ZZSN Dokumentacja Wstępna

Projekt 12. Analiza wydźwięku depesz giełdowych za pomocą modelu typu BERT.

Skład zespołu:

- Robert Ostoja-Lniski
- Dawid Brzozowski

Opis zadania

Celem projektu jest wykorzystanie polskiego modelu typu BERT (np. Herbert albo Polbert) do oceny wydźwięku (*negatywny/neutralny/pozytywny*) depesz na polskiej Giełdzie Papierów Wartościowych.

Przewidywany sposób rozwiązania problemu

Pierwszym etapem rozwiązania problemu będzie pobranie danych finansowych z systemu *ESPI*, na którym znajdują się treści depeszy giełdowych oraz z platformy *Quandl*, która udostępnia API do pobierania kursów akcji spółek. Utworzone zostaną skrypty, które uruchomione zostaną na platformie *Google Cloud Platform* w celu pozyskania danych.

Sposób generowania zbioru danych:

1. Pobieranie danych finansowych
    a. **ESPI**: Na podstawie udostępnionych depesz na platformie *ESPI* planujemy zaimplementować *API* służącego do pobierania raportów dla dowolnej spółki z zadanego okresu. Na wejściu podana będzie nazwa spółki i przedział czasowy, a zwrócony będzie plik zawierający treści depesz dla tej spółki dla każdego dnia z zadanego przedziału (o ile w tym dniu opublikowano raport). Plik będzie w formacie JSON.
    b. **Quandl**: Dla poszczególnych wierszy uzyskanych z systemu ESPI pobrane zostaną ceny kursów akcji spółek z dnia poprzedzającego oraz następującego po dniu udostępnienia depeszy. Następnie zmiana kursu akcji zostanie porównana do zmiany kursu indeksu WIG aby ustalić wydźwięk dla depeszy względem szerokiego rynku.
    c. Z treści depesz przechowywanych w zbiorze usunięte zostaną nazwy spółek.
2. Ustalanie wydźwięku dla pobranych danych
    a. Do ustalenia wartości wydźwięku posłuży nam wzór:
    $$m = \left[\frac{a2 - a1}{a1} - \frac{w2 - w1}{w1} \right]$$, gdzie:
        i. *a1* jest ceną akcji spółki podczas poprzedniego zamknięcia
        ii. *a2* jest ceną akcji spółki podczas zamknięcia kolejnego notowania po wydaniu depeszy
        iii. *w1* jest ceną indeksu WIG podczas poprzedniego zamknięcia

iv. *w2* jest ceną indeksu WIG podczas zamknięcia kolejnego notowania po wydaniu depeszy

b. Podczas eksperymentów wyznaczymy dokładne wartości parametrów *p i n (p > n),* takich, że
   i. dla wydźwięku pozytywnego $m \geq p$
   ii. dla wydźwięku neutralnego $n \leq m < p$
   iii. dla wydźwięku negatywnego $m < n$

3. **PolElmo2.0**. Do wstępnego trenowania modelu dla zadania analizy wydźwięku posłużą nam dane PolElmo.

Model BERT:

1. Za pomocą biblioteki transformers wykorzystany zostanie model *HerBERT* (base) ze względu na wysokie wyniki uzyskane w benchmarkach (*https://klejbenchmark.com/leaderboard/)*. Wybrana zostanie wersja bazowa ze względu na krótszy czas treningu w porównaniu do modelu w wersji *large*.
2. Model ten zostanie wstępnie wytrenowany na zbiorze *PolElmo* oraz zmierzona zostanie jego jakość na zbiorze testowym z tej samej dziedziny jak na zmienionej (odpowiednio *PolElmo2.0-IN, PolElmo2.0-OUT*).
3. Następnie, model zostanie dotrenowany na przygotowanych przez nas danych finansowych. Porównana zostanie jakość predykcji modelu po dotrenowaniu z modelem uczonym tylko za pomocą danych *PolElmo*. Ostateczna weryfikacja działania modeli zostanie przeprowadzona na niezależnym zbiorze testowym zawierającym treści depesz giełdowych.

Wybrane zbiory danych

W projekcie skorzystamy ze zbioru depesz opublikowanych na portalu: http://infostrefa.com/infostrefa/pl/raporty/espi/firmy/. Zawiera on raporty dla wybranych spółek (od października 2004 roku do bieżącego kwietnia 2021 roku).

Wartość cen akcji spółki w danym dniu, oraz wysokość notowania WIG (w celu zbadania korelacji między szerokim rynkiem) zostanie pobrana na dzięki *API* udostępnionemu przez platformę *Quandl*.

Narzędzia

- Język programowania: *Python 3.8*.
- Instalacja pakietów za pomocą narzędzia *Anaconda*.
- Biblioteki oraz inne narzędzia wykorzystane w projekcie:
  - *requests* - w celu pozyskiwania stron html zawierających depesze giełdowe.
  - *BeautifulSoup4* - w celu przetwarzania stron html zawierających treść depesz giełdowych wraz z nazwą spółki oraz datą opublikowania.
  - Quandl - do pobierania kursów akcji spółek.
  - PyTorch oraz transformers - do stworzenia modelu opartego o BERT dla zadania analizy sentymentu.
  - Matplotlib - do wizualizacji osiągniętych rezultatów.

- Scikit - learn - do operacji na zbiorach danych (podział na zbiór trenujący, walidujący i testowy itp.).
- Google Cloud Platform: do pobierania danych za pomocą wielu maszyn oraz trenowania modelu na wydajnej maszynie z GPU.

Zakładana funkcjonalność

Model po procesie uczenia na danych finansowych potrafi wskazać sentyment (pozytywny/neutralny/negatywny) dla treści depeszy giełdowej.