Analysis of TripAdvisor Data for Restaurants in Thessaloniki

Contents

Introduction

- Web scraping and spefically Selenium were used for the extraction of the necessary data from the Tripadvisor site.
- Our team was tasked to analyse the reviews written for all the restaurants listed in Tripadvisor for the city of Thessaloniki.

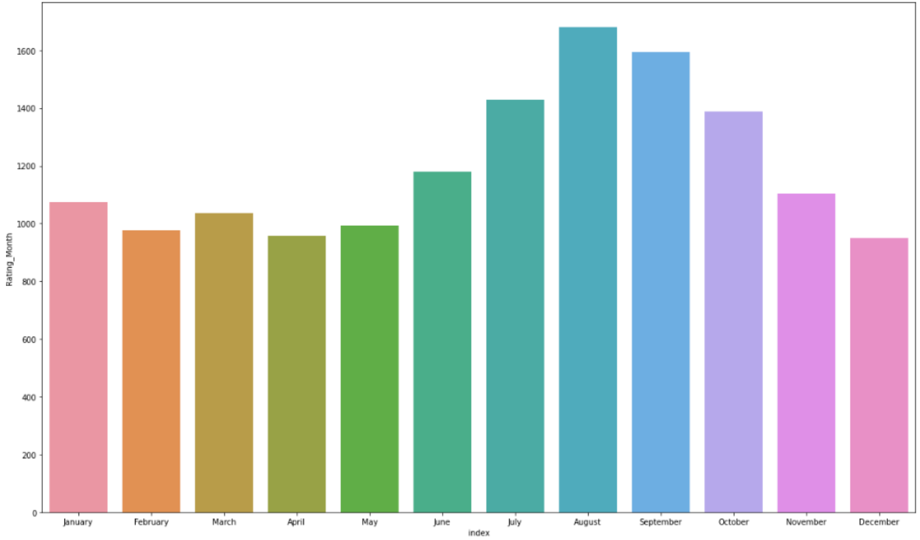
The informations we decided to collect belong to the following categories.

- Username of the reviewer
- Name of the business
- Review Date
- Visit date
- Title of the review
- Text of the review
- Rating of the review
- Reviewer’s Age
- Reviewer’s Gender
- Reviewer’s Location
- Reviews distribution

Number of Monthly Reviews

Before analysing further the reviews, an interesting observation

★ Most of the reviews as seen, were made during the summer indicating that a lot of reviewers were probably tourists.

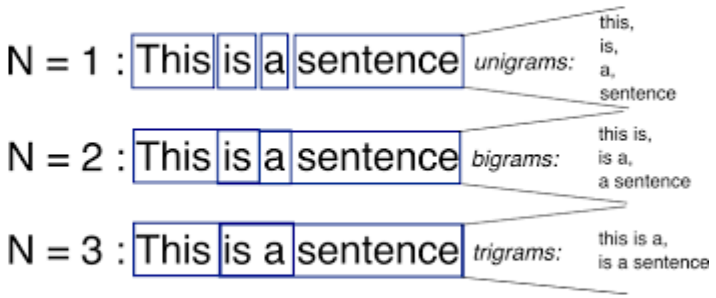


Most Common Words, Bi-grams, Tri-grams

- Bi-grams and Tri-grams are correspondingly series of 2 or 3 words.

Steps:

1. FreqDist function of NLTK was used. Takes as input the list of texts and the ngram type you are interested in.
    2. Create a dictionary with the information concerning each type of n-grams.
    3. Create a dataframe for each type of n-gram storing the word(s) and their frequencies.
    4. Use word cloud for the presentation of the results.



Most Common uni-grams

- Food
- Good
- Great
- Place
- Servic
- Restaur

Little amount of context.

Most Common bi-grams

- good_Food
- one_best
- service_good
- staff_friendli
- reason_price

More context. Generally giving a good feeling about the overall quality of the restaurant.

Most Common tri-grams

- food_great_servic
- one_best_restaur
- good_valu_money
- definit_come_back
- food_good_price

Even more context. Different topics are distinguishable. Price, overall experience, thoughts.

One-star Reviews

The same 3 wordclouds were created but this time only taking into consideration 1 star or 5 star reviews respectively.

