NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "HSE and University of London Double Degree Programme in Data Science and Business Analytics"

Software project report
"Sentiment Analysis based on Glamping Reviews from TripAdvisors: Developing a Decision-Making Assistant"

Student: Prokhorova Anastasiia

Supervisor: Chuvilina Anna

28.05.2022

Table of contents

Abstract

Nine out of ten users, when thinking about buying a product or service, do not know which company to contact. Then people begin to look for the necessary information on the Internet and read the reviews of those who have already interacted with the company. Such a study was conducted by Uberall. Reviews are a strong motivator to buy: 88% of users trust them in the same way as the recommendations of friends.

With the help of feedback, the client can explain why he will contact your company again or, on the contrary, will not come again. It is important to understand why brands fall in love with some, and do not return to others, and how to build a business strategy in such a way as to develop and improve the service. To do this, you need to assess customer's level of excitement.

Text analytics and opinion mining find numerous applications in e-commerce, marketing, advertising, politics, market research and any other research.

Sentiment analysis is part of the Natural Language Processing (NLP) techniques that consists in extracting emotions related to some raw texts. This is usually used on social media posts and customer reviews in order to automatically understand if some users are positive or negative and why.

1 The object of research

Reviews of the TripAdvisors website users, who did glamping in Russia.

2 The result

The main result of the project is an attempt to create the recommendation system. This project proposes to develop a sentimental analysis using machine learning tools based on real up-to-date data received by the guests. We should parse the website in order to extract the data needed, highlight emotionally colored words and expressions and analyze them with the respect to the tonality of the words. Finally, we need to develop the frequency analysis of the texts and visualize to the user the "word cloud."

3 Basic keywords

Glamping – a style of camping with amenities and resort-style services not usually associated with ”traditional” camping. Glamping has become particularly popular with 21st century tourists seeking the luxuries of hotel accommodation alongside ”the escapism and adventure recreation of camping”

Sentiment analysis is a method of recognizing and classifying opinions from a piece of text in order to establish whether the writers' attitude toward a given topic or product is positive, negative, or neutral.

A machine learning model is a computer software that has been taught to identify specific patterns. The model is trained on a specific data set using an algorithm that allows the data to be analyzed and the findings to be memorized.

Recommendatory system – a software that uses machine learning models to forecast which things will be of interest to the user based on his personal information.

A dataset is a collection of raw data that will be used to train and test a machine learning algorithm.

Text semantics – a direction that considers the substantive side of the text, structuring meanings expressed explicitly and implicitly.

Implicit information – information that does not constitute the direct meaning of the text components (words, etc.). Implicit information serves to express the implicit meaning of a text or statement, the hallmark of which is the non-necessity of receiving it on understanding, its incomplete restoration to the listeners.

Explicit information – information derived from the meaning of words in the text or statement that are presented in the dictionary and therefore understandable to the recipient. Explicit information in the text can be presented in the form of explicit statements, which mean statements that carry information directly derived from the dictionary meanings of words used in the statement, that is, those whose content can be established from the superficial form of the statement without directly carrying out additional semantic transformations.

Tokenization – dividing phrase into a different set of statements/dividing a statement into different set of words.

Cleaning the data – removing the special characters/words that do not create any value to the analytics part.

$$\sqrt{\tfrac{1}{n}(a_1^2+a_2^2+\cdots+a_n^2)}$$

$$\textstyle\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\textstyle\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

$$\textstyle\int_0^2 \frac{\frac{n(n+1)+\frac{f(x+h)-f(x)}{h}}{2}}{\sqrt{\frac{1}{n}(a_1^2+a_2^2+\cdots+a_n^2)}}$$

4 Introduction

4.1 Subject area

Subject area: sentiment analysis

4.2 Relevance

Sentiment analysis of hotels is now used [4] in the creation of recommendation systems by most hotels around the world. At present, given the current geopolitical situation, the analysis of user reviews from Russia is the most relevant for tourism businesses. Due to limited flights, nature trips within the country are becoming more popular among consumers.

The project will allow to make an optimal decision on the choice of the place to stay for glamping, making the process of surfing more comfortable. It will also give students participating in the project a lot of experience in data science to use in the future.

4.3 Goal

Develop a recommendatory system for glamping tourism based on customer's feedback.

4.4 Tasks

- Examine possible approaches to the task, study tourism market
- Find and collect reviews of the customers of places they have stayed in
- Find the main components of ratings using uncontrolled learning to process natural language.
- Predict ratings with controlled training based on specific feedback.
- Develop a recommendatory system for glamping tourism

5 Comparative analysis of sources and analogues

There exists quite many sentiment analysis of hotel reviews from different parts of the world [11]. For instance, sentiment analysis of hotel reviews from Booking.com [5]. The dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe with the geographical location of hotels provided.

Despite the fact that there are no direct analogues of our work, there are similar works on hotels in Europe. The principle of sentiment analysis based on hotel reviews and glamping reviews differ. It is important to keep in mind differences between hotel rooms and apartments outside. These differences may influence the thematic vocabulary of users.

Due to the novelty of the glamping tourism there are no direct analogues of sentimental analysis of glamping tourism in Russia.

Let's compare our project with the sentiment analysis on tweeter reviews.

