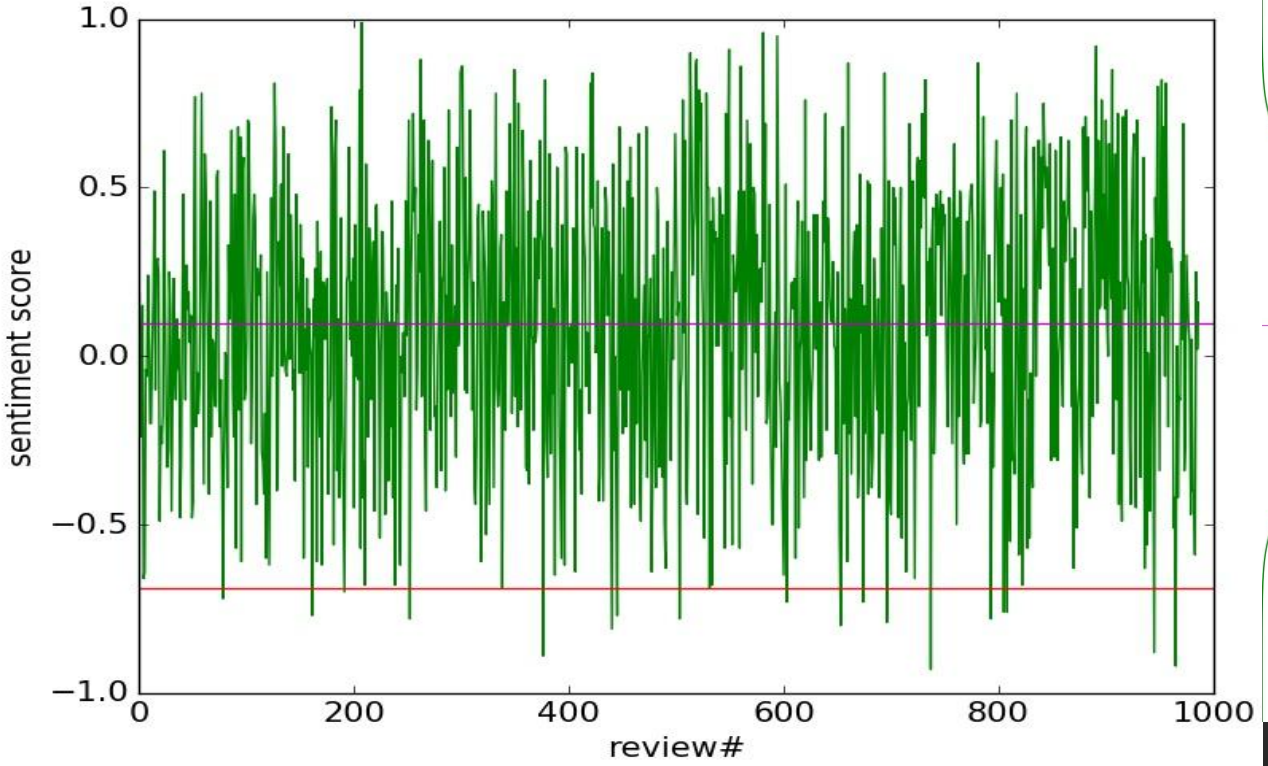
Top business categories



Most popular business



Statistical Analysis