- Bad_service , recommend_place_anyone (didn’t quite catch the content right) , never_come_back

Five-Star Reviews

- friendly_staff , great_food , definetely_come_back , would_highly_recommend

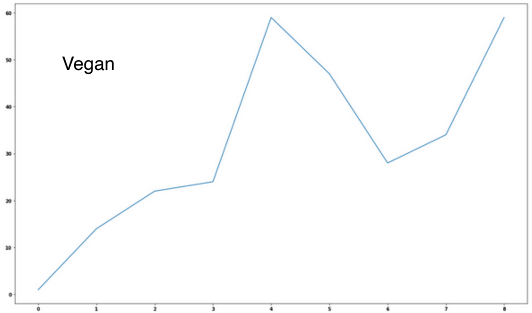
Fastest Growing and Shrinking Words

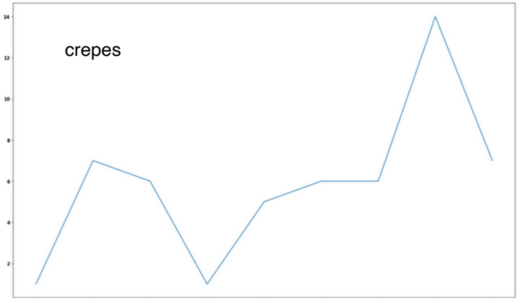
Top 10 growing and shrinking words based on the sum of the percentage of differences

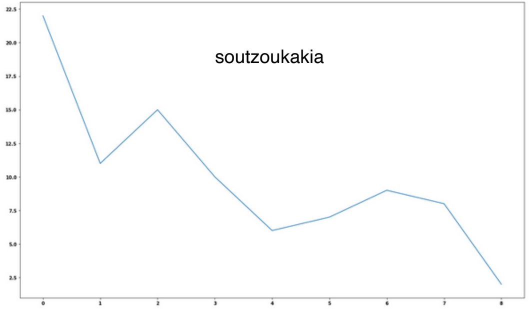
specialti	-0.978968
aristotel	-1.032368
meze	-1.037670
room	-1.061111
famou	-1.067434
window	-1.084149
werent	-1.151007
soutzoukakia	-1.278427
youll	-1.520704
histori	-1.865385

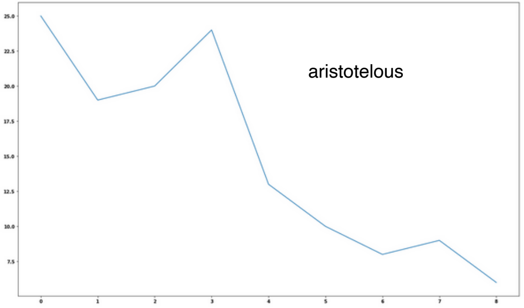
vegan	15.462606
apolog	12.472222
decad	11.555556
appetis	11.352453
thin	11.242424
becam	10.955556
warmli	10.750000
request	10.473485
carpaccio	10.164502
crepe	10.057143

