3 Analiza tweetów z okresu pandemii COVID-19

Pandemia koronawirusa SARS-COVID-19, trwająca na świecie w latach 2019-2023, przyczyniła się do zmiany trendów i zachowania społeczeństw na całym świecie. Wirus zmusił rządy do podjęcia restrykcji, by ograniczyć skalę zachorowań i hospitacji. Sformułowane nakazy wymuszały pobyt w domu pracowników wszystkich branż rynkowych, uruchamiały specjalne godziny zakupowe dla osób starszych oraz wnosiły obowiązek noszenia maseczek ochronnych w miejscach publicznych. Przedsiębiorstwa musiały zwiększyć wydatki na środki dezynfekcyjne dla swoich pracowników i klientów. Ponadto ustalono maksymalną liczbę osób przebywających w budynkach użyteczności publicznej. Niejasnym w tamtym czasie dla społeczeństw była również kwestia nowopowstałych szczepionek. Wówczas w szczególności uwidoczniła się niechęć do ich przyjmowania, niezależnie od producenta.

Zmiany, jakie wywołał koronawirus na świecie, przełożyły się w dużym stopniu na ukształtowanie opinii publicznej w stosunku do tej choroby. W szczególności platformy społecznościowe zdominowane były wypowiedziami użytkowników na temat pandemii. Jedną z takich platform był Twitter.com, gdzie pojawiły się niezliczone ilości wpisów na ten temat. Takie wypowiedzi stanowią bardzo dobre źródło danych do analizy. Należy natomiast mieć świadomość, że takie hurtownie danych są w postaci nieustrukturyzowanej, więc wymagają procesu strukturyzacji (ang. *preprocessing),* aby móc wyciągnąć trafne wnioski. Wówczas wartym jest wykorzystanie platformy Apache Spark, gdyż jest dostosowana do danych typu Big Data oraz względnie przyspieszy przetwarzanie danych. Ważnym aspektem jest również wybór odpowiedniego środowiska programistycznego. Na uwagę zasługuje język R, w którym możliwe jest efektowne procesowanie danych w celach analitycznych.

W podjętej pracy zostanie przeprowadzona analiza tekstowego zbioru danych wraz z wizualizacją wyników. Następnie możliwym będzie przedstawienie wniosków końcowych. Cele pracy dyplomowej można podzielić na dwa zasadnicze działy tematyczne. Pierwszym z nich jest wykorzystanie silnika Apache Spark w środowisku języka R tak, by ukazać zalety użycia platformy dla Big Data. Drugim celem jest zbadanie opinii i sentymentu użytkowników platformy społecznościowej Twitter.com na temat

pandemii COVID-19 w okresie jej początków. Dane do analizy zostaną pobrane z platformy Kaggle.com, będącej częścią naukowego działu korporacji Google LLC. Firma ta bowiem udostępnia zebrane przez siebie hurtownie danych na różne tematy do badań dla naukowców i analityków.

Obrana hurtownia danych będzie dotyczyć wypowiedzi użytkowników platformy Twitter.com na temat pandemii COVID-19. Okres zebranych obserwacji waha się pomiędzy 25 lipca, a 29 sierpnia 2020 roku, czyli z okresu początku pandemii. Struktura danych jest następująca:

- obserwacje dzielą się na 13 zmiennych, z czego jedna z nich dotyczy komentarzy użytkowników Twittera;
- Pozostałe z nich zawierają daty, liczby oraz lokalizacje użytkowników;
- Liczba wszystkich nieprzetworzonych obserwacji dla każdej zmiennej jest równa 179 108.

Do analizy Big Data istotnym jest przedstawienie każdej zmiennej tak, aby lepiej poznać posiadane dane oraz móc zaplanować wstępne badania. Poniżej znajduje się opis trzynastu zmiennych z posiadanych danych:

- **user_name** – nazwa lub pseudonim użytkownika zamieszczającego wypowiedź (zmienna tekstowa);
- **user_location** – nazwa miejscowości, w której przebywał użytkownik w trakcie pisania wypowiedzi (zmienna tekstowa);
- **user_description** – opis użytkownika (zmienna tekstowa);
- **user_created** – data stworzenia konta w serwisie społecznościowym (zmienna typu DateTime);
- **user_followers** – popularność użytkownika w społeczności internetowej (zmienna ilościowa);
- **user_friends** – liczba znajomych danego użytkownika (zmienna ilościowa);
- **user_favourites –** liczba polubień konta na platformie Twitter.com (zmienna ilościowa);
- **user_verified** – zweryfikowanie konta przez użytkownika (zmienna binarna typu TRUE or FALSE)*;*
- **date** – data zamieszczenia wypowiedzi (zmienna typu DateTime);

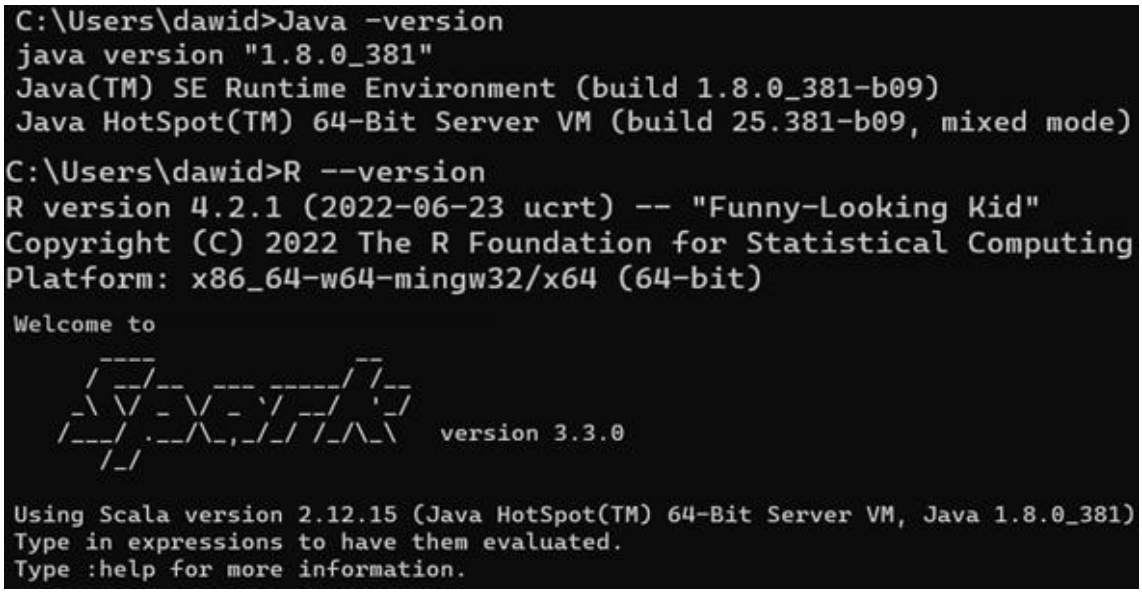
- **text** – wypowiedź zamieszczona przez użytkownika platformy Twitter (zmienna tekstowa);
- **hashtags** – hasztagi tematyczne zamieszczone do wypowiedzi (zmienna tekstowa ze znacznikiem „#”);
- **source –** oprogramowanie urządzenia, z jakiego została zamieszczona wypowiedź (zmienna tekstowa);
- **is_retweet** – zmienna określająca, czy wypowiedź została udostępniona przez innego użytkownika (zmienna binarna typu TRUE or FALSE).

Przedstawione powyższe informacje umożliwiają przejście do planowania strukturyzacji posiadanych danych. Omawiając poszczególne kroki, na samym początku hurtownia przejdzie proces oczyszczania obserwacji z wybrakowanych danych oraz znaków interpunkcji. Ujednolicona zostanie wielkość liter oraz rozpocznie się podział zdań na osobne tokeny (zastosowanie tokenizacji). Następstwem będzie dotarcie do rdzenia każdego wyrazu (czyli jego podstawowej formy) dla danych tekstowych (stemming i lematyzacja). Wówczas można podjąć usunięcie słów zatrzymania, zaś ostatnim etapem będzie wektoryzacja danych tekstowych na liczbowe. W przypadku pozostałych zmiennych, zostaną one użyte jako wsparcie analizy sentymentalnej.

Początek procesowania i analizy danych to skompletowanie wymaganych narzędzi. Stąd też niezbędna jest najnowsza wersja języka R wraz z środowiskiem RStudio. Pobrana dodatkowo zostaje platforma Apache Spark ze strony głównej rozszerzenia *sparklyr*. Następnie, w systemie Windows tworzona jest ścieżka w panelu zmiennych środowiskowych, tak aby język R mógł wykryć pakiet *sparklyr*. Po tym procesie należy upewnić się, czy posiadana wersja środowiska Java jest kompatybilna z silnikiem Sparka. Wynika to z faktu transponowania kodu przez Java Virtual Machine.

Paleta narzędzi, w podjętej pracy, posiada następujące wersje oprogramowania:

- Java version: 1.8.0_381 for 64-Bit Server VM;
- R language version: 4.2.1 (2023-06-23 ucrt);
- RStudio IDE version: 2023.06.2+561;
- Spark version 3.3.0 using Scala version 2.12.15 [Java 1.8.0_381].



Rysunek 26. Wersje kontrolne technologii wykorzystanych do procesowania i analizy danych; Źródło: Opracowanie własne.

Pierwszym w kolejności, uruchomionym narzędziem jest środowisko języka R wraz z RStudio. W nim bowiem pobiera się i instaluje platformę Apache Spark. Dzięki stworzonej ścieżce zmiennych środowiskowych, proces odbywa się za pośrednictwem dwóch wierszy: instalującego środowisko z CRAN-u języka R oraz wczytującego w obszar RStudio.

```
#Instalacja oraz wczytanie silnika do R
install.packages("sparklyr")
library("sparklyr")
```

Technologia Java, odpowiadająca za wirtualną maszynę dla Apache Spark, jest automatycznie włączana po wgraniu platformy do języka R. Posiadając skompletowane narzędzia, można rozpocząć połączenie w Apache Spark oraz wczytać dane do środowiska roboczego.

Dla platformy Apache Spark w omawianej pracy początkowo ustawiane jest połączenie lokalne, czyli bez wykorzystania zewnętrznego menedżera klastra. W tym wypadku działa wbudowany system zarządczy – Standalone. Również liczba węzłów wykonawczych jest określona na wartość jednego; domyślnie dla połączenia lokalnego. Przejście na połączenie URL i większą ilość węzłów wykonawczych zostanie

ustanowione w momencie pojawienia się błędów natury optymalizacyjnej. Istotna z perspektywy prawidłowego działania Sparka jest odpowiednia konfiguracja wstępna. Wykonuje się ją przed rozpoczęciem połączenia. Dla rozpoczęcia prowadzonych badań wstępnym parametrem jest ustawienie zużycia pamięci programu sterującego na wartość 4 GB. Następnie możliwe jest nawiązanie połączenia Apache Spark z środowiskiem języka R. Od tego momentu użytkownik ma wgląd do monitoringu Spark UI, dostępnego poprzez adres połączenia lokalnego.

```
# ** TWORZENIE POŁĄCZENIA LOKALNEGO SPARKA W R **
config <- spark_config()
config$spark.driver.memory <- "4g"
sc <- spark_connect(master = "local", config = config)
spark_connection_is_open(sc)
```

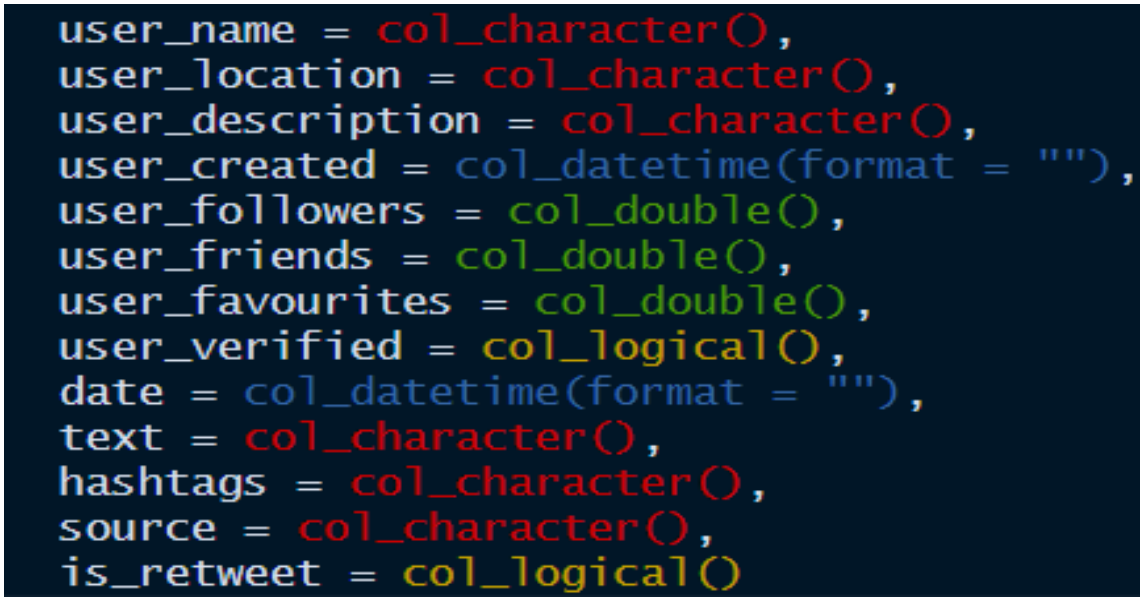
Po ustanowieniu połączenia, posiadaną hurtownie danych należy dostarczyć do środowiska programistycznego. Z powodu występowania błędów kompilacyjnych dla Spark SQL, wczytanie posiadanych danych nie jest możliwe poprzez silnik Sparka. Należy najpierw wczytać dane do środowiska RStudio, gdzie przyjmą formę ramki danych. Następnie takie dane należy przetransportować do wnętrza silnika Apache Spark. Posłuży do tego funkcja *copy_to(...)*, która umieszcza dane w stworzonym klastrze.

```
# ** WCZYTANIE TWEETÓW DO R-STUDIO Z ODPOWIEDNIM TYPEM**
tweety_Spark2 <- read_csv(choose.files(),
                          col_types = cols(user_name = col_character(),
                                           user_location = col_character(),
                                           user_description = col_character(),
                                           user_created = col_datetime(format = ""),
                                           user_followers = col_double(),
                                           user_friends = col_double(),
                                           user_favourites = col_double(),
                                           user_verified = col_logical(),
                                           date = col_datetime(format = ""),
                                           text = col_character(),
                                           hashtags = col_character(),
                                           source = col_character(),
                                           is_retweet = col_logical()))

#Przetrasportowanie danych do Apache Spark
tweety_Spark <- copy_to(sc, tweety_Spark2, "tweety_Spark")
```

Rysunek 27. Widok wczytania hurtowni danych do środowiska RStudio wraz z jego przeniesieniem do klastra Apache Spark; Źródło: Opracowanie własne.

Po wczytaniu danych warto sprawdzić, czy wszystkie dane zostały ustawione prawidłowo. Pozwoli to uniknąć ewentualnych błędów w dalszych etapach pracy.



Rysunek 28. Kontrola typów wczytanych zmiennych do środowiska Apache Spark; Źródło: Opracowanie własne.

Na końcu konfigurowania narzędzi analitycznych można skupić się na wczytaniu pierwszych rozszerzeń. Jednym z przykładów jest użycie biblioteki *dplyr*, kompatybilnej z Apache Spark oraz powszechnie stosowanej do procesowania danych w R.

