Raport z projektu

“Analiza wpisów o wirusie COVID-19 na portalu Twitter z wykorzystaniem narzędzi Data Science”

Nazwisko i Imię	Tyran Borys
Uczelnia	Politechnika Łódzka
Numer indeksu	216410
Wydział	Wydział Mechaniczny
Kierunek	Inżynieria Kosmiczna
Specjalność	Konstrukcje i Materiały
Semestr	V
Rok akademicki	2019/2020

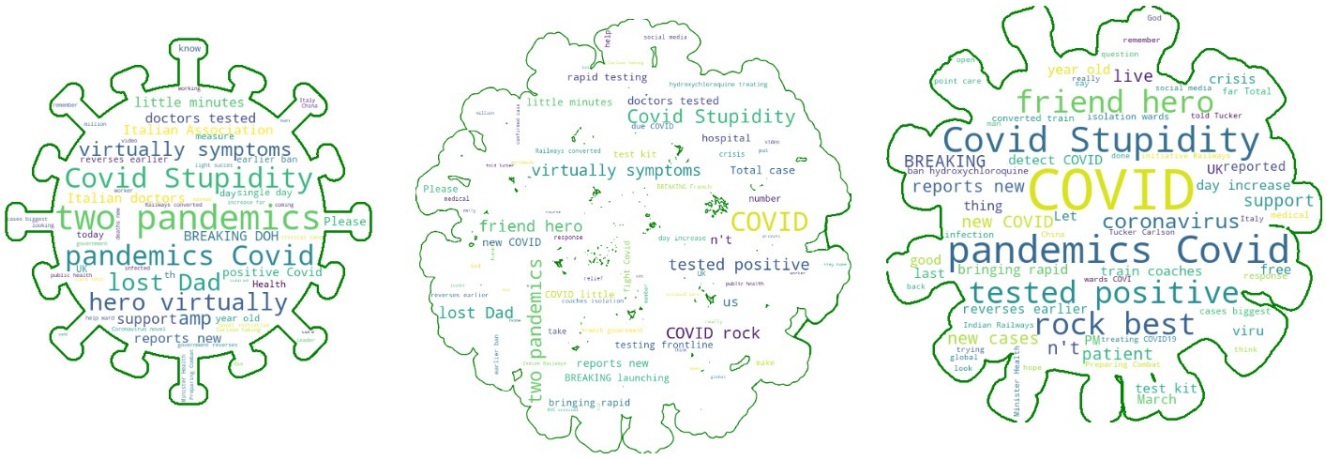
Streszczenie projektu

Analizie poddano wpisy zawierające kluczowe słowa o COVID-19, z których wykonano tzw. Wordcloudy (ang. Chmura słów) – graficzne reprezentacje najęczęściej powtarzających się słów, w kształcie wirusa COVID-19. Na wykresach przedstawiono główne metody używania serwisów oraz wykonano analizę sentymentu tweetów wykorzustując modele nauczania maszynowego.

Wyniki

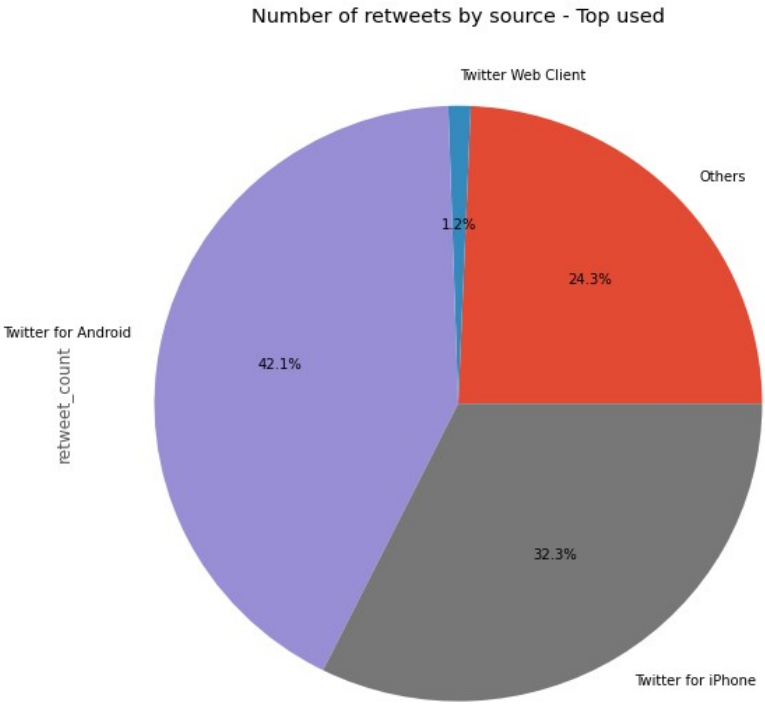
Analizę wykonano na 1803 tweetach z dnia 27.03.2020 (22:26) – 28.03.2020 (09:56).