Fastest Growing and Shrinking Words





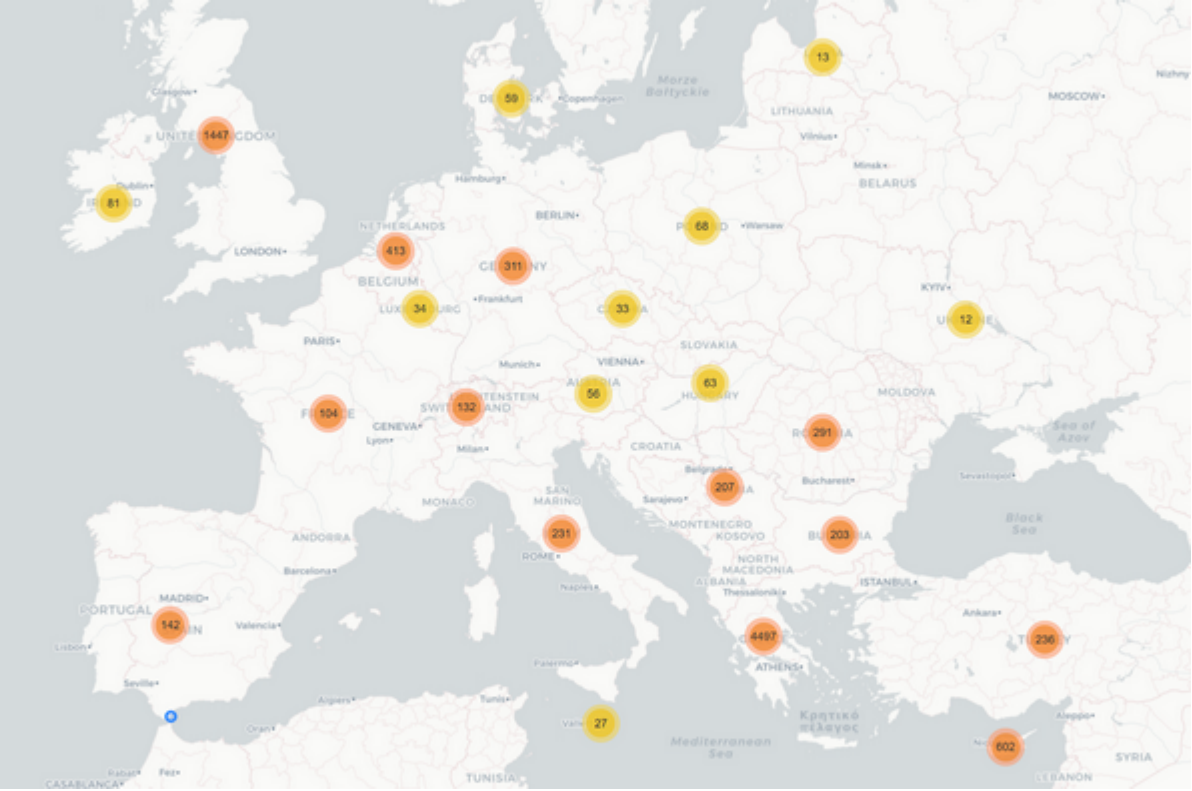




Map of Locations

The world map where users reviewed from

Map of Locations

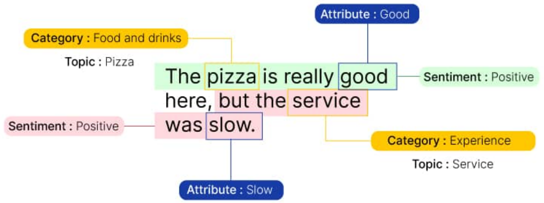


Here we can see the europe region

and the distribution of the reviewers

in the said region

Sentiment Analysis



Sentiment analysis is important as it helps with the extraction, scoring, classification and visualization of the feelings and opinions the customers display for the business.

Simplest method to get the general feeling of the customers

- Get a look at the total score of the restaurant! (Can be misleading)

Calculated the polarity score:

- Per review
- Per restaurant

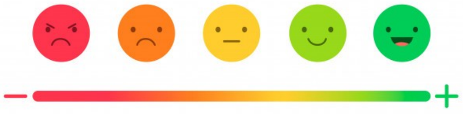
Basic Visualizations using most common words

- Positive
- Negative

```
Business_Name
MINIMAL ALL DAY COFFEE HOUSE    50.0
MIMIS                           50.0
MILONGA MEZE CLUB               50.0
MIKRI MARMITA                   50.0
MEZEDOPOLEIO ZAIR 79            50.0
```

```
Business_Name
LOS HERMANOS                         10.000000
ASIAN HOUSE TOUMPA                   10.000000
MODIANO RETSINADIKO                  10.000000
TAVERNA SOTIRIS                      10.000000
TOMMY COFFEE HOUSE & EATERY          10.000000
```

Sentiment Analysis



The polarity score is a float within the range [-1.0, 1.0] .

Per Review

Negative Polarity Score

654	eat worst thing vacat seafood frozen wait 15 h..
8715	rudenessaw servic treat u like anim credit car..
11432	eat hungri believ one worst place ever eat nev..
5011	dirti place terribl servic waiter behav improp..
12822	worst pizza ever eaten ask crispi pepperoni go..

654	-1.0
8715	-1.0
11432	-1.0
5011	-1.0
12822	-1.0

Positive Polarity Score

9771	visit takada even read tripadvisor review disa..
13562	that best gyro ever life visit soon mani choic..
890	realli best valu money steak hous town bavett ..
3076	restaur feel comfort eat perfect foodth person..
4508	impress qualiti food tri mani differ dish reco..