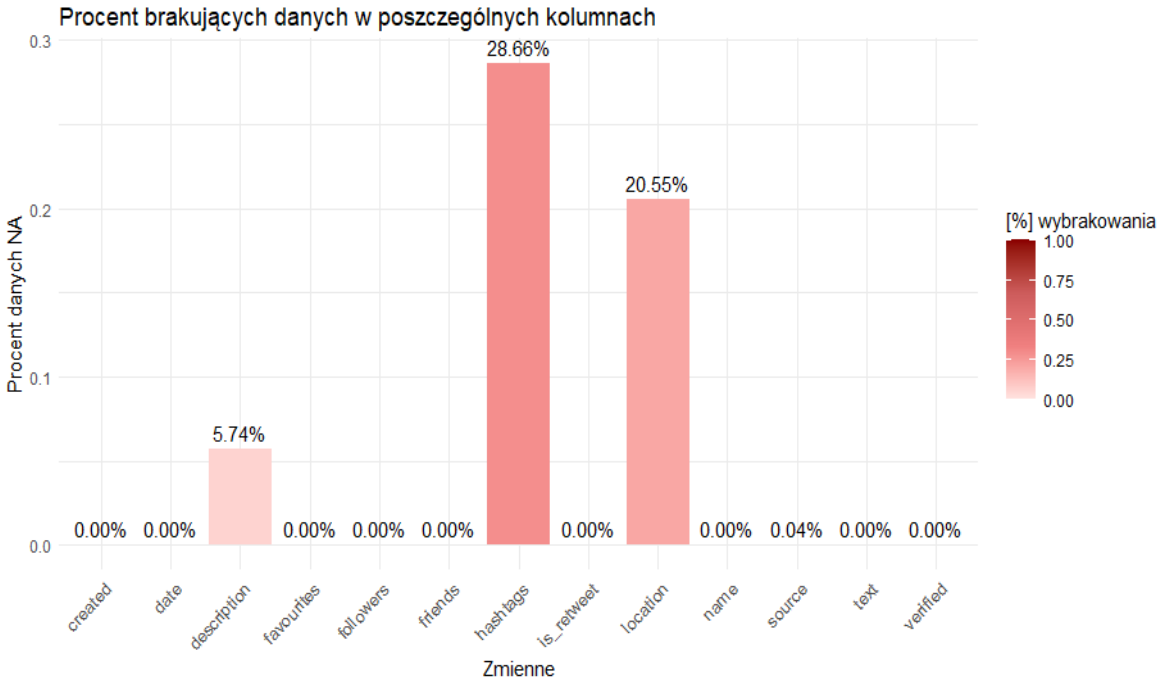
3.1 Opis preprocessing-u zbioru danych

Posiadając odpowiednio ustawione narzędzia, można przejść do procesu oczyszczania danych. Wartym uwagi jest to, że każda zmienna w hurtowni zostanie w tym etapie ustrukturyzowana, niezależnie od jej użyteczności. W następnych bowiem krokach zostaną wybrane najważniejsze zmienne z perspektywy prowadzonych analiz. W szczególności ważna będzie zmienna *text*, zawierająca opinie użytkowników na temat COVID-19.

W pierwszej kolejności rozpoznana zostanie struktura wybrakowanych danych względem całości hurtowni. Jest to problematyczny proces na wzgląd wykorzystania Sparka w środowisku R. W podstawowym podejściu i przy stosunkowo niedużej bazie danych, można skorzystać z wywołania *summarise_all(...)*, celem szybkiego otrzymania struktury wybrakowanych danych. Apache Spark w tym wypadku generuje problemy natury transponowania kodu z języka Scala i ukazuje się związany z nim błąd. Próbą zaradzenia może być przetransportowanie danych z klastra Sparka do ramki danych R i procesowanie danych bez platformy Apache Spark. Jednakże w przypadku Big Data może wiązać się to z długotrwałym procesem obliczeniowym. Innym, korzystniejszym rozwiązaniem z perspektywy obliczeniowej, jest pozostawienie danych w klastrze Sparka. Następnie, przy pomocy biblioteki *dplyr* można wyeliminować problem transponowania kodu źródłowego poprzez napisanie własnej funkcji, zliczającej braki w danych. Spark ma możliwość szybkiego stosowania nowych funkcji do danych, o ile są one oparte o kompatybilne rozszerzenia.

```
# Funkcja zliczająca
count_na_values <- function(dataframe, column) {
  na_count <- dataframe %>%
    filter(is.na(rlang::sym(column))) %>%
    sdf_nrow()
  na_count
}
```

Wywołanie funkcji jest następnie stosowane do każdej ze zmiennych. Pozwala to na uzyskanie informacji o wybrakowanych danych.



Rysunek 29. Wykres słupkowy procentu wybrakowanych obserwacji w zmiennych; Źródło: Opracowanie własne.

Za pośrednictwem stworzonej wizualizacji (rys. 29) możliwym jest wyciągnięcie informacji o procentowym odsetku brakujących danych w hurtowni:

- najmniejszym wybrakowaniem cechuje się zmienna *source* (0,04% obserwacji brakujących);
- największym wybrakowaniem cechuje się zmienna *hashtags* (28,66% obserwacji brakujących);
- dziewięć zmiennych nie posiada wybrakowanych obserwacji;
- średnie wybrakowanie w całości hurtowni danych wynosi ~ 4%.

W ukazanej strukturze na rysunku 29. należy podjąć próbę estymacji z przedstawionych braków. Obranie właściwej strategii jest kluczowe z perspektywy informacji końcowych. Dodatkowo, aby ukazać możliwości optymalizacyjne silnika Apache Spark w połączeniu z RStudio, każda zmienna będzie posiadać inną metodę oszacowania nieznanych danych.

Zmienna *source* jest zmienną kategoryczną. Oznacza to, że cechy w niej istniejące posiadają określone propozycje wartości. Pierwotną strategią oszacowania wybrakowanych danych była estymacja poprzez regresje liniową. Natomiast podczas próby jej zastosowania, kompilator wywołuje błąd o zbyt dużej liczebności wag – powyżej 3 000 kategorii. Próby podjęcia innych technik szacunkowych, opartych na losowaniu zmiennych tekstowych wraz z podzespołem platformy MLlib do uczenia maszynowego, kończyły się błędem transponowania kodu R na Spark SQL. Wynika to z faktu braku kompatybilności Sparka z metodami losowania, dostępnymi w RStudio. Ostatecznym rozwiązaniem całości problemu było oparcie estymacji o pakiet *dplyr*. Obrano wskaźnik, jakim jest dominanta, czyli najczęściej występująca kategoria w całości danych. Uzasadniając decyzję wyboru tej wartości, warto nadmienić iż, nie zaburzy ona realnych wyników końcowych ze względu na małe wybrakowanie w zmiennej *source*. Wynosi ono bowiem 0,04% w stosunku do całości danych, czyli 77 brakujących obserwacji.

Zmienna *description* jest zmienną tekstową i zawiera około 6% brakujących danych w stosunku do całości (dokładna liczba: 10 283 obserwacji nieznanych). Strategia estymacyjna to uzupełnienie opisu użytkownika, poprzez sklasyfikowanie go jako niezdefiniowanego (ang. *undefined*). Uzasadnieniem jest fakt, iż szacowanie takiej wartości jest procesem złożonym i niejednoznacznym. Zmienna ta bowiem w dużej mierze zależy od indywidualnej predyspozycji każdego użytkownika do autocharakterystyki. Prawdopodobnym faktem jest również niewykorzystanie tejże zmiennej w kontekście założonych celów badawczych dla prowadzonej pracy. Jednak potrzeba estymacji wynika z faktu próby ustrukturyzowania całej hurtowni danych oraz z możliwości przedstawienia platformy Apache Spark w tematyce optymalizacji kodu kompilowanego w języku R.

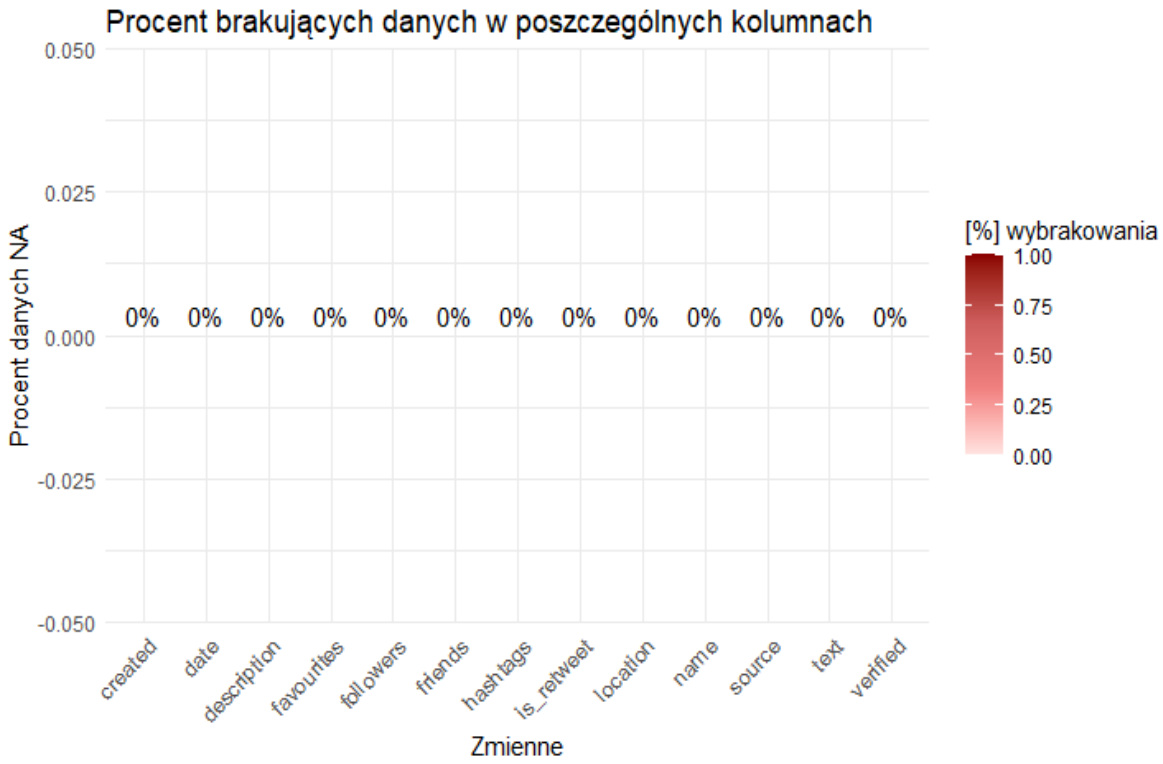
Zmienna *location* to zmienna tekstowa, gdzie użytkownicy mogą wpisywać dowolne frazy. Posiada ona 21% brakujących danych (dokładna liczba: 36 804 obserwacji nieznanych). Początkowo owe dane zostaną zmienione na ramkę danych R. Działanie to jest motywowane, ze względu na istotę tej zmiennej w kontekście późniejszych analiz, a także mając na uwadze częste problemy z transponowaniem kodu w silniku Apache Spark. Jednak po zakończonym procesie oczyszczania braków powrócą do Sparka jako Spark DataFrame. Strategią estymacji w środowisku R jest podział danych na grupy po

10 obserwacji, na podstawie których zostanie oszacowana brakująca lokalizacja. Należy przy tym podjąć kryteria szacunkowe tak, aby:

- uniknąć przypadku stworzenia grupy posiadającej same wybrakowane informacje;
- uniknąć przypadku stworzenia grupy z przeważającą liczbą nieznanych lokalizacji.

Pomocnicze będą obserwacje najczęściej występujące oraz odpowiednie rozłożenie stosunku braków w grupach szacunkowych.

Zmienna *hashtags* to zmienna tekstowa, gdzie użytkownicy mogą wybierać dowolne słowa jako oznaczenie tematu wypowiedzi. Zawiera ona najwięcej brakujących danych – na poziomie ~ 29% (dokładna liczba: 51 334 obserwacji nieznanych). Jest to również problematyczna zmienna, podobnie jak wcześniej wspomniane. Opiera się bowiem na indywidualnej tendencji do wyboru słów przez użytkownika. Usunięcie takich zmiennych byłoby jednak zbyt drastycznym posunięciem, ze względu na dużą redukcję informacji z pozostałych zmiennych. Dlatego należy rozważyć użyteczność takiej zmiennej w kontekście założonych celów badawczych. W przypadku analizy tekstowej może zostać wykorzystana jako podstawa do badania tematyki komentarzy użytkowników. Jednak w opozycji do tego stwierdzenia jest fakt, iż znana jest tematyka danych (pandemia koronawirusa) oraz metody tworzenia hasztagów na platformie społecznościowej Twitter.com. Późniejsze wykorzystanie wspomnianej zmiennej może być znikome w stosunku do pozostałych danych. Stąd też obrana strategia to wykorzystanie słów kluczowych jako uzupełnienie braków. Słowa te będą pochodzić z podliczenia najczęściej przejawianych tematów, dla których następnie zostanie stworzona lista słów kluczowych. Uzupełnienia będą się odbywać poprzez losowanie ważone (im większa częstotliwość występowania danego tematu, tym częstsze prawdopodobieństwo jego wylosowania). Podjęcie tej strategii wiąże się z potrzebą przeniesienia zmiennej do ramki danych R. Po dokonanym losowaniu powrócą jednak do Spark DataFrame.



Rysunek 30. Wykres słupkowy procentu wybrakowanych danych po procesie ich uzupełnienia; Źródło: Opracowanie Własne.

Wykres, ukazujący procentowy udział wybrakowanych danych (rys. 30), jednoznacznie sugeruje sukces w procesie uzupełniania zmiennych. Takie dane są gotowe do przeprowadzenia analizy właściwej. Niemniej jednak główna zmienna *text* wymaga dalszego procesowania, celem otrzymania prawidłowej struktury wyrazów do analizy sentymentalnej. Stąd też następne kroki będą opisywać przekształcenia tej zmiennej.

Opracowanie zmiennej *text* rozpocznie się od procesu tokenizacji. Zdania zostaną podzielone na listy wyrazów, w których to dojdzie do eliminacji słów małoistotnych (proces usuwania słów zatrzymań). Pomocne na tym etapie będą wbudowane narzędzia Apache Spark do procesowania danych tekstowych. Za pośrednictwem funkcji *ft_tokenizer()* oraz *ft_stop_words_remover()* można przekształcić dane do właściwej struktury.

	text	word_list	wo_stop_words
1	If I smelled the scent of hand sanitizers today on someone i...	list("if", "i", "smelled", "the", "scent", "of", " [...]	list("smelled", "sce
2	Hey @Yankees @YankeesPR and @MLB wouldn t it have m...	list("hey", "@yankees", "@yankeespr", "and", "@mlb [...]	list("hey", "@yank
3	@diane3443 @wdunlap @realDonaldTrump Trump never o...	list("@diane3443", "@wdunlap", "@realdonaldtrump", [...]	list("@diane3443"
4	@brookbanktv The one gift COVID19 has give me is an app...	list("@brookbanktv", "the", "one", "gift", "", "co [...]	list("@brookbankt
5	25 July Media Bulletin on Novel CoronaVirusUpdates CO...	list("25", "july", "", "", "media", "bulletin", "o [...]	list("25", "july", "",
6	coronavirus covid19 deaths continue to rise It s almost as...	list("", "coronavirus", "", "covid19", "deaths", " [...]	list("", "coronaviru
7	How COVID19 Will Change Work in General and recruiting ...	list("how", "", "covid19", "will", "change", "work [...]	list("", "covid19", "
8	You now have to wear face coverings when out shopping t...	list("you", "now", "have", "to", "wear", "face", " [...]	list("wear", "face",
9	Praying for good health and recovery of @ChouhanShivraj ...	list("praying", "for", "good", "health", "and", "r [...]	list("praying", "go
10	POPE AS GOD Prophet Sadhu Sundar Selvaraj Watch here ...	list("pope", "as", "god", "", "", "prophet", "sadh [...]	list("pope", "god",

	word
1	smelled
2	scent
3	hand
4	sanitizers
5	today
6	someone
7	past
8	think
9	intoxicated
10	that...

Rysunek 31. Prezentacja przemiany zdań w pojedyncze, wyczyszczone słowa za pośrednictwem procesu tokenizacji i usunięcia słów zatrzymiana; Źródło: Opracowanie własne.