	Sentiment analysis on glamping reviews	Sentiment analysis on tweeter reviews [14]
Availability of a ready-made dataset	Need to collect data	Already has the dataset
Importing libraries	NLTK, Scikit-learn, WordCloud, Dostoevsky, RuSentiment, Pandas, NumPy	NLTK, Scikit-learn, WordCloud, Pandas, Numpy
Simple preprocessing step	Has	Has
Lemmatization step	Yes	Yes
Stemming step	No	No
Visualization through WordClouds	Yes	Yes
Evaluation function	Yes	Yes
Modeling	Yes	Yes
User interface	Yes	No

Characteristics	**Sentiment analysis based on Glamping reviews from TripAdvisors: a decision-making assistant**	**Twitter Sentiment Analyzer** [1]
Paid/free	Free	Paid
Visualization	Option to get sentiment via mind maps, wordclouds and graphs	Option to get sentiment expressed via short video clips
Function of calculating the rating of an object based on reviews	No	Yes
	Extremely simple and intuitive UI	
Used for	Analyzing glamping places in Russia before travelling	Analyzing the product reviews before buying, analyze trends at stock market, get sentiments on trending

As we can see the inner methods of sentiment analysis on tweeter and glamping reviews are mostly the same. The significant difference is in data preprocessing and data collection steps. [12]

6 Description of functional and non-functional requirements

6.1 Functional requirements

- An ideal glamping market portfolio is made up of the recommendation system, which includes machine learning algorithms.
- The recommendation system should allow to consider the present status of your portfolio and provide solutions for changing it.

6.2 Non-functional requirements

- The system should be easily scalable:
  - Uploading the dataset
  - Choosing places, that analysis the user wants to receive

7 Roles of other members of the project

The project is implemented by two members. Another participant beside me is Morozova Milena. She is also a 2nd year bachelor degree student at HSE university on faculty of computer science. Our educational program is called “Data science and business analytics”. The role of Milena was dedicated to the creation of the user-friendly interface of the recommendation system. While my task in the project was dedicated to the backend development. In particular, I was to prepare the ready-to-use dataset and collect the data needed for doing the analysis. Moreover, tokenization, cleaning the data, lemmatization, and classification were also in my competence. Together we discussed and implemented machine learning methods.

8 Models, algorithms and methods

The main purpose of sentiment analysis is to find opinions in a text. [18]

One-dimensional emotive space is most typically utilized in current systems for automatically detecting the emotional assessment of a text: positive or negative. The fundamental goal of sentiment analysis is to determine if a document's or sentence's polarity is positive, negative, or neutral.

The polarity of the text can be determined in several ways, the main ones are: classification on a binary scale, on a multi-band scale, the use of scaling systems, identification of subjectivity/objectivity.

In our work we rely on an automated sentiment analysis. Since the project is to evaluate the comments of vacation spots, I think it will be most effective to use supervised machine learning. It lies in the fact that at the first stage, a machine classifier is trained on pre-marked texts, and then the resulting model is used when analyzing new documents. There are a number of dictionaries needed by computer programs for sentiment analysis. Among them are WordNet-Affect, SentiWordNet, SenticNet.

MeaningCloud Sentiment Analysis API works with unstructured texts to determine the tone. The API identifies the tonality locally for individual phrases, establishes a relationship between them to determine the general tonality of the text.

Solves problems such as determining irony in a sentence, determines whether a sentence is a subjective opinion or an objective fact, distinguishes between opposite [4] or ambiguous points of view. Tonality analysis can occur both at the level of an entire document and on its individual aspects in any of the ten supported languages. Despite the fact, that the library does not work with Russian language, we have studied the way it works.

For steps of tokenization, cleaning data, lemmatization, and stemization I have used the most widely used NLTK library.

The Python package for NLP approaches is NLTK (Natural Language Toolkit). NLTK is a popular Python programming language for working with human language data. It includes a set of text processing tools for categorization, tokenization, stemming, tagging, parsing, and semantic reasoning, as well as wrappers for industrial-strength NLP libraries and easy-to-use interfaces to over 50 corpora and lexical resources like WordNet.

For text vectorization and convenient work with machine learning models, we used the Scikit-learn library.

Scikit-learn is a Python package for machine learning. Scikit-learn is a Python-based machine learning package that is available for free. It includes support-vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy. [16]

TextBlob is another great library to word with. It is quiet intuitive and has a lot of tools to operate with. I have studied it additionally in order to apply in this project, however, it works only with English texts. Moreover, it has translation on many different languages, but I got an error while translating all reviews from Russian to English. Probably, this happened because google translate does not support the library anymore. Anyway, knowledge about the work of this library were useful. [7]