Wordcloudy

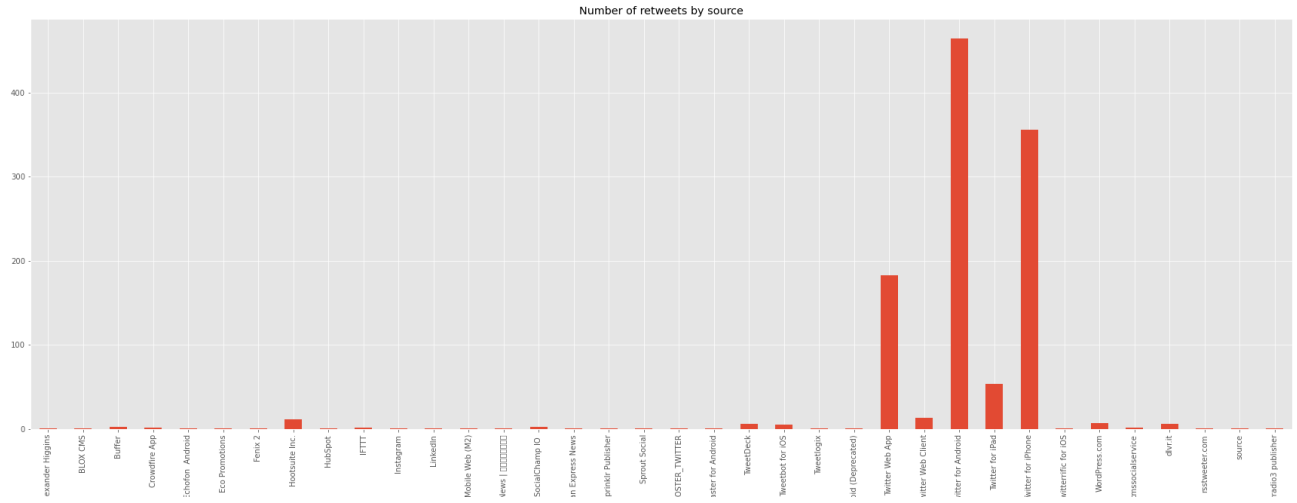


Ilustracja 1: Wygenerowane wordcloudy

Najczęściej używane platformy

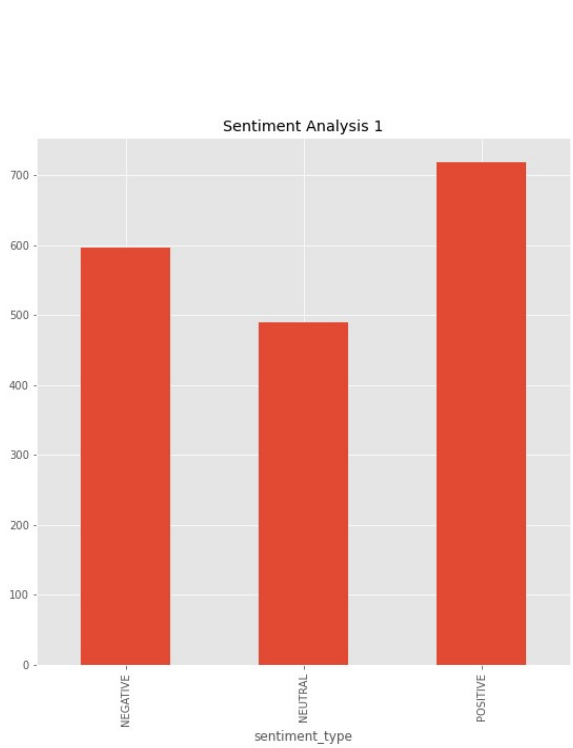


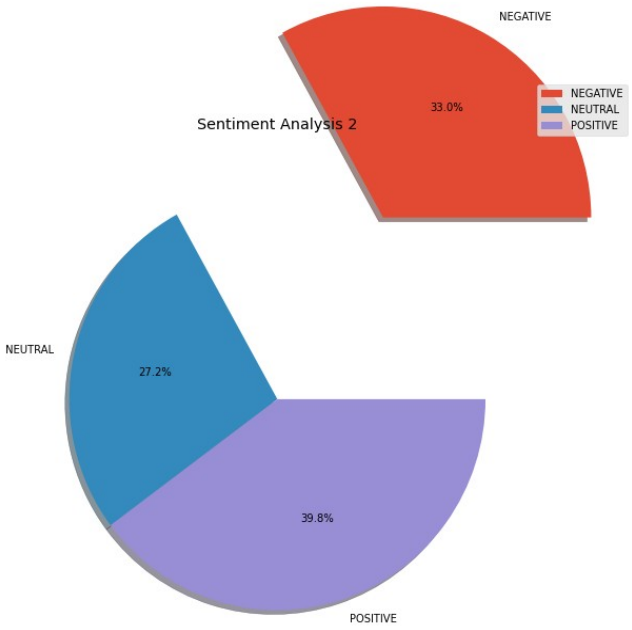
Wykres 1: Stosunek ilości "retweetów" do platformy, na której wpisy zostały wykonane – najczęściej używane



