Языковые технологии в гостиничном бизнесе: анализ описаний, отзывов и заголовков

Омельянчук Ольга

План презентации

1. Задача
2. Этапы проекта
3. Сложности
4. Парсинг
5. Отзывы
6. Заголовки
7. Описания
8. Аналоги
9. Идеи на будущее

Задача

* **Практика полученных навыков Python**

- Scraping
- Sentiment analysis
- Frequency of n-grams
- Clustering

Этапы проекта

- Сбор данных (отели в Дублине)
- Заголовки для определения самых частотных n-grams
- Отзывы для Sentiment analysis и language detection
- Описания для TF-IDF и clustering

Источник данных: Booking.com

Сложности

- Airbnb.com

- Большой объем парсинга → о-о-очень долго в Colab

Парсинг

Описания

Ссылки

Названия отелей

Отзывы

Заголовки отзывов

Библиотеки

- Pandas
- BeautifulSoup
- Regex
- Urllib
- Requests

Описания

```
# Работа с df
import pandas as pdimport pandas as pd
# Парсинг
import requests as rq
from bs4 import BeautifulSoup as bs
import time
import re
```

```
# Работа с df
import pandas as pd
# Парсинг
from bs4 import BeautifulSoup as bs
import urllib
import re
import requests as rq
import time
```

Отзывы

Парсинг

Data.json

- Получаем ссылки на все страницы с отелями по поисковому запросу с городом (Дублин) через offset (смещение)
- Получаем ссылки и названия 419 отелей через bs
- Получаем описания для каждого отеля

Парсинг

Reviews.json & reviews_titles.json

- Получаем уникальные названия отелей для ссылки через регулярные выражения
- Получаем ссылки на первую страницу отзывов для каждого отеля через offset
- Получаем из каждой ссылки заголовки и отзывы через bs
- Удаляем дубли

Парсинг

Ссылка, название, описание

	link	title	description
0	https://www.booking.com/hotel/ie/the-gresham.en-gb.html?label=FTUAirBnBAlt&sid=c8d4684eb847d49d0a1fb6350659016c&aid=385205&ucfs=1&arphpl=1&dest_id=-1502554&dest_type=city&group_adults=2&req_adults=2&no_rooms=1&group_children=0&req_children=0&hpos=1&hapos=1&sr_order=popularity&srpvid=a977878f5e0a01a0&srepoch=1675279009&from_sustainable_property_sr=1&dcs_click=1&from=searchresults#hotelTmpl	Riu Plaza The Gresham Dublin	Situated in the heart of Dublin city centre in a historic building, The Gresham Hotel benefits from its own restaurant 'Toddy's', and a bar. The hotel offers free WiFi and spacious rooms overlooking O’Connell Street. Dublin 3Arena is 1 miles away.Each bedroom features an LCD TV, a safe, iron and ironing board and tea and coffee making facilities. The majority of bedrooms overlook the rear of The Gresham Hotel, at neighbouring buildings.The Gallery Restaurant serves breakfast each morning. Toddys Bar and Brasserie and Writers Lounge serve a wide variety of food and beverages throughout the day.The hotel has its own gym with 24 hour access located on the first floor.Extensive car parking is available next to the hotel, at a surcharge.The River Liffey, Temple Bar, and the shopping districts are a few minutes’ walk away. Dublin Airport is 6.2 miles away and the port is 1.6 miles from the hotel. Connolly train station is a 5-minute walk away.

Результаты

Парсинг

Корпусы:

- Ссылки, описания, названия: 419
- Отзывы: 5321
- Заголовки отзывов: 2905

Выводы:

- Полезно освоить дополнительные инструменты помимо BS
- Не использовать Colab (желательно)

Отзывы

Этапы

- Определение языков отзывов
- Подсчет количества отзывов на каждом языке
- Выделение отзывов на английском языке
- Анализ отзывов