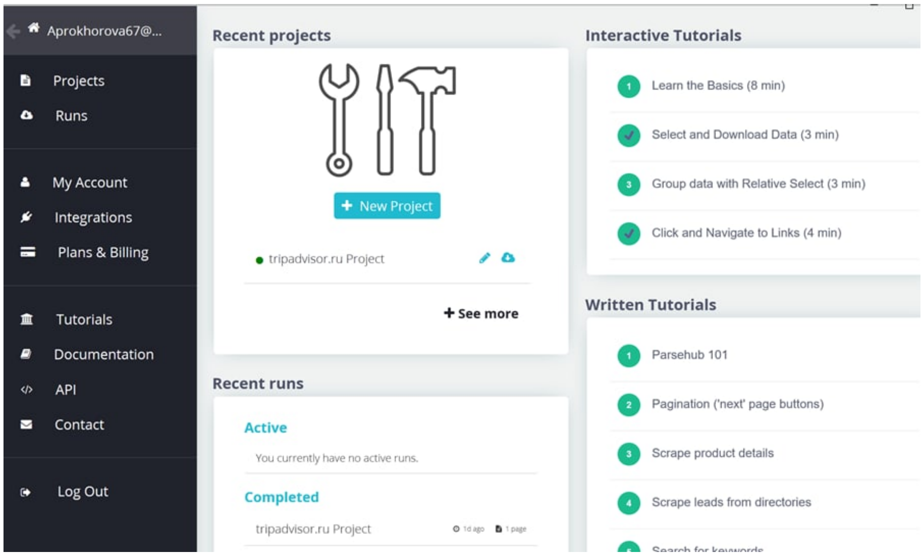
9 Project implementation

9.1 Preparation step

Includes: scraping, gathering data and creating database

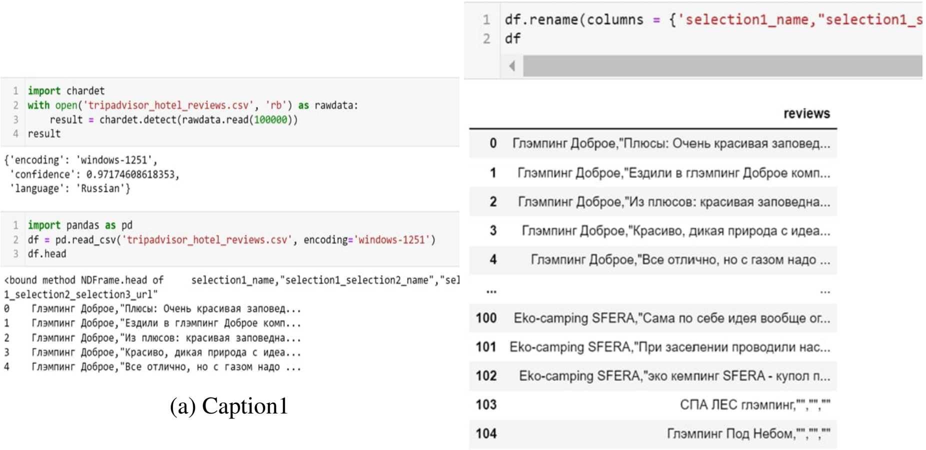
The data was gathered from the TripAdvisors website. For this purpose, I used special application called ParseHub, that has a great user-friendly interface and is easy to work with.

This is how the ParseHub looks like:



The collected data is a plain text with symbols {} and sections 'places', 'section_1', and 'collecting_comments' to divide text on readable blocks.

Here is the way received data looks like:



```python
import chardet
with open('tripadvisor_hotel_reviews.csv', 'rb') as rawdata:
    result = chardet.detect(rawdata.read(100000))
result
```

```
{'encoding': 'windows-1251',
 'confidence': 0.97174608618353,
 'language': 'Russian'}
```

```python
import pandas as pd
df = pd.read_csv('tripadvisor_hotel_reviews.csv', encoding='windows-1251')
df.head
```

```
<bound method NDFrame.head of     selection1_name,"selection1_selection2_name","sel
1_selection2_selection3_url"
0    Глэмпинг Доброе,"Плюсы: Очень красивая заповед...
1    Глэмпинг Доброе,"Ездили в глэмпинг Доброе комп...
2    Глэмпинг Доброе,"Из плюсов: красивая заповедна...
3    Глэмпинг Доброе,"Красиво, дикая природа с идеа...
4    Глэмпинг Доброе,"Все отлично, но с газом надо ...
```

(a) Caption1

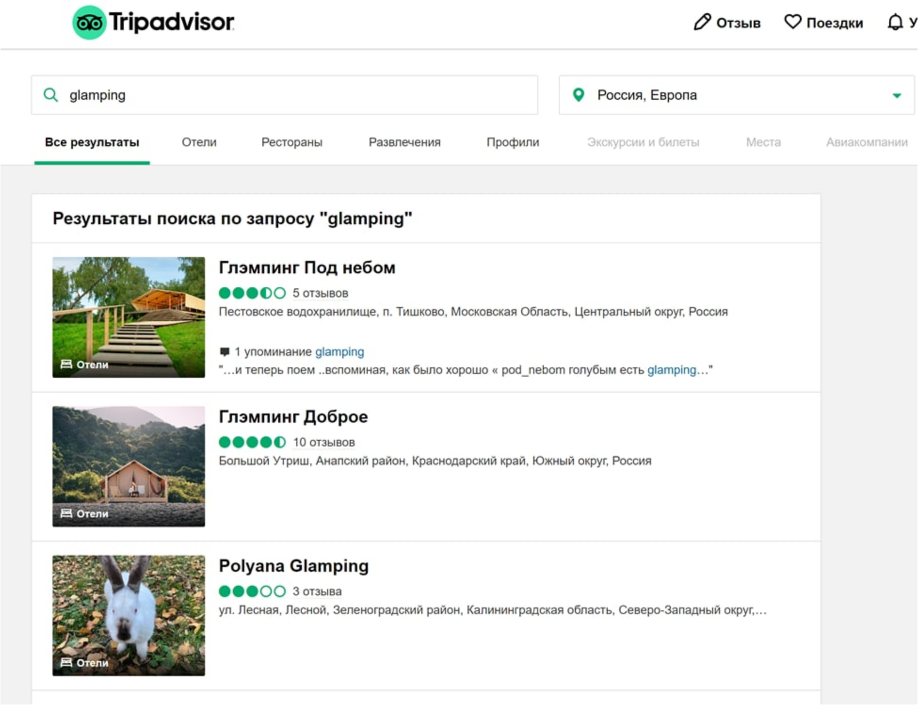
```python
df.rename(columns = {'selection1_name,"selection1_s
df
```

```
                                              reviews
  0  Глэмпинг Доброе,"Плюсы: Очень красивая заповед...
  1   Глэмпинг Доброе,"Ездили в глэмпинг Доброе комп...
  2   Глэмпинг Доброе,"Из плюсов: красивая заповедна...
  3    Глэмпинг Доброе,"Красиво, дикая природа с идеа...
  4     Глэмпинг Доброе,"Все отлично, но с газом надо ...
...                                                 ...
100  Eko-camping SFERA,"Сама по себе идея вообще ог...
101 Eko-camping SFERA,"При заселении проводили нас...
102   Eko-camping SFERA,"эко кемпинг SFERA - купол п...
103                             СПА ЛЕС глэмпинг,"","",""
104                          Глэмпинг Под Небом,"","",""
```

(b) Caption 2

Figure 1: Data after the **”preparation step”**

Dataset was collected from the Tripadvisors website, filtered with ‘glamping in Russia’ pattern. The link: `https://www.tripadvisor.ru/Search?q=glamping&searchSessionId=C72C15DF0E5D73ACFC82EB365AF1EDF61644256081930ssid&sid=D9D482CDC580403DAD25CDAEEFAF63BC1644256153605&blockRedirect=true&ssrc=a&geo=294459`



Результаты поиска по запросу "glamping"