9771	1.0
13562	1.0
890	1.0
3076	1.0
4508	1.0

Sentiment Analysis

Per Restaurant

- No maximum or minimum polarity score
- Top restaurants could be described as very good according to the score
- Negative ~ 0 polarity score means the restaurants create a neutral to mildly negative emotion

Positive Polarity Score

Business_Name	
KAFENEION DIEFNES	0.700000
ANOPOLIS	0.700000
LARRY'S COFFEE-WINE-DRINKS	0.700000
CHICKEN BAR	0.700000
AVANTI PIZZA	0.716667
ATHENA BAKERY CAFE	0.750000
TO STEKI AFYTOS	0.750000
MER KA BA	0.800000
R2 STORE	0.800000
MUZIK BAR CAFE	1.000000

Negative Polarity Score

Business_Name	
MARMELADA	-0.291667
MYSTIKOS DIPNOS	-0.250000
MARCHE BISTRO	-0.250000
OVELISTIRIO TO PAGGAIO	-0.212727
PEINIRLADIKO "TO ROLOI"	-0.125000
OMBRELLO	-0.125000
10 TO KALO	-0.124444
RALENTATO	-0.100000
KOTOPARADOSI	-0.091667
DESKATI	-0.089315

Sentiment Analysis

But wait!

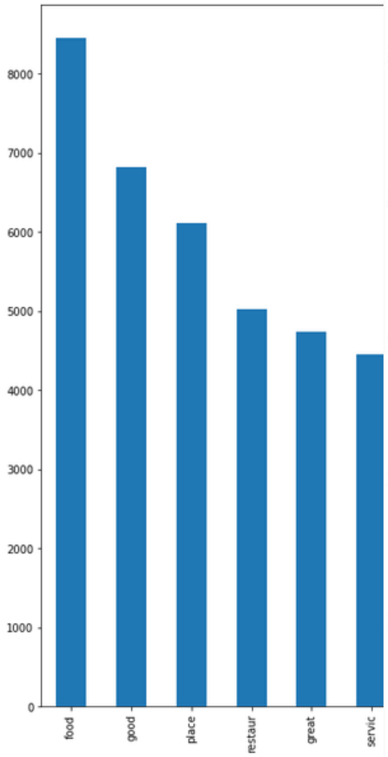
- We need to consider the number of reviews.
- One negative or positive review can greatly impact the outcome of the polarity score, if there is only a handful amount of reviews to analyze.
- Now we don't see that big of a difference!

Give it a go!