Результат

Sentiment	Count
Positive	2980
Negative	1568

Отзывы

Визуализация

Review	Sentiment
Location, breakfast, cleanliness etc	POSITIVE
Staff were super helpful with getting me checked in early after a long distance flight. Buffet breakfast was also delicious. Great value hotel in a central location only moments from a Luas stop.	POSITIVE
Breakfast was not included and for the price paid it would have been nice if it was, location was perfect for Grafton Street and all the places of interest, room was ideal although bed was a bit weird as it was pushed up against the window.	NEGATIVE
Absolutely loved the hotel! The staff members were friendly, good location, excellent facilities, good wifi connection, I totally recommend it!	POSITIVE
beautifully designed and in an excellent location	POSITIVE
The cleaners never came on our 3rd day there!	NEGATIVE
clean, tidy and friendly	POSITIVE
it’s about 10minutes away from main attractions	POSITIVE
clean good for the purpose	POSITIVE
Oriol and Jaime (the Spanish staff) were really good and kind with us.	POSITIVE

Отзывы

Библиотеки

- Pandas
- Langdetect
- Flair

```
# Работа с csv и df
import pandas as pd
import csv
# Language detection
pip install langdetect
from langdetect import detect
# Sentiment analysis
pip install flair
from flair.models import TextClassifier
from flair.data import Sentence
```

Заголовки

Этапы

- Препроцессинг
  - Нижний регистр
  - Пунктуация
  - Токенизация
  - Стоп-слова
  - Стемминг
- Анализ n-grams
- Облако слов

Заголовки

Библиотеки

- Pandas
- Nltk
- Wordcloud
- Matplotlib

```
# Работа с df
import pandas as pd
# Препроцессинг
import nltk
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from nltk import download
download('punkt')
download('stopwords')
# N-grams
from nltk import FreqDist
from nltk.util import ngrams
# Визуализация
#pip install wordcloud
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

Заголовки

N-grams

Word_frequency

Unigrams	Bigrams	Trigrams	Quatrograms
('good') 733	('great', ‘locat') 120	('would', 'high', 'recommend') 14	('locat', 'good', 'valu', ‘money') 5
('stay') 422	('place', 'stay') 53	('good', 'valu', 'money') 12	('good', 'locat', 'good', 'valu') 3
('great') 408	('great', 'stay') 50	('great', 'valu', 'money') 12	('stay', 'one', 'night', 'good') 3
(‘except') 362	('good', 'locat') 41	('locat', 'good', 'valu') 9	('love', 'stay', 'citi', 'centr') 3
('locat') 324	('valu', 'money') 39	('great', 'place', 'stay') 9	('afford', 'place', 'stay', 'dublin') 3
('superb') 268	('good', 'valu') 34	('definit', 'come', 'back'),8	('good', 'valu', 'good', 'locat') 3
('dublin') 163	('high', 'recommend') 32	('place', 'stay', 'dublin') 8	('great', 'valu', 'great', 'locat') 3
('place') 16	('great', 'valu') 32	('hotel', 'great', 'locat') 8	('locat', 'heart', 'templ', 'bar') 3
(‘staff') 149	('love', 'stay') 32	('good', 'place', 'stay') 7	('great', 'locat', 'good', 'valu') 3

Результаты

Заголовки и отзывы

- Все частые N-grams в заголовках - положительные:
- Положительные отзывы: 2980
- Негативные отзывы: 1568

Выводы:

- Люди склонны писать положительные отзывы в 2 раза чаще
- Даже если пишут о негативном опыте, то в название не это не выносят
- Возможно, в Дублине очень хорошие отели ☺

Описания

Этапы

- Препроцессинг
  - Нижний регистр
  - Пунктуация
  - Цифры в слова
  - Стоп-слова
- TF-IDF
- Кластеризация
- Анализ характеристик кластеров
  - Самые частотные слова
  - Объем кластера
  - Средняя длина текста
- Визуализация