Глэмпинг Под небом
●●●◐○ 5 отзывов
Пестовское водохранилище, п. Тишково, Московская Область, Центральный округ, Россия
1 упоминание glamping
"...и теперь поем ..вспоминая, как было хорошо « pod_nebom голубым есть glamping..."
Отели

Глэмпинг Доброе
●●●●◐ 10 отзывов
Большой Утриш, Анапский район, Краснодарский край, Южный округ, Россия
Отели

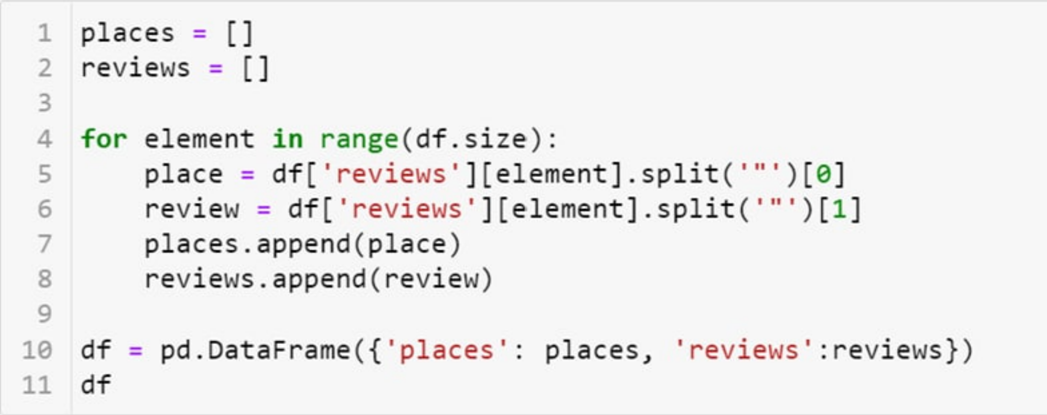
Polyana Glamping
●●●○○ 3 отзыва
ул. Лесная, Лесной, Зеленоградский район, Калининградская область, Северо-Западный округ,...
Отели

9.2 Preprocessing of data

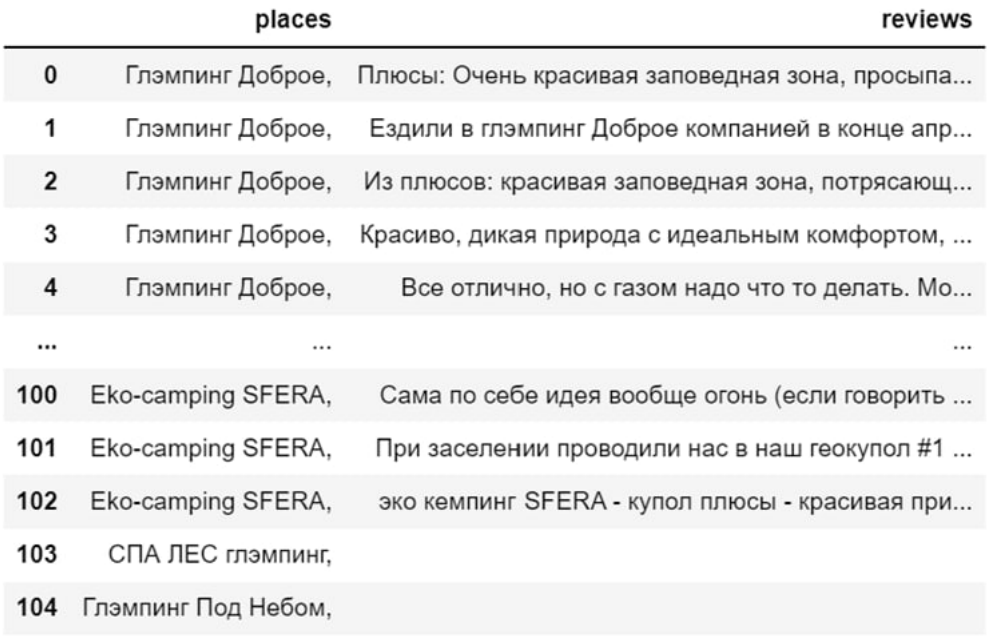
Preprocessing of data is the stage of data mining, which includes the transformation of the source data into an understandable format.

Starting from this stage I have worked with Jupyter Notebook [2] [8] as it is a convenient [4] way to debug and visualize the problem. The collected data I transformed in the dataframe creating two columns: 'places' and 'collected_reviews', where each place has the corresponding [10] review cell with all collected reviews dedicated to the particular place. There are 30 places in total scrapped from the first page of the website. Each place has review length of about 1800 characters.

The resulting view of the dataframe is as follows:

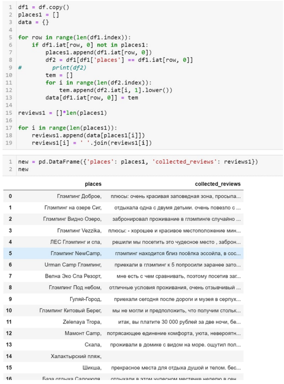


```
places = []
reviews = []

for element in range(df.size):
    place = df['reviews'][element].split('"')[0]
    review = df['reviews'][element].split('"')[1]
    places.append(place)
    reviews.append(review)

df = pd.DataFrame({'places': places, 'reviews':reviews})
df
```



	places	reviews
0	Глэмпинг Доброе,	Плюсы: Очень красивая заповедная зона, просыпа...
1	Глэмпинг Доброе,	Ездили в глэмпинг Доброе компанией в конце апр...
2	Глэмпинг Доброе,	Из плюсов: красивая заповедная зона, потрясающ...
3	Глэмпинг Доброе,	Красиво, дикая природа с идеальным комфортом, ...
4	Глэмпинг Доброе,	Все отлично, но с газом надо что то делать. Мо...
...	...	...
100	Eko-camping SFERA,	Сама по себе идея вообще огонь (если говорить ...
101	Eko-camping SFERA,	При заселении проводили нас в наш геокупол #1 ...
102	Eko-camping SFERA,	эко кемпинг SFERA - купол плюсы - красивая при...
103	СПА ЛЕС глэмпинг,	
104	Глэмпинг Под Небом,	

