METODY EKSPLORACJI DANYCH
Laboratorium
Klasyfikacja postów na podstawie naiwnego klasyfikatora bayesowskiego

Zadanie 3

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)."