Описания

Библиотеки

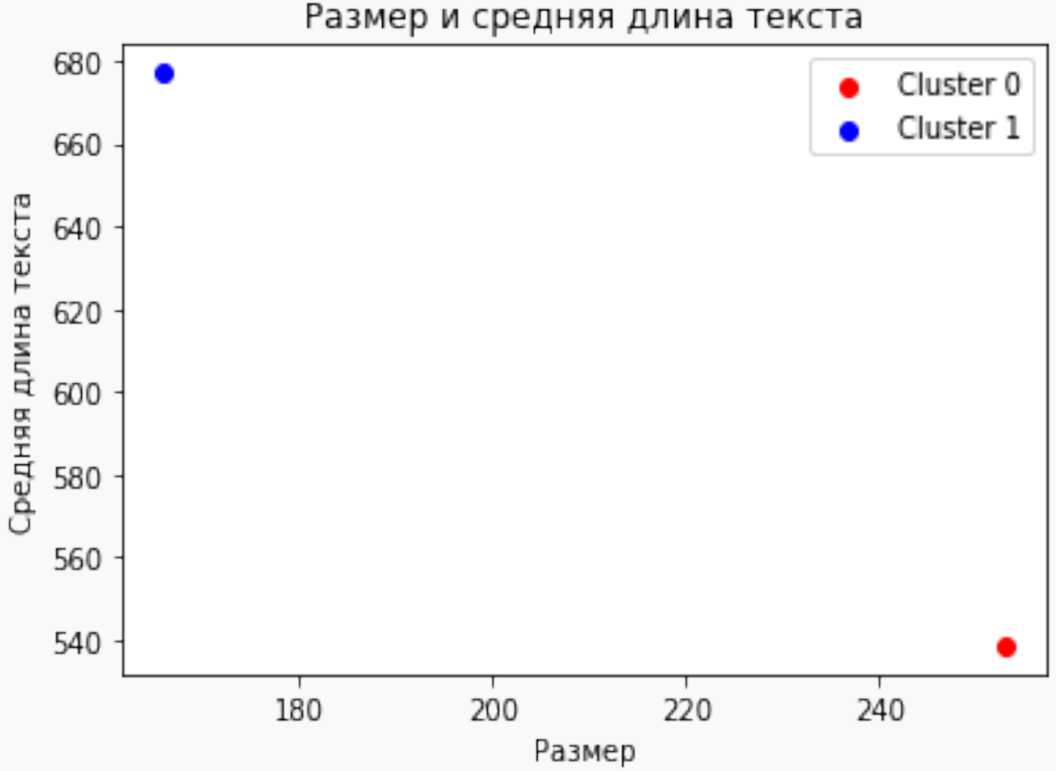
- Pandas
- Num2words
- Nltk
- Numpy
- Scikit-learn
- Matplotlib

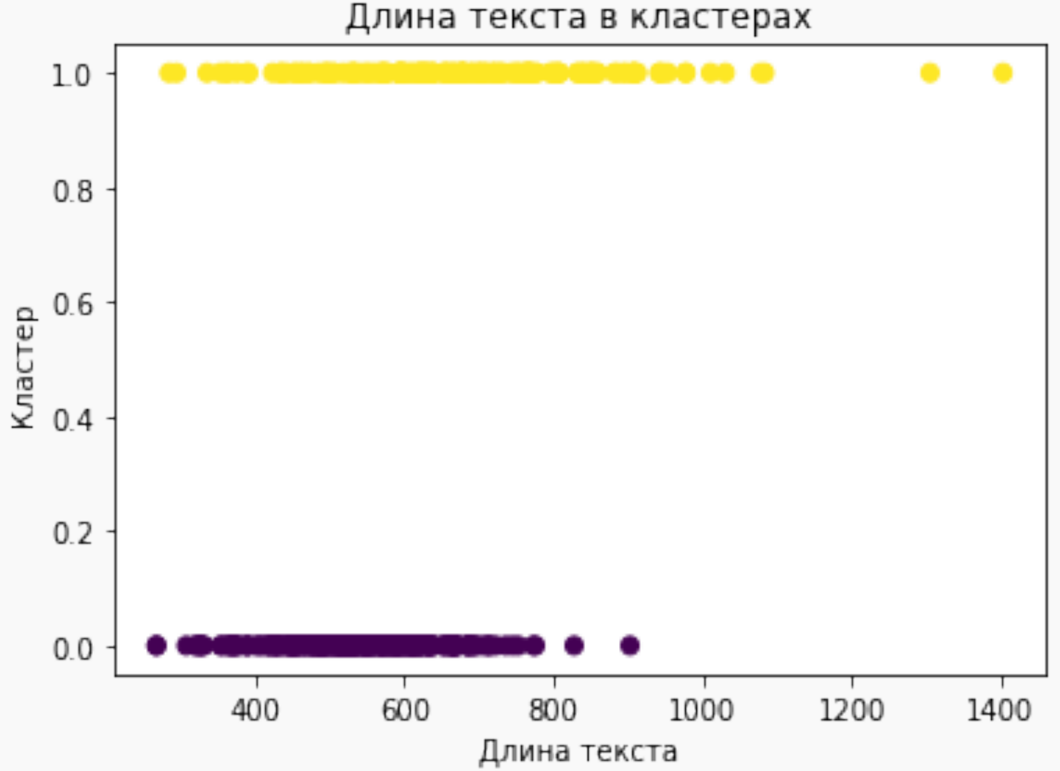
```
# Работа с df
import pandas as pd
# Препроцессинг
pip install num2words
import num2words
import string
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
stop_words.add('and')
# Кластеризация
import numpy as np
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from collections import Counter
# Визуализация
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
```

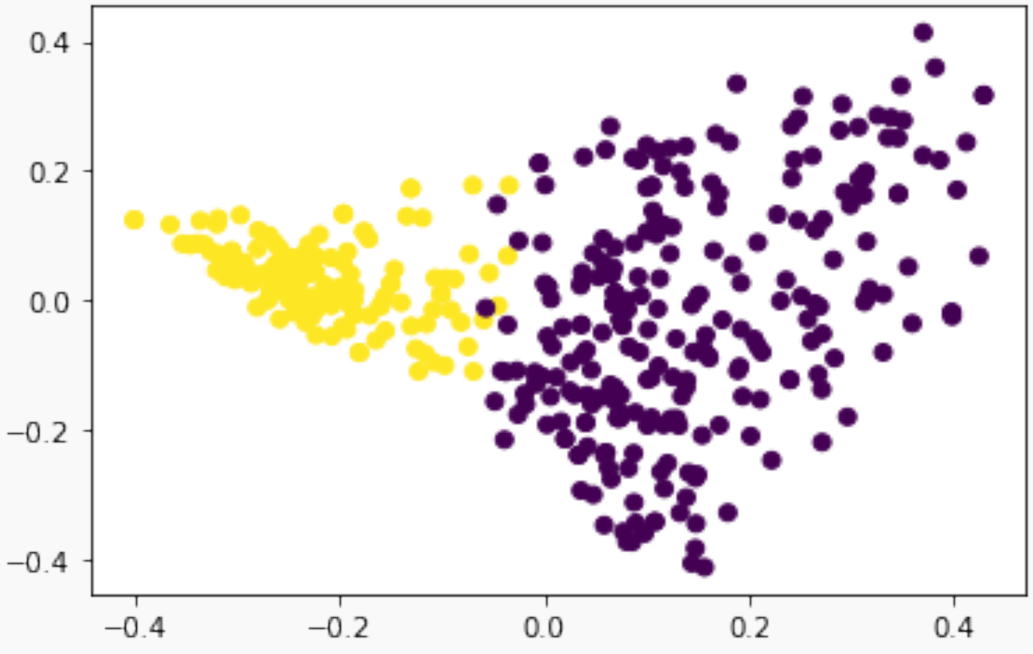
Описания

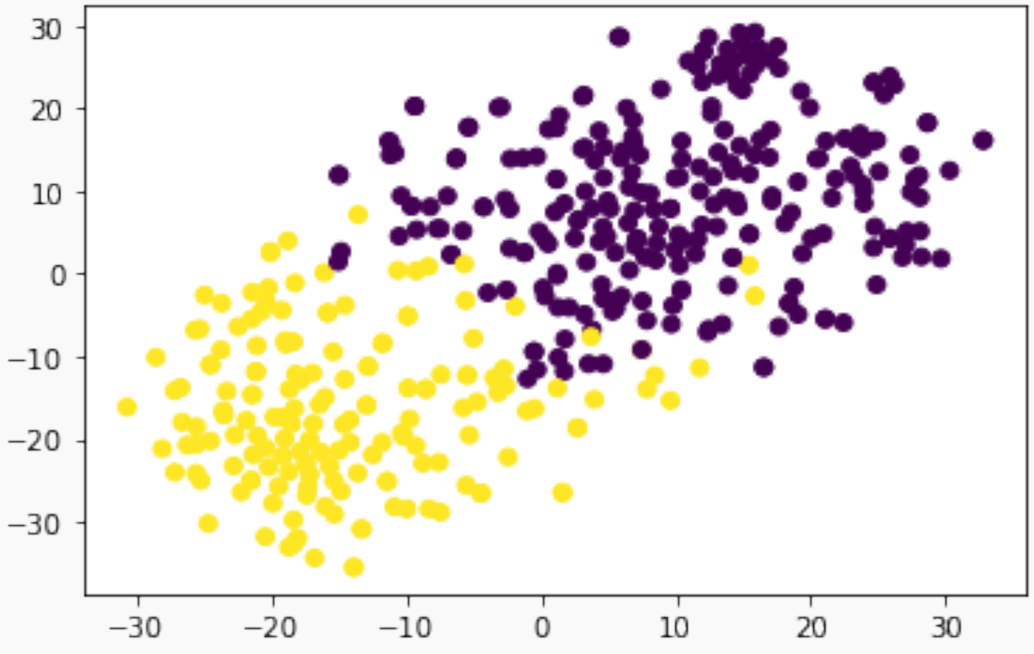
Визуализация

	Text	Length	Cluster	Most Important Word
0	situated heart dublin city centre historic building gresham hotel benefits restaurant toddys bar hotel offers free wifi spacious rooms overlooking o’connell street dublin 3arena one miles awayeach bedroom features lcd tv safe iron ironing board tea coffee making facilities majority bedrooms overlook rear gresham hotel neighbouring buildingsthe gallery restaurant serves breakfast morning toddys bar brasserie writers lounge serve wide variety food beverages throughout daythe hotel gym twentyfour hour access located first floorextensive car parking available next hotel surchargethe river liffey temple bar shopping districts minutes’ walk away dublin airport sixtytwo miles away port sixteen miles hotel connolly train station 5minute walk away	748	1	hotel
1	dcu rooms hallows located dublin property situated seven miles croke park stadium property eleven miles dcu dublin city university sixteen miles convention centre dublinat hostel room fitted desk includes private bathroom shower rooms also wardrobea buffet continental breakfast available morning propertydcu hallows offers terracewhen guests need guidance visit reception happy provide advice3arena eighteen miles accommodation nearest airport dublin airport thirtyseven miles dcu rooms hallows	495	1	hallows