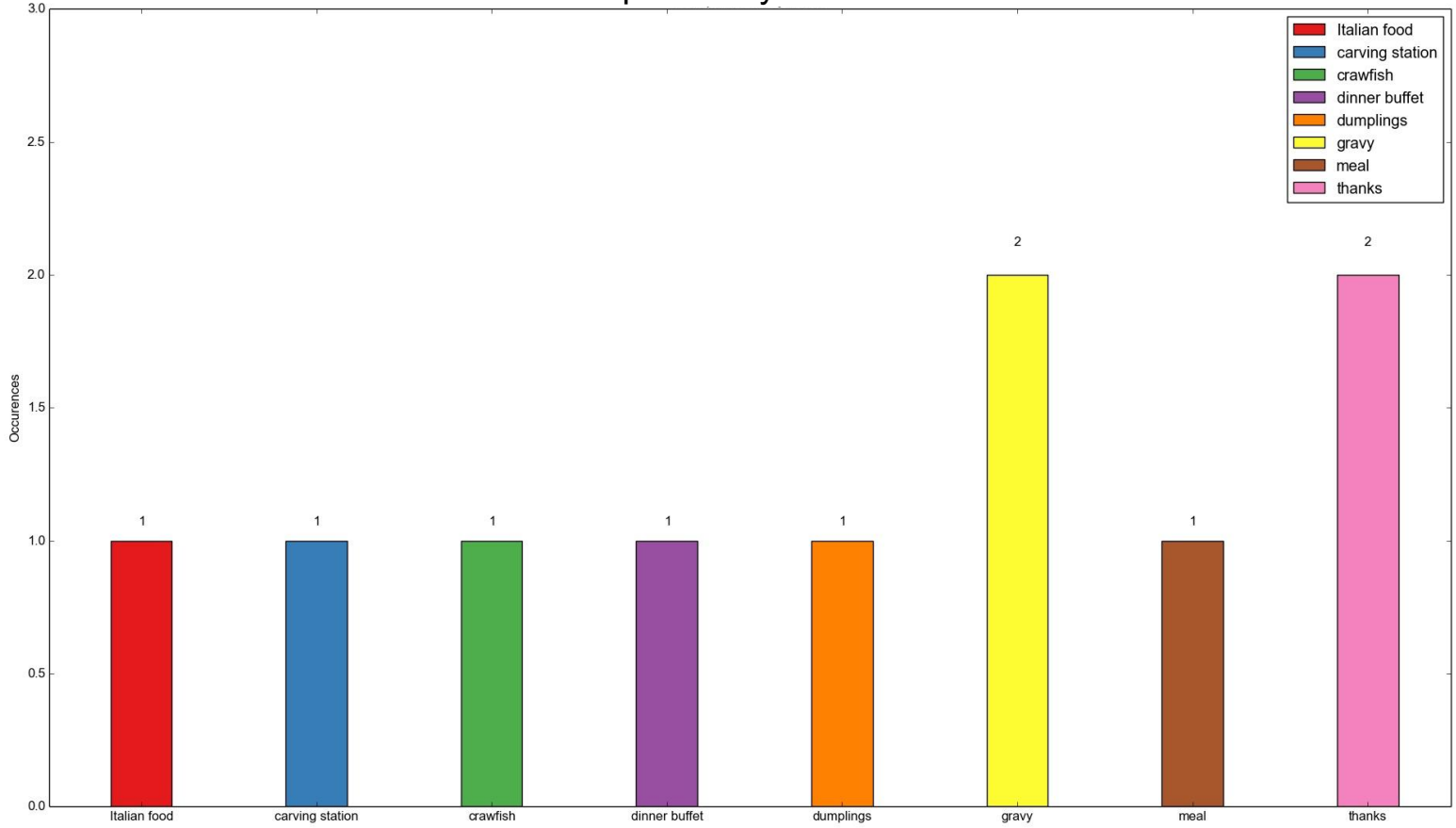
Sentiment Score Time Series for a business.

Normally Distributed Sentiment Score.