Wykres 2: Stosunek ilości "retweetów" do platformy, na której wpisy zostały wykonane – wszystkie zestawione platformy

Analiza sentymentu wpisów





Wykres 3: Analiza sentymentu wpisów, sprawdzająca jaka część wpisów miała charakter pozytywny, negatywny lub neutralny

Metodologia

Zebranie wpisów z Twittera

Wpisy z twittera zgromadzono przy użyciu API Twittera, wykorzystując w tym bibliotekę *Tweepy* (*http://www.tweepy.org/*) . Do otrzymania dostępu do danych Twittera, konieczne było założenie konta *Twitter developer* w wersji studenckiej. Ogranicznie zakłada ok max 2000 twittów dziennie, które można zebrać lub ok 200 przy pojedynczym zapytaniu (co pół godziny). Uzyskano 1803 tweetach z dnia 27.03.2020 (22:26) – 28.03.2020 (09:56).

Literatura i źródła wykorzystane w wykonaniu kodu:
[1] - https://bullseyestock.wordpress.com/2018/02/09/creating-a-twitter-application/
[2] - http://docs.tweepy.org/en/v3.5.0/cursor_tutorial.html

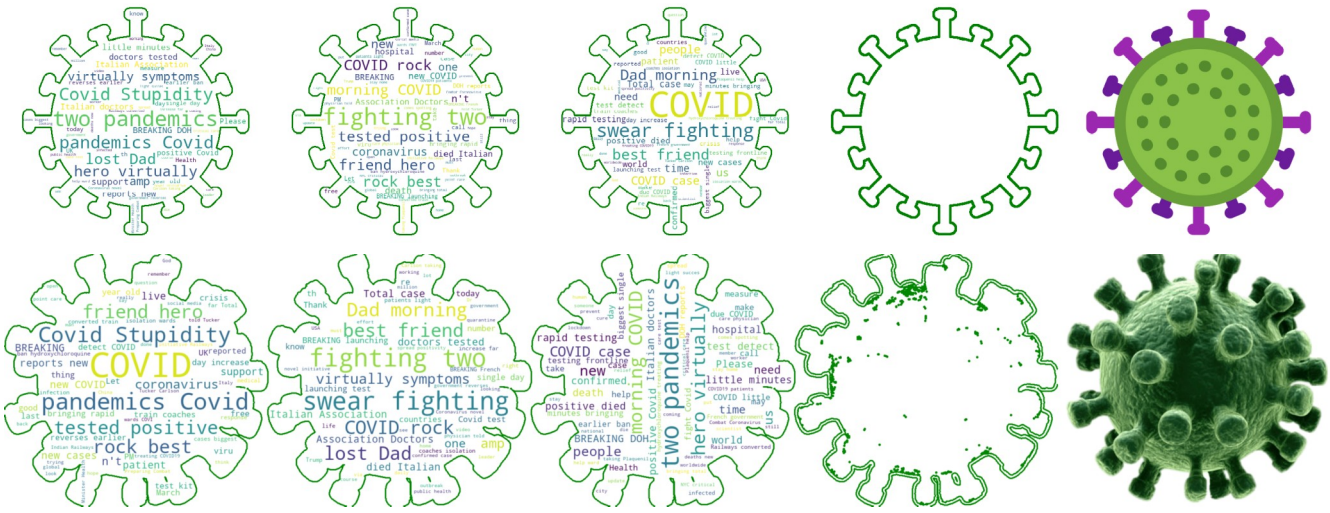
Zapisanie danych w postaci .csv ich pre-processing

