Sprawozdanie projektu: Analiza tweetów dotyczących filmu „Avatar”

Anna Ledworowska

nr indeksu: 266376

Wstęp

Celem projektu była analiza i porównanie tweetów dotyczących pierwszej i drugiej części filmu „Avatar” reżyserii Jamesa Camerona.

Twitter

Twitter jest to serwis społecznościowy założony w 2006 roku, który daje użytkownikom możliwość publikowania krótkich wpisów nazywanych tweetami. Twitterzy dzielą się tam swoimi przemyśleniami na dowolne tematy, również na związane z rozrywką.

„Avatar”

Pierwszy film o tytule „Avatar” miał premierę w 2009, a na kolejny fani musieli czekać, aż 13 lat. Obie części zostały stworzone przy użyciu zaawansowanych technologii, przez co wzbudzają duże zainteresowanie widzów. Aktywni użytkownicy twittera komentowali swoje spostrzeżenia na jego temat.

Więcej o projekcie

Tweety zostały pozyskane za pomocą snscrape. To wpisy w języku polskim oraz angielskim umieszczone miesiąc przed i miesiąc po premierze filmu zarówno dotyczące pierwszej części filmu oraz drugiej. Porównywane zostaną statystyczne własności tweetów. Sprawdzone zostanie jak zmieniał się sentyment wpisów przed i po premierze. Cały projekt został napisany w języku Python w Jupyter Notebook.

Potrzebne paczki

Do stworzenia projektu potrzebne było zaimportowanie poniższych paczek.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Pozyskiwanie tweetów

By pozyskać tweety potrzebne do analizy została stworzona funkcja twitterScraper w której za pomocą TwitterSearchScraper wyszukiwane są tweety:

- związane z „avatar”

- napisane w przedziale danych dat

- wybranym języku.

Ilość pobranych tweetów ogranicznona jest przez podanie maksymalnej ilości wpisów, którą funkcja ma zbierać. Funkcja upewnia się czy w tweetcie jest wzmianka o avatarze.

Zbierane informacje z każdego tweeta to:

- czas jego udostępnienia,

- treść,

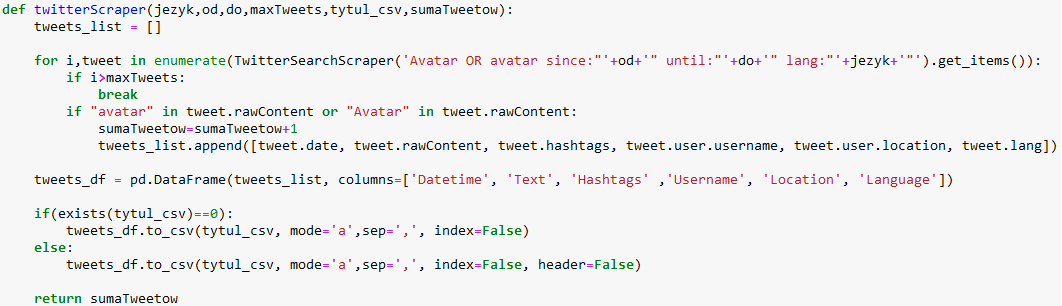
- użyte hashtagi,

- nazwa użytkownika,

- lokalizacja osoby piszącej

- język w którym został napisany.

Następnie zebrane dane są zapisywane w pliku csv. Funkcja zwraca ilość zebranych tweetów.



Kolejna funkcja została stworzona by wywoływać twitterScraper dla kolejnych dni, tak by dane były zbierane z każdego dnia w zadanym przedziale dat.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Poniżej zaprezentowane jest wywołanie funkcji tak by uzyskać tweety:

- o pierwszej części w języku polskim, miesiąc przed i po premierze ( premiera: 25.12.2009 )

- o pierwszej części w języku angielskim, miesiąc przed i po premierze ( premiera: 10.12.2009 )

- o drugiej części w języku polskim, miesiąc przed i po premierze ( premiera: 16.12.2022 )

- o drugiej części w języku angielskim, miesiąc przed i po premierze ( premiera: 14.12.2022 ).