Jak przedstawiono na rysunku 31. zastosowane procesy przyniosły oczekiwany skutek. Wynikiem jest lista wszystkich istotnych słów z perspektywy analizy sentymentu. Ze 179 108 zdań powstało 2 231 682 słowa w procesie tokenizacji, zaś po selekcji ważnych słów pozostało 2 040 324 wyrazy. Wartym uwagi jest również sam silnik Apache Spark i jego wyniki procesowania w czasie. Przedstawia to poniższa tabela 2, która prezentuje wycinek interfejsu Spark UI dla poszczególnych zadań oraz ich czas wykonania.

Tabela 2. Wycinek tabeli kontrolnej w Spark UI odnośnie czasu pracy Sparka w poszczególnych zadaniach; Źródło: Opracowanie własne.

91	count at <unknown>:0 count at <unknown>:0	**Proces zliczenia wszystkich słów**	2023/09/17 16:45:42	11 s
90	collect at utils.scala:26 collect at utils.scala:26	**Usuwanie stop word'ów**	2023/09/17 16:45:00	0,2 s
89	collect at utils.scala:26 collect at utils.scala:26	**Tokenizacja zdań**	2023/09/17 16:43:29	0,2 s

Środowisko RStudio w trakcie pracy na danych tekstowych oferuje gotowe zestawienie funkcji, potrzebnych w dalszym etapie przetwarzania danych. Mowa tu o procesie stemmingu oraz lematyzacji. Uzyskane wyrazy po tokenizacji można więc od razu sprowadzić do znaczenia podstawowego (nieodmienionego).

token_id	token	lemma
7	has	have
8	is	be
9	me	I
10	things	thing

Rysunek 32. Przemiana tokenów w procesie stemmingu i lematyzacji; Źródło: Opracowanie własne.

Powyższy wynik transformacji tokenów do form podstawowych (rys. 32) sugeruje powodzenie całości procesu. Zmienna *text* jest przygotowana do procesów analitycznych, związanych z sentymentem. Warto podjąć ostatni krok procesowania, jakim jest wektoryzacja tej zmiennej. Nie we wszystkich funkcjach do metod analitycznych znajdzie ta zamiana zastosowanie, lecz pozwoli to zrealizować niektóre założenia na ramki danych w metodach słownikowych. Ponadto przejście na wartości liczbowe umożliwi przeprowadzenie analiz pomocniczych wspomnianych w podrozdziale 1.1.3 (analiza n-gramów oraz analiza skupień). Wektoryzacja zmiennej *text* zostanie przeprowadzona za pomocą metody Częstotliwości Słowa (TF-IDF). Wybrana metoda bowiem, oprócz zamiany tekstu pisanego na formę numeryczną, bada częstotliwość występowania danego słowa w całości tekstu.

Wszystkie podjęte dotąd czynności, sprowadzają hurtownie danych do możliwości przeprowadzenia właściwych analiz. Od tego momentu możliwe będzie przeprowadzenie weryfikacji danych dla celów prowadzonej pracy:

- przedstawienie zastosowania i korzyści platformy Apache Spark w kontekście analizy danych tekstowych typu Big Data;
- stworzenie analizy komentarzy z okresu pandemii koronawirusa COVID-19 i wyciągnięcie wniosków końcowych.

3.2 Przeprowadzona analiza danych wraz z wynikami

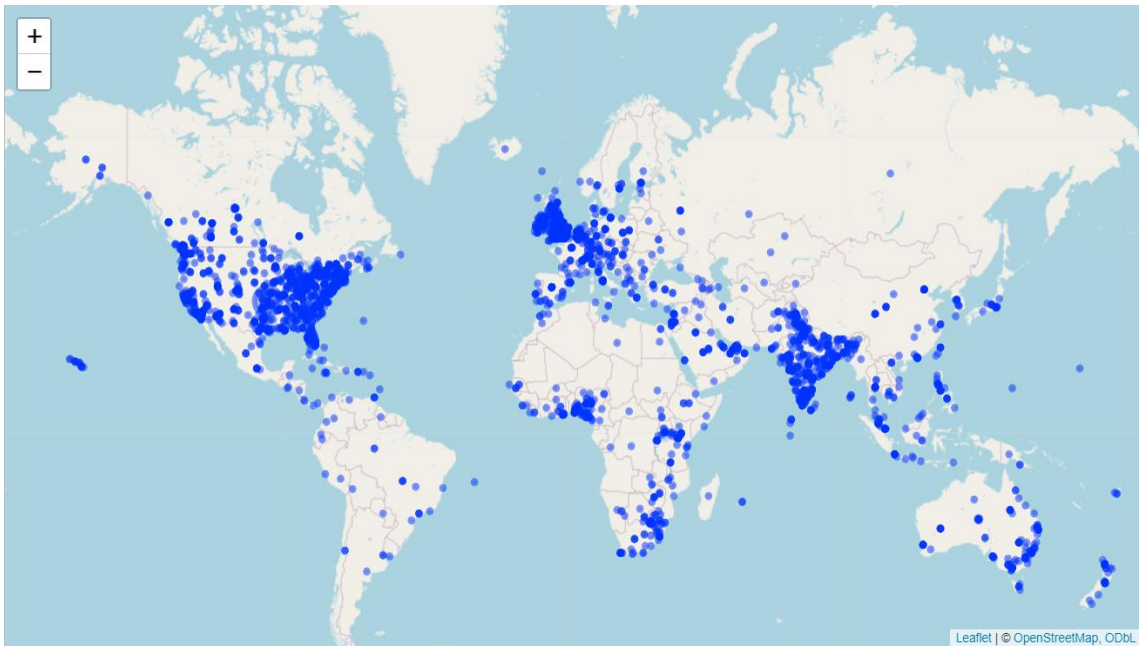
Analiza Geolokalizacji

W prowadzonej pracy badawczej analiza geolokalizacji jest możliwa już po oczyszczeniu danych z wybrakowanych obserwacji. Polega ona na sprawdzeniu, z jakiego miejsca na świecie najczęściej pisane były komentarze oraz jak zmieniała się ich liczba w czasie. Ważnymi elementami są więc zmienne *user_location* oraz *date*.

Poniżej ukazana mapa świata (rys. 33) oraz tabela wynikowa (tabela 3). Pierwsza z nich przedstawia najczęściej występujące, pojedyncze lokalizacje komentarzy. Oparta została na 5 tysiącach najczęstszych miejsc tworzenia opinii. Tabela natomiast prezentuje sumę wszystkich stworzonych opinii dla państwa pojawiającego się najczęściej w całym zbiorze danych.

Tabela 3. Tabela wynikowa: 10 najczęstszych lokalizacji tweetów oraz ich liczebność; Źródło: Opracowanie własne z wykorzystaniem RStudio.

Lokalizacja	United States	India	United Kingdom	Worldwide	Nigeria	Australia	Canada	South Africa	Kenya	Switzerland
Liczebność	23921	16357	11135	7741	4346	3489	3072	2752	2365	851



Rysunek 33. Mapa świata. Lokalizacje występujących opinii; Źródło: Opracowanie własne z wykorzystaniem RStudio oraz OpenStreetMap.

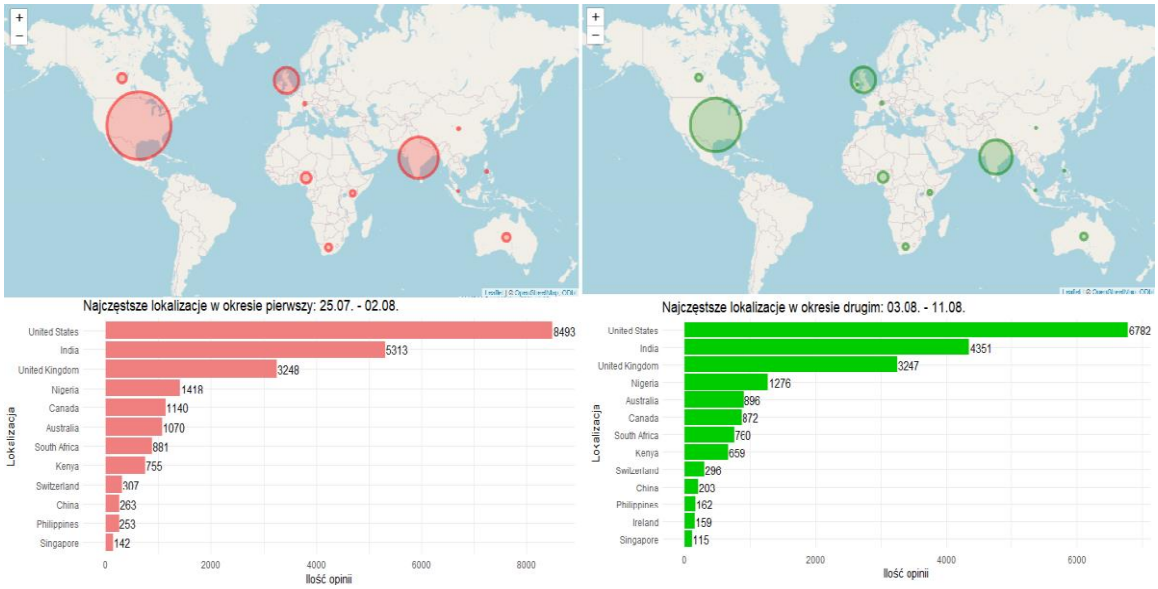
Stworzona mapa świata (rys. 33) pokazuje obszary o największej częstotliwości tworzenia opinii. Są to: Wielka Brytania, Europa Środkowa, Indie, wschodnie oraz zachodnie wybrzeże Stanów Zjednoczonych, obszary afrykańskie Nigerii, Kenii i Republiki Południowej Afryki, wschodnie wybrzeże Australii oraz wyspiarskie obszary Filipin i Nowej Zelandii. Obserwacje te pokrywają się z wynikami końcowymi przedstawionymi w tabeli 3. Liczba wszystkich komentarzy w danym państwie odpowiada najbardziej zapunktowanym miejscom na wspomnianej mapie.

Pierwszą nasuwającą się myślą po przeanalizowaniu powyższych danych, jest fakt, że temat pandemii COVID-19 wzbudził globalne zainteresowanie społeczeństw, niezależnie od ich lokalizacji. Choroba dotknęła nie tylko słabiej zurbanizowane obszary świata, lecz również te uznawane za rozwinięte. Wszyscy ludzie na Ziemi byli pod wpływem groźby zachorowania, przez co oddziaływały na nich wszelkie rządowe regulacje dotyczące życia społecznego. Owe restrykcje wzbudzały duże emocje oraz miały realny wpływ na kształtowanie się trendów społecznych. W ukazanych danych można odnieść wrażenie, że platformy społecznościowe (takiej jak m.in. Twitter.com) w tamtym czasie skupiały się na tematyce pandemii. Ludzie z różnych krajów dzielili się za ich pośrednictwem obawami, przemyśleniami oraz informacjami odnośnie COVID-19.

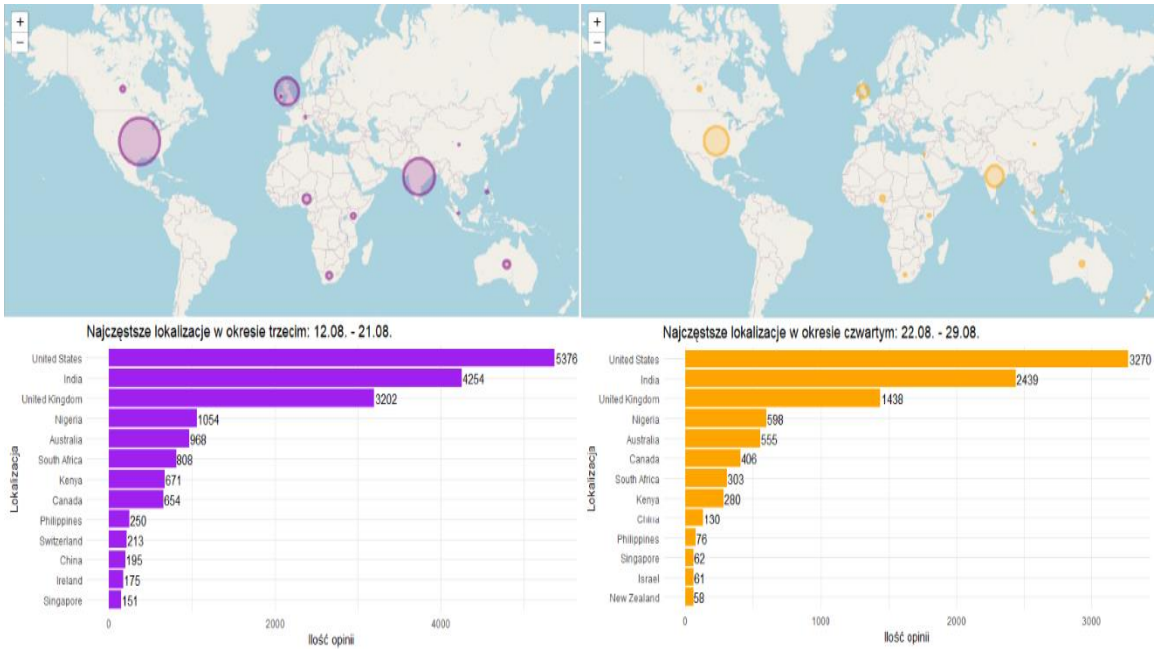
Wartym zauważenia jest również to, iż większość komentarzy pochodzi z krajów, w których używany jest (lub był) język angielski. Ma się tu na myśli zarówno kraje uznające owy język za urzędowy (takie, jak np.: Stany Zjednoczone, Wielka Brytania, Australia) oraz kraje, które w przeszłości stanowiły kolonie zamorskie dawniej dominujących państw (takie, jak np.: Nigeria – była kolonia brytyjska, Indie – byłe dominium brytyjskie, Filipiny – były protektorat Stanów Zjednoczonych). Mówi to wiele o strukturze danych, które będą analizowane w formie tekstowej. Prawdopodobna tematyka, pojawiająca się w danych (oprócz SARS COVID-19), będzie związana z wydarzeniami i obawami z tamtych regionów świata. Dominującym zaś językiem będzie język angielski.

Za pośrednictwem zmiennej czasowej *date*, możliwym jest zbadanie zmienności w występowaniu komentarzy na świecie. Dane zostaną podzielone na cztery, równe

okresy czasowe, w których następnie zostanie przeanalizowana częstotliwość wystąpienia opinii względem lokalizacji.



Rysunek 34. Mapy badanego okresu pierwszego (po lewej) oraz drugiego (po prawej), ukazująca nowych komentarzy w danym okresie; Źródło: Opracowanie własne.



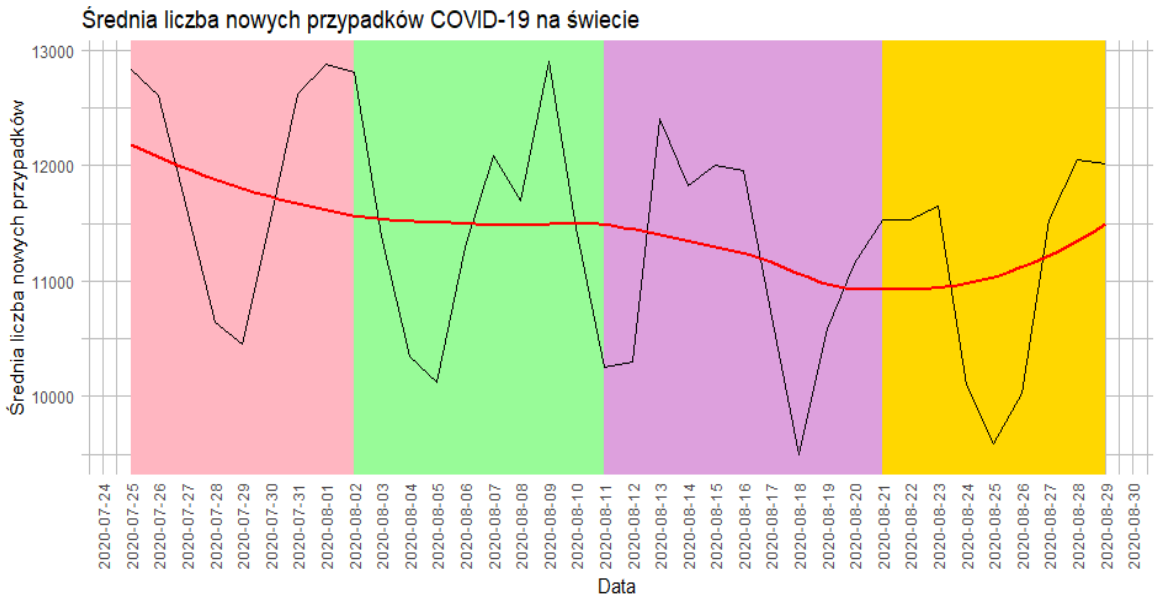
Rysunek 35. Mapy badanego okresu trzeciego (po lewej) oraz czwartego (po prawej), ukazująca nowych komentarzy w danym okresie; Źródło: Opracowanie własne.

