Sprawozdanie

Laboratorium 3

Zadanie 3

GRUPA WCY21IJ1N1

Opracowali:

Karczewski Paweł

Tarkowski Adam

# Treść zadania

Zadanie odnosi się do praktyki całkiem częstego zastosowania naiwnego klasyfikatora bayesowskiego do klasyfikacji dokumentów. Przykładem zastosowania tego klasyfikatora jest klasyfikacja wiadomości e-mail jako nas interesującą lub spam. Inne przykłady. Odpowiadamy na pytanie czy dany post wyraża zadowolenie, obojętność, złośliwość lub agresję piszącego. Czy przechwycona widomość powinna zostać przekazana Policji? Itp.

Dane załączone do zadania pochodzą z postów w serwisie Twitter dotyczące portalu mandrill.com firmy MailChimp. Portal służy do przesyłania informacji handlowych za pośrednictwem e-mail i jest przeznaczony dla programistów, którzy piszą aplikacje do wysyłania zindywidualizowanych widomości, powiadomień, faktur, wezwań do zapłaty, itd.

Zadanie polega na stworzeniu modelu, który odróżnia intersujące nas posty od postów nieinteresujących, a które traktujemy jako szum informacyjny. Interesuje nas aplikacja Mandrill, tzn. chcemy zakwalifikować opublikowane posty w serwisie Twitter odnoszące się tylko do aplikacji Mandrill jako „Mandrill”, a te które nie odnoszą się do niej, ale odnoszą się do innych rzeczy związanych z rzeczownikiem „mandrill” zakwalifikujemy jako „inne”.

Zadanie jest namiastką przetwarzania języka naturalnego (ang. NLP – Neuro Linguistic Programming). W takim przypadku prawie zawsze należy przygotować treść napisaną przez użytkownika (w naszym przypadku postów opublikowanych w serwisie Twitter) do przetworzenia przez model.

W załączonym do zadania pliku w formacie .xlsx „MED-lab-3-Zad 3-Mandrill-Dane.xlsx” znajdują się dwa arkusze zawierające posty odnoszące się do aplikacji Mandrill oraz do „innych rzeczy”. Proszę zwrócić uwagę na wielojęzyczność postów.

Uwaga. W zadaniach polegających na przetwarzaniu języka naturalnego zamiast odrzucenia wszystkich krótkich słów usuwa się tylko te słowa, które wchodzą w skład słów przystankowych danego języka (w wiadomościach napisanych w j. angielskim są to słowa pochodzące z tzw. „stop list”. Są to słowa charakteryzujące się niską zawartością leksykalną. Z uwagi na to, że przedstawione dane zawierają posty w j. angielskim prześledźmy to na przykładzie. W języku angielskim przykładami takich słów są „because” lub „instead”, które mogą się występować w wielu grupach postów. Jednak większość słów o niskiej zawartości leksykalnej jest krótka lub bardzo krótka – są to na przykład „a”, „an”, „the”, itp. Wobec tego proszę w zadaniu uprościć proces przetwarzania postów i usunąć z nich słowa o niskiej zawartości leksykalnej. Innymi słowy podzielić na leksemy (znaczenie patrz niżej).

Słownik PWN: „leksem - wyraz lub wyrażenie traktowane jako jednostka słownikowa”

Encyklopedia PWN: „leksem [gr. léxis ‘wyraz’], wyraz jako abstrakcyjna jednostka systemu językowego, wyraz słownikowy; na leksem składają się: określone znaczenie leksykalne, zespół wszystkich funkcji gramatycznej oraz ogół form językowych reprezentujących w tekście l. w jego poszczególnych funkcjach; np. pol. formy obraz, obrazami, obrazie reprezentują l. obraz w jego 3 różnych funkcjach gramatycznych (Obraz jest wystawiony w muzeum; Krytyk zachwycił się obrazami ekspresjonistów; Na obrazie widać krajobraz górski); w szczególnych wypadkach l. może być reprezentowany w tekście przez jedną i tę samą formę, np. miło, wczoraj, natomiast (wyrazy nieodmienne).

# Scenariusz postępowania

Na początku warto abyśmy znali jakie słowa zawarte są w postach zarówno tych dotyczących mandrilla jak i pozostałych. Aby to zrobić warto jest znormalizować posty tzn dla nas tym samym jest słowo które zaczyna się z wielkich liter jak i tych z małej. Do tego jest wiele niepotrzebnych słów, które możemy wyrzucić, bo nie naprowadzają nas na informacje czy są to posty dotyczące mandrilla czy nie. Usunięte zostaną słowa z języka angielskiego głównie dlatego że jest ich przeważająca ilość, oraz znamy ten język. Usunięte zostaną również znaki interpunkcyjne.

Jak już będziemy mieli słowa to chcemy znać prawdopodobieństwo pod warunkiem że są to posty dotyczące Mandrilla jak i prawdopodobieństwo, że są w postach o innej tematyce.

Gdy już mamy prawdopodobieństwo każdego słowa to możemy nasze znormalizowane posty sprawdzając prawdopodobieństwo każdego słowa mnożąc je przez siebie a na końcu pomnożymy przez prawdopodobieństwo czy jest bądź nie postem o tematyce Mandrilla.

Jeżeli nie chcemy, aby wyszły nam głupoty, bo w danych testowych mogą być słowa, których nie ma w sprawdzanym poście, to musimy zrobić wygładzanie Laplace’a.

Interesuje nas taki iloczyn:

Gdzie:

Sprawdzamy dla 2 klas jakie może przyjąć C i wybieramy większy iloczyn.

