**ZUM\_NLP: SPECYFIKACJA PROJEKTU**

Celem projektu jest stworzenie analizatora sentymentu do analizy polaryzacji społeczeństwa w Polsce na podstawie tweetów.

Pełen projekt składa się z 4 etapów, ale w zależności od oczekiwanej oceny końcowej wystarczy wykonać wskazane poniżej etapy:

3 – etap 1 i 2

4 – etapy 1 – 3

5 – wszystkie etapy

ETAP 1: POZYSKANIE DANYCH

Pozyskanie danych dotyczy zebrania tweetów. Każda osoba scrapuje tweety (ok. 500 k), żeby stworzyć dataset do dalszej obróbki. Dla ułatwienia wykorzystamy zbiór dot. Covida, który można pobrać tutaj: <https://zenodo.org/record/5090588#.YZDZdmDMKUn>

Dołączony jest plik z id tweetów dla języka polskiego. (Na podanej stronie można pobrać najnowszą wersję, która jest uzupełniania co 2 tygodnie).

Następne kroki wykonujemy pojedynczo lub w grupach (do 3 os):

1. Dodanie etykiet klasy: dane pozyskane nie są oznaczone jako pozytywne/negatywne/neutralne. Na 3 wystarczą 2 klasy, na 4 i 5 już po 3 klasy.
   1. Wybieramy ilość klas docelowych (2 lub 3)
   2. Usuwamy ze zbioru stopwords, żeby odszumić dane
   3. Tworzymy word embeddings do wektorowej reprezentacji słów o podobnym znaczeniu // LUB korzystamy z gotowego rozwiązania dla języka polskiego (pretrained)
   4. Wykorzystujemy K-MEANS do stworzenia clustrów i ustawiamy k=2 lub k=3 w zależności od ilości oczekiwanych klas końcowych
   5. Na podstawie clustrów etykietyjemy dane i ewentualnie ręcznie poprawiamy skład clustrów
2. Czyszczenie danych: dane normalizujemy, usuwamy znaki specjalne, interpunkcyjne, URL, email, powtórzenia, zamieniamy na małą literę, jeżeli pracujemy na pełnych danych to odfiltrowujemy stop words i wybieramy typ tokenizacji. UWAGA: ten etap jest konieczny już przez stworzeniem word embeddings.

ETAP 2: KLASYCZNY ML

Wybieramy 3 modele ML, na których trenujemy dane i pokazujemy wyniki wykorzystując confusion matrix i roc curve.

ETAP 3: MODEL NN

Wybieramy rodzaj modelu NN do wytrenowania, i poprzez walidację staramy się uzyskać jak najlepsze wyniki (nie wystarczy raz zbudować modelu i pokazać danych).

ETAP 4: MODEL OPARTY O MODEL JĘZYKOWY

Końcowy etap to wykorzystanie BERTa lub innego modelu języka do stworzenia klasyfikatora do sentiment analysis.

TERMIN ODDANIA PROJEKTU: DO POTWIERDZENIA

FORMA ODDANIA: repo na githubie udostępnione dla [s18070@pjwstk.edu.pl](mailto:s18070@pjwstk.edu.pl) (mój mail konta na githubie), jeżeli repo jest prywatne, a jeżeli nie to tylko wysłać link do repo i listę autorów na [dwnuk@pjwstk.edu.pl](mailto:dwnuk@pjwstk.edu.pl)

UWAGA! Projekt na 4 i 5 można robić w grupach, ale nie większych niż 3 osoby i z założeniem bardziej skomplikowanych modeli, jak np. pokazanie dwóch sieci neuronowych (np. RNN i CNN), czy wykorzystanie więcej niż jednego modelu językowego.

Repozytorium powinno:

- zawierać skrypty/pliki .py, plik README.md z opisem repo, zapisane modele nn/z modelem języka, które można wczytać (load) – jeżeli jednak okażą się za duże do wgrania można to pominąć

- zawierać korpus wykorzystanych danych, np. w formacie csv

- można krótko opisać otrzymane wyniki w pliku README wraz opisem założeń projektu (ogólnie, bez wchodzenia w zbędne szczegóły).