Biorąc pod uwagę oszacowanie ukazane na czterech mapach (rys. 34 i rys. 35), można zaobserwować, iż lokalizacje z najczęściej pojawiającymi się komentarzami pozostają takie same. Zmiennym wskaźnikiem natomiast jest ilość nowych opinii.

Na przestrzeni przedstawionych czterech okresów czasowych można zaobserwować spadek jej liczebności w każdym państwie świata. Sugerować to może spadek zainteresowania tematyką COVID-19 na świecie w okresie lipca i sierpnia 2020r. Natomiast określenie przyczyny spadku zaangażowania w dyskusje na temat pandemii koronawirusa jest niejasne. Może być to spowodowane „oswojeniem się" społeczeństw globalnych z sytuacją chorobową na świecie. Aczkolwiek istnieje również możliwość występowania korelacji ilości nowych przypadków choroby z zaangażowaniem społecznym w dyskusje. Aby to zweryfikować, zostanie użyta pomocnicza baza danych, informująca o ilości nowych przypadków COVID-19 w badanym okresie, tj. pomiędzy 25 lipca, a 29 sierpnia 2020r. Hurtownia danych wspomagających pochodzić będzie z oficjalnej strony WHO.

Prosta Analiza Szeregów Czasowych

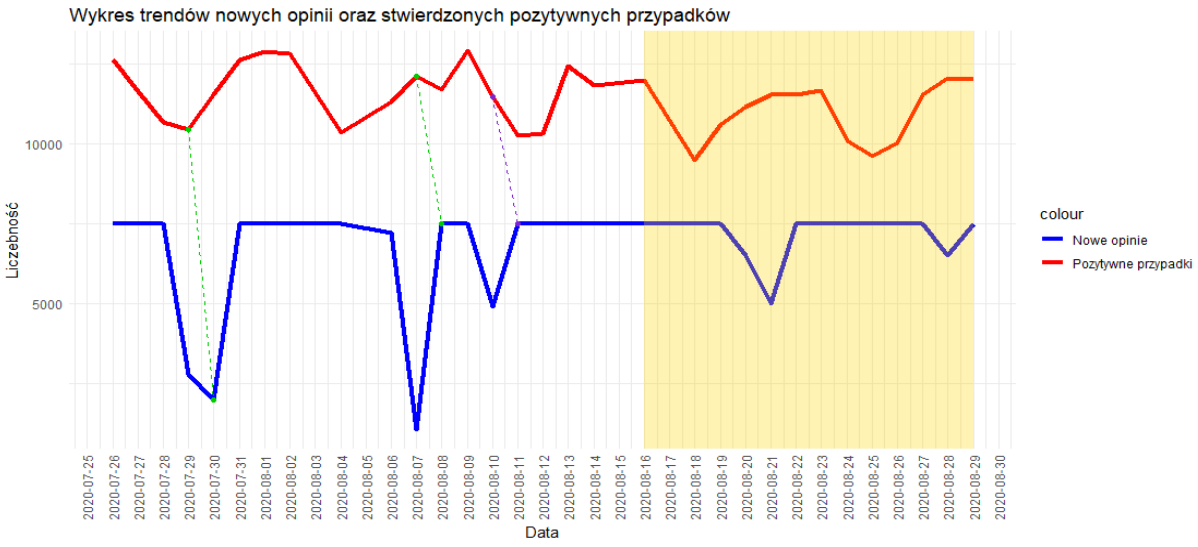
Aby zweryfikować tezę współzależności danych należy przebadać trend nowych przypadków z okresu zbierania danych testowych. W tym celu zostanie zastosowana pomocnicza hurtownia danych liczbowych, wskazująca dzienny przyrost zachorowań podczas pandemii z oficjalnej strony Światowej Organizacji Zdrowia (ang. World Health Organization, WHO).



Rysunek 36. Średnia liczba nowych, pozytywnych przypadków COVID-19 wraz z linią trendu z podziałem na 4 okresy dla całego świata; Źródło: Opracowanie własne na podstawie danych nt. ilości nowych przypadków zarażenia koronawirusa WHO.

Otrzymując wykres średniej liczby nowych przypadków (rys. 36), można ocenić skalę pozytywnych testów na COVID-19 dla ustalonych czterech okresów czasowych. Zestawiając go z mapami świata i ilościami nowych komentarzy (rys. 34 i rys. 35), można wstępnie zauważyć, iż spadek liczby nowych przypadków choroby istotnie posiada współzależność w stosunku do ilości nowych komentarzy. Natomiast nie jest możliwym nazwanie tej obserwacji tezą o silnej korelacji tych zmiennych. Sugeruje to przede wszystkim okres czwarty, w którym to uzyskano najmniejszą ilość nowych komentarzy na platformie Twitter.com, zaś liczba pozytywnych przypadków COVID-19 we wstępnej fazie tegoż okresu wyrównuje się, a następnie rozpoczyna się jej wzrost.

Dokładny obraz całości danych należy więc porównać w ujęciu ilościowym. Zestawienie wartości wykrytych przypadków COVID-19 wraz z ilością ukazujących się nowych komentarzy, pozwoli ocenić stan korelacyjny pomiędzy wspomnianymi danymi. Takiego przeglądu danych dokonano na poniższym wykresie (rys. 37). Ponadto ocenę współzależności pomoże zbadać wskaźnik, jakim jest korelacja krzyżowa (ang. *cross-correlation*). Polega ona na stworzeniu funkcji wartości współczynnika korelacji Pearsona dla dwóch szeregów czasowych przesuniętych o pewien moment w czasie względem siebie[118].



Rysunek 37. Zestawienie wartości wykrytych przypadków COVID-19 z ilością ukazujących się nowych komentarzy w ujęciu dziennym; Źródło: Opracowanie własne

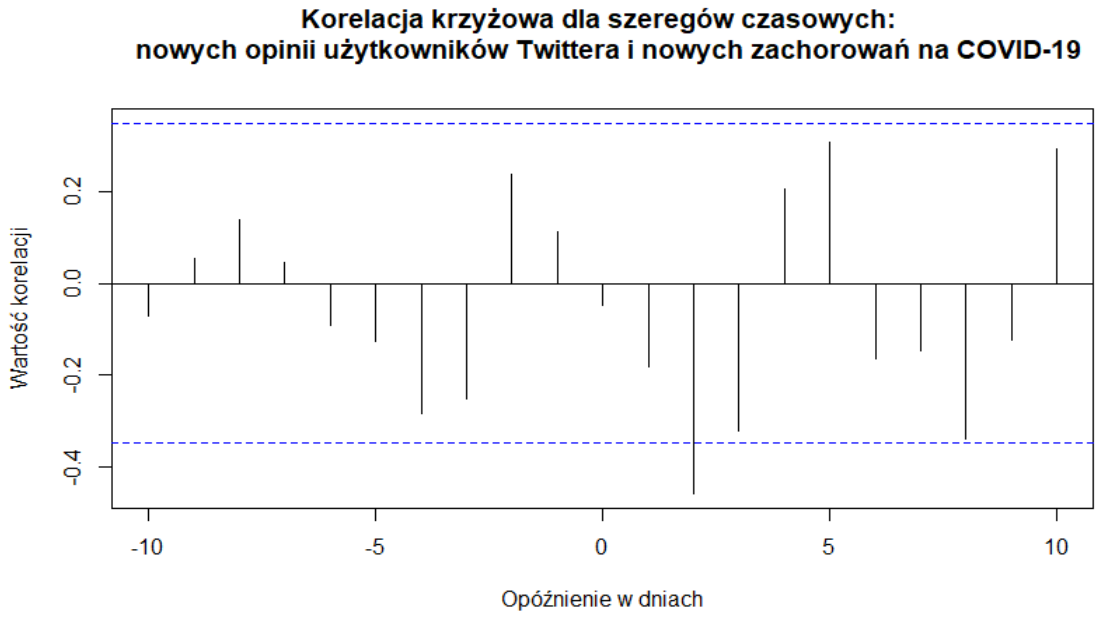
[118] Wikipedia: Wolna Encyklopedia (2023), *Korelacja wzajemna (krzyżowa)*, Dostęp: 17 września 2023.

Uzyskany wykres ilości nowych przypadków do ilości nowych komentarzy (rys. 37) jest niejednoznaczny w interpretacji. Wynika to z faktu, iż dane z hurtowni opinii użytkowników Twittera były zbierane do maksymalnej wartości 7 500 komentarzy na dzień. Stąd też komplikacje w analizie powyższego wykresu. Warto jednak zwrócić uwagę na dni, w których nie uzyskano maksymalnych wartości. Pokazują one zmienności ilości komentarzy, które można ocenić. Dodatkowo można założyć, iż reakcja na publikowane dane nt. zachorowań na COVID-19 jest opóźniona w czasie, co wskazują zaznaczone punkty (zielone i fioletowy) oraz żółty obszar na części wykresu.

Jednak w przypadku istniejącej niejasności zostanie przeprowadzone badanie korelacji krzyżowej. Dla sprawdzenia przedstawionych danych na rysunku (rys. 37) moment przesunięcia w czasie zostanie przeanalizowany do 10 dni względem każdej ze zmiennych. Moment opóźnienia został dobrany w oparciu o cztery, równe okresy zbierania danych. Taka struktura czasowa pozwoli zaobserwować zarówno krótkoterminowe reakcje (do 5 dni) oraz te długoterminowe (do 10 dni) na ilość nowych, pozytywnych przypadków zachorowania na COVID-19 wśród społeczności.

Tabela 4. Korelacja krzyżowa szeregu czasowego nowych opinii z nowymi przypadkami COVID-19; Źródło: Opracowanie własne.

Opóźnienie	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
Wartość_korelacji	-0.07	0.05	0.14	0.05	-0.09	-0.13	-0.28	-0.25	0.24	0.11	-0.05	-0.18	-0.46	-0.32	0.2	0.31	-0.16	-0.15	-0.34	-0.12	0.29



Rysunek 38. Korelacja krzyżowa dla dwóch hurtowni danych czasowych: nowych opinii oraz nowych przypadków zachorowania na COVID-19; Źródło: Opracowanie własne.

Zastosowanie korelacji krzyżowej przynosi efekt w postaci uzyskania wartości korelacji Pearsona w opóźnieniu czasowym do 10 dni dla każdego szeregu (rys. 38, tab. 4). Na powyższych danych można zaobserwować następujące fakty:

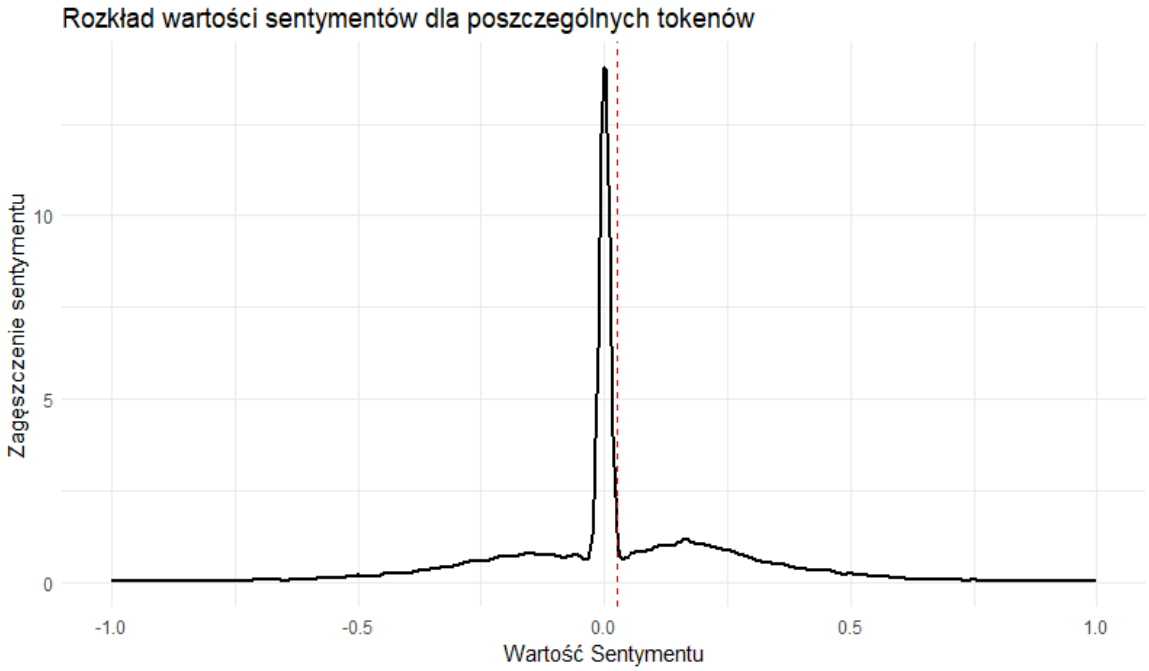
- największa bezwzględna wartość korelacji równa 0.46 występuje w drugim dniu opóźnienia, zaś znak ujemności wskazuje na korelacje ujemną – sugeruje to, że wraz ze spadkiem średniej liczby zachorowań pojawia się opóźnione dwoma dniami skorelowanie umiarkowane poprzez wzrost ilości komentarzy na Twitterze;
- największa dodatnia wartość korelacji równa 0.31 pojawia się w piątym dniu opóźnienia – sugeruje to, że wraz ze wzrostem średniej liczby zachorowań pojawia się po pięciu dniach słabe skorelowanie poprzez wzrost liczby komentarzy na Twitterze;
- moment zerowego przesunięcia czasowego sugeruje przypuszczenie, iż publikacja ilości przypadków w tym samym dniu nie wpływa na pojawienie się nowych komentarzy (wartość korelacji = - 0,05, oznacza jej praktyczny brak);
- największe stężenie reakcji występuje pomiędzy opóźnieniem o dwa dni i o trzy dni – uzyskują odpowiednio -0.46 i -0,32.

Analizując ogólną tendencję, wyniki wskazują na zmiany kierunku korelacji w zależności od opóźnienia czasowego. Najbardziej prawdopodobnym faktem jest umiarkowane skorelowanie wzrostu ilości komentarzy po dwóch dniach od publikacji spadkowych wyników nt. nowych przypadków koronawirusa. Motywacją do przyjęcia tej tezy jest fakt, że zbiór danych lepiej pokazuje spadek niż wzrost, ze względu na ograniczenia maksymalnej liczby nowych komentarzy w danym dniu.

Warto również mieć na uwadze, że korelacja ta nie występuje w każdym przypadku. Jest to umiarkowana współzależność, co oznacza, że może być zakłócana np. poprzez inne ważne wydarzenia, panujące na świecie w tamtym okresie. Natomiast opóźniona reakcja społeczeństw o dwa dni jest całkiem możliwa, gdyż spowodowana jest utwierdzeniem opinii o niskiej zmienności, z jaką ma się do czynienia (np. dwudniowy spadek liczby zachorowań).

