| Assignment 1 | | Project Summary | |
| --- | --- | --- | --- |
| Course | | **Distributed Machine Learning with Applications in Robotics and IoT** **- 2021** | |
|  | | | |
| Project author | | | |
| № | Name | | FN |
| 1  2  3 | Мирослав Стефанов  Георги Димитров  Кристиан Димитров | | 62282  62256  62260 |

| Project name | Sentiment analysis on hotel reviews |
| --- | --- |

| 1. Short project description (Business needs and system features) |
| --- |
| Идеята е да разработим уеб приложение, което да позволява регистриране на хотели и добавяне на отзиви като коментари към тях. Тези отзиви ще бъдат оценявани числово от Deep Learning модел. За обучение на нашия модел ще използваме отзивите в Booking, които освен коментар, съдържат и рейтинг. Платформата съдържа търсачка по името на хотела. При избор на даден хотел се зареждат отзивите към него с опция за добавяне на нов. |

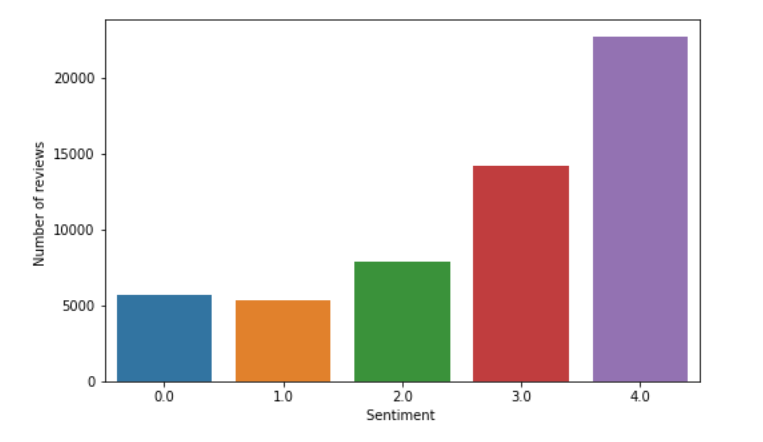
| 1. Project Goals |
| --- |
| 1. Python приложение, което да извлича структурирани ревюта с тяхната оценка от уебсайта Booking.com. Приложението ще генерира dataset със оценени ревюта от 1 до 10 и дали даден коментар е позитивен или негативен. 2. Deep Learning модел за класификация на ревюта в една от следните категории:    1. Негативен    2. Отчасти негативен    3. Неутрален    4. Отчасти позитивен    5. Позитивен 3. Приложение с графичен web интерфейс, в което да се добавят хотели, към които могат да бъдат въвеждани ревюта. Приложението ще използва модела за оценка на тези ревюта и ще класира профилите на базата на оценките. |