Before the step of cleaning the received data I split it into two tables: place and reviews. Afterwards, I have united all the reviews to the corresponding places.
Grouping reviews by places:



```
df1 = df.copy()
places1 = []
data = {}

for row in range(len(df1.index)):
    if df1.iat[row, 0] not in places1:
        places1.append(df1.iat[row, 0])
        df2 = df1[df1['places'] == df1.iat[row, 0]]
#         print(df2)
        tem = []
        for i in range(len(df2.index)):
            tem.append(df2.iat[i, 1].lower())
        data[df1.iat[row, 0]] = tem

reviews1 = []*len(places1)

for i in range(len(places1)):
    reviews1.append(data[places1[i]])
    reviews1[i] = ' '.join(reviews1[i])
```

```
new = pd.DataFrame({'places': places1, 'collected_reviews': reviews1})
new
```

	places	collected_reviews
0	Глэмпинг Доброе,	плюсы: очень красивая заповедная зона, просыпа...
1	Глэмпинг на озере Сиг,	отдыхала одна с двумя детьми. очень повезло с ...
2	Глэмпинг Видно Озеро,	забронировал проживание в глэмпинге случайно ...
3	Глэмпинг Vezzika,	плюсы: - хорошее и красивое местоположение мин...
4	ЛЕС Глэмпинг и спа,	решили мы посетить это чудесное место , заброн...
5	Глэмпинг NewCamp,	глэмпинг находится близ посёлка эссойла, в сос...
6	Urman Camp Глэмпинг,	приехали в глэмпинг к 5 попросили заранее зато...
7	Велна Эко Спа Резорт,	мне есть с чем сравнивать, поэтому посетив заг...
8	Глэмпинг Под небом,	отличные условия проживания, очень отзывчивый ...
9	Гуляй-Город,	приехали сегодня после дороги и музея в серпух...
10	Глэмпинг Китовый Берег,	мы не могли и предположить, что получим стольк...
11	Zelenaya Tropa,	итак, вы платите 30 000 рублей за две ночи, бе...
12	Мамонт Camp,	потрясающее единение комфорта, уюта, невероятн...
13	Скала,	проживали в домике с видом на море. ощутил пол...
14	Халактырский пляж,	
15	Шикша,	прекрасное места для отдыха душой и телом. бес...
16	База отдыха Сапокюля,	отдыхали в этом чудесном местечке неделю в сен...

9.3 Cleaning the data

Cleaning the data – remove the special characters/words that do not create any value to the analytics part. [15]

Going through each word I filtered them with respect to punctuation in strings, special characters and Russian alphabet, because all reviews from the dataset are in Russian language. Thus all the used libraries and methods should be working with Russian texts. [13]

9.4 Tokenization

Tokenization – dividing phrase into a different set of statements/dividing a statement into different set of words.

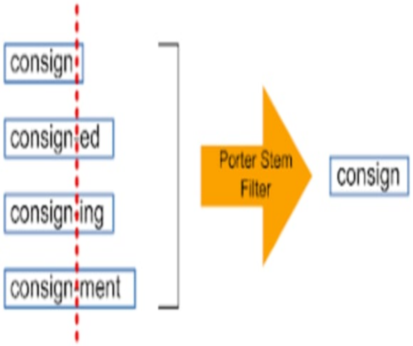
I used the NLTK library, which is a popular open package of libraries used for various kinds of NLP tasks. Before splitting reviews into words I transformed all the reviews into lower case texts.

9.5 Removing stop words

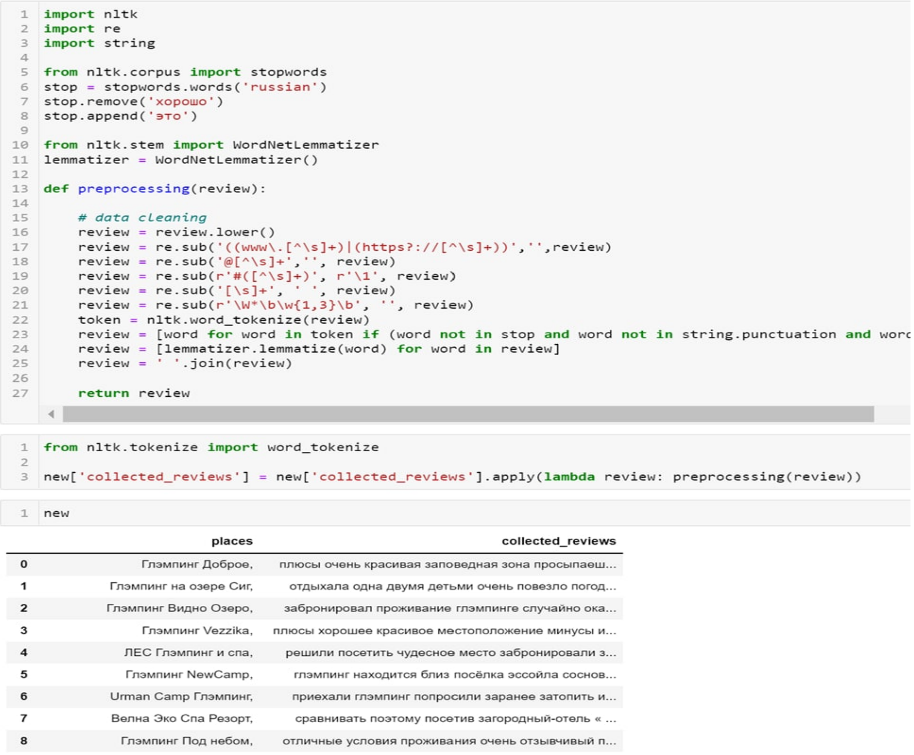
Stop words are frequently used words that do not add any additional information to the text. Words like "а "но "и" have no value and only add noise to the data.

