# Sprawozdanie z laboratorium 10

Krzysztof Makieła

Tematem zadania było przetwarzanie języka naturalnego oraz jego wektoryzacja.
Z powodów technicznych nie udało mi się pobrać dużego korpusu tekstów więc pracowałem na kilkunastu dokumentach, chociaż działanie jest skalowalne dla większej ilości (chociaż zastosowane rozwiązanie jest naiwne i ilość słów wyznacza wielkość macierzy i wektorów).
Przykładowe działanie programu:
Dla podsumowań artykułów (wikipedia.summary) o tytułach:

```
["Knife", "Otter", "Epica", "Shakespeare", "Tiger", "Julius Ceasar", "Sun", "Bridge", "Poker", "Sony"]
```

Wyniki zapytań wyglądają następująco:

```
Best matches for query: i like card games very much
Poker

Best matches for query: Why are all symphonic metal bands dutch?
Epica
Shakespeare

Best matches for query: Aggressive, carnivorous animals are better than passive herbivores
Otter
Tiger
Julius Ceasar


Process finished with exit code 0
```

Kod zostanie dołączony do dokumentacji, a odpowiadając na najważniejsze pytania:

- Reprezentacja dokumentów wygląda tak:

| | I | love | dogs | hate | and | knitting | is | my | hobby | passion |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc 1 | 0.18 | **0.48** | 0.18 | | | | | | | |
| Doc 2 | 0.18 | | 0.18 | **0.48** | 0.18 | 0.18 | | | | |
| Doc 3 | | | | | 0.18 | 0.18 | **0.48** | **0.95** | **0.48** | **0.48** |

- Zapytanie jest reprezentowane w postaci wektora, gdzie wartościami jest liczba wystąpień danego słowa w zapytaniu
- Zastosowana metryka podobieństwa (cosinusowa):

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|} = \frac{\sum_{i=1}^{|V|} x_i y_i}{\sqrt{\sum_{i=1}^{|V|} x_i^2} \sqrt{\sum_{i=1}^{|V|} y_i^2}}$$

gdzie x to wektor zapytania a y to wektor pojedynczego dokumentu

- Do reprezentacji dokumentu został wykorzystany algorytm tf - idf, w celu wyznaczenia właściwych słów kluczowych