Sprawozdanie

Laboratorium 3

Zadanie 2

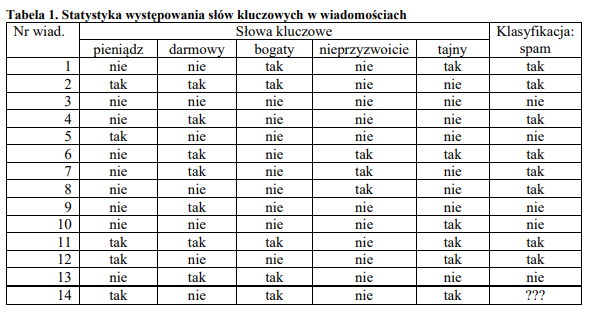
Opracowali:

Karczewski Paweł

Tarkowski Adam

# Treść zadania

Załóżmy, że chcemy zbudować algorytm odfiltrowania spamu z otrzymanych wiadomości pocztą lub za pomocą czatu. W spamie występują treści wg. określonych słów kluczowych, których statystyka występowania została zawarta w tabeli wraz z dokonaną klasyfikacją wg. pewnego1 algorytmu.



W zadaniu

1. opierając się na klasyfikacji bayesowskiej proszę określić status wiadomości/posta nr 14 na podstawie występowania lub niewystępowania słów kluczowych.
2. Przedyskutować wynik otrzymanego rezultatu. Przedyskutować, czy naiwna klasyfikacja bayesowska jest dobrą metodą wykrywania spamu. Proszę uzasadnić swoje wnioski.
3. wyniki swojej pracy proszę zawrzeć w sprawozdaniu. Do sprawozdania proszę dodać jako załączniki wszystkie pliki z obliczeniami.

# Scenariusz postępowania

Dla danych prawdopodobieństw:

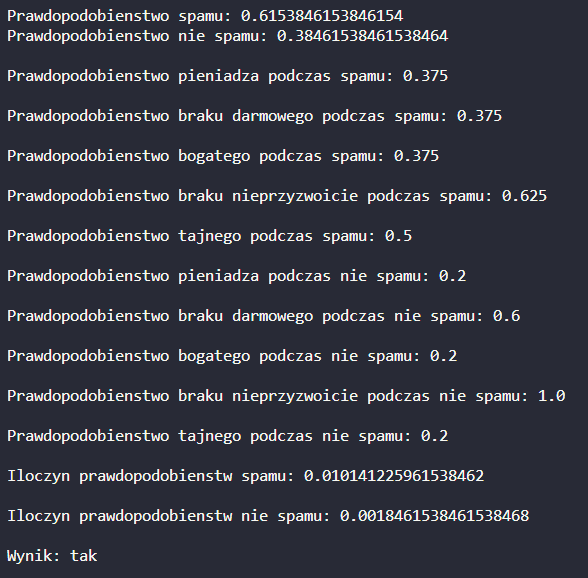
1. Spam P1, a w nim:
   1. Występuje pieniądz P1\_1
   2. Nie występuje darmowy P1\_2
   3. Występuje bogate P1\_3
   4. Nie występuje nieprzyzwoicie P1\_4
   5. Występuje tajny P1\_5
2. Normalna wiadomość P2, a w nim:
   1. Występuje pieniądz P1\_1
   2. Nie występuje darmowy P1\_2
   3. Występuje bogate P1\_3
   4. Nie występuje nieprzyzwoicie P1\_4
   5. Występuje tajny P1\_5

musimy rozwiązać zadanie optymalizacyjne:

# Obliczenia

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

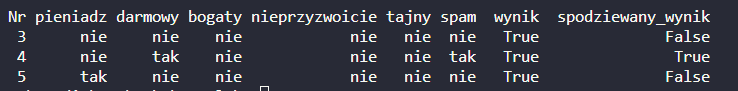
Wyniki obliczeń:



Jak widać wiadomość 14 została oznaczona jako spam.

# Ocena modelu

Ogólnie to ciężko jest przetestować model na tak małej liczbie danych. Mimo to próba została podjęta. Wybrane zostały 3 wpisy do oceny działania modelu (3,4,5). Przy takim doborze danych trenujących i testowych mamy wynik, który odbiega od wartości których się spodziewaliśmy:



Pokazuje to, że albo potrzebowalibyśmy więcej danych testowych albo sposób użycia naiwnej klasyfikacji bayesowskiej jest nie trafiony.

Trzeba zaznaczyć, że przy ocenie modelu zastosowaliśmy wygładzanie Laplace’a.

# Podsumowanie

Udało nam się stworzyć klasyfikator. Na podstawie wyników ciężko jest to ocenić czy to dobry model. Ocena nie poszła po naszej myśli.