Nałożono ograniczenie 250 tweetów na dzień.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Następnie ukazany jest wydruk z konsoli, gdzie zostało zaprezentowane ile tweetów udało się zebrać i ile czasu to zajęło.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Po mimo możliwości zebrania dziennie 250 tweetów nie każdego dnia udało się tyle zebrać, gdyż widocznie nie zostało umieszonych tyle wpisów na badany temat. Nie jest zaskakujące to, że tweetów w języku angielskim jest zdecydowanie więcej niż w polskim. Chociaż widać wzrost liczby wpisów w języku polskim, co może sugerować większą popularność Twittera w 2022 wśród polskich użytkowników niż to było w 2009.

Poniżej kilka zebranych tweetow oraz jeden z nich wyszukany na Twiterze.

Obraz zawierający tekst

Opis wygenerowany automatycznie



Część praktyczna

Przed analizą tweetów sprawdzono czy słowo „avatar” istnieje w WordNecie – leksykalnej bazie danych języka angielskiego.

Obraz zawierający tekst

Opis wygenerowany automatycznie

„Avatar” znajduje się w WordNecie, lecz pod znaczeniami: wcielenie i hinduskie bóstwo. Są one różne od znaczenia słowa „avatar”, który jest tematem projektu.

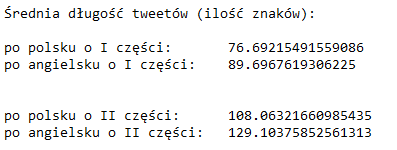
Analiza długości tweetów

Funkcja srednia\_długosc\_tweetow sprawdza ile średnio znaków miały tweety związane z „Avatarem”.

Obraz zawierający tekst, wewnątrz, zrzut ekranu

Opis wygenerowany automatycznie

Poniżej zaprezentowany jest wynik wywołania funkcji dla zebranych tweetów.



Tweety o drugiej części są zdecydowanie dłuższe, zarówno w języku polskim jak i angielskim. Może to świadczyć o tym, że kiedyś użytkownicy Twittera pisali krótsze tweety niż teraz lub rzeczywiście druga część wzbudziła większe emocje i sprawiła, że użytkownicy mieli więcej do napisania na jej temat.

Występowanie wzmianki o technologii 3D

Filmy 3D w 2009 roku były czymś nowym i niezwykłym dla wielu widzów. Wydaje się, że w dzisiejszych czasach ludzie przywykli do tej technologii i nie robi to już takiego dużego wrażenia na odbiorcach. Dlatego sprawdzono czy w tweetach jest różnica w ilości wzmianek o 3D.

Funkcja szukanie\_3D wyszukuje w każdym tweetcie frazy „3D” i zlicza ilość jej występowania.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Poniżej wynik wywołania funkcji dla zebranych tweetów.

Obraz zawierający tekst

Opis wygenerowany automatycznie

W drugiej części filmu ilość wzmianek o technologii 3D jest mniejsza. Różnicę dobrze widać w tweetach w języku angielskim. W pierwszym i drugim filmie zebrano zbliżoną liczbę tweetów w tym języku, a ilość wzmianek w części z 2022 roku jest o około 1 tysiąc mniejsza. W polskich tweetach również widać różnicę, gdyż wzmianek o 3D jest ok 115 i w jednej i drugiej części, jednak na temat młodszego filmu było około 3 razy więcej tweetów.

Tłumaczenie

By móc przeanalizować sentyment tweetów w języku polskim zostały one przetłumaczone na język angielski przy użyciu googletrans.

Funkcja tłumaczenie tłumaczy treść tweetów i zapisuje do pliku csv.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Wykorzystywana funkcja sprawdzała ile czasu zajęło przetłumaczenie zebranych tweetów w języku polskim.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Chmury tagów

Chmura tagów to graficzne zobrazowanie zawartości tekstu, jest to ciekawe ukazanie najczęściej występujących wyrazów.

W projekcie za tworzenie chmur tagów odpowiedzialna jest funkcja tworzenie\_wordClouds. By stworzyć taki rezultat dokonuje się tokenizacji, czyli rozbicia tekstu na pojedyncze wyrazy. Następnie w funkcji usuwane są słowa ze stop listy. Są to popularne słowa które mają małe znaczenie dla tekstu. Potem robiona jest lematyzacja, czyli sprowadzenie słowa do jego formy podstawowej. Informacyjnie wyświetlana jest lista 10 najczęściej występujących w danych tweetach słów. Kolejnym krokiem jest stworzenie i zapisanie chmury tagów.

Obraz zawierający tekst

Opis wygenerowany automatycznie

By uatrakcyjnić wygląd chmury tagów zostaje jej nadany kształt jak na poniższej grafice (główni bohaterzy filmu).



Wywołanie funkcji skutkuje poniższym wydrukiem.

Obraz zawierający stół

Opis wygenerowany automatycznieObraz zawierający stół

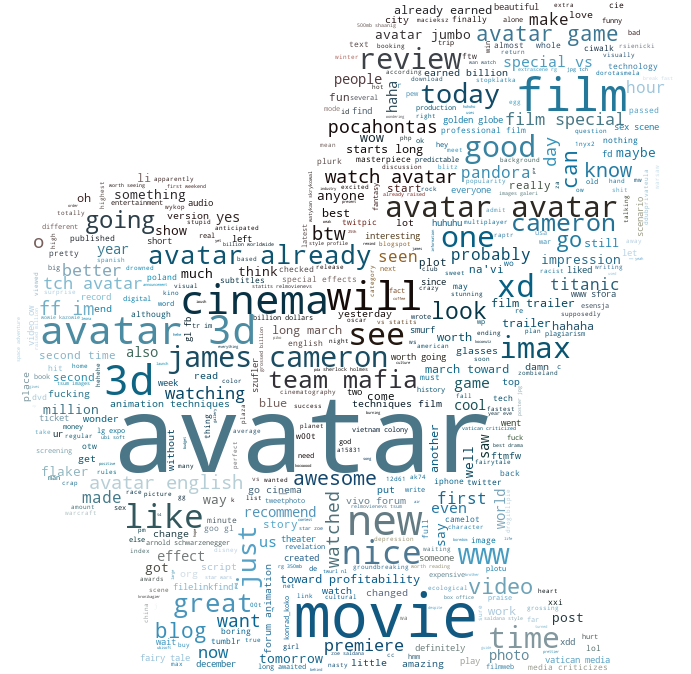
Opis wygenerowany automatycznie

Wywołanie funkcji

Poniżej znajdują się stworzone chmury tagów.

Dla pierwszej i drugiej części w języku polskim (tweety zostały przetłumaczone wcześniej na język angielski)

Obraz zawierający tekst, gazeta, paragon

Opis wygenerowany automatycznie

Dla pierwszej i drugiej części w języku angielskim.



Obraz zawierający tekst, gazeta, paragon

Opis wygenerowany automatycznie

Analiza sentymentu

Najciekaszą statystyką jest analiza sentymentu, np. czy zmieniały się nastroje w tweetach przed i po premierze filmu. W projekcie tą analizą zajmuje się funkcja analizaSentymentu. Sprawdzany jest sentyment dla kolejnych dni i z wyników tworzony jest średni wynik dla każdego dnia. Sentyment zostaje badany za pomocą vaderSentiment oraz textblob, by można było porównać dane narzędzia. Zliczana jest liczba pozytywnych i negatywnych tweetów na temat filmu.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Wydruk po wywołaniu funkcji dla zebranych tweetów o „Avatarze”

Obraz zawierający stół

Opis wygenerowany automatycznie

Porówanie działania vaderSentiment oraz textblob dla jednego konkretnego tweeta.

Treść badanego tweeta: „Avatar - The creature of water Wonderful special effects create a great spectacle worth the money spent. It's good to watch from start to finish, but the content of the film is too simple and that's why I left the cinema with a slight unsatisfaction. I recommend. This is a must watch movie.”

Treść tweeta jest raczej pozytywna, autor zachwala film lecz ma pewne zastrzeżenia co do filmu.

Obraz zawierający tekst

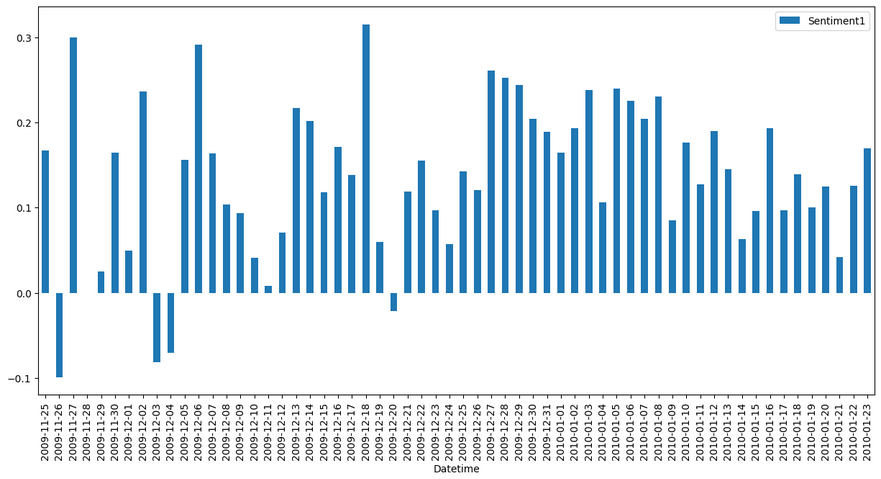
Opis wygenerowany automatyczniePoniżej widizmy jak z jego oceną poradziły sobie badane narzędzia.