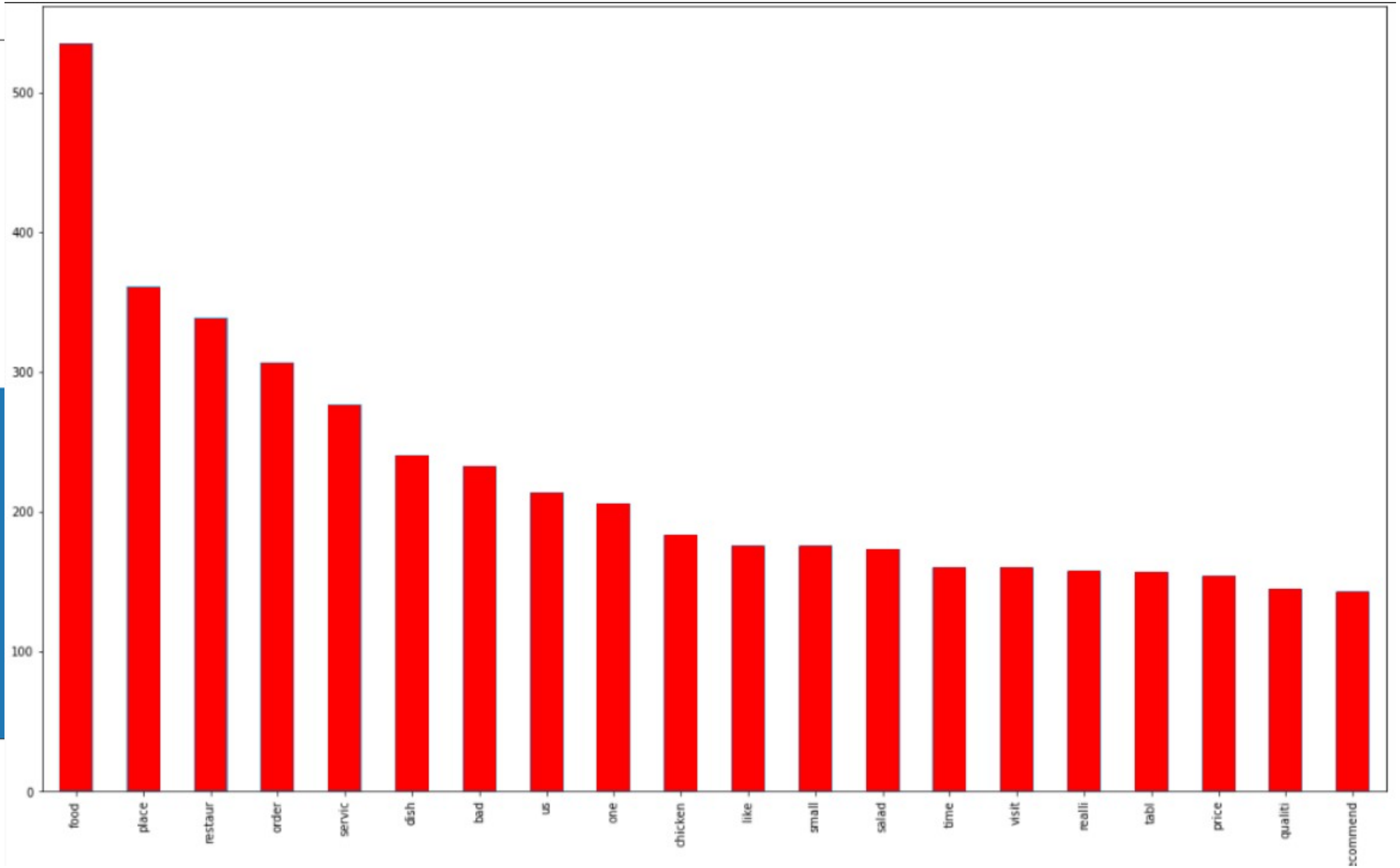
Maybe once in a while…

Business_Name	polarity_score	Count_of_reviews
BOUGATSA BANTIS	0.420580	75.0
MENU ME...NOU	0.422440	33.0
THE BACKROOM	0.435356	33.0
VOGATSIKOU 3	0.461843	53.0
DA VINCI	0.467343	26.0
EVORA EATERY	0.467901	47.0
ON THE ROAD	0.471508	23.0
FALAFEL HOUSE	0.475996	22.0
BEER O' CLOCK	0.494002	21.0
LA PASTERIA	0.510186	21.0

Business_Name	polarity_score	Count_of_reviews
NARGIS INDIAN RESTAURANT	0.114171	52.0
OLIVE OIL & OREGANO	0.207828	26.0
SOLOMONIDIS FILETO	0.209954	24.0
NOODLE BAR	0.217656	56.0
H KOUZINA TIS AGLAIAS	0.224221	25.0
ESTRELLA	0.230627	122.0
OUZERI TSINARI	0.236954	50.0
TO MIKRAKI	0.238385	30.0
NEA FOLIA	0.239944	98.0
OUZO STON PINAKA	0.244421	63.0



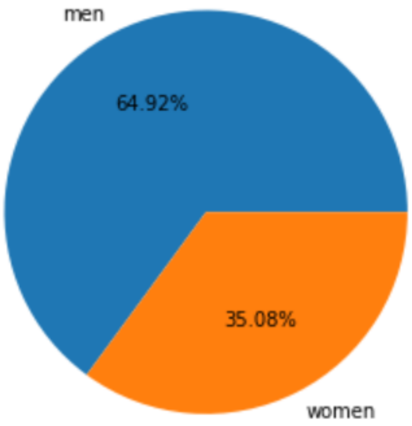
0
1000
2000
3000
4000
5000
6000
7000
8000
food
good
place
restaur
great
servic



