Text mining
Pracownia 4
Pierwsza część pracowni jest pierwszych zajęciach po 5 maja

Uwaga: podczas pierwszej części należy oddać co najmniej jedno zadanie za 7 punktów.

Zadanie 1. (7p) Napisz prostą wyszukiwarkę do (małej) Wikipedii, w której użytkownik podaje tytuł i otrzymuje w wyniku treść artykułu. Wyszukiwarka powinna korygować błędy użytkownika, a czas korekty powinien być mniejszy niż 1 sekunda. Przykładowe zapytania:

- uniweraytrer wroclawaski
- progamownaienie lfogiczne
- programwowanielogiczne
- to nie jest kraej dla styartrych liddzo

W przypadku, gdy użytkownik wpisze (być może z błędami) podciąg wielu tytułów, powinieneś zwrócić artykuł o najkrótszym tytule, remisy rozstrzygając za pomocą numeru artykułu (preferując wcześniejsze). W ostatnim przypadku mamy dwa artykuły (książka i film), które pasują tak samo. W takich sytuacjach, w których tytuły różnią się słowem (słowami) w nawiasie, powinieneś wyświetlać wszystkie trafienia (jest to wszakże zachowanie opcjonalne, warte bonusowe +1).

Zadanie 2. (8p) W tym zadaniu i kolejnych interesuje nas wyszukiwanie tekstów podobnych do danego. Przy czym testować je będziemy używając naszego zbioru pytań, z których znaczna część przypomina poniższe:

> Jak nazywa się pojazd z siodełkiem napędzany siłą mięśni nóg?

czyli przedstawia pewną definicję i pyta o nazwę. Mając własny zbiór definicji, możemy je porównywać i wybierać obiekt, którego definicja jest najbardziej podobna do tej z pytania.

Rozważamy dwa źródła definicji:

1. Definicje z Wikisłownika (będą dostępne na stronie kursu w wygodnym formacie tekstowym)
2. Definicje z Wikipedii (zakładamy, że definicją jest początkowy fragment artykułu w Wikipedii, przy czym można dowolnie wybrać sposób wyznaczania tego fragmentu)

Twoim zadaniem jest:

a) Napisać funkcję, która wybiera z pytania część definiującą („pojazd (...) nóg"), zwracającą pusty napis, jeżeli pytanie nie jest pytaniem o nazwę.

b) Napisać wyszukiwarkę podobnych definicji, traktując w tym zadaniu podobieństwo jako cosinus rzadkich wektorów TF-IDF. Obsługa pojedynczego pytania powinna zająć mniej niż sekudnę (dla szybkich komputerów istotnie mniej niż sekundę). Otrzymany ranking nie musi mieć własności *safe*.

c) Połączyć funkcjonalność z tego zadania z Twoim aktualnym systemem odpowiadania na pytania. Przetestować osobno nowy moduł (w wariancie z Wikipedią i Wikisłownikiem) oraz jakąś formę integracji obu źródeł ze starym kodem

Zadanie 3. (7p) Kontynuujemy zadanie poprzednie. Napisz wyszukiwarkę podobnych definicji korzystającą z zanurzeń wektorów. Możesz skorzystać z dowolnych wektorów nauczonych przez Ciebie, zarówno dla słów, jak i form bazowych, możesz również wczytać wektory nauczone przez kogoś innego[1]. Podobnie, jak w poprzednim zadaniu, przetestuj swoje rozwiązanie w systemie odpowiadania na pytