Laboratorium 1 - 10 marca 2025

Przetwarzanie Języka Naturalnego

Zadanie 1.(2 pkt) - Odległość edycyjna

Rozważmy następujące operacje na ciągach:

- **insert**(x, i, a) - wstawienie a pomiędzy i-tym i $(i+1)$-szym elementem $x - a$;
- **delete**(x, i) - usunięcie i-tego elementu $x - a$;
- **replace**(x, i, a) - zastąpienie i-tego elementu $x - a$ przez a.

Jak łatwo zauważyć, dla każdych dwóch ciągów x i y istnieją sekwencje powyższych operacji przekształcające x w y. Jeśli każdej operacji przypiszemy koszt (nieujemną liczbę rzeczywistą), możemy mówić o minimalnym koszcie przekształcenia x w y.
(1 pkt) Ułóż algorytm, który dla danych dwóch ciągów znajdzie ich minimalny koszt.
(1 pkt) Jak się zmieni algorytm, kiedy dopuścimy operację zamiany sąsiednich znaków miejscami?

Zadanie 2.

Jaki jest koszt dla słów $algorytm$ i $aforyzm$? Pokaż działanie algorytmu dla tych dwóch słów.

Zadanie 3.

Zaimplementuj algorym z **Zadania 1.**, który opcjonalnie będzie wypisywał macierz przekształceń dla zadanych słów. Koszty operacji to 1 dla usuwania i wstawiania, 2 dla zamiany. Wykorzystaj go do wygenerowania przykładowych słów, których koszt względem słowa *zadanie* wynosi 2.
Parametr verbose będzie odpowiedzialny za wypisanie na standardowym wyjściu macierzy kosztów.

```
def solution(word1, word2, verbose = True/False):
    ...
```

Zadanie 4.

Rozwiń rowiązanie **Zadania 3.** generując tylko te słowa, która są w słowniku. Dodaj funkcjonalność, która pozwoli na generowanie wszystkich słów o zadanym koszcie względem słowa wejściowego. Słownik można pobrać: https://sjp.pl/sl/growy/

Zadanie 5. Porównaj SVM i Naive Bayes na zbiorze **Spam**:

- Zaimplementuj oba klasyfikatory
- Zmierz czas treningu i F1-score
- Wyjaśnij wyniki w kontekście założeń modeli

Zadanie 6. (2 pkt)

Pokaż działanie **Multinomial Naive Bayes (MNB)** w kontekście prostego problemu klasyfikacyjnego w jednowymiarowej przestrzeni cech. Załóżmy, że mamy zbiór danych, w którym każda próbka składa się z pojedynczej cechy całkowitej:

Rozważmy zbiór, w którym każda próbka reprezentowana jest przez jedną cechę o wartości całkowitej. Załóżmy, że mamy dwie klasy:

- **Klasa** C_A: $\mathbf{x} = \{1, 1, 2, 2, 3\}$,
- **Klasa** C_B: $\mathbf{x} = \{3, 4, 4, 5, 5\}$.

Zdefiniujmy słownik cech jako zbiór unikalnych wartości:

$$V = \{1, 2, 3, 4, 5\}.$$

Dla powyższych danych:

1. **Oblicz funkcję częstości cech:**
   Dla każdej klasy $C_k \in \{C_A, C_B\}$ oblicz liczbę wystąpień każdej wartości cechy $x_i \in V$, czyli wyznacz funkcję:

   $$f(x_i, C_k), \quad \text{dla } x_i \in V.$$

2. **Wyznacz warunkowe prawdopodobieństwa cech:**
   Przyjmując wartość parametru wygładzania Laplace'a $\alpha = 1$, oblicz warunkowe prawdopodobieństwa

   $$P(x_i \mid C_k) = \frac{f(x_i, C_k) + \alpha}{N_{C_k} + \alpha \cdot |V|},$$

   gdzie

   $$N_{C_k} = \sum_{x_i \in V} f(x_i, C_k)$$

   oznacza całkowitą liczbę wystąpień cech w klasie C_k.

3. **Określ granicę decyzyjną:**
   Granica decyzyjna wyznaczana jest na podstawie równości iloczynów prawdopodobieństw warunkowych:

   $$P(x \mid C_A) = P(x \mid C_B).$$

   Podstawiając wzór na prawdopodobieństwa dla MNB, otrzymujemy wzór na granicę decyzyjną:

   $$\frac{f(x, C_A) + \alpha}{N_{C_A} + \alpha \cdot |V|} = \frac{f(x, C_B) + \alpha}{N_{C_B} + \alpha \cdot |V|}.$$

   Rozwiązując to równanie względem wartości cechy x, otrzymujemy punkt (lub punkty) graniczne, w których model nie wykazuje preferencji między klasami.

Odpowiedz na poniższe pytania:

- Dlaczego stosujemy wygładzenie Laplace'a? Co by się stało, gdybyśmy nie zastosowali wygładzania, zwłaszcza przy zerowych wartościach liczebności?
- Jakie zmiany w obliczeniach i decyzji klasyfikacyjnej nastąpiłyby, gdyby w zbiorze danych pojawiły się dodatkowe, nowe wartości cechy, których nie ma w obecnym słowniku? Rozważ dodanie jednej nowej wartości.
- Jak wpływa liczba unikalnych wartości cech na klasyfikację w MNB?

Docs