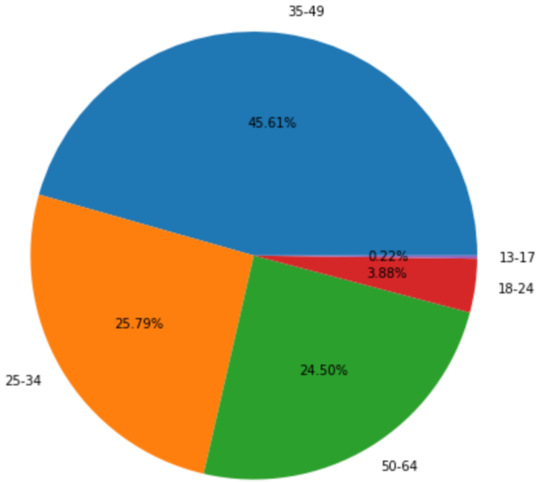
0
100
200
300
400
500
food
place
restaur
order
servic
dish
bad
us
one
chicken
like
small
salad
time
visit
realli
tabl
price
qualiti

User Profiling

- Most reviewers are men



- Most reviewers are middle-aged
- Few minors



User Profiling

Features:

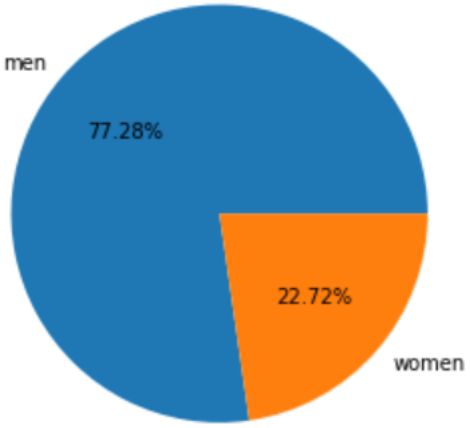
- Review Text
  - TF-IDF
  - Polarity
- Review Distribution
- Score