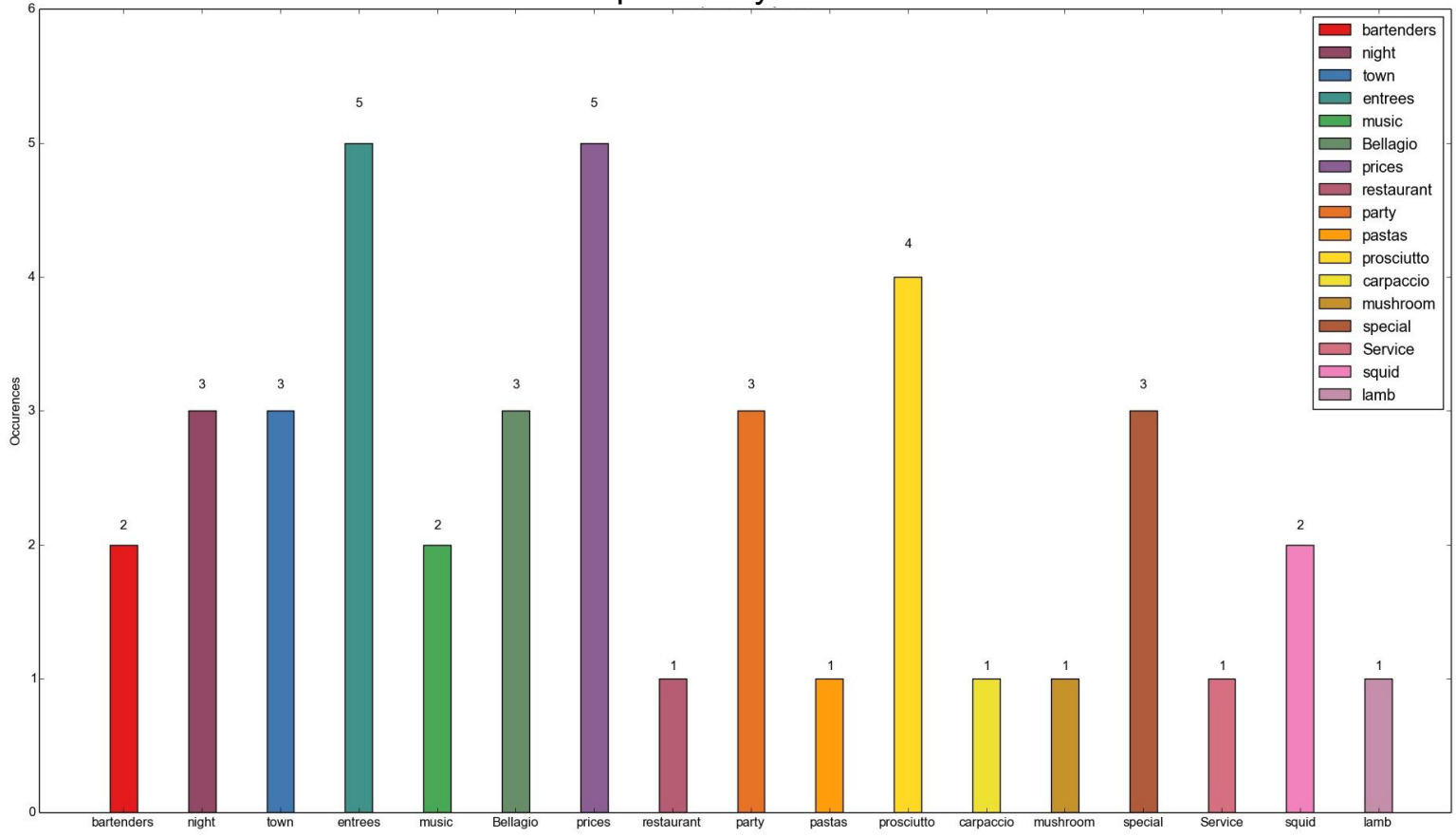
Significance Testing

- Use Normality test to identify non random review cluster that correspond to business problems
- Check Significance of their keyword sentiment scores against overall mean and standard deviation.
- Keep only negative significant keywords