Analiza Sentymentalna – wskaźniki sentymentu

Początkiem analizy sentymentalnej danych tekstowych będzie rozpoznanie wskaźników sentymentu dla zmiennej *text*. Pierwszym z nich jest średni sentyment.



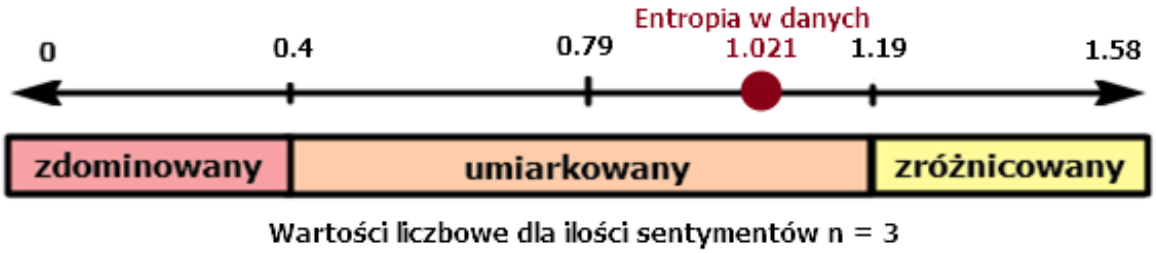
Rysunek 39. Rozkład wartości sentymentów dla poszczególnych tokenów; Źródło: Opracowanie własne.

Rozkład wartości sentymentu dla każdej obserwacji jest przedstawiony na rysunku 39. Widać na nim, iż większość wartości sentymentu oscyluje wokół wartości 0. Dodatkowo średnia wartość sentymentów w danych (czerwona linia) wynosi 0,04. Obserwacja oraz wartość wskazują, że kierunek sentymentu tekstów w hurtowni jest neutralny. Oznacza to, że ilość sentymentów pozytywnych równoważy kierunek sentymentów negatywnych.

Kolejnym istotnym wskaźnikiem jest wartość średniej polaryzacji dla sentymentów. W przypadku sentymentu pozytywnego uzyskuje się wartość 0.48, zaś dla sentymentu negatywnego jest to 0.34. Niesie to informacje o podzielności opinii i emocji pomiędzy dwoma przeciwnymi kierunkami. Ponad 48% analizowanych tekstów jest bardziej pozytywnych, niż negatywnych, czy neutralnych. Z drugiej strony 34% danych tekstowych jest bardziej negatywnych, niż pozytywnych czy neutralnych. Sugeruje to, że baza danych odznacza się większą liczbą pozytywnych emocji i reakcji. Należy jednak

sprawdzić, czy te reakcje nie zdominowały pozostałych sentymentów (neutralnych, negatywnych). Owe oszacowanie kolejny wskaźnik.

Ostatnim wskaźnikiem pomocniczym do analizy sentymentalnej jest entropia sentymentu. Jest to stopa niepewności, która mierzy, jak niejednoznaczne są etykiety sentymentu. Owe zróżnicowanie dla opracowywanej hurtowni danych nt. COVID-19 wynosi 1.021. Sugeruje to, że dane tekstowe są zróżnicowane pod względem sentymentów – żaden z nich nie dominuje nad innymi. Występuje więc duża różnorodność emocji i reakcji na temat pandemii koronawirusa w danych.



Rysunek 40. Skala wynikowa dla entropii danych; Źródło: Opracowanie własne.

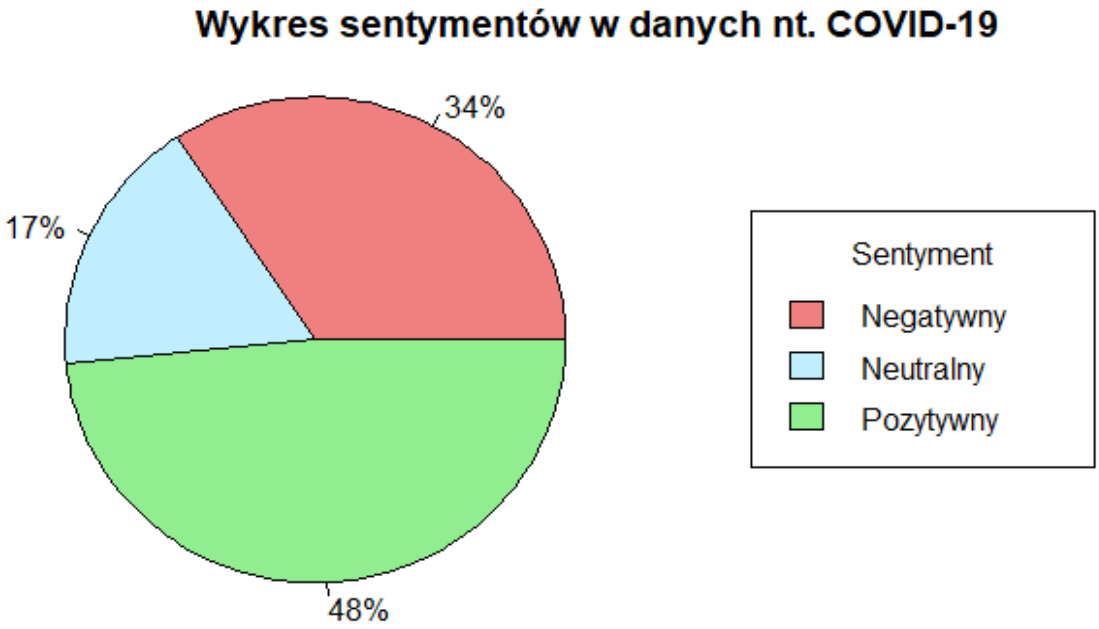
Analiza Sentymentalna – Metoda słownikowa

Metoda słownikowa w analizie sentymentalnej to badanie leksykalnej strony wypowiedzi. Jest to możliwe dzięki zastosowaniu odpowiednich ram decyzyjnych oraz słowników.

W omawianej pracy pojawia się jedna rama decyzyjna. Objęła one jedno słowo - „positive”. Obserwacja ta bowiem mogłaby zawyżać wynik pozytywnych sentymentów, zaś samo słowo jest dwuznaczne w kontekście zachorowań. Bowiem zarówno może reprezentować stan emocjonalny, jak i pozytywny wynik testu COVID-19. Zostało to ujęte również we wcześniejszej analizie wskaźników. Słowniki natomiast dzieli się na proste i złożone.

Proste analizują dane pod kątem trzech głównych sentymentów: negatywnego, neutralnego i pozytywnego. Dlatego opracowywana hurtownia danych tekstowych

skorzysta ze słownika „Bing”, który jest zintegrowany ze środowiskiem języka R. Słownik przeznaczony jest do danych napisanych w języku angielskim. Przypisuje on odpowiedni sentyment zdaniom, a następnie transponuje je do trzech kategorii sentymentów. Stąd też dane, jakie należy użyć, powinny nie być po procesie tokenizacji.



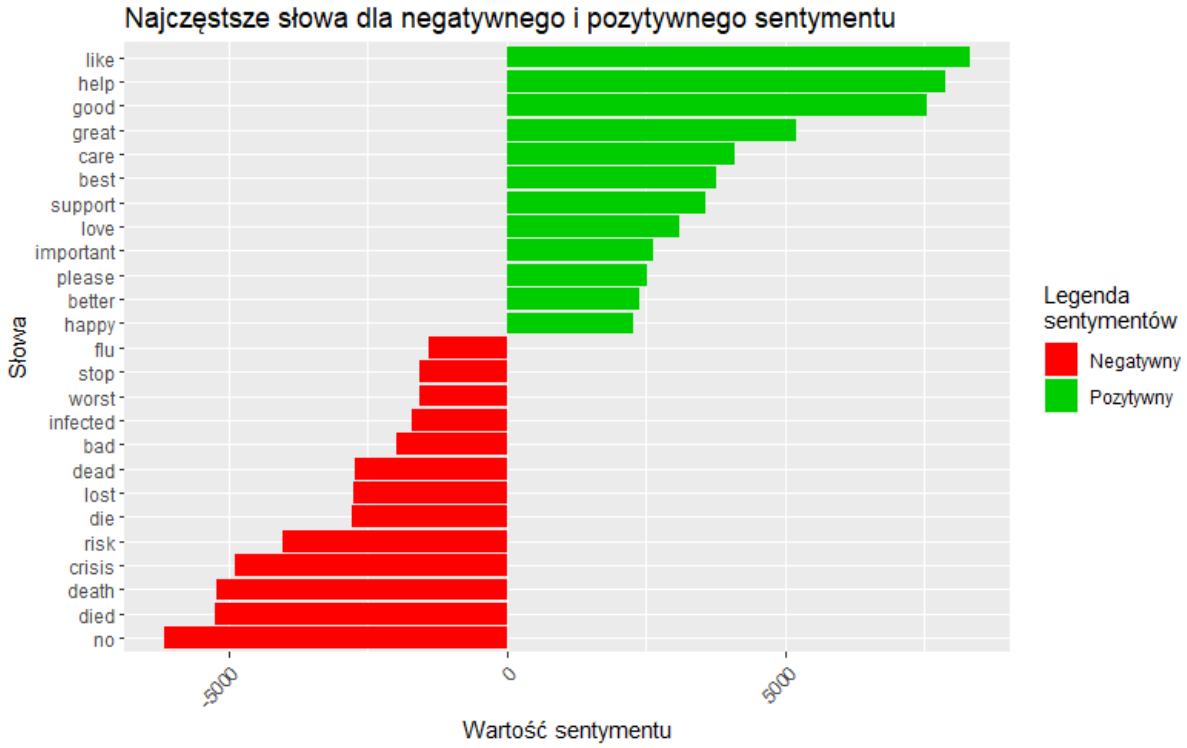
Rysunek 41. Diagram kołowy ilości sentymentów w zdaniach, będących częścią hurtowni danych nt. COVID-19; Źródło: Opracowanie własne.

Za pośrednictwem słownika „Bing” udało się nadać zdaniom sentyment, a następnie zostały one podliczone we wszystkich trzech kategoriach. Odpowiednio dla:

- pozytywnego, liczebność nacechowanych zdań wyniosła 86 774;
- neutralnego, liczebność nacechowanych zdań wyniosła 30 857;
- negatywnego, liczebność nacechowanych zdań wyniosła 61 477.

Ukazany wynik na diagramie kołowym (rys. 41) pokrywa się z wcześniejszą analizą wskaźników sentymentu. Większość danych stanowią obserwacje pozytywne, jednak nie dominują całości zbioru danych; równoważą je sentymenty neutralny i negatywny. Jednak, aby wyciągnąć wnioski z danych należy zastosować również bardziej złożone słowniki oraz ramy decyzyjne do analizy leksykalnej.

W tym celu zostanie użyty zintegrowany z środowiskiem R słownik „Afinn", mający różne wersje językowe. Pomoże on z ekstrakcją bardziej rozbudowanej palety emocji i reakcji użytkowników platformy Twitter nt. pandemii COVID-19. Ponadto działa on na tokenach (osobnych wyrazach), a nie na całości zdań.

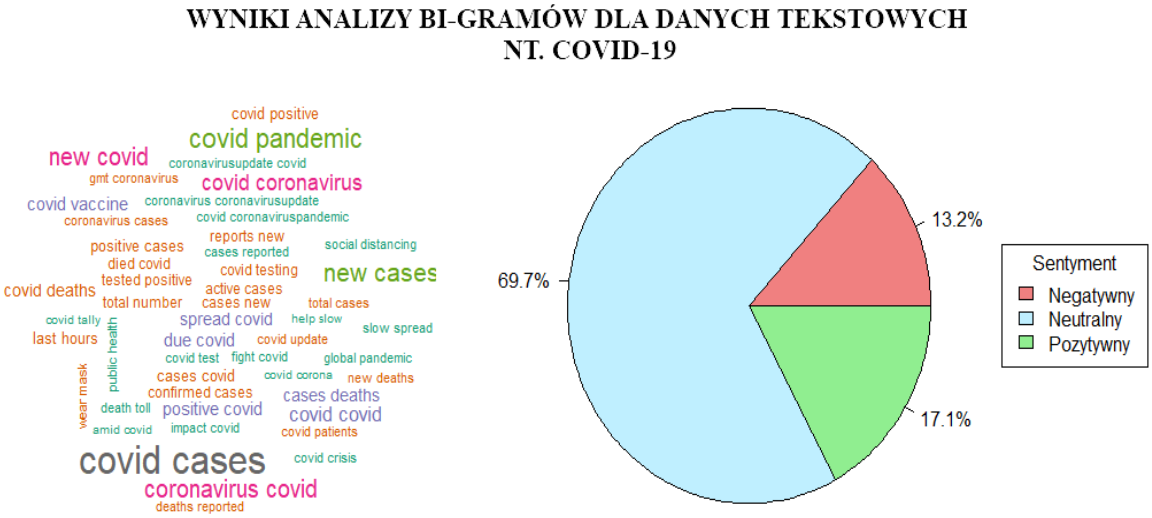


Rysunek 42. Najczęstsze słowa dla negatywnego i pozytywnego sentymentu; Źródło: Opracowanie własne.

Estymacja wielkości używanych przymiotników ukazana na rysunku (rys. 42) pokazuje przewagę sentymentów pozytywnych. Wśród nich widać takie emocje jak: dobre lub poprawiające się samopoczucie, potrzeba opieki i wsparcia. Przeciwnym kierunkiem są sentymenty negatywne, w których dominuje przekonanie o panującym kryzysie, dużym odsetku śmierci, poczuciu straty (zapewne bliskich lub związanej z sytuacją zawodową). Są również głosy wskazujące na ryzyko zachorowania lub o samym odbywaniu choroby. Istnieje też ciekawe odniesienie do grypy (ang. *flu*). Wyraz ten może wskazywać na mylne nazewnictwo COVID-19 lub wskazywać na objawy grypopodobne (ang. *flu-like*).

Widoczna jest tutaj problematyka dwuznaczności słów w kontekście tematu COVID-19. Próba stworzenia założeń poprzez ramy decyzyjne do słowników, przynosi efekt

w postaci zbyt dużego zapotrzebowania na moc obliczeniową przez działający kod. Ponadto trudnością pojawiającą się, jest liczba takich dwuznaczności w całości danych i próba ich ujęcia. Rozwiązaniem tego problemu w prowadzonej pracy będzie zastosowanie pomocniczej analizy N-Gramów. Jest ona możliwa, dzięki zastosowaniu wektoryzacji zmiennej *text*. Analiza ta pomoże zbadać najczęściej występujące ze sobą słowa w danych i będzie się opierać o bi-gramy (analiza częstości występowania dwóch słów w ciągu, n = 2) oraz tri-gramy (analiza częstości występowania trzech słów w ciągu, n = 3).

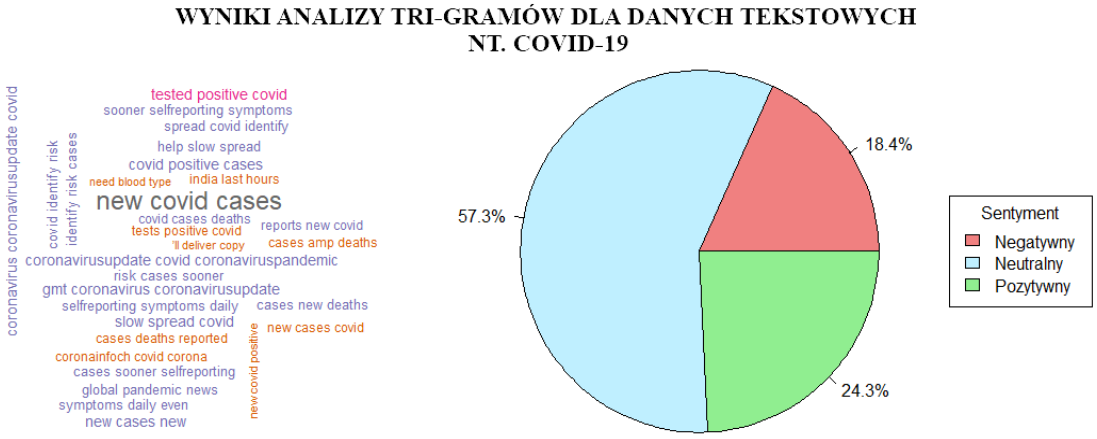


Rysunek 43. Chmura słów najczęstszych bi-gramów (po lewej) oraz diagram kołowy rozkładu sentymentów w bi-gramach (po prawej); Źródło: Opracowanie własne.