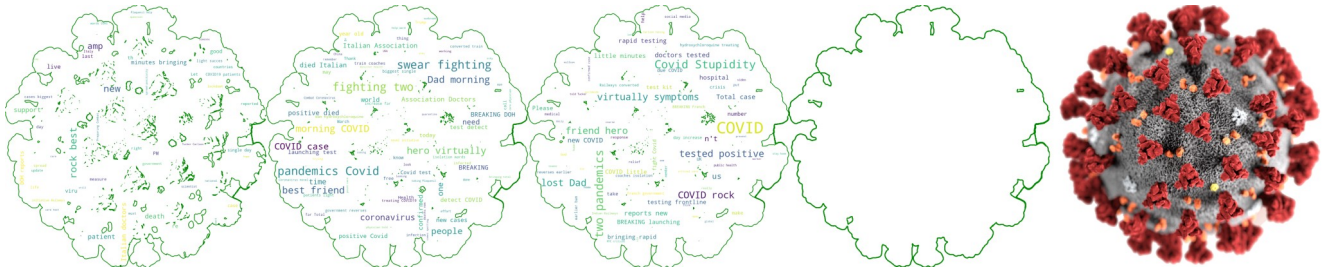
Wpisy do ułatwienia pracy z postaci .json zapisano do formatu .csv co pozwoliło na stworzenie struktury *dataframe* przy użyciu biblioteki *pandas.* Następnie, na tak stworzonej strukturze, przeprowadzono pre-processing – czyli “oczyszczono” np. teksty z niepotrzebnych w analizie słów tzw. *stopping words* lub z linków URL wyciagnięto platformy, na których zrobiono „retweety” świadczących o platformie z jakiej użytkownik korzystał. Wykorzystano biblioteki *nltk (natural language tool kit), preprocessor* oraz *textblob*.

[3] - https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf

Wordcloudy

Do wygenerowania wordcloudów, wykorzystano bibliotekę *wordcloud* oraz *numpy* do stworzenia „masek” czyli kontur, w których dane wordcloudy miały się stworzyć.





Ilustracja 2: Wygenerowane wordcloudy koronawirusa - CoronaClouds

Generowano od razu 3 wordcloudy – co jest raczej w tym przypadku „bug'iem". Załadowane pliki .png koronawirusa do postaci macierz przy pomocy biblioteki *numpy* odczytywane były w trzech wymiarach. Znaleziono rozwiązanie na *stackoverflow* jak rozwiązać problem z transformacją macierzy ale poskutkowało to uzyskaniem jednocześnie trzech wordcloudów jednocześnie. Dla porównania dano obrazek wykorzystany przy generacji.

[4] - https://www.datacamp.com/community/tutorials/wordcloud-python

Zrzucenie na wykres najbardziej popularnych platform Twitter'a

Główne metody używania serwisów przedstawiono w postaci wykresu w stosunku ilości "retweetów" do platformy, na której wpisy zostały wykonane. Na wykresie 1 znajdują się wszystkie wykorzystane platformy, z kolei na wykresie 2 zgrupowane zostały pozostałe, o mniejszym użytku serwisy (np. Instagram). Do wygenerowania wykresów wykorzystano bibliotekę *matplotlib.*

[5] - http://queirozf.com/entries/pandas-dataframe-plot-examples-with-matplotlib-pyplot

Analiza sentymentu wpisów

Wykonano również analizę sentymentu wpisów wykorzustując modele nauczania maszynowego biblioteki *scikit*. Na podstawie tej analizy stworzono dwa wykresy, pokazujące procentowy odsetek wpisów o charakterze pozytywnym, negatywnym lub neutralnym.

[6] - https://datascienceplus.com/twitter-analysis-with-python/

Wnioski

Główne słowa, które padają w wordcloudach oprócz COVID czy wirus to frazy o testach, śmierci, powiązanych z Włochami, lekarzach, symptomach, szpitalach ale nawet i o prezydencie stanów zjednoczonych, Donaldzie Trumpie.

Sama forma przedstawienia najczęściej padających słów w postaci wordcloud'a, jest już rzadziej stosowana w artykułach (mniej modna na aktualny czas) ale ciekawa do wykonania.

Odnosząc się do analizy platform, spodziewano się, że najpopularniejsze będą oficjalne metody użytkowania serwisu – czyli aplikacje na Androida (telefon), Iphone’a oraz oficjalna strona. Można jednak zauważyć, że inne metody (Instagram, tablety) to wciąż duży odsetek użytkowników i konieczne jest wspieranie tych platform.

W przypadku analizy sentymentu wpisów można zauważyć, że pomimo pandemii, przeważa część pozytywna wypowiedzi. Należy jednak wziąć pod uwagę, że negatywny tweet może odnosić się nie do wirusa a np. prezydenta danego kraju i jego działania na rzecz wirusa. Z kolei pozytywny może odnosić się do pozytywnego wpływu wirusa np. na oczyszczenie powietrza lub euforii z pracy zdalnej.

Podsumowując, należy jednak pamiętać, że analiza danych została wykonana na dość niskiej populacji (1803) wpisów. Większa i bardziej dokładne zobrazowanie rzeczywistości wymagałoby dostępu do większej ilości danych – znacznie przewyższającej udostępniane zasoby studentom. Konieczna wtedy byłaby współpraca z Twitterem lub pobranie danych w sposób nie korzystających z udostępnianego API (np. *web scrapping)* – lecz takie działanie mogłoby być blokowane przez Twittera.