Sprawozdanie z projektu

Temat: 8 - Predykcja zainteresowania postami w social media z użyciem metod NLP
Grupa: Arkadiusz Trojanowski, Łukasz Kisielewski, Wiktor Gaworek
Język: MATLAB
Kod źródłowy projektu: https://github.com/greemote/aghust-mio
Data: 15.06.2022

Cel zadania

Celem zadania jest oszacowanie liczby retweetów danego tweeta na podstawie jego treści. Dane, na których operujemy, to posty Donalda Trumpa zebrane do pliku *csv*, które można znaleźć pod adresem https://www.kaggle.com/datasets/austinreese/trump-tweets.

Opis danych

Przykładowe wiersze

id	*link*	*content*	*date*	*retweets*	*favourites*	*mentions*	*hashtags*
611 994 465 931 264 000	https://twitter.com/realDonaldTrump/status/611994465931264000	. @ patrickbuchanan thx 4 your great article. Our country is in big trouble w/the dopey politicians that are running ithttp://buchanan.org/blog/the-anti-politician-16164 …	2015-06-19 15:30:25	111	206	@patrickbuchanan	
611 994 802 289 307 000	https://twitter.com/realDonaldTrump/status/611994802289307648	Our country is in a major crisis of incompetent leadership. We cannot continue to go on with these politicians who do nothing but talk.	2015-06-19 15:31:45	1170	1645		
611 998 039 700 578 000	https://twitter.com/realDonaldTrump/status/611998039700578304	@ stevenkirk @ realDonaldTrump @ PatrickBuchanan great article .. At least # thedonald is finally saying what us average working people know!	2015-06-19 15:44:37	93	215	@stevenkirk,@realDonaldTrump,@PatrickBuchanan	#thedonald

Tabela 1: przykładowe wiersze.

Opis kolumn

- **id** - identyfikująca wartość tweeta, nie będziemy jej używać w naszym programie, bo nie jest istotna,
- **link** - link do tweeta na twitterze, nie będziemy go używać w naszym programie, bo nie jest istotny,

- **content** - treść tweeta; wszystko, co zostało napisane,
- **date** - data opublikowania tweeta rok-miesiąc-dzień i godzina:minuta:sekunda,
- **retweets** - liczba retweetów, czyli ile razy dany tweet został udostępniony przez ludzi. To jest to, co chcemy oszacować,
- **favourites** - liczba polubień tweeta, nie będziemy z tego korzystać w naszym programie, ponieważ publikując tweeta nie wiedzielibyśmy ile osób go polubi,
- **mentions** - wspomnienia nazw innych kont w tweecie, poprzedzone znakiem @,
- **hashtags** - tematy wspominane w tweecie poprzedzone znakiem #.

Schemat projektu

Nasz projekt składa się z trzech plików: *prepareData.m*, *learn.m i analyzeResults.m*. Do poprawnego działania *prepareData.m* potrzebny jest dołączony do projektu plik *realdonaldtrump.csv*. Wygenerowane już dane umieszczone są w pliku *preppedData.mat*.

Plik *prepareData*

Uruchamiając ten plik tworzymy dane wejściowe do sieci, która będzie użyta w kolejnym z nich.

Początkowe operacje

Korzystając z funkcji *preprocessText* oczyszczamy nieco nasze dane dzieląc string zawierający cały tweet na słowa i usuwając znaki punktuacyjne oraz nieznaczące wyrazy. Ostatecznie zamieniamy wszystkie duże litery na małe, żeby nie traktować tych samych słów jako różnych.

Tworzenie danych uczących

Zastanawiając się nad tym, co wpływa na popularność tweeta wybraliśmy kilka czynników.

- treść tweeta,
- sentyment,
- długość,
- data opublikowania,
- liczba wzmianek innych kont,
- hasztagi,
- udostępnienia.

Niestety, nasza baza danych nie posiada statystyk użytkownika w momencie pisania tweeta, które z pewnością poprawiłyby wynik (w szczególności liczba obserwujących). Bardzo przydatna byłaby również informacja, czy tweet miał podpięty element graficzny, taki jak film, zdjęcie, czy GIF.