W powstałych bi-gramach na wykresie chmury słów (rys. 43) widać, że najczęstszymi słowami współistniejącymi w opiniach użytkowników Twittera są wyrazy nacechowane sentymentem negatywnym. Wykres kołowy, przedstawiający rozkład sentymentów w bi-gramach, ukazuje względnie duży odsetek danych neutralnych. Natomiast, co ważniejsze, pokazuje, iż procent sentymentów pozytywnych przewyższa procent tych negatywnych. Utwierdza to wnioski płynące ze wcześniejszych analiz. Patrząc jednak na chmurę słów i przyrównując ją do najczęściej pojedynczych słów, widać problem ujęcia słów dwuznacznych dla tej tematyki danych. Przykładowo, słowo *help* zaliczone do słów nacechowanych pozytywnie, w estymacji bi-gramów ujęte jest wraz ze słowem *slow*. Zbitek wyrazowy *slow help* posiada wydźwięk negatywny (dotyczy powolnego niesienia pomocy). Stąd też ukazana jest trudność w prawidłowym

określeniu sentymentu w danych. Sugeruje to również, iż dodatkowe analizy pomogą dostrzec nieoczekiwane obserwacje, które mogą wpłynąć na interpretacje końcowe.

W celu lepszego rozpoznania struktury występowania danych zostanie przeprowadzona ponownie analiza n-gramów, lecz tym razem dla związków trójwyrazowych – tzw. tri-gramów.



Rysunek 44. Chmura słów najczęstszych tri-gramów (po lewej) oraz diagram kołowy rozkładu sentymentów w tri-gramach (po prawej); Źródło: Opracowanie własne.

W powstałych tri-gramach ma wykresie chmury słów (rys. 44) widać, powtórzenie trendu słów w odbiorze negatywnym, jak na wykresie chmury słów bi-gramów (rys. 43). Różnica natomiast uwidoczniona została na wykresie kołowym rozkładu sentymentów w tri-gramach. Neutralny sentyment, ciągle dominujący, został zredukowany na poczet sentymentów negatywnego i pozytywnego. W strukturze sentyment pozytywny przewyższa w dalszym ciągu sentyment negatywny, co również utwierdza wcześniejsze badania sentymentów dla całości zbioru danych (rys. 41).

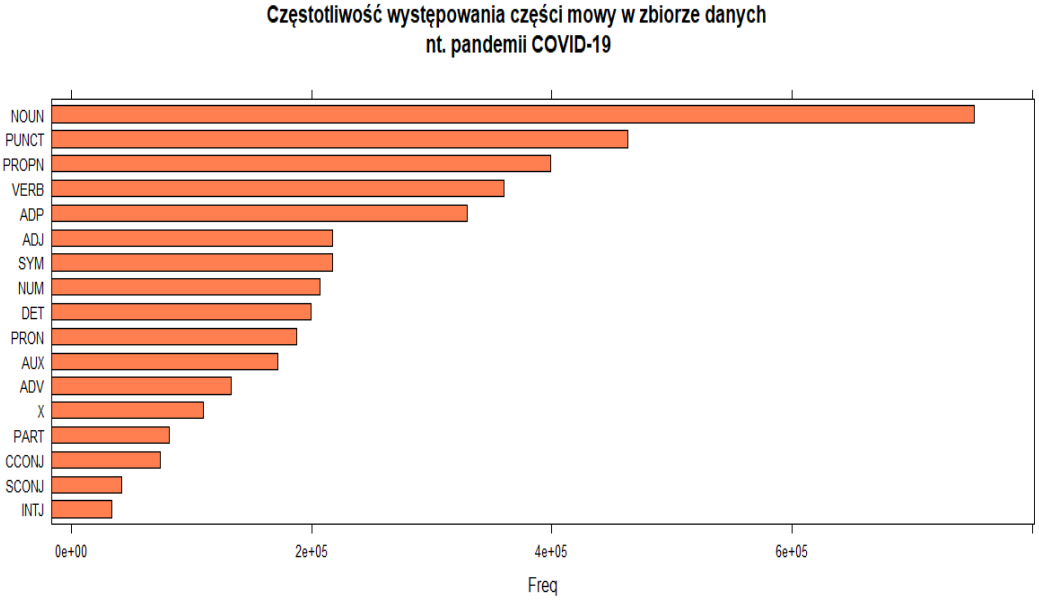
Analiza N-Gramów w swoich wynikach pokazała różnorodność danych oraz ich nieoczekiwane wyniki w zależności od doboru metod i badań analitycznych. Potwierdziła również wynik wskaźnika sentymentalnego, jakim była entropia danych, odnośnie zróżnicowania sentymentów w danych (rys. 40). Wizualizacje chmury słów w bi-gramach oraz w tri-gramach pokazują, że najczęściej użytkownicy Twittera tworzyli swoje wypowiedzi w oparciu o słowa z negatywnym sentymentem. Natomiast wartym zauważenia jest, że ich wykorzystanie niekoniecznie zawsze było w ujęciu

pesymistycznym. Pokazują to wyniki, w których w całości zbioru danych oraz w poszczególnych N-gramach mimo wszystko dominuje sentyment pozytywny (w porównaniu tylko do negatywnego). Całość pokazuje, że temat COVID-19 był częstym punktem rozmów o różnorodnym nacechowaniu emocjonalnym w analizowanej hurtowni danych.

Analiza Sentymentalna – Metoda statystyczna

Metoda statystyczna w przedstawionym badaniu będzie bazować na pojedynczych słowach. W metodzie słownikowej na rysunku 42, ukazano strukturę najczęściej używanych słów sentymentalnych. Warto jednak sprawdzić, jakie słowa najczęściej wystąpiły w rozrachunku ogólnym oraz to, jaki mają stosunek do słów sentymentalnych. Stąd też, biorąc pod uwagę proces tokenizacji oraz wektoryzacji danych tekstowych zmiennej *text*, można dokonać analizy współwystępowania słów sentymentalnych z innymi częściami mowy.

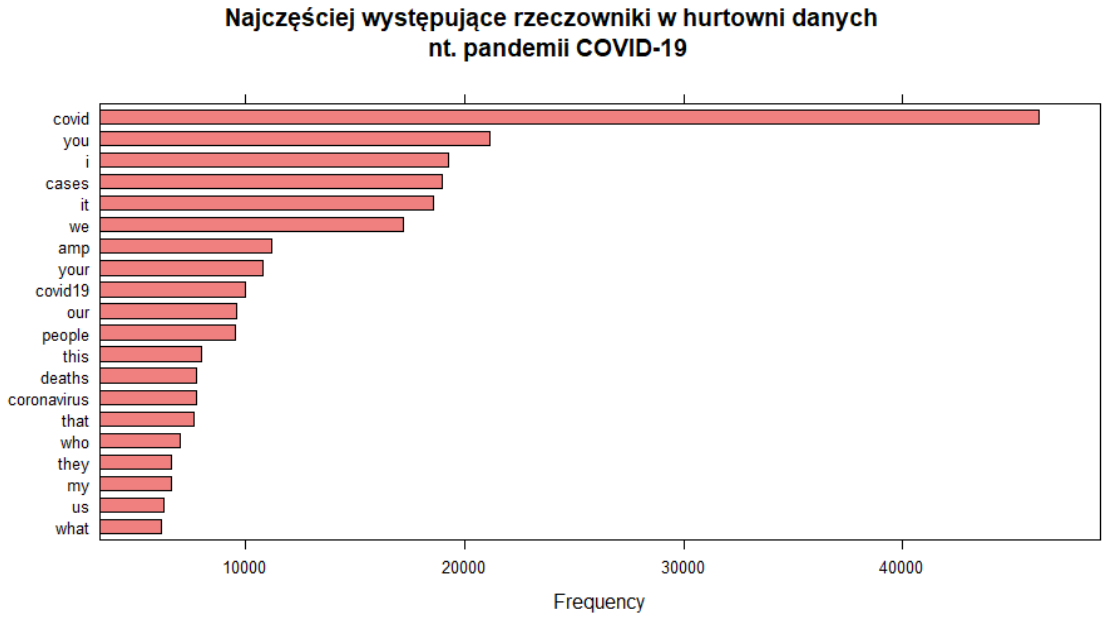
Pierwszym badaniem w tej analizie będzie rozkład części mowy w zależności od ich wystąpienia w tekście. Zostanie to zaprezentowane na wykresie poniżej (rys. 45).



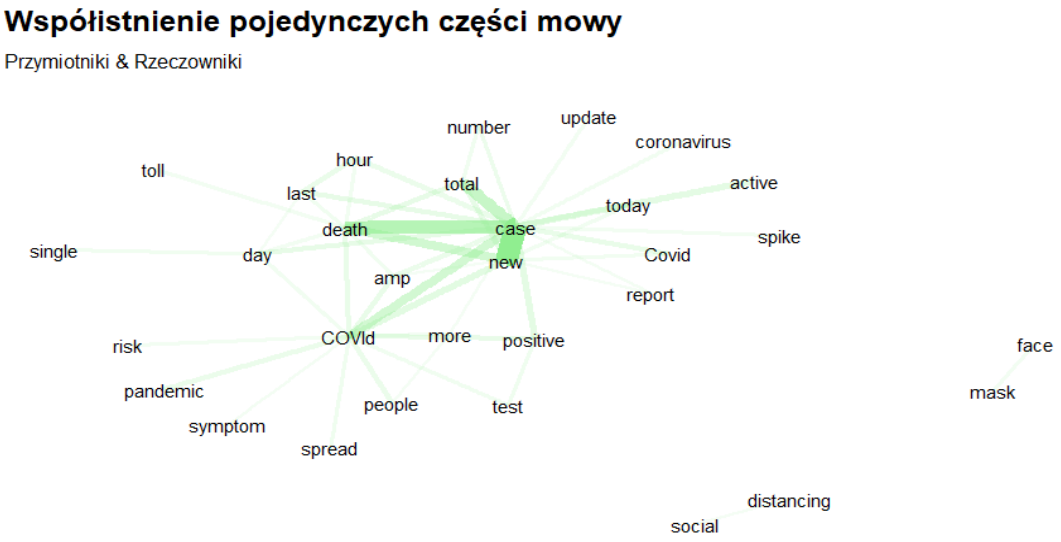
NOUN	PUNCT	PROPN	VERB	ADP	ADJ	SYM	NUM	DET	PRON	AUX	ADV	X	PART	CCONJ	SCONJ	INTJ
751405	463251	398691	360284	329707	217605	217077	206732	199621	187524	171880	133213	109992	81646	73847	42276	33669

Rysunek 45. Wykres słupkowy wraz z tabelą częstotliwości występowania poszczególnych części mowy w hurtowni danych; Źródło: Opracowanie własne.

Wykres słupkowy części mowy informuje, że największym odsetkiem z kategorii części mowy są rzeczowniki. Dlatego też zostaną one wykorzystane w celu sprawdzenia współzależności ze słowami sentymentalnymi (na wykresie *ADJ*). W dalszej części trzeba określić, jakie konkretne rzeczowniki występowały najczęściej. Pomoże to powiązać słowa sentymentu. Ostatecznym wynikiem będzie uzyskanie wyników w postaci wizualizacji sieci słów odnośnie tego, które sentymenty stanowią najczęstszą kombinacje z rzeczownikami.



Rysunek 46. Najczęściej występujące rzeczowniki w hurtowni danych nt. pandemii COVID-19; Źródło: Opracowanie własne.



Rysunek 47. Najczęściej współistniejące, pojedyncze części mowy w rozważaniu przymiotników (jako słów sentymentu) oraz rzeczowników; Źródło: Opracowanie własne.

Przeprowadzone badanie współistnienia najczęstszych przymiotników z rzeczownikami (rys. 47) pokazało najczęściej stosowane kombinacje wyrazów w tym kontekście. Analizując oszacowane dane, najczęstszymi określeniami cieszyły się następujące kombinacje słów: ilość nowych przypadków, ilość śmiertelnych przypadków, dzienne raporty chorobowe, społeczne zdystansowanie. Jest również możliwe wyczytanie większej ilości takich współzależności. Globalny charakter pandemii opierał się więc o dyskusje społeczności związane z: ilością śmiertelnych i nowych przypadków choroby, społecznym zdystansowaniem, czy ryzykiem zachorowania.

Podsumowanie statystyczne z wykonanych analiz

W trakcie analizy wypowiedzi użytkowników platformy społecznościowej Twitter.com powstało wiele interesujących zmiennych pośrednich, bądź wynikowych, w których dane zostały zliczone. Poniżej zostały przedstawione intrygujące liczby z opracowanych danych:

- ilość wszystkich zmiennych w danych była równa 13, zaś w skład każdej z nich wchodziło 179 108 obserwacji;
- całkowita ilość tokenów ze zmiennej *text* wyniosła 2 231 682 słowa;
- ilość tokenów po usunięciu słów zatrzymania ze zmiennej *text* wyniosła 2 040 324 słowa;
- liczebność sentymentu pozytywnego dla zdań wyniosła 86 774 zdania;
- liczebność sentymentu negatywnego dla zdań wyniosła 61 477 zdań;
- liczebność sentymentu neutralnego dla zdań wyniosła 30 857 zdań;
- najczęstszym krajem komentującym zdarzenia było USA, posiadające ponad 23 921 komentarzy użytkowników;
- liczebność wszystkich uwzględnionych kombinacji bi-gramów wyniosła 1 277 766 podwójnych słów;
- liczebność wszystkich uwzględnionych kombinacji tri-gramów to 1 672 721 potrójnych słów;
- maksymalna dzienna ilość zbieranych obserwacji przez stronę Kaggle.com wyniosła 7 500.

Analiza wydajności pracy Apache Spark w środowisku RStudio

Przy opracowanym podsumowaniu wszystkich liczb warto zestawić działanie głównego oprogramowania wspomagającego, jakim jest silnik Apache Spark, w RStudio.

Pierwszym aspektem będzie porównanie czasu pracy wykonywania zadania. Zestawienie będzie się opierać o pracę środowiska RStudio przed i po zastosowaniu rozwiązań Sparka. Funkcję zostały dobrane tak, aby działały w obu narzędziach i były identyczne.

Tabela 5. Zestawienie czasu pracy RStudio przed i po zastosowaniu silnika Apache Spark dla wybranych operacji na danych; Źródło: Opracowanie własne.

	Oczyszczanie danych	Tokenizacja zmiennych	Zliczanie ilości słów	Analiza N-gramów
RStudio	4.25 s	1.5 s	12.17 s	35 min 40 s
Apache Spark	1.44 s	0.75 s	0.79 s	9 min 13 s

W przedstawionej tabeli 5. porównano ze sobą czas wykonywania poszczególnych czynności w każdej technologii. Wynikiem otrzymanym jest wskazanie zastosowania Apache Spark jako szybszego rozwiązania. W przypadku prostych manipulacji na danych różnice mogą być znikome. Natomiast owa przewaga uwidacznia się w procesach, których analizowane są związki między wyrazami, gdzie potrzeba większej mocy obliczeniowej. Spark w tym wypadku wygrywa z przewagą czasową niemal równą czterokrotności. Można więc stwierdzić, iż zastosowanie platformy Apache Spark, przyspiesza pracę wykonywanych algorytmów.