The NLTK library has a built-in stopword list that can be used to remove stopwords from text, but I slightly modified it with regard to the working dataset.

9.6 Lemmatization & Stemization



Stemization – the process of bringing a word to its root/stem. It reduces various variations of a word to its initial form, removes all word appendages and leaves only the base of the word.
Lemmatization is similar to stemization in that it returns the word to its original form, but there is one difference: the root of the term in lemmatization is a word that already exists in the language.
In our project I applied lemmatization approach as it takes a word into its original lemma, not only the linguistic root of the word. It is also much easier to interpret.
Cleaning data, tokenization, lemmatization and reducing stop words I have made in one function called ‘preprocessing’:



```python
import nltk
import re
import string

from nltk.corpus import stopwords
stop = stopwords.words('russian')
stop.remove('хорошо')
stop.append('это')

from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()

def preprocessing(review):

    # data cleaning
    review = review.lower()
    review = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','',review)
    review = re.sub('@[^\s]+','', review)
    review = re.sub(r'#([^\s]+)', r'\1', review)
    review = re.sub('[\s]+', ' ', review)
    review = re.sub(r'\W*\b\w{1,3}\b', '', review)
    token = nltk.word_tokenize(review)
    review = [word for word in token if (word not in stop and word not in string.punctuation and wor
    review = [lemmatizer.lemmatize(word) for word in review]
    review = ' '.join(review)

    return review
```

```python
from nltk.tokenize import word_tokenize

new['collected_reviews'] = new['collected_reviews'].apply(lambda review: preprocessing(review))
```

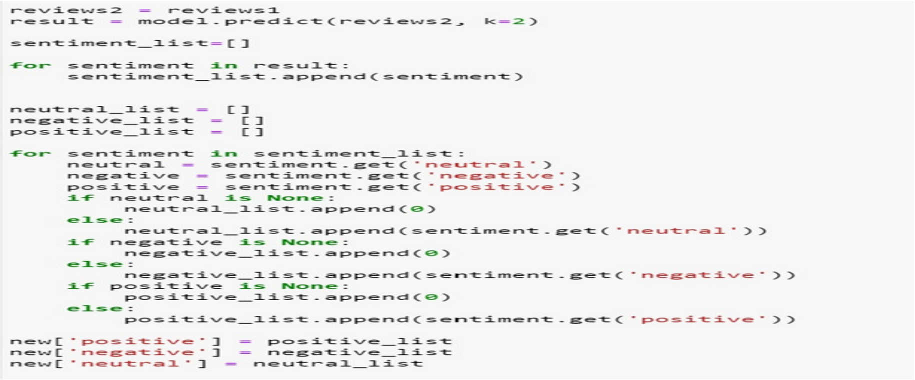
```python
new
```

	places	collected_reviews
0	Глэмпинг Доброе,	плюсы очень красивая заповедная зона просыпаеш...
1	Глэмпинг на озере Сиг,	отдыхала одна двумя детьми очень повезло погод...
2	Глэмпинг Видно Озеро,	забронировал проживание глэмпинге случайно ока...
3	Глэмпинг Vezzika,	плюсы хорошее красивое местоположение минусы и...
4	ЛЕС Глэмпинг и спа,	решили посетить чудесное место забронировали з...
5	Глэмпинг NewCamp,	глэмпинг находится близ посёлка эссойла соснов...
6	Urman Camp Глэмпинг,	приехали глэмпинг попросили заранее затопить и...
7	Велна Эко Спа Резорт,	сравнивать поэтому посетив загородный-отель « ...
8	Глэмпинг Под небом,	отличные условия проживания очень отзывчивый п...

9.7 Natural language processing

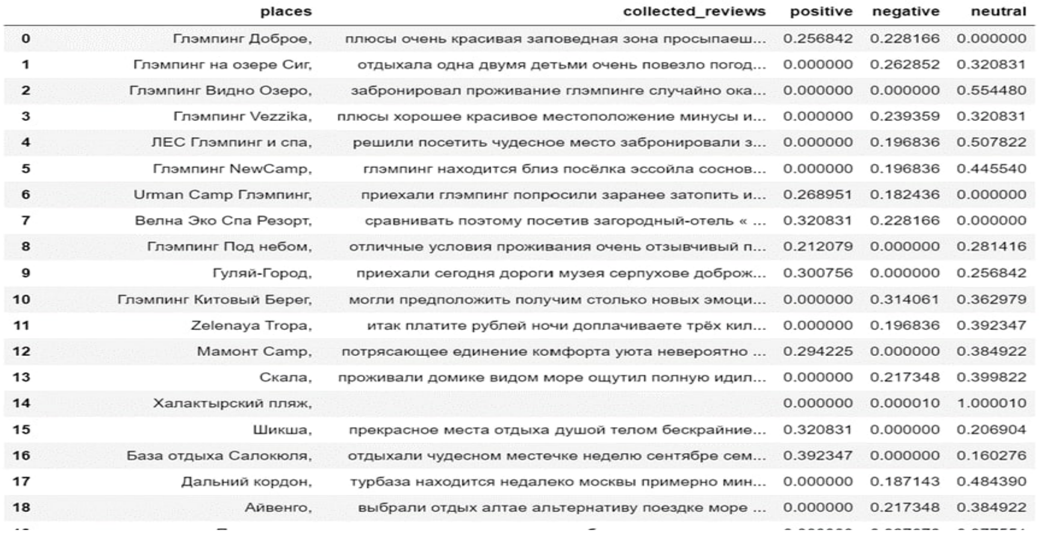
For the natural language processing I have used the library called Dostoevsky [17]. It is an NLP library mostly developed on data from the VK platform. It is based on Russian texts, which is exactly what is analyzed in the project. Moreover, the development team is a native speaker of the Russian language, which accordingly guarantees its best support. [9] The library is simple.