Размер и средняя длина текста
Cluster 0
Cluster 1
Средняя длина текста
Размер
680
660
640
620
600
580
560
540
180
200
220
240


Длина текста в кластерах
Кластер
Длина текста
1.0
0.8
0.6
0.4
0.2
0.0
400
600
800
1000
1200
1400


0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4


30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

Результаты

Получили общее представление после анализа кластеров:

По тематике:

Cluster 0 связан с **географией отелей**, особенно близостью к аэропорту. Наиболее распространенные уникальные слова относятся к особенностям расположения в Дублине, например, 'miles', 'airport', 'nearest'.

Cluster 1 связан с **характеристикой отелей**. Наиболее распространенные уникальные слова относятся к внутренним особенностям пребывания в отелях в Дублине, например, 'rooms', 'bar', 'walk', 'offers'.

По размеру: *Cluster 0 > Cluster 1.* Чаще информация о расположении отеля, а не о его инфраструктуре.

По средней длине текста: Длина текстов в *Cluster 0* < чем в *Cluster 1.* Описания с фокусом на географические особенности гораздо компактнее.

Выводы:

- Полезно разбить на большее количество групп
- Отели чаще стараются давать компактную информацию о расположении в описании

Аналоги

Другие интересные библиотеки:

Scraping

- Использована - BeautifulSoup
- Selenium, Scrapy

Sentiment analysis

- Использована Flair (imdb)
- TextBlob, VADER sentiment

Language detection

- Использована Langdetect
- Cld2-cffi, Langid

Аналоги

Другие интересные библиотеки:

Preprocessing

- Использована - NLTK
- Gensim, Stanford NLP

N-grams

- Использована - NLTK
- PyNLPI, TextBlob

TF-IDF

- Использована - Scikit-learn
- Gensim, NLTK, PyNLPI

Аналоги

Другие интересные библиотеки:

Clustering

- Использована - Scikit-learn
- PyClustering

Visualisation

- Использована - Matplotlib
- Seaborn, Plotly

Идеи на будущее

- Генерация описания/отзыва
- Саммаризация описаний
- Адаптировать для Airbnb

Спасибо за внимание