Wydaje się, że lepszej oceny dokonało narzędzie vaderSentiment.

Analiza sentymentu na przestrzeni miesiąca przed i po premierze

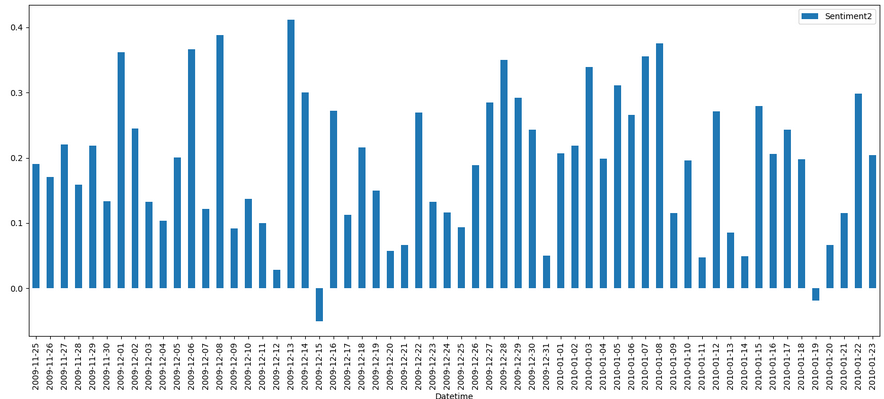
Analiza sentymentu tweetów polskich dotyczących pierwszej część filmu .

Poniżej znajduje się wykres obrazujący analizę sentymentu stworzoną przy pomocy narzędzia textblob



Analizując wykres widoczne jest, że przez większość dni średnia sentymentu tweetów przyjmuje pozytyne wyniki. Zauważalne jest również, że po premierze filmu (25.12.2009) średnia sentymentu przyjmuje bardzo zbliżone wyniki, gdzie przed premierą wachania wyników między dniami były bardzo duże.

Poniżej znajduje się wykres obrazujący analizę sentymentu stworzoną przy pomocy narzędzia vaderSentiment



Na tym wykresie również widoczne jest, że przez większość dni średnia sentymentu tweetów przyjmuje pozytyne wyniki. Niestety nie ma znaczących zmian związanych z datą premiery filmu.

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie Poniżej chmura tagów dla tweetów ocenionych jako pozytywne i negatywne.