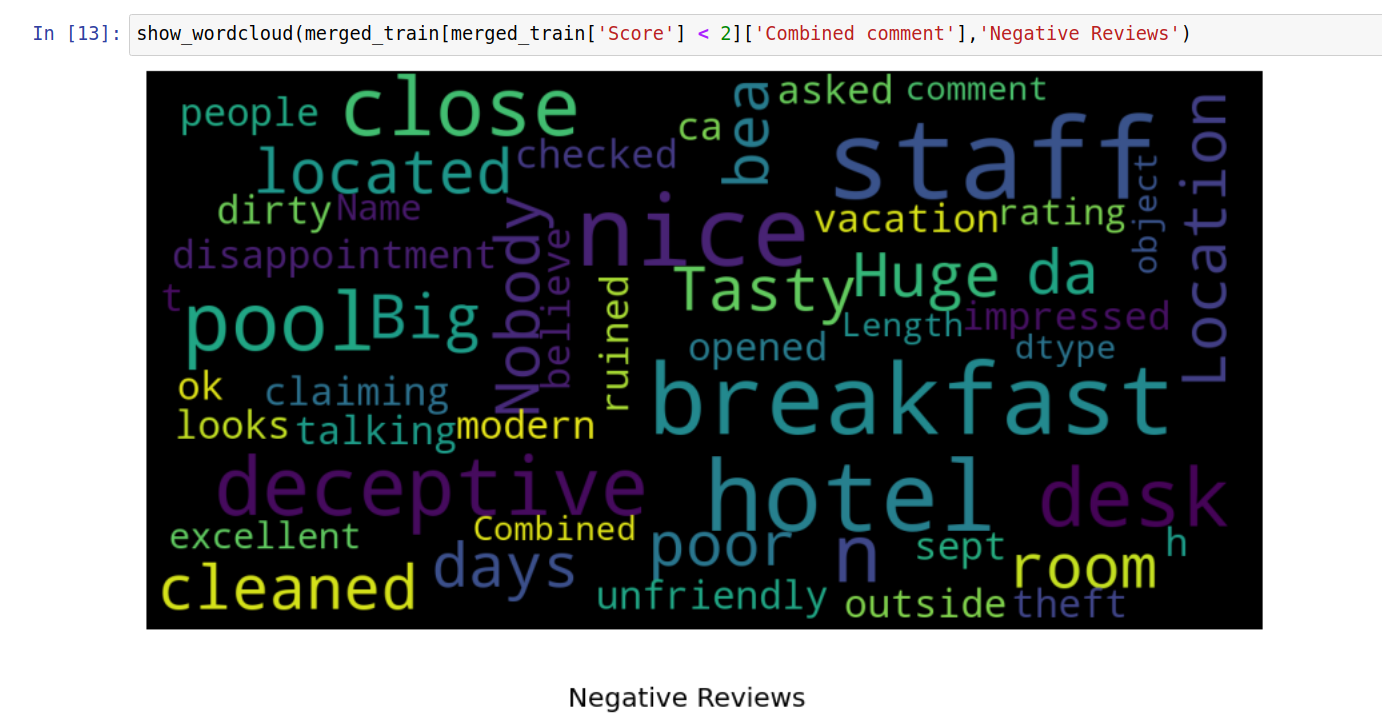
| 1. Implementation Methodology |
| --- |
| Данните за обучение са комбинация от съществуващ dataset с 20 хил. Лейбълирани ревюта с оценки от 0 до 4 и още 35 хил. Лейбълирани ревюта сбърани чрез автоматизиран python скрипт. Скриптът използва Selenium библиотеката за python за scraping на данни от уеб страници, който беше приложен върху сайтът Booking.com.  За модели за оценка на ревю използваме библиотеката tensorflow. След изследване на данните и експериментиране с различни модели: N-Grams method, LSTM model и CNN mode, се спряхме на модела, който постигна най-добър резултат - N-Grams. Накрая запазихме обучения модел и съответния vectorizer. Запазеният модел се зарежда от python приложение, което приема текст и id на ревю, оценява ревюто и изпраща направената оценка.  Уеб платформата за ревюта е реализирана с помощта на рамката за разработка на уеб приложения с Java - Spring Framework. По-точно използваме Spring MVC за предоставяне на функционалността като ресурси в уеб. Spring Security използваме за автентикация и оторизация на потребителите на платформата. Данните за хотели, ревюта и профили на потребители се съхранява в MySQL база данни и се интегрират с помощта на Spring JPA. Интерфейсът на платформата е комбинация между статични web страници изградени с помощта на Thymeleaf и динамични елементи, които използват JavaScript за комуникация със сървъра.  Комуникацията между уеб платформата и приложението за оценка на ревюта е реализирана чрез размяна на съобщения върху Apache Kafka. Оценяващото приложение консумира съобщения с новодобавени ревюта, оценява ги и продуцира съобщения със съответстващите им оценки. От другата страна уеб приложението продуцира съобщението за новодобавено ревю и консумира съобщението с оценката на ревюто.  Цялата система се композира с помощта на Docker Compose като двете приложения са build-нати върху Docker контейнери. |

| 1. Use-cases |
| --- |
|  |

| 1. Used Technologies |
| --- |
| 1. *Selenium with Python* - <https://selenium-python.readthedocs.io/> 2. *Tensorflow -* [*https://www.tensorflow.org/*](https://www.tensorflow.org/) 3. *Docker -* [*https://www.docker.com/*](https://www.docker.com/) 4. Spring - <https://spring.io/> 5. Apache Kafka - <https://kafka.apache.org/> |

| 1. Training and Testing Data (Links to external datasets) |
| --- |
| Данните са взети, чрез scrape-ване на сайта [https://www.booking.com/](https://www.booking.com/index.en-gb.html?aid=397594;label=gog235jc-1DCAEoggI46AdIM1gDaBeIAQGYAQm4ARfIAQzYAQPoAQGIAgGoAgO4Ao_uyowGwAIB0gIkYmUyYTA3YWEtMDlhMi00NTM0LWEyNDktOWU1MTg4ZjFkNzlm2AIE4AIB;sid=c4f78565f7e55e289e3ef5bd4ca9768c;keep_landing=1;sb_price_type=total;sig=v1Z1q67VGB&). В него са представени оферти за хотели, както и ревюта на гости. С помощта на библиотеката [selenium](https://pypi.org/project/selenium/) на python сме свалили над 35000 ревюта преведени на английски. Всяко ревю съдържа коментар и крайна оценка. За да подобрим точността на нашия модел добавихме още 20000 ревюта, които намерихме на този линк: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews/.  Следващата стъпка в процеса на извличане на данните е тяхното почистване. В Booking не всички ревюта са в еднакъв формат и не всички хотели имат еднакъв формат на страниците, но тези проблеми се решават още при scrape-ването. Трябва да сме сигурни, че няма да обучаваме модела с невалидни данни, затова трябва да премахнем пунктуационни знаци и празни коментари.  В Booking оценката на хотел е от 1 до 10, но след изследване на ревютата забелязваме, че средните оценки на хотелите са между 7 и 10, както и че почти няма ниски оценки. Оценките между 1 и 6 са толкова малко, че взехме решението да ги обединим в обща оценка, която да е най- ниската от възможните. След което намалихме с 6 всички оценки, за да ги вкараме в диапазона от 0 до 4. Тези данни използвахме за да обучим модел, който предсказва оценка при входен коментар. |





****



| 1. Resources and References |
| --- |
| 1. Sentiment Analysis using 7 different techniques - <https://www.kaggle.com/parth05rohilla/sentiment-analysis-using-7-different-techniques> 2. Sentiment Analysis on Movie reviews - <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/overview> |