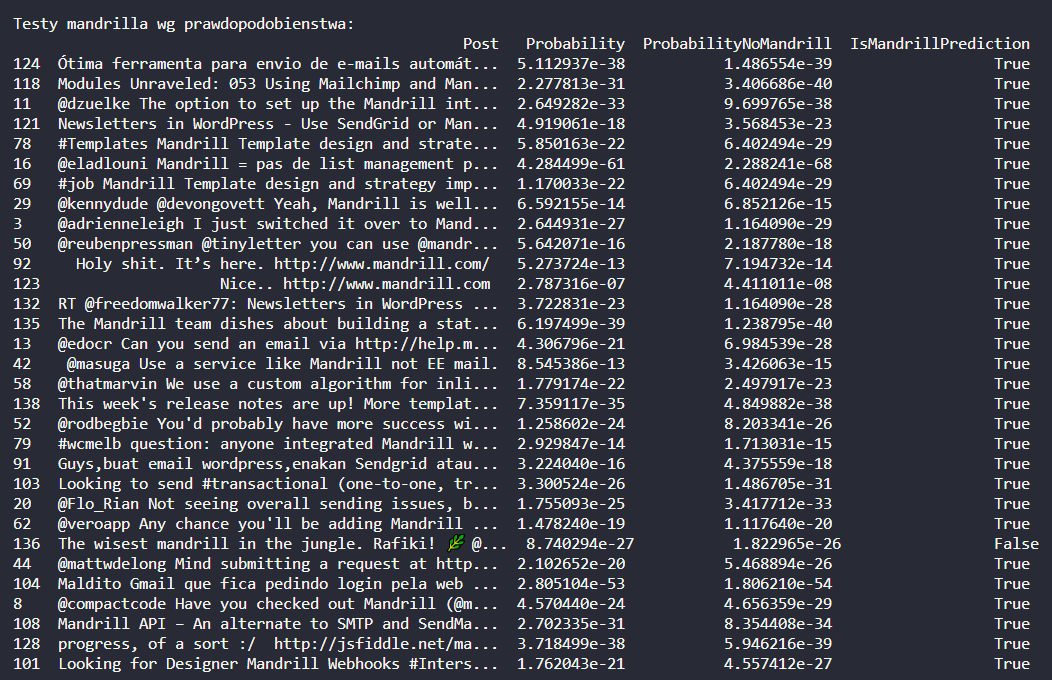
Każde z prawdopodobieństw jeszcze poprawiamy o wygładzenie Laplace’a według następującego wzoru:

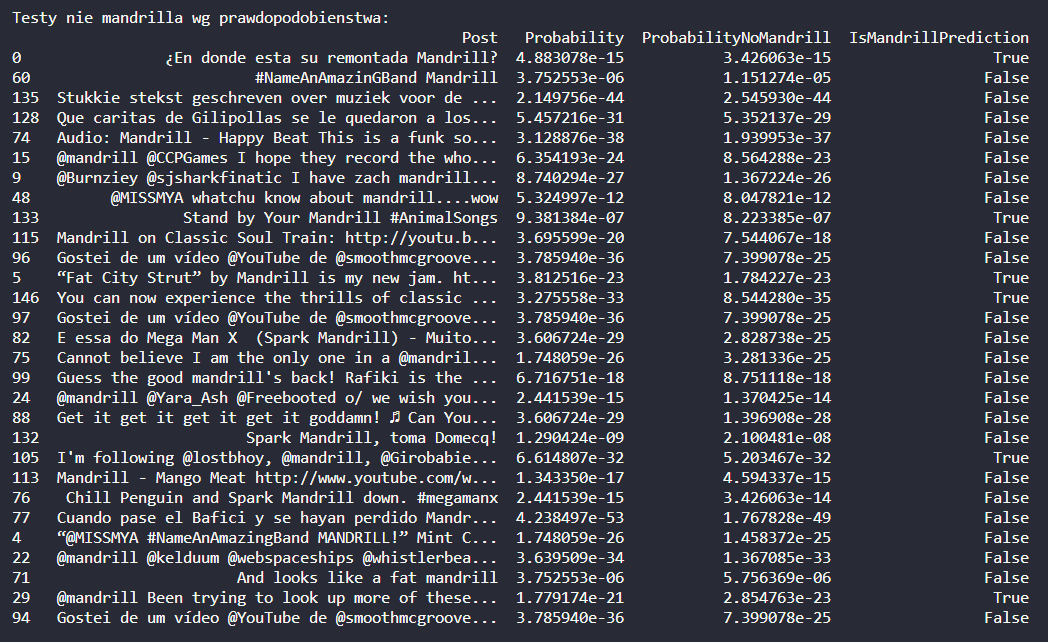
Gdzie:

# Obliczenia

Obliczenia zostały zrobione przy użyciu języka python w wersji 3.11.5. Kod jest dołączony do sprawozdania “CalculationsBayes.py”.

Wyniki nauki oraz testów:





Wychodzi, że prawdopodobieństwo wykrycia postu o Madrillu pod warunkiem, że nim właśnie jest wynosi ok 97% a wykrycie innego postu pod warunkiem, że nie dotyczy Mandrilla wynosi ok 79%.  
  
Oczywiście są to przykładowe wyniki, bo dane są dzielone na zbiór uczący i testujący w losowy sposób.

Jeżeli jednak usunęlibyśmy część z usuwaniem słów i znaków dalej mamy nie najgorsze wyniki, ale nie tak dobre. Odpowiednio ok 97% i 76%.

# Podsumowanie

Jak widać klasyfikacja Bayesa działa w tego typu zadaniach. Mimo to warto, aby przy opracowywaniu, pełnego algorytmu uczestniczyli również osoby związane z filologią, bo usunięcie paru słów mogłoby dać nam kilka punktów procentowych więcej.