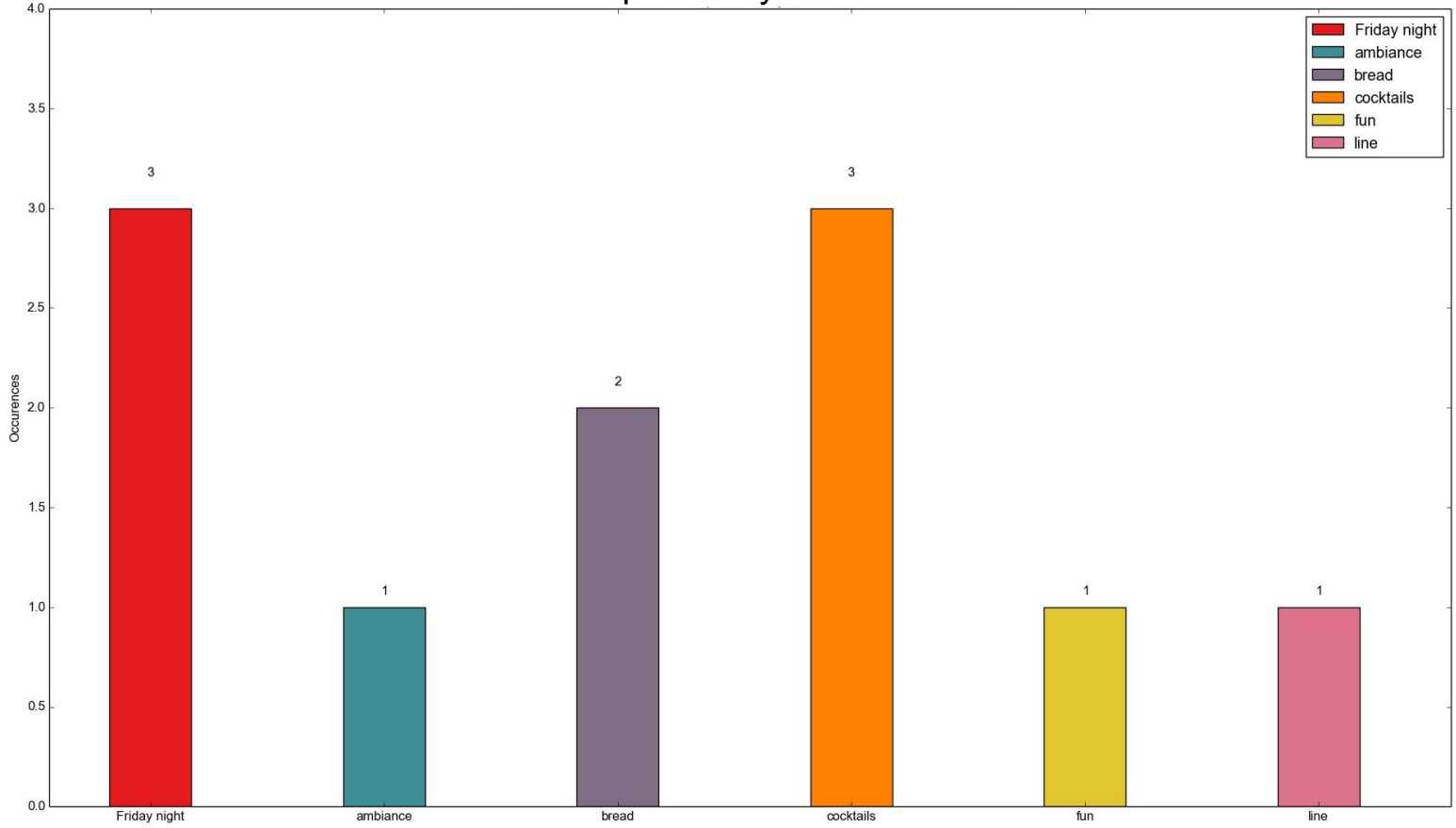
Complaint Keywords



Complaint Keywords



Complaint Keywords



Conclusions

- We can succesfully identify business issues
- We can describe them
- We can use the last 20 reviews to point out current issues