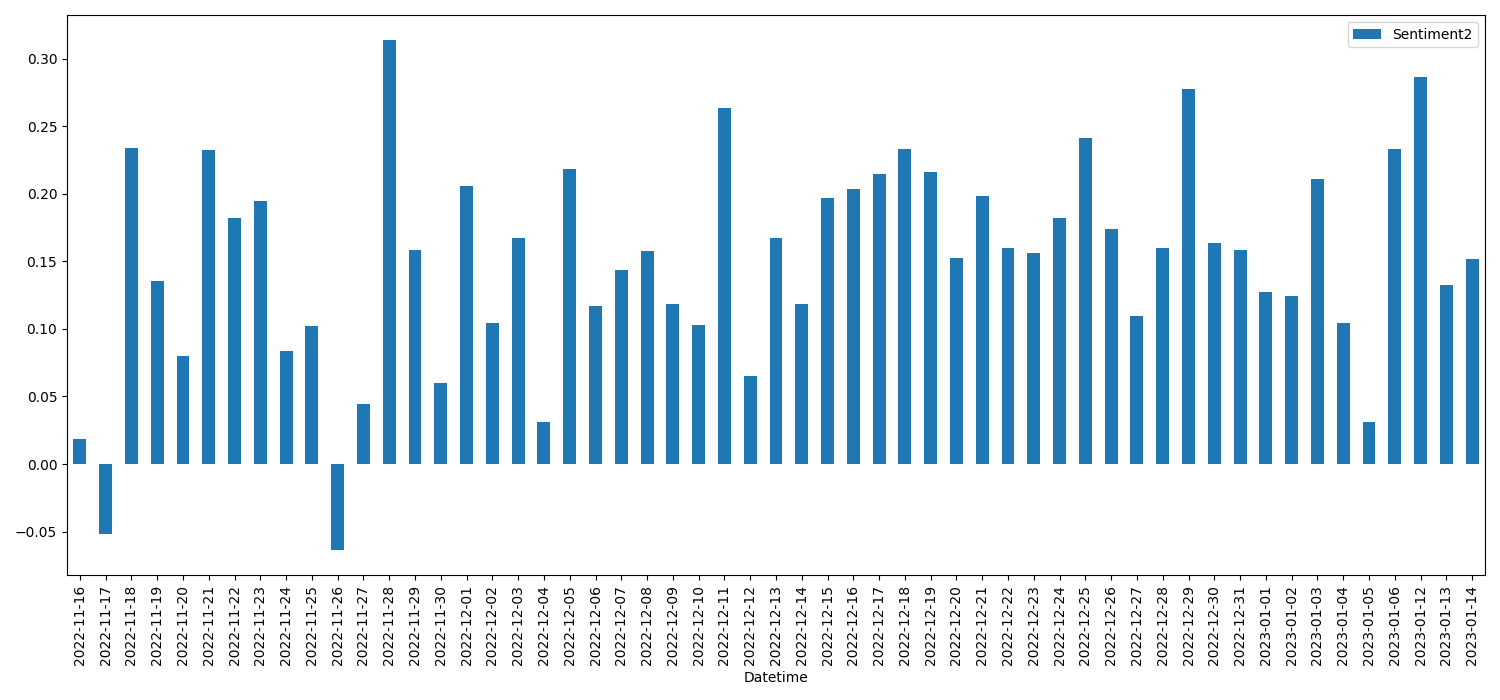
Obraz zawierający tekst

Opis wygenerowany automatycznie

Zauważalne jest, że w „pozytywnej chmurze” znajdują się słowa aprobaty co do filmu takie jak „good”, „awesome”, „great”, „nice”, których brakuje w „negatywnej chmurze”.

Analiza sentymentu tweetów polskich dotyczących drugiej część filmu .

Dla kolejnych przykładów zostanie omówiona tylko ocena syntementu stworzona dzięki vaderSentiment, gdyż narzędzie wydaje bardziej wiarygodne zwracając uwagę na wcześniejszą ocenę pojedyńczego tweetu.



Ponownie przez większość dni średnia sentymentu tweetów przyjmuje pozytyne wyniki. Po premierze filmu (16.12.2022) widzimy, że wyniki sentymentu są na ogół wyższymi wartościami niż te przed premierą oraz wachania wyników między dniami są mniejsze.

Obraz zawierający tekst, gazeta, paragon

Opis wygenerowany automatyczniePoniżej chmura tagów dla tweetów ocenionych jako pozytywne i negatywne.

Obraz zawierający tekst

Opis wygenerowany automatycznie

W chmurze stworzonej z tweetów ocenionych jako pozytywne znajdują się słowa takie jak „like”, „love”, „great”, których nie ma w chmurze stworzonej z tweetów negatywnych, lecz np. znajduje się tam słowo „long”, co może świadczyć, że użytkownicy narzekali na długość trwania filmu (ponad 3h).

* + Jak wygląda baza danych, co oznaczają kolumny, co jest klasą? Czy trzeba  
    byłą ją przetworzyć/zedytować/wyczyścić. Co oznaczają kolumny?
  + Jak były pozyskiwane dane? Np. jakie tweety ściągaliśmy, jakim  
    narzędziem?

1. Część eksperymentalna/praktyczna. Odpalamy kod pythonowy i patrzymy co  
   się dzieje. Przeprowadzamy eksperymenty na wybranych przez nas instancjach  
   problemów, testujemy klasyfikatory
2. Wnioski, interpretacja, podsumowanie

Korzystałam z - <https://betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af>

https://py-googletrans.readthedocs.io/en/latest/