Applying methods from Dostoevsky and RuSentiment libraries I have obtained three sentiments: positive, negative, neutral. [6]



```
reviews2 = reviews1
result = model.predict(reviews2, k=2)

sentiment_list=[]

for sentiment in result:
    sentiment_list.append(sentiment)


neutral_list = []
negative_list = []
positive_list = []

for sentiment in sentiment_list:
    neutral = sentiment.get('neutral')
    negative = sentiment.get('negative')
    positive = sentiment.get('positive')
    if neutral is None:
        neutral_list.append(0)
    else:
        neutral_list.append(sentiment.get('neutral'))
    if negative is None:
        negative_list.append(0)
    else:
        negative_list.append(sentiment.get('negative'))
    if positive is None:
        positive_list.append(0)
    else:
        positive_list.append(sentiment.get('positive'))

new['positive'] = positive_list
new['negative'] = negative_list
new['neutral'] = neutral_list
```

The data now is as follows:

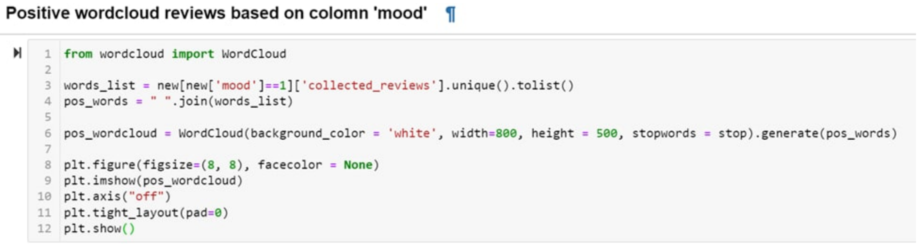


	places	collected_reviews	positive	negative	neutral
0	Глэмпинг Доброе,	плюсы очень красивая заповедная зона просыпаеш...	0.256842	0.228166	0.000000
1	Глэмпинг на озере Сиг,	отдыхала одна двумя детьми очень повезло погод...	0.000000	0.262852	0.320831
2	Глэмпинг Видно Озеро,	забронировал проживание глэмпинге случайно ока...	0.000000	0.000000	0.554480
3	Глэмпинг Vezzika,	плюсы хорошее красивое местоположение минусы и...	0.000000	0.239359	0.320831
4	ЛЕС Глэмпинг и спа,	решили посетить чудесное место забронировали з...	0.000000	0.196836	0.507822
5	Глэмпинг NewCamp,	глэмпинг находится близ посёлка эссойла соснов...	0.000000	0.196836	0.445540
6	Urman Camp Глэмпинг,	приехали глэмпинг попросили заранее затопить и...	0.268951	0.182436	0.000000
7	Велна Эко Спа Резорт,	сравнивать поэтому посетив загородный-отель « ...	0.320831	0.228166	0.000000
8	Глэмпинг Под небом,	отличные условия проживания очень отзывчивый п...	0.212079	0.000000	0.281416
9	Гуляй-Город,	приехали сегодня дороги музея серпухове доброж...	0.300756	0.000000	0.256842
10	Глэмпинг Китовый Берег,	могли предположить получим столько новых эмоци...	0.000000	0.314061	0.362979
11	Zelenaya Tropa,	итак платите рублей ночи доплачиваете трёх кил...	0.000000	0.196836	0.392347
12	Мамонт Camp,	потрясающее единение комфорта уюта невероятно ...	0.294225	0.000000	0.384922
13	Скала,	проживали домике видом море ощутил полную идил...	0.000000	0.217348	0.399822
14	Халактырский пляж,		0.000000	0.000010	1.000010
15	Шикша,	прекрасное места отдыха душой телом бескрайние...	0.320831	0.000000	0.206904
16	База отдыха Салокюля,	отдыхали чудесном местечке неделю сентябре сем...	0.392347	0.000000	0.160276
17	Дальний кордон,	турбаза находится недалеко москвы примерно мин...	0.000000	0.187143	0.484390
18	Айвенго,	выбрали отдых алтае альтернативу поездке море ...	0.000000	0.217348	0.384922

9.8 Word cloud

Another interesting pattern I have developed was word clouds. With the help of the WordClouds library.

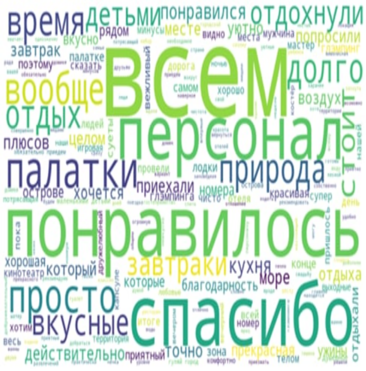
Here is the word cloud created for reviews that were positively colored. The column ‘mood’ has two values: 1 and 0, where 1 denotes that the review is more positive than negative or neutral, and 0 – otherwise (negative or neutral).