Treść tweeta

Początkowo użyliśmy samego sentymentu, ale wyniki nie były dla nas wystarczająco satysfakcjonujące, więc zdecydowaliśmy się też na użycie konkretnych słów. Próba skorzystania z wszystkich wyrazów znaczących użytych w tweetach pokazała, że przygotowanie danych trwa zabójczo długo. Ostatecznie zdecydowaliśmy się na bardziej symboliczne użycie tej części. Sprawdziliśmy treści 5 tweetów, które były najczęściej udostępniane i wyciągnęliśmy wszystkie słowa, które się w nich pojawiły. Są to między innymi tak charakterystyczne i silne zwroty jak "*asap*", "*impeachment*", czy składniki słynnego już sloganu "*Make America Great Again*". Następnie stworzyliśmy wektory opisujące dla każdego z tweetów, czy w nich poszczególne słowa z tego zbioru się pojawiają.

Sentyment

Doszliśmy do wniosku, że silne naładowanie emocjonalne jest czynnikiem, który wpływa na chęć podzielenia się postem. Ludzie chętniej będą się dzielić informacją bardzo pozytywną lub bardzo negatywną.

W związku z tym wytrenowaliśmy sieć *fitcknn* (*fit k-nearest neighbor classifier* - klasyfikacja k najbliższych sąsiadów) na danych zebranych w leksykonie opinii.
Sieć ta umie przewidzieć, czy nowe słowo jest pozytywne czy negatywne.
Pozwoliło nam to sklasyfikować wszystkie słowa znaczące pojawiające się w tweetach. Mając tę informację, zmieniliśmy wektor sentymentów słów na nieujemną wartość liczbową określającą naładowanie emocjonalne. Ponieważ silniejsze emocje wzbudzają w ludziach negatywne sentymenty, obliczaliśmy je podwójnie.

Długość

Na popularność tweeta może wpływać jego długość. Ludzie, jako istoty z natury leniwe, nie mają ochoty czytać zbyt wielu słów. Dlatego zdecydowaliśmy się uwzględnić długość posta jako jeden z czynników. Uzyskujemy ją zwyczajnie sprawdzając rozmiar wektora słów danego tweeta, więc nie jest bezpośrednim przełożeniem oryginalnej liczby słów, tylko jest liczbą słów znaczących pojawiających się w nim.

Data opublikowania

Z uwagi na specyfikę zbioru, na którym mamy przewidywać, zdecydowaliśmy się uwzględnić tę informację. Donald Trump stał się nagle znacznie bardziej interesującą osobistością, gdy pojawiła się możliwość, że zostanie prezydentem. Ogłosił to 16.06.2015. Wtedy liczba retweetów w kolejnych miesiącach zaczęła zauważalnie rosnąć. Zdecydowaliśmy się więc na podzielenie naszego zbioru na dwa przedziały dat.

Liczba wzmianek innych kont

Wspominając innych użytkowników, zwiększa się szansę na otrzymanie od nich odpowiedzi. Gdy osoba napotyka post, który jest właśnie taką reakcją, może sprawdzić, na którego tweeta odpowiada. A to zwiększa szansę, że się tą treścią podzieli z innymi. W związku z tym zliczyliśmy liczbę innych kont wspomnianych w tweecie uznając jako istotną do oceny popularności.

Hashtags

Hasztagi oznaczają tematykę posta. W związku z tym wyszukując informacje z danej tematyki człowiek trafia na tweety nimi oznaczone. Może w ten sposób znaleźć interesującą informację na koncie, którego normalnie nie śledzi. W ten sposób część ludzi mogła trafić na posty Trumpa i je udostępnić, więc zdecydowaliśmy się uwzględnić liczbę hasztagów w naszym modelu predykcyjnym.

Retweets

Ostatnia, ale niepomijalna część naszego zbioru uczącego to retweety - czyli to, co chcemy oszacować. Wybranie ich do przewidywania popularności zamiast liczby polubień podyktowane jest przez jest fakt, że użytkownicy klikając w serduszko z reguły zgadzają się z treścią tweeta, natomiast retweet jest neutralny, ponieważ tweety można podawać dalej nie tylko z aprobatą, ale również w celach krytyki. Ze względu na często olbrzymie różnice w ilości retweetów, zgodnie z sugestią osoby prowadzącej zajęcia przekonwertowaliśmy je do skali logarytmicznej.

