

## Instrukcja

Repozytorium znajduje się pod adresem <https://github.com/AgnesGru/Ocena-komentarzy-NLP> i składa się z dwóch głównych folderów: `web_scraping` oraz `Flask`.

- **Folder `web_scraping`**

W folderze `web_scraping` znajdują się dwa pliki jeden `web_scraping.py` a drugi `mediamarket.py` oba służą do ściągnięcia komentarzy i ocen jeden z `trust_pilot` a drugi z `mediamarket`. Tu jest kod i biblioteki których użyłam (standardowo `Request` i `Beautiful Soup`), żeby mieć na czym szkolić model. Razem około 1200 komentarzy. Wyniki zapisałam w `allegro_scraped1`.

### Folder `Flask`

Znacznie obszerniejszy jest folder `Flask`.

Zacznij od pliku `model_testing_comments_from_internet.py` tam jest stworzona macierz słów i wykorzystany algorytm `MultinomialNaiveBayes`, żeby nauczyć model które komentarze są pozytywne, a które są negatywne. Dokładność czyli `Accuracy` wynosi powyżej 90% przy testowaniu. Wyszkolony algorytm i macierz na koniec spiklowałam.

Następnie przechodzimy do `loaded_pickle_main.py`

Ten plik zawiera dwie funkcje `get_string` i `change_into_string` (wydaje mi się że nazwy mówią same co robi każda funkcja, choć ta pierwsza nie tylko pobiera ale i analizuje tekst) obie są potrzebne w pliku `minimal_app.py` czyli sercu web developmentu. W pliku znajduje się funkcja, która po wpisaniu opinii, wykorzysta dwie poprzednie funkcje i zwróci szablon `answer.html` z wynikiem. (W tej części zakomentowanej, zwracany jest `JSON`)

No i oczywiście jest też część frontendowa (foldery `static` i `templates`).

### `preprocessing.ipynb`

Znacznie ciekawsza jest część rozwijająca projekt a mianowicie pliki `EDA.ipynb` i `preprocessing.ipynb`. Jeśli chodzi o `EDA.ipynb` to tam z ciekawostek zostało użyte `WordClouds`, fajne żeby poszpanować i tyle (wykresy kolumnowe są o wiele bardziej czytelne), natomiast plik `preprocessing.ipynb` zawiera ciekawą funkcję (`clean_text`) do czyszczenia tekstu, w której użyłam biblioteki `re`. Niestety gdy zaczęłam tekst czyścić to Dokładność spadła poniżej 90%, więc część funkcji została zakomentowana i zostało tylko usuwanie części interpunkcji, liczb i cudzysłówów.

Dalej proponuję zajrzeć do `SQL_Alch.py`

Plik `csv` został tu przerobiony na bazę danych z opiniami i ocenami. Pomógł mi w tym `StackOverflow` :-)

Na koniec plik `Questions.py`, baza dała mi odpowiedź na trzy pytania:

Które opinie są średnio dłuższe, pozytywne czy negatywne?

W których opiniach jest więcej przekleństw, w pozytywnych czy w negatywnych?

(przepraszam jeśli naruszam czyjąś wrażliwość językową ale byłam po prostu ciekawa)

Ile razy wystąpiła fraza Polecam a ile Nie polecam?

Na koniec zajrzyj do **AgnesGru.pythonanywhere.com** tam znajdziesz działającą aplikację , i możesz sobie zagrać jak chcesz. Miłej zabawy!

Aga 😊