Dataset

- About ⅓ of data was labeled

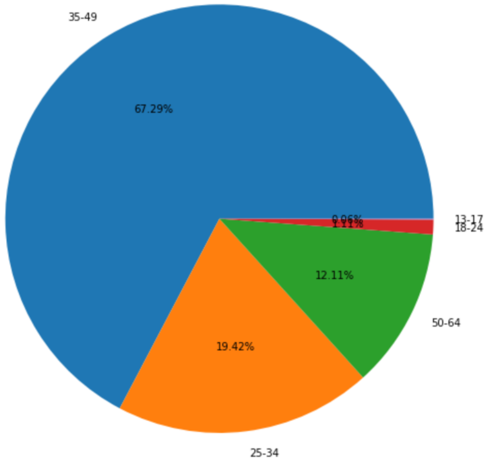
Random Forest

Accuracy = 0.70



Logistic Regression

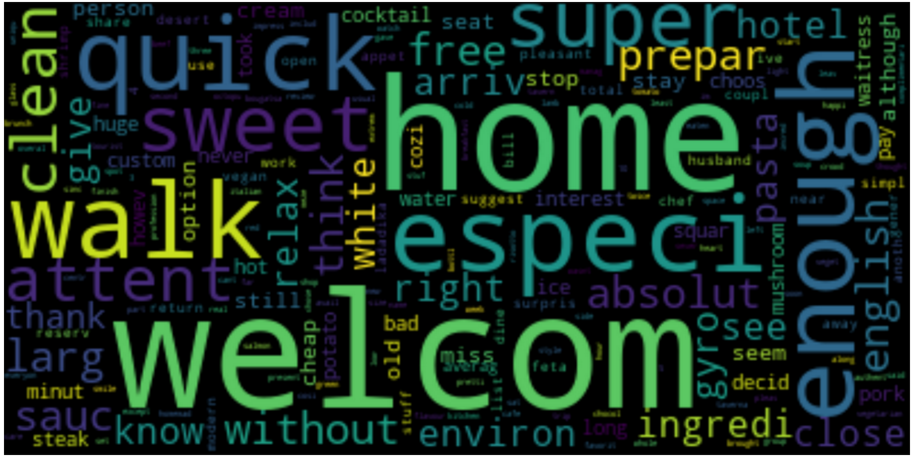
Accuracy = 0.47



User Profiling - Gender

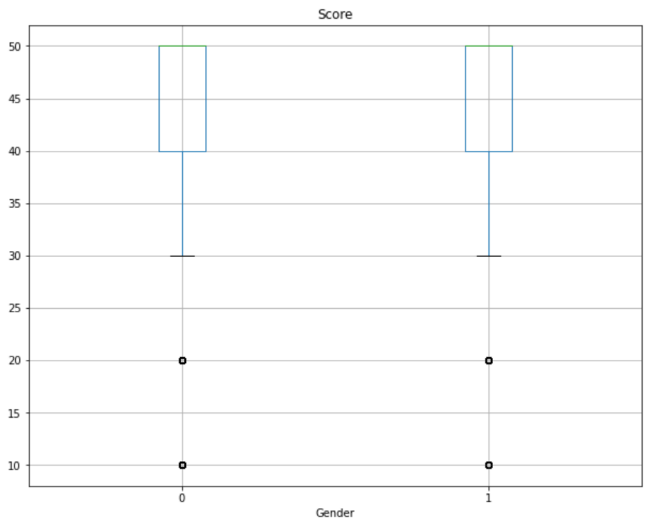
- Wordcloud for Men

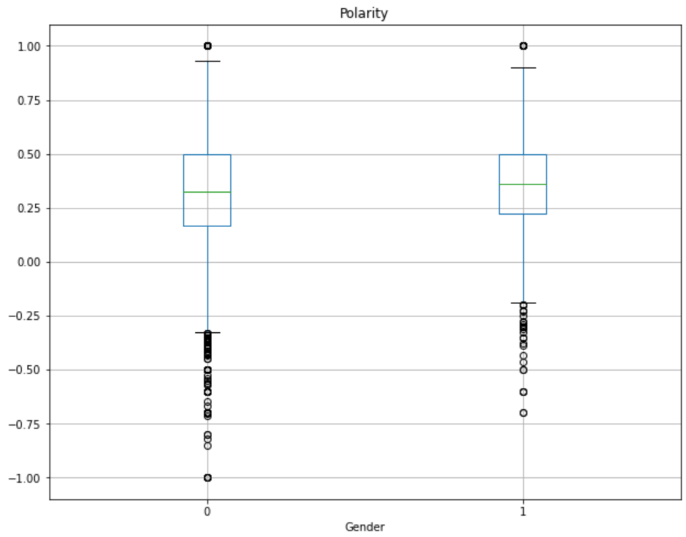
- Wordcloud for Women



User Profiling - Gender

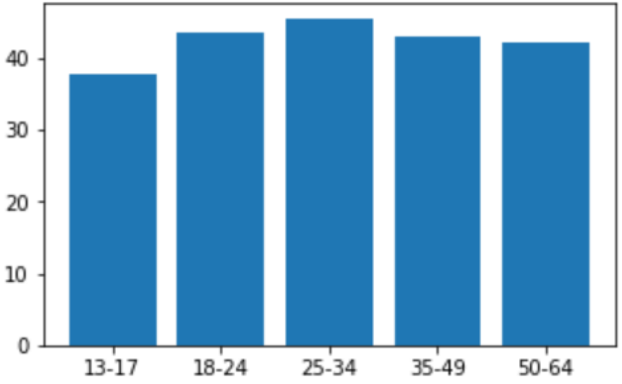
Comparison in Score and Polarity between men and women

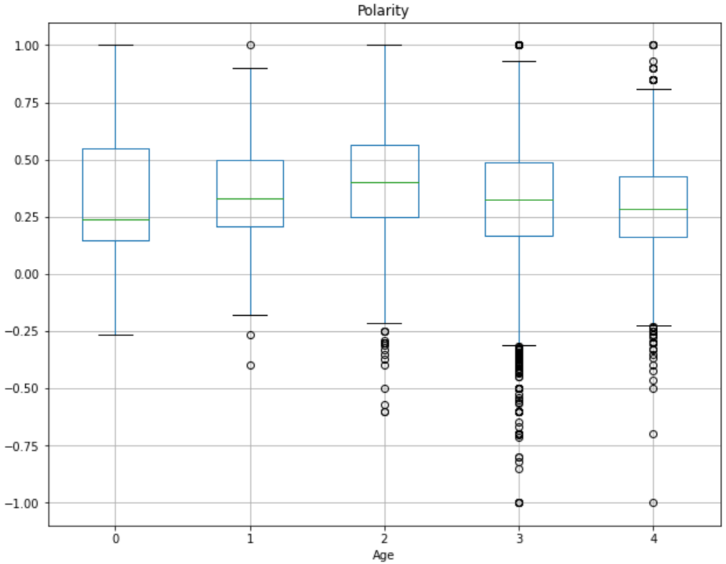




User Profiling - Age

Comparison in mean Score and Polarity between each age group





Thank you!
Any Questions?