Projekt PK4: Platforma NLP

Dawid Głąb

3 kwietnia 2024

1 Wstęp

Projekt do wykonania w tym semestrze to platforma nlp. Jest to akronim od przetwarzania języka naturalnego *(eng. natural langugage processing)*. W ramach projektu zaimplementowane zostanie tłumaczenie tekstu z jednego języka na drugi oraz badanie sentymentu, czyli ocena tekstu pod względem użytego słownictwa

2 Opis projektu

Projekt będzie się składał z trzech głównych części:

- Core - napisany w C++, główna część programu analizująca sentyment i tłumacząca tekst.
- Middleware - serwer API, łączący komunikację między programem, a interfejsem sieciowym.
- Frontend - strona internetowa korzystająca z API aby komunikować się z programem, główna część odpowiedzialna za komunikację użytkownika z programem.

Założenie projektu jest bycie modularnym i dość prostym w rozbudowie. W szczególności, iż projekt jest dość złożony, i wymaga korzystania z wielu języków i bibliotek.

3 Podział zadań

Planowany jest następujący ogólny podział zadań. Jeśli chodzi o rdzeń programu, Dawid Głąb zajmie się badaniem sentymentu podawanego tekstu, natomiast Michał Czyż zajmie się tłumaczeniem podanego tekstu na wybrane języki. Każda z implementacji będzie spełniać wyamagania projektu, jakie były przewidziane na początku semestru. Jeśli chodzi o wartwę graficzną oraz komunikację pośrednią między interfejsem użytkownika, a samym programem w C++, Michał Czyż skupi się na bardziej na interfejsie od strony użytkownika, natomiast Dawid Głąb bardziej skupi swą uwagę nad stworzeniem poprawnego interfejsu API, między frontendem, a middlewarem. Planowane jest wykorzystanie różnych bibliotek w trakcie łączenia programu w C++ z Pythonem.

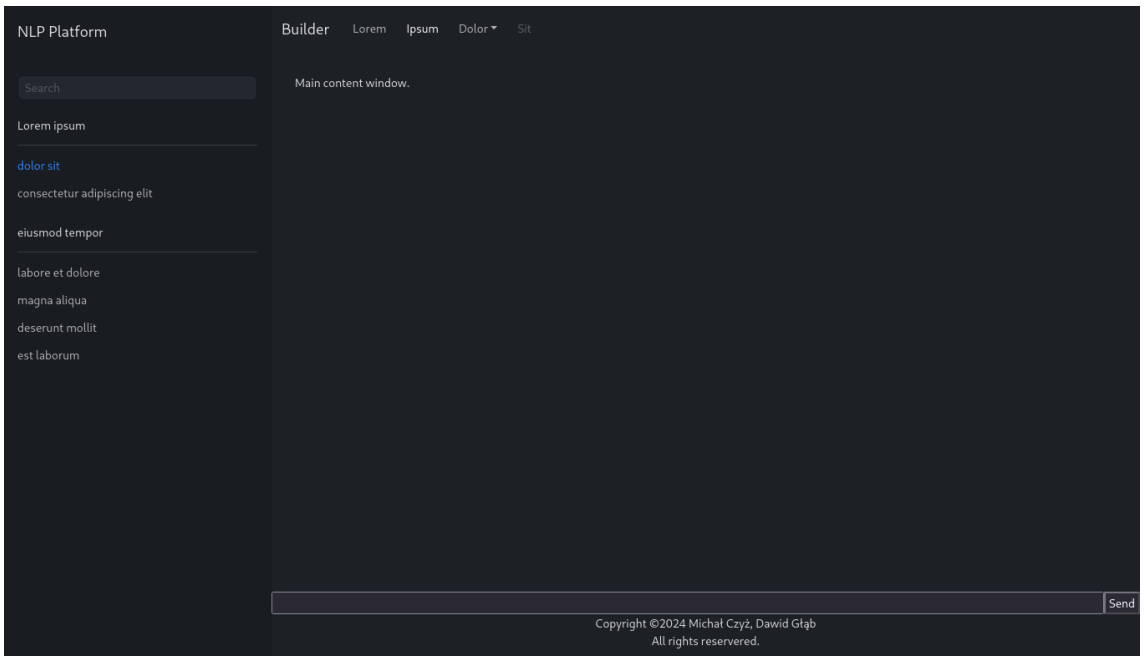
4 Opis problemu - Analiza sentymentu

Analiza sentymentu umożliwia wykrycie nacechowania zdania. Moim aktualnym planem jest użycie gotowego datasetu recenzji z rottentomato, jego serializacja oraz ocena sentymentu przy użyciu algorytmu Naive Bayes. Serializacja zaimplementowana będzie jako osobny moduł przekazujący dane do części wykonawczej algorytmu Naive Bayes. Dane przychodzące z Frontendu przy pomocy API, będą przekierowywane do programu przy użyciu protokołu RPC.

5 Wykorzystane technologie

Projekt będzie wykorzystywał trzy języki programowania.

- Głowna część projektu będzie stworzona w C++ z wykorzystaniem gRPC do komunikacji. Część C++ będzie stworzona obiektowo i zawierać będzie większość wymagań projektu.
- Część pośrednia (middleware backend), będzie napisana w Pythonie, z wykorzystaniem Flaska, jako frameworka serwerowego.



Rysunek 1: Propozycja interfejsu użytkownika.

- Część użytkownika (frontend) będzie napisana w JavaScripcie, a w szczególności wykorzystać będzie framework React wraz z Next.JS. Warstwa wizualna będzie wykorzystywała framework CSS *halfmoon*.

6 Wykorzystane zagadnienia z laboratorium

Na chwilę obecną użytymy zagadnieniami z laboratorium będą:

- Async - Obsługa kominikacji warstw aplikacji
- Wątki - Użycie jednego wątku spowoduje szybką śmierć laptopa którego używam jako debug enviroment
- Regex - Detekcja, manipujacja i testy danych wejściowych
- FileSystem - Serwer config i logi

7 Propozycja interfejsu użytkownika

Przygotowany został początkowy mockup UI. Interfejs składa się z nagłówka i panelu bocznego z poziomu którego będą dostępne tryby aplikacji wraz z ich ustawieniami. W głównym oknie będą pojawiać się konkretne elementy UI do interakcji z programem.

Serwer python oraz program do analizy sentymentu nie będzie zawierał ui. Jedyna komunikacja z tymi warstwami będzie możliwa poprzez API(JS<->Python) lub RPC (Python<->C++)