Plik *learn.m*

Ten plik odpowiada za uczenie sieci na podstawie przygotowanych wcześniej danych i zobrazowanie ich w przystępnej dla ludzkiego oka formie.

Operacje

Ponieważ przyjęliśmy skalę logarytmiczną, w przypadku, gdy tweet nie został ani razu podany dalej, uzyskalibyśmy -∞, dlatego wyrzuciliśmy takie wiersze z tabeli. Podobnie jak te z pojedynczym retweetem, gdyż nie jest ich dużo, a zera przeszkadzałyby w obliczaniu precyzji wyników.

Użyliśmy uczenia nadzorowanego. Przygotowane dane zostały losowo podzielone w 80% na część trenującą, a pozostała część posłużyła do testów. Skonfigurowaliśmy jednokierunkową sieć MATLABA (*feedforwardnet*) o domyślnej liczbie warstw.

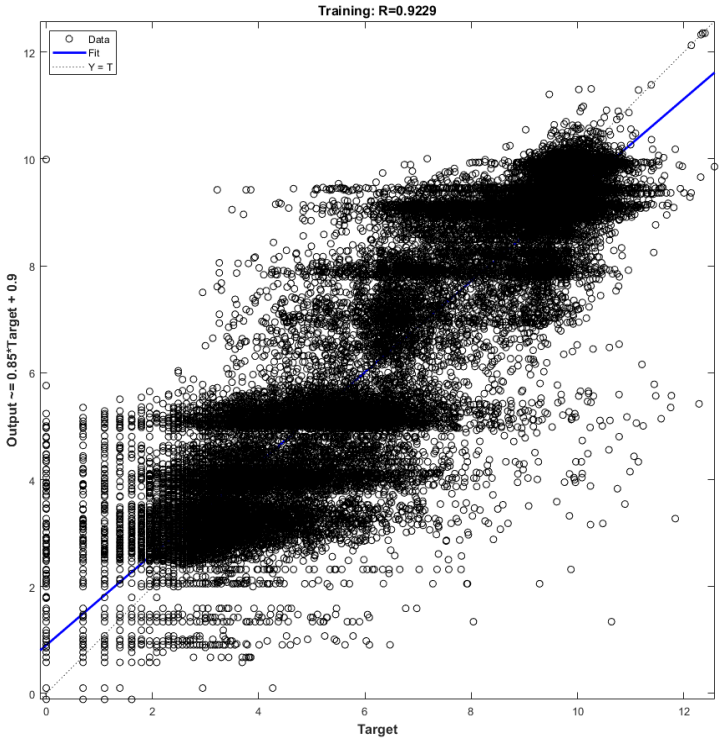
Aby wykluczyć bardzo niewielką ilość zaburzeń, wybraliśmy przewidziane wartości większe niż trzykrotność bezwzględnego odchylenia mediany (*MAD*) i usunęliśmy odpowiednie rekordy.

Plik *analyzeResults.m*

Tutaj analizujemy rezultaty uczenia. Temat jest rozwinięty w rozdziale *Wyniki*.