Platforma Apache Spark dostarcza środowisko Spark UI do monitorowania zasobów, ilości wykonawców oraz stanu aktualnie wykonywanych zadań. Wszystkie przeprowadzone analizy w silniku posiadają swoje odzwierciedlenie, dzięki czemu możliwa jest analiza całokształtu pracy platformy. Spark w trakcie omawianej analizy:

- wykonał 114 zadań, z czego wszystkie zostały zakończone pomyślnie;
- zapisał w pamięci łącznie 147 megabajtów wyników z wszystkich prowadzonych badań przy pracy na jednym węźle wykonawczym;
- nie miał potrzeby przejścia na większą ilość węzłów wykonawczych do procesowania 178 108 obserwacji bazowych;

- prowadził działania analityczne w sposób stosunkowo szybki, o czym informuje wizualizacja w tabeli 2. oraz 5.

Procesowanie i manipulowanie na danych w opisywanej technologii działało bez zarzutu. Spark bowiem posiada pełną kompatybilność z pakietem *dplyr*. Połączenie obu tych narzędzi, tworzy bardzo potężne oprogramowanie do przetwarzania danych typu Big Data, niezależnie od ich struktury. Zdecydowaną również zaletą jest wspomniany czas przetwarzania danych oraz gotowe funkcje do analiz danych liczbowych oraz tekstowych, które opierają się w dużej mierze na komponencie Sparka, jakim jest MLlib.

Z perspektywy kompatybilności funkcjonalnej, Apache Spark nie jest jednak zawsze najlepszym wyborem. W trakcie dokonywania preprocessingu danych, czy przeprowadzania analiz, dochodziło niejednokrotnie do momentów, kiedy Spark SQL nie mógł przetransponować kodu do Java Virtual Machine. Błędy polegały na niekompatybilności używanych bibliotek, przenoszeniu algorytmu do innego środowiska, bądź na niewspieranych już funkcjach Sparka w RStudio. Nie ma również dużo opracowań dotyczących błędów, pojawiających się po stronie kompilatora. Niekiedy dane należało przekształcać z Spark DataFrame do zwykłej ramki danych R, aby wyeliminować pojawiające się błędy. Stąd też należy dokładnie zaplanować proces pracy ze Sparkiem oraz przemyśleć ewentualne alternatywy, kiedy pojawią się błędy.

3.3 Wnioski końcowe z przeprowadzonej analizy danych

W roku 2020 świat stanął w obliczu jednej z największych i najpoważniejszych pandemii współczesnej historii - pandemii COVID-19, wywołanej przez koronawirusa SARS-CoV-2. To wydarzenie zmieniło nasze życie i miało ogromny wpływ na różne aspekty naszego społeczeństwa, gospodarki i opieki zdrowotnej. Reakcje na tę pandemię były bardzo zróżnicowane i stały się głównym tematem debat w opinii publicznej na całym świecie.

Przeprowadzona w tej pracy analiza danych dotyczących COVID-19 koncentruje się na wczesnym okresie pandemii, który miał miejsce w sierpniu 2020 roku. To właśnie wtedy można było zaobserwować wiele interesujących zjawisk. Przede wszystkim warto zwrócić uwagę na to, że choroba dotknęła praktycznie każdego zakątka globu, niezależnie od stopnia zurbanizowania danego regionu. Była to sytuacja bezprecedensowa, która skupiła uwagę milionów ludzi na całym świecie. Przykład Twittera pokazuje, iż różne fora internetowe i platformy społecznościowe stały się areną aktywnych dyskusji na temat nowych przypadków COVID-19. Z niecierpliwością oczekiwano na nowe raporty dotyczące dziennych statystyk pandemii.

Analiza danych pozwoliła również na zidentyfikowanie dominującej roli użytkowników z państw anglojęzycznych w tych debatach. Ta obserwacja była ważna z punktu widzenia przeprowadzonej analizy sentymentalnej i analizy słownictwa. Pozwoliła bowiem zbadać odczucia ludzi z różnych regionów świata, ze względu na popularność języka angielskiego. Analiza sentymentalna wskazała, że przeważały odczucia pozytywne, takiej jak wzajemna pomoc, ozdrowienie, czy nadzieja na zakończenie choroby poprzez wynalezienie lekarstwa. Nie zdominowały one jednak wszystkich emocji użytkowników. Istotny odsetek stanowił sentyment negatywny, który informował o poczuciu straty, panującej śmierci, czy ryzyka zachorowania. Nie brakowało również porównań COVID-a do grypy. Na skutek wprowadzonych restrykcji w społeczeństwach odczuwano zdystansowanie i alienacje nie tylko w stosunku do drugiego człowieka, ale również znajomych, rodziny i bliskich. Wskazywano również na niewydolność służby zdrowia i powolne niesienie pomocy potrzebującym.

Ponadto, zestawienie głównych zmiennych z pomocniczą hurtownią danych, która wskazywała ilości nowych przypadków COVID-19, pozwoliło zrozumieć, w jakich okolicznościach pojawiały się komentarze na platformie Twitter.com. Warto zauważyć, że przez większość badanego okresu ilość nowych przypadków koronawirusa spadała, co wzbudzało zainteresowanie użytkowników. Badania wykazały także obecność umiarkowanej korelacji odwrotnej pomiędzy ilością nowych komentarzy, a ilością pozytywnych przypadków COVID-19. To oznaczało, że wzrost liczby komentarzy występował głównie po dwóch dniach od opublikowania raportu o nowych przypadkach.

Podsumowując, analiza sentymentalna i analizy pomocnicze dostarczyły cennych informacji na temat reakcji społeczeństwa na pandemię COVID-19. To zrozumienie emocji, jakie towarzyszyły temu nowemu na ówczesny czas, globalnemu wydarzeniu.

Podsumowanie końcowe

Stworzone badania dla hurtowni danych nt. COVID-19 dotknęły różnych aspektów tematycznych. Zostały podjęte kroki w celu ustrukturyzowania zmiennych opisowych oraz ich dalszego procesowania, gdzie osiągnęły gotowość do poddania ich analizie. W skutek czego udało się dokładnie poznać strukturę badanych danych i wyciągnąć z nich wnioski końcowe. Te ukazały sentyment panujący wśród użytkowników Twittera w początkach pandemii koronawirusa. Ponadto dostarczyły ciekawych obserwacji na temat najczęściej występowanych lokalizacji w zmiennych, czy pozwoliły na zastosowanie łączonych metod analizy danych, by uzyskać potwierdzenie uzyskanych wyników.

Nie bez znaczenia w całości analiz było wykorzystanie duetu technologii Apache Spark z językiem R, w środowisku RStudio. Stanowiło to cenny element wsparcia pracy analitycznej. Spark wykazał się dużą szybkością w przetwarzaniu zmiennych oraz skutecznie zarządzał zleconymi zadaniami, co stanowiło niemałe ułatwienie w czasie pracy. Wskazują to również zestawienia, ukazane w pracy, w których silnik Apache Spark przewyższał możliwościami obliczeniowymi podstawowe funkcje języka R. Niemniej jednak nie obyło się bez wyzwań. W całości analizy pojawiło się wiele komplikacji natury współpracy silnika wraz z dodatkowymi rozszerzeniami. Często bywały również komunikaty o błędach w transformacji kodu języka R przez komponent Spark SQL. Powodowało to żmudne przenoszenie danych do podstawowej formy ramki danych, co wydłużało proces analityczny.

Podsumowując, praca z technologią Apache Spark była kluczowa do przeprowadzenia analizy zbioru danych tekstowych. Jej zalety, takie jak szybkość przetwarzania i bogata funkcjonalność, uczyniły ją wartościowym narzędziem do pracy z danymi typu Big Data. Jednak podczas wyboru tego narzędzia trzeba być świadomym pewnych wyzwań, takich jak brak kompatybilności, czy dobrze opisanych rozwiązań do występujących błędów. Wartość technologii Apache Spark jako narzędzia do analizy danych jest niezaprzeczalna, ale wartość ta zależy od umiejętności i zaangażowania osób pracujących z nią.

Bibliografia

1. Plewiak. W. (2020). Analiza Sentymentu. Encyklopedia Zarządzania. Pobrano 23 kwietnia 2023. Lokalizacja: https://mfliles.pl/pl/index.php/Analiza_sentyment

2. Asurion (2022). The New Normal: Phone Use is Up Nearly 4-Fold Since 2019, According to Tech Care Company Asurion. Asurion. Pobrano 23 kwietnia 2023. Lokalizacja: https://www.asurion.com/connect/news/tech-usage/
3. Gupta. S. (2018). Sentiment Analysis: Concept, Analysis and Applications. Medium. Pobrano 10 maja 2023.

   Lokalizacja: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
4. Wikipedia: The Free Encyclopedia (2023). Unstructured data. Pobrano 10 maja 2023. Lokalizacja: https://en.wikipedia.org/wiki/Unstructured_data

5. NapSaga (2023). Chat GPT: Achieving 100 Milion Users in Just 2 Month – A Deep Analysis. Medium. Pobrano 10 maja 2023. Lokalizacja: https://ai.plainenglish.io/chat-gpt-achieving-100-million-users-in-just-2-month-a-deep-analysis-a453e6f85acf

6. PwC Malta (2022). The rising popularity of Low-Code App Development. Pobrano 10 maja 2023. Lokalizacja: https://www.pwc.com/mt/en/publications/technology/the-rising-popularity-of-low-code.html

7. Willrobotstakemyjob (2023). Statisticians. Pobrano 10 maja 2023. Lokalizacja: https://willrobotstakemyjob.com/statisticians

8. MonkeyLearn (2023). What Is Text Mining? A Beginner's Guide. Pobrano 18 maja 2023. Lokalizacja: https://monkeylearn.com/text-mining/

9. Oracle (2014). Czym jest hurtownia danych?. Pobrano 13 sierpnia 2023. Lokalizacja: https://www.oracle.com/pl/database/what-is-a-data-warehouse/

10. Nieinformatyk (2021). Hurtownia danych – co to jest i jak działa? Porównanie z relacyjną bazą danych SQL. Youtube. Pobrano 13 sierpnia 2023. Lokalizacja: https://www.youtube.com/watch?v=K5kUfaeFMLk&ab_channel=nieinformatyk

11. ETL-Tools.Info (n.d.). Architektura Gwiazdy (Star schema). Pobrano 13 sierpnia 2023. Lokalizacja: https://etl-tools.info/pl/bi/hurtownia_danych_schemat-gwiazdy.htm

12. Wikipedia: Wolna Encyklopedia (2022). Bazy danych. Pobrano 13 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/w/index.php?title=Specjalna:Cytuj&page=Baza_danych&id=68745919&wpFormIdentifier=titleform

13. KlasterIT (n.d.). Hurtownie danych. Pobrano 13 sierpnia 2023. Lokalizacja: https://www.klasterit.pl/oferta/oprogramowanie/hurtownie-danych/

14. Lula. P. (2005). Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych. StatSoft. Pobrano 23 czerwca 2023. Lokalizacja: http://media.statsoft.nazwa.pl/_old_dnn/downloads/text_mining_jako_narzedzie_pozyskiwania.pdf

15. Sydow. M. (n.d.). Eksploracja Danych – Wstępne przetwarzanie danych. Polsko-Japońska Akademia Technik Komputerowych. Pobrano 13 lipca 2023. Lokalizacja: http://users.pja.edu.pl/~msyd/ewd/preprocessing.pdf

16. Tomanek. K. (2014). Analiza sentymentu – metoda analizy danych jakościowych. Pobrano 13 lipca 2023. CEJSH. Lokalizacja: http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-f457ca52-41e7-4a37-a655-d157381b3a02

17. Wachnicki. J. (2021). Text Mining w 5 krokach. Predictive Solutions. Pobrano 13 lipca 2023. Lokalizacja: https://predictivesolutions.pl/text-mining-w-5-krokach

18. Bosko. E. (2022). Czym jest NLP?. Pogromcy Kodu. Pobrano 13 lipca 2023. Lokalizacja: https://pogromcykodu.pl/czym-jest-nlp/

19. Garcia. C. (2022). Czym jest normalizacja danych?. AppMaster. Pobrano 13 lipca 2023. Lokalizacja: https://appmaster.io/pl/blog/co-to-jest-normalizacja-danych

20. Sypytkowski. B. (2012). [NLP] Stemming i lematyzacja. Simple Solution. Pobrano 17 lipca 2023. Lokalizacja: http://horusiath.blogspot.com/2012/08/nlp-stemming-i-lematyzacja.html

21. Dobrysłownik (n.d.). Lematyzacja. Pobrano 17 lipca 2023. Lokalizacja: https://dobryslownik.pl/slowo/lematyzacja/224219/

22. Ganesan. K. (2014). What are Stop Words?. Pobrano 17 lipca 2023. Lokalizacja: https://kavita-ganesan.com/what-are-stop-words/

23. Piątkowska. K. (2020). Klasyfikacja tekstu, czyli „text classification". Blog Statystyczny. Pobrano 17 lipca 2023. Lokalizacja: https://www.statystyczny.pl/klasyfikacja-tekstu-text-classification/

24. Bosko. E. (2021). Czym jest NLP?. Pogromcy Kodu. Pobrano 17 lipca 2023. Lokalizacja: https://pogromcykodu.pl/czym-jest-nlp/

25. (…)

26. M. Mamczur. (2019). Jak wizualizować word embedding?. Pobrano 11 września 2023. Lokalizacja: https://miroslawmamczur.pl/jak-zwizualizowac-word-embedding/

27. Liu. B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. Pobrano 17 lipca 2023. Lokalizacja: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

28. Data Science Team (2020). Zrozumienie BERT. Data Science. Pobrano 17 lipca 2023. Lokalizacja: https://datascience.eu/pl/przetwarzanie-jezyka-naturalnego/zrozumienie-bert/

29. Zhang. L. & Wang. S. & Liu. B. (2018). Deep Learning for Sentiment Analysis: A Survey. Cornell University. Pobrano 23 lipca 2023. Lokalizacja: https://arxiv.org/ftp/arxiv/papers/1801/1801.07883.pdf

30. (…)

31. Scaramozzino. R. & Cerchiello. P. & Aste. T. (2021). Information Theoretic Causality Detection between Financial and Sentiment Data. MDPI. Pobrano 23 lipca 2023. Lokalizacja: https://www.mdpi.com/1099-4300/23/5/621

32. Czapiewski. B. (2014). Dziewięć zasad skutecznego użycia koloru. Skuteczne Porady. Pobrano 27 lipca 2023. Lokalizacja: https://skuteczneraporty.pl/blog/dziewiec-zasad-skutecznego-uzycia-koloru/

33. Santus. M. (2020). Jak korzystać z kolorów do prezentacji danych. Data Wizards. Pobrano 27 lipca 2023. Lokalizacja: https://datawizards.pl/blog/jak-korzystac-z-kolorow-do-prezentacji-danych/

34. Kochanowska. M. (2015). Wpływ kolorów na emocje i zdolności. NeuroSkoki. Pobrano 27 lipca 2023. Lokalizacja: https://neuroskoki.pl/wplyw-kolorow-na-emocje-i-zdolnosci/

35. Wikipedia: Wolna Encyklopedia (2022). Analiza skupień. Pobrano 27 lipca 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Analiza_skupie%C5%84

36. Wikipedia: Wolna Encyklopedia (2023). Grupowanie hierarchiczne. Pobrano 27 lipca 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Grupowanie_hierarchiczne

37. Wikipedia: Wolna Encyklopedia (2023). Analiza skupień. Pobrano 27 lipca 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Analiza_skupie%C5%84

38. Chojnacki. R. (2021). Czym jest analiza N-Gram słów kluczowych i jak ją wykorzystać?. Pobrano 27 lipca 2023. SpaceAds.Digital. Lokalizacja: https://spaceads.pl/blog/czym-jest-analiza-n-gram-slow-kluczowych-i-jak-ja-wykorzystac/

39. Tomanek. K. (2014). Analiza sentymentu – metoda analizy danych jakościowych. CEJSH. Pobrano 12 stycznia 2023. Lokalizacja: http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-f457ca52-41e7-4a37-a655-d157381b3a02