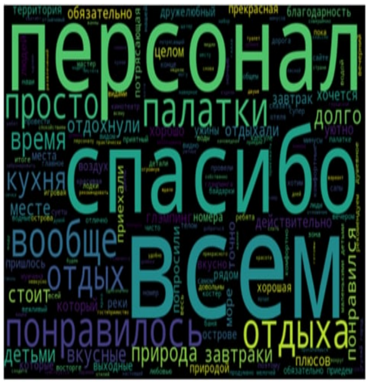


Positive wordcloud reviews based on colomn 'mood' ¶

```
from wordcloud import WordCloud

words_list = new[new['mood']==1]['collected_reviews'].unique().tolist()
pos_words = " ".join(words_list)

pos_wordcloud = WordCloud(background_color = 'white', width=800, height = 500, stopwords = stop).generate(pos_words)

plt.figure(figsize=(8, 8), facecolor = None)
plt.imshow(pos_wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```

As we compare the results for ‘positive’ >0 and ‘mood’ == 1, we can say that the assumption underlying the mood column creation was close to results given by the ‘positive’ criteria.





Summing up the output above, the widely used words in describing the place people liked, were words: ‘понравилось’, ‘спасибо’, ‘вкусные’, ‘отдохнули’, ‘прекрасное’, ‘дружелюбный’, ‘хорошая’, ‘плюсов’, ‘потрясающая’, ‘вежливый’, ‘приятный’, ‘довольны’, ‘персонал’.

Whereas, more neutral and negative words that do not have any emotional patterns are: ‘персонал’, ‘просто’, ‘отдыха’, ‘глэмпинг’, ‘природа’, ‘море’, ‘вода’, ‘минусов’, ‘пляж’, ‘домик’, ‘территория’, ‘ресторан’, ‘мангал’, ‘шатер’.

9.9 Text data vectorization

Text data vectorization the process of converting text into numbers.

Now after text preprocessing, we need to represent the text in numerical form. As an introductory part of machine learning and natural language processing I have tried two ways of vectorising the data: the ‘bag of words’ (BOW) method and TF-IDF method of text vectorization. In BOW’s logic two sentences are called similar if they consist of the identically the same set of words.

BOW creates a dictionary of unique words in a corpus. Now we can create a table where the columns correspond to the unique words in the corpus and the rows to the sentences. Set the value to 1 if the word is in the sentence, and 0 if it is not there.

Term Frequency Inverse Document Frequency (TF IDF). The concept is to weight rare words greater than more frequently used and common words.

TF calculates the probability of finding a certain word in a document.

IDF determines the uniqueness of the word in the entire set of reviews.

9.10 Building model

After implementing data vectorization we need to build models and check how well and reliably did the model learn. So we built naive Buyers model, naïve Buyes TF-IDF model, and also looked at logistic regression TF-IDF model with the corresponding confusion matrixes.

10 Conclusion

In this software project I have acquired many useful skills and knowledge especially in such fields as preprocessing data, tokenization, cleaning data, lemmatization, stemization and machine learning processes on an example of Dostoevsky and RuSentiment toolchains. I also got acquainted with NLTK, WordCloud, Skitic-learn libraries, MeaningCloud Sentiment Analysis API. While surfing the net [3] and reading through different researches on sentiment analysis of texts and comments, especially works dedicated to the Russian language texts, I found out that there are not so many libraries and trained models for sentiment analysis in Russian language. Also the process of learning the model is time-consuming and mostly conducted without using already existing word libraries. While doing the project and diving into machine learning techniques, predictions and model construction I have learned of various ways of data vectorization and prediction models application and implementation. Furthermore, the work and skills that have been acquired on this project will be really useful for me in the upcoming academic year and a good bonus in a portfolio for the interview.

All the materials and source code are stored here: `https://github.com/Anastation67/glamping_project_2022.git`

References

[1] Waqar Ahmad and Maryam Edalati. Urdu speech and text based sentiment analyzer. *arXiv e-prints*, pages arXiv–2207, 2022.

[2] Eli Bressert. Scipy and numpy: an overview for developers. 2012.

[3] Elisa Claire Alemán Carreón, Hirofumi Nonaka, and Toru Hiraoka. Relation analysis between hotel review rating scores and sentiment analysis of reviews by chinese tourists visiting japan. *arXiv preprint arXiv:2110.00821*, 2021.

[4] Ran Duchin and Haim Levy. Markowitz versus the talmudic portfolio diversification strategies. *The Journal of Portfolio Management*, 35(2):71–74, 2009.

[5] Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. Hotel arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent natural language processing: Trends and applications*, pages 35–52. Springer, 2018.

[6] Tal Feldman. Behind the covid curtain: Analyzing russia's covid-19 response on twitter using natural language processing and deep learning. *Intersect: The Stanford Journal of Science, Technology, and Society*, 14(3), 2020.

[7] J Praveen Gujjar and HP Kumar. Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends*, 7(2):1097–1099, 2021.

[8] Matt Harrison and Theodore Petrou. *Pandas 1. x Cookbook: Practical recipes for scientific computing, time series analysis, and exploratory data analysis using Python*. Packt Publishing Ltd, 2020.

[9] O Yu Koltsova, Svetlana Alexeeva, and Sergei Kolcov. An opinion word lexicon and a training dataset for russian sentiment analysis of social media. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE*, 2016:277–287, 2016.

[10] Alla Kravets, Maxim Shcherbakov, Marina Kultsova, and Olga Shabalina. *Creativity in Intelligent Technologies and Data Science: First Conference, CIT&DS 2015, Volgograd, Russia, September 15-17, 2015. Proceedings*, volume 535. Springer, 2015.

[11] George Markopoulos, George Mikros, Anastasia Iliadi, and Michalis Liontos. Sentiment analysis of hotel reviews in greek: a comparison of unigram features. In *Cultural tourism in a digital era*, pages 373–383. Springer, 2015.

[12] Anurag Patel, Bhavya Chheda, Bhavik Jain, and Manya Gidwani. Sentiment analysis of customers opinions on hotel stays using voted classifier. *International Journal of Engineering Research & Technology (IJERT)*, 9(05):827–833, 2020.

[13] AG Pazelskaya and A Solovyev. A method of sentiment analysis in russian texts. In *Proceedings of the Dialog 2011 the 17th International Conference On Computational Linguistics, Moscow region, Russia*, 2011.

[14] Kahlil Philander, Y Zhong, et al. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, 55(2016):16–24, 2016.

[15] Saurav Pradha, Malka N Halgamuge, and Nguyen Tran Quoc Vinh. Effective text data preprocessing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering (KSE)*, pages 1–8. IEEE, 2019.

[16] Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[17] Anton Sysoev, Andrei Linchenko, Vladimir Kalitvin, Daniil Anikin, and Oksana Golovashina. Studying comments on russian patriotic actions: Sentiment analysis using nlp techniques and ml approaches. In *2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, pages 494–499. IEEE, 2021.

[18] Thang Tran, Hung Ba, and Van-Nam Huynh. Measuring hotel review sentiment: An aspect-based sentiment analysis approach. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 393–405. Springer, 2019.