Wyniki

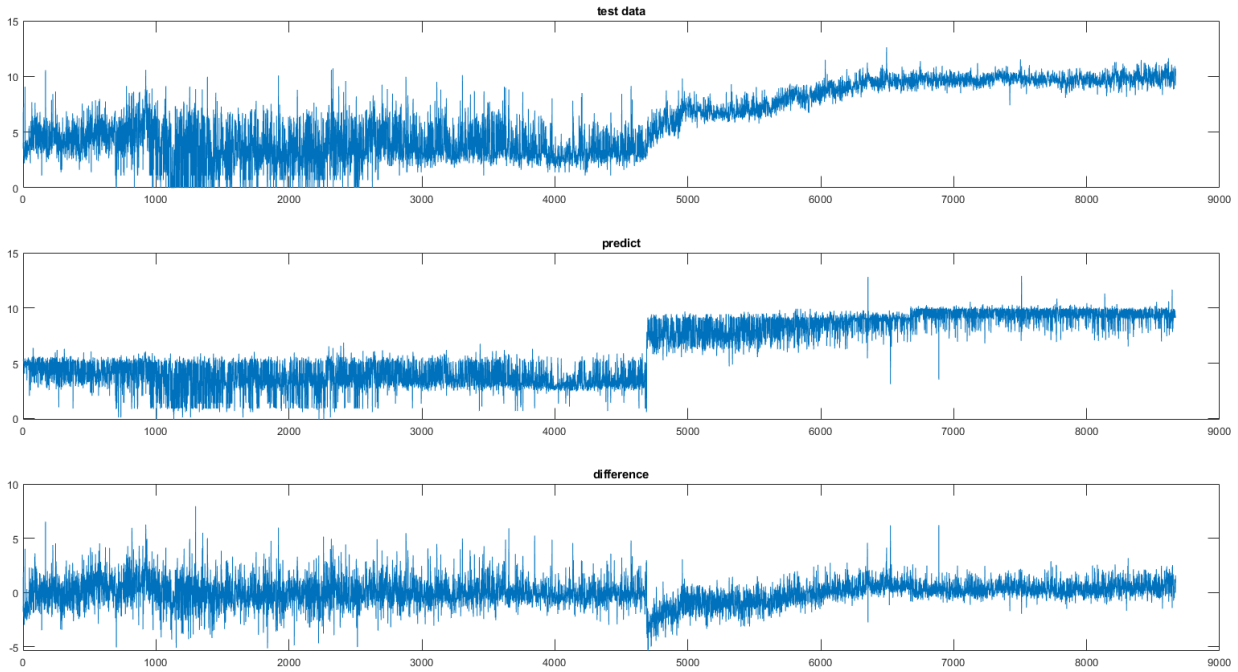
Po około czternastu godzinach przygotowywania danych (z czego praktycznie całość zajęło uzupełnianie wektora ze słowami kluczowymi), sieć zakończyła uczenie po osiągnięciu ustalonej liczby epok (1000).



Rysunek 1: wykres przedstawiający regresję liniową.

Na *rysunku 1.* wyraźnie widać efekt podzielenia danych na dwa zbiory zależne od daty zapostowania tweeta. Nasza regresja liniowa ma wzór:

$$output = 0,85 * target + 0,9.$$



Rysunek 2: wykresy obrazujące liczbę retweetów w skali logarytmicznej, kolejno od góry: dane testowe, dane przewidziane przez sieć, różnice między nimi. Wyraźnie widoczne załamanie spowodowane podzieleniem danych w zależności od daty.

W związku z ogłoszeniem startu w wyborach prezydenckich 16.06.2015 liczba retweetów zaczęła spektakularnie rosnąć w czasie, przez co same zawartości tweeta straciły na znaczeniu, co widać na *rysunku 2*. Brak podzielenia danych na dwie części mógł znacznie zaburzyć wyniki dla okresu, w którym Donald nie kandydował na stanowisko - jest to jednocześnie czas, w którym najbardziej widoczna jest korelacja między wartościami przewidzianymi, a prawdziwymi. Na środkowym wykresie widać jeszcze trzeci nagły, ale niewielki skok, z którym związane tweety nie zostały ręcznie oznaczone. Jest to moment, w którym Trump został prezydentem, przez co nastąpiła stabilizacja.

Wykres różnic między danymi wyjściowymi a testowymi załamuje się w miejscu podzielenia danych na daty, ale generalnie, zgodnie z oczekiwaniami, trzyma się on wartości zerowej. Średnia różnic wynosi około 0,0032.

Aby obliczyć precyzję nauczonej sieci, obliczyliśmy błąd względny dla każdego tweeta w tabeli:

$$error = \left|\frac{x'-x}{x}\right|$$, gdzie x' - wartość przewidziana, x - wartość prawdziwa.