40. Turek. T. (2017). Możliwości wykorzystania analizy sentymentu w procesach konsumenckich. Politechnika Częstochowska. Pobrano 14 stycznia 2023. Lokalizacja: https://depot.ceon.pl/bitstream/handle/123456789/19403/Turek_Mo%20liwosci_wykorzystania_analizy_sentymentu_w_procesach_prosumenckich.pdf?sequence=1

41. Reinsel. D. & Gantz. J. & Rydning. J. (2018). The Digitization of the World From Edge to Core. Data Age. Pobrano 14 stycznia 2023. Lokalizacja: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

42. Digital 2021 (2022). Korzystanie z Internetu: 2020 rok – raport ze świata. FunkyMedia. CEJSH. Pobrano 15 stycznia 2023. Lokalizacja: https://funkymedia.pl/korzystanie-z-internetu-2020-rok-raport-ze-swiata-digital-2021.html

43. Tomanek. K. (2014). Analiza sentymentu – metoda analizy danych jakościowych. CEJSH. Pobrano 15 stycznia 2023. Lokalizacja: http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-f457ca52-41e7-4a37-a655-d157381b3a02

44. Sawka. K. (2020). Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow. Helion. Gliwice.

45. Wikipedia: Wolna Encyklopedia (2021). Uczenie nienadzorowane. Pobrano 16 stycznia 2023.

    Lokalizacja: https://pl.wikipedia.org/wiki/Uczenie_nienadzorowane

46. Plewiak. W. (2020). Analiza sentymentu. Encyklopedia Zarządzania. CEJSH. Pobrano 16 stycznia 2023. Lokalizacja: https://mfiles.pl/pl/index.php/Analiza_sentymentu

47. Wikipedia: Wolna Encyklopedia (2023). Języki fleksyjne. Pobrano 16 stycznia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/J%C4%99zyki_fleksyjne

48. Wikipedia: Wolna Encyklopedia (2023). Aglutynacyjność. Pobrano 16 stycznia 2023.

    Lokalizacja: https://pl.wikipedia.org/wiki/Aglutynacyjno%C5%9B%C4%87

49. (…)

50. Brand24 Team (2022). Czym jest analiza sentymentu i jak ją przeprowadzić. Pobrano 16 stycznia 2023. Lokalizacja: https://brand24.pl/blog/co-to-jest-analiza-sentymentu-oraz-jak-mozesz-ja-wykorzystac/

51. Tomanek. K. (2014). Analiza sentymentu – metoda analizy danych jakościowych. Pobrano 16 stycznia 2023. Lokalizacja: http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-f457ca52-41e7-4a37-a655-d157381b3a02

52. Data Science Team (2020). Analiza sentymentalna. Pobrano 16 stycznia 2023. Lokalizacja: https://datascience.eu/pl/matematyka-i-statystyka/analiza-sentymentalna/

53. (…)

54. (…)

55. Marszal. A. (2022). O czym mówią rynki finansowe? Ocena sentymentu. Forsal.pl. Pobrano 15 lipca 2023. Lokalizacja: https://forsal.pl/finanse/artykuly/8611695,o-czym-mowia-rynki-finansowe-ocena-sentymentu.html

56. Srividhya. V. & Meenakshi. Raja. G. (2018). Comparison of Sentiment Analysis of Goverment of India Schemes using Tweets. JCSE. Pobrano 15 lipca 2023. Lokalizacja: https://www.ijcseonline.org/pdf_paper_view.php?paper_id=2288&155-IJCSE-03865.pdf

57. StatSoft Polska (n.d). Naiwna metoda Bayesa. Pobrano 15 lipca 2023. Lokalizacja: https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstnaiveb.html

58. Statista (2023). Most popular social networks worldwide as of January 2023, ranked by numer of monthly active users. Pobrano 21 lipca 2023. Lokalizacja: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

59. Arun. K. & Srinagesh. A. & Makala. R. (2017). Twitter Sentiment Analysis on Demonetization tweets in India Using R language. ResearchGate. Pobrano 21 lipca 2023. Lokalizacja: https://www.researchgate.net/publication/358915759_Twitter_Sentiment_Analysis_on_Demonetization_tweets_in_India_Using_R_language

60. Słownik Języka Polskiego PWN (n.d.). Demonetyzacja. Pobrano 21 lipca 2023. Lokalizacja: https://sjp.pwn.pl/sjp/demonetyzacja;2554639.html

61. Łomanowski. A. (2017). Indie: 100 dni bolesnej reformy. Rzeczpospolita. Pobrano 21 lipca 2023. Lokalizacja: https://www.rp.pl/polityka/art10634721-indie-100-dni-bolesnej-reformy

62. Wikipedia: Wolna Encyklopedia (2022). N-gram. Pobrano 21 lipca 2023. Lokalizacja: https://pl.wikipedia.org/wiki/N-gram

63. Arun. K. & Srinagesh. A. & Makala. R. (2017). Twitter Sentiment Analysis on Demonetization tweets in India Using R language. ResearchGate. Pobrano 21 lipca 2023. Lokalizacja: https://www.researchgate.net/publication/358915759_Twitter_Sentiment_Analysis_on_Demonetization_tweets_in_India_Using_R_language

64. (…)

65. Wikipedia: Wolna Encyklopedia (2023). Big data. Pobrano 5 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Big_data

66. Wikipedia: Wolna Encyklopedia (2021). Apache Spark. Pobrano 5 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Apache_Spark

67. Gonzalez. S. (2023). Co to jest licencja BSD?. AppMaster. Pobrano 5 sierpnia 2023. Lokalizacja: https://appmaster.io/pl/blog/co-to-jest-licencja-bsd

68. Wikipedia: Wolna Encyklopedia (2022). Apache License. Pobrano 5 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Apache_License

69. Laskowski. J. (2015). Czym jest Apache Spark i w jaki sposób go wykorzystać?. iTWIZ. Pobrano 5 sierpnia 2023. Lokalizacja: https://itwiz.pl/czym-jest-apache-spark-jaki-sposob-wykorzystac/

70. Microsoft Learn (2019). What is Apache Spark. Pobrano 10 sierpnia 2023. Lokalizacja: https://learn.microsoft.com/pl-pl/previous-versions/dotnet/spark/what-is-spark#apache-spark-architecture

71. Apache Spark (n.d.). Cluster Mode Overview. Pobrano 10 sierpnia 2023. Lokalizacja: https://spark.apache.org/docs/latest/cluster-overview.html

72. Microsoft Learn (2019). What is Apache Spark?. Pobrano 10 sierpnia 2023. Lokalizacja: https://learn.microsoft.com/pl-pl/previous-versions/dotnet/spark/what-is-spark#apache-spark-architecture

73. (…)

74. Apache Spark (n.d.). Cluster Mode Overview. Pobrano 10 sierpnia 2023. Lokalizacja: https://spark.apache.org/docs/latest/cluster-overview.html

75. (…)

76. (…)

77. Wikipedia: The Free Encyclopedia (2023). Apache Spark. Pobrano 13 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Apache_Spark

78. (…)

79. Data Flair (2016). Spark RDD Operations - Transformation & Action with Example. Pobrano 13 sierpnia 2023. Lokalizacja: https://data-flair.training/blogs/spark-rdd-operations-transformations-actions/

80. Dancuk. M. (2022). Resilient Distributed Datasets (Spark RDD). PhoenixNAP. Pobrano 13 sierpnia 2023. Lokalizacja: https://phoenixnap.com/kb/resilient-distributed-datasets

81. Wikipedia: The Free Encyclopedia (2023). Apache Spark. Pobrano 14 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Apache_Spark

82. Microsoft Learn (2023). Tworzenie aplikacji ucznia maszynowego i analizowanie zestawu danych przy użyciu biblioteki MLlib platformy Apache Spark. Pobrano 14 sierpnia 2023. Lokalizacja: https://learn.microsoft.com/pl-pl/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython

83. Wikipedia: The Free Encyclopedia (2023). Apache Spark. Pobrano 14 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Apache_Spark

84. Apache Spark (n.d). Spark Streaming Program Guide. Pobrano 14 sierpnia 2023. Lokalizacja: https://spark.apache.org/docs/latest/streaming-programming-guide.html

85. Wikipedia: The Free Encyclopedia (2023). Apache Spark. Pobrano 14 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Apache_Spark

86. Horzyk. A. (n.d.). Podstawy Informatyki – Grafy. Pobrano 14 sierpnia 2023. Lokalizacja: https://home.agh.edu.pl/~horzyk/lectures/pi/ahdydpiwykl9.html

87. Rukavytsia. A. (2017). Introduction to Spark GraphX. Medium. Pobrano 15 sierpnia 2023. Lokalizacja: https://medium.com/@rukavitsya/introduction-to-spark-graphx-748f5bbcddc5

88. Microsoft Learn (2022). Przetwarzanie wsadowe. Pobrano 15 sierpnia 2023. Lokalizacja: https://learn.microsoft.com/pl-pl/azure/architecture/data-guide/big-data/batch-processing

89. Wikipedia: The Free Encyclopedia (2023). Apache Spark. Pobrano 15 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Apache_Spark

90. Apache Spark (n.d.). Tuning Spark – Data Serialization. Pobrano 15 sierpnia 2023. Lokalizacja: https://spark.apache.org/docs/latest/tuning.html

91. Strona główna Apache Spark: https://spark.apache.org/

92. Naveen (2023). Spark Deploy Modes – Client vs Cluster Explained. SparkByExamples. Pobrano 15 sierpnia 2023. Lokalizacja: https://sparkbyexamples.com/spark/spark-deploy-modes-client-vs-cluster/

93. Apache Spark (n.d.). Spark Configuration. Pobrano 16 sierpnia 2023. Lokalizacja: https://spark.apache.org/docs/latest/configuration.html

94. Sparklyr RStudio (n.d.). Configuring Spark Connections. Pobrano 16 sierpnia 2023. Lokalizacja: https://spark.rstudio.com/guides/connections

95. DalleMule. L. & Davenport. H. T. (2017). What's Your Data Strategy. Pobrano 10 sierpnia 2023. Lokalizacja: https://hbr.org/2017/05/whats-your-data-strategy

96. Ko S. & Wo J-H. (2016). Processing Large-Scale Data with Apache Spark. Pobrano 10 sierpnia 2023. Lokalizacja: https://won-j.github.io/326_621a-2018fall/hw/hw4/spark16.pdf

97. Kumar. A. (2017). Problems with Hadoop Map Reduce. Pobrano 10 sierpnia 2023. Lokalizacja: https://www.linkedin.com/pulse/problems-hadoop-map-reduce-abhinav-kumar/

98. PWN (n.d.). Monte Carlo – wyznaczamy przybliżenie liczby Pi. Pobrano 11 sierpnia 2023.

Lokalizacja: https://it.pwn.pl/plain_site/layout/set/print/Artykuly/Programowanie/Monte-Carlo-wyznaczamy-przyblizenie-liczby-Pi

99. Bhadani. N. (2021). Apache Spark Structured Streaming – First Streaming Example (1 of 6). Pobrano 11 sierpnia 2023.

    Lokalizacja: https://medium.com/expedia-group-tech/apache-spark-structured-streaming-first-streaming-example-1-of-6-e8f3219748ef

100. Chaffai A. & Hassouni L. & Anoun H. (2017). Real-Time Analysis of Students' Activities on an E-Learning Platform based on Apache Spark. Pobrano 12 sierpnia 2023.

    Lokalizacja: https://thesai.org/Downloads/Volume8No7/Paper_15-Real_Time_Analysis_of_Students_Activities.pdf

101. Hazelcast (n.d.). What is Stream Processing?. Pobrano 12 sierpnia 2023. Lokalizacja: https://hazelcast.com/glossary/stream-processing/

102. Babatunde. A. (2022). Why Spark Structured Streaming Could Be The Best Choice. NetGuru. Pobrano 12 sierpnia 2023. Lokalizacja: https://www.netguru.com/blog/spark-streaming

103. Wikipedia: The Free Encyclopedia (2023). Learning analytics. Pobrano 12 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/Learning_analytics

104. Kondas. A. (2016). Algorytm K-Średnich – Uczenie Nienadzorowane. ITCraftsMan. Pobrano 12 sierpnia 2023. Lokalizacja: http://itcraftsman.pl/algorytm-k-srednich-uczenie-nienadzorowane/

105. Witan. K. (2020). Analiza Skupień. Segmentacja rynku za pomocą grupowania metodą k-średnich. RPubs. Pobrano 12 sierpnia 2023. Lokalizacja: https://rpubs.com/katarzynawitan/664935

106. Codenga (2023). Język R – gdzie się go używa?. Pobrano 19 sierpnia 2023. Lokalizacja: https://codenga.pl/artykuly/poradniki/jezyk_r_gdzie_sie_go_uzywa

107. Wikipedia: The Free Encyclopedia (2023). R (programming language). Pobrano 19 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/R_(programming_language)

108. Wikipedia: Wolna Encyklopedia (2023). R (język programowania). Pobrano 19 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/R_(jezyk_programowania)

109. Wikipedia: Wolna Encyklopedia (2023). GNU General Public License. Pobrano 19 sierpnia 2023. Lokalizacja: https://pl.wikipedia.org/wiki/GNU_General_Public_License

110. Wikipedia: The Free Encyclopedia (2023). RStudio. Pobrano 20 sierpnia 2023. Lokalizacja: https://en.wikipedia.org/wiki/RStudio

111. Sparklyr (n.d.). R interface to Apache Spark. Pobrano 20 sierpnia 2023. Lokalizacja: https://spark.rstudio.com/

112. Sparklyr (n.d.). Manipulating Data with dplyr. Pobrano 20 sierpnia 2023. Lokalizacja: https://spark.rstudio.com/guides/dplyr.html

113. Suffyan. A. (2023). Beginner's Guide to Spark UI: How to Monitor and Analyze Spark Jobs. Medium. Pobrano 21 sierpnia 2023. Lokalizacja: https://medium.com/@suffyan.asad1/beginners-guide-to-spark-ui-how-to-monitor-and-analyze-spark-jobs-b2ada58a85f7

114. Cecchini. D. (2023). Boost Your NLP Results with Spark NLP Stemming and Lemmatizing Techniques. Medium. Pobrano 21 sierpnia 2023. Lokalizacja: https://medium.com/john-snow-labs/boost-your-nlp-results-with-spark-nlp-stemming-and-lemmatizing-techniques-8d734081264d

115. Harmon M. (n.d.). Sentiment Analysis, Part 2. Pobrano 21 sierpnia 2023. Lokalizacja: http://michael-harmon.com/blog/SentimentAnalysisP2.html

116. J. Snow Labs (2020). Bringing Spark NLP to R | NLP Summit 2020. Youtube. Pobrano 21 sierpnia 2023.

Lokalizacja: https://www.youtube.com/watch?v=QGZWddmzeag&ab_channel=JohnSnowLabs

117. Rinker. T. (2022). Calculate Text Polarity Sentiment. CranR. Pobrano 21 sierpnia 2023.

Lokalizacja: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

118. Wikipedia: Wolna Encyklopedia (2023). Korelacja wzajemna (korelacja krzyżowa). Pobrano 17 września 2023. Lokalizacja: https://pl.wikipedia.org/wiki/Korelacja_wzajemna

Spis Tabel

Spis Rysunków

Źródła Danych

Hurtownia danych 1. Dane tekstowe użytkowników platformy Twitter.com na temat pandemii koronawirusa COVID-19 wraz ze zmiennymi pomocniczymi. Dane uzyskano na stronie Kaggle.com, będącej własnością firmy Google LCC. Dostęp: 06 grudnia 2022r.
Lokalizacja: https://www.kaggle.com/datasets/gpreda/covid19-tweets

Hurtownia danych 2. Dane liczbowe odnośnie wielkości zachorowań i śmierci w każdym dniu pandemii. Dane uzyskano na stronie The Humanitarian Data Exchange, własność danych należy do Światowej Organizacji Zdrowia (WHO). Dostęp: 09 września 2023r. Lokalizacja: https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths/resource/2ac6c3c0-76fa-4486-9ad0-9aa9e253b78d#