Następnie uzyskaliśmy dokładności w postaci procentowej:

$$accuracy = 100\% - error * 100\%.$$

Sieć uczona była 5 razy, otrzymując następujące precyzje po uśrednieniu:

numer próby	średnia dokładność
1	78,6226%

2	79,1847%
3	79,2209%
4	78,8756%
5	77,9990%

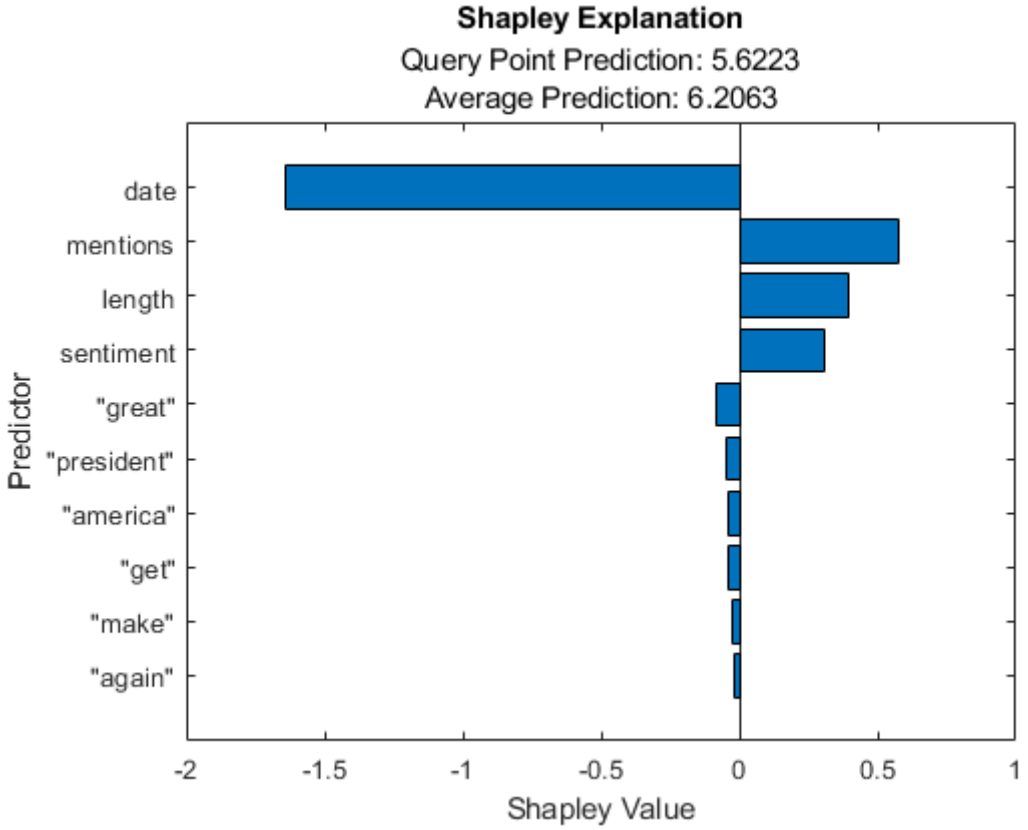
Tabela 2: uzyskane średnie precyzje nauczonych sieci.

Uśredniając te wartości, otrzymujemy w przybliżeniu 78,7806%. Można więc uznać, że precyzja naszego modelu jest na poziomie około 79%.

Analiza *SHAP*

SHAP oznacza *Shapley Additive Explanations*. Celem analizy jest wyjaśnienie przewidywania modelu uczenia maszynowego przez obliczenie wkładu każdej cechy w proces przewidywania, i polega na wydobyciu tak zwanych wartości Shapley'a wskazujących na udział odpowiednich im cech w procesie.

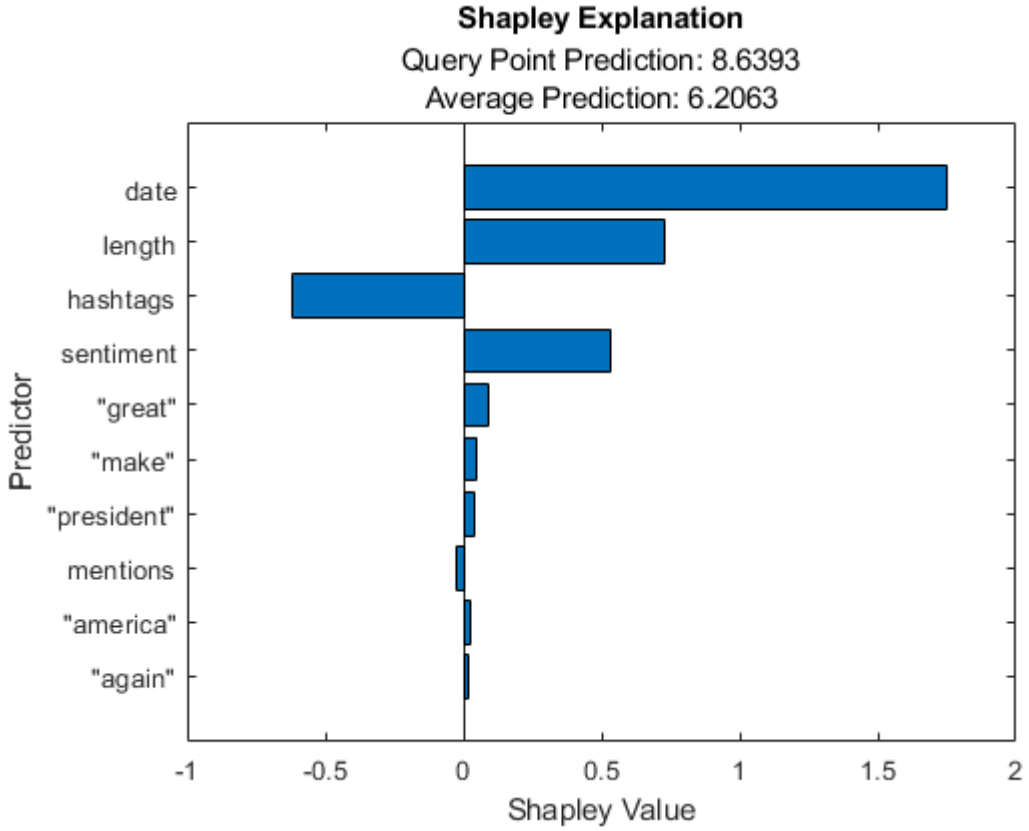
Na początku tworzymy model regresji liniowej za pomocą funkcji *fitrkernel* podając na wejściu dane testowe i zlogarytmowaną liczbę retweetów jako odpowiedź. Następnie tworzymy obiekt *shapley* i jako punkt zapytania wybieramy pierwszy tweet z bazy. Obliczamy odpowiednie dla niego wartości Shapley'a i rysujemy wykres.



Rysunek 3: wykres przedstawiający najistotniejsze wartości Shapley'a dla pierwszego tweeta.

Widzimy, że największy, i jednocześnie negatywny wpływ na ilość retweetów, ma podział na daty, co ma sens, bo Donald Trump nie kandydował wtedy jeszcze na

prezydenta i nie miał względnie dużej popularności. Pozytywnie na ilość podań dalej wpływają wzmianki, długości tweetów oraz sentyment, choć na wykres nie załapała się ilość hashtagów, której wartość jest zbyt mała, by być znacząca. Słowa kluczowe mają od nich nieco większy udział i brak przykładowo słowa "*great*" negatywnie wpływa na popularność tweeta.



Rysunek 4: wykres przedstawiający najistotniejsze wartości Shapley'a dla jednego z ostatnich tweetów.

Na *rysunku 4.* widać, że data zapostowania tweeta ponownie ma najwięsze znaczenie, tym razem pozytywne (Trump był wtedy prezydentem). Na znaczeniu zyskują hashtagi, jest to jednak wpływ negatywny. Może być to spowodowane faktem, że oznaczane są nimi tweety dotyczące jakiegoś nieciekawego procesu, natomiast tweety bez hashtagów mogły zawierać śmiałe i impulsywne stwierdzenia Donalda, które ludzie chętniej retweetowali. Wzmianki mają teraz dużo mniejszy udział, być może dlatego, że większość ludzi, których pan prezydent oznaczał, była już w tamtym momencie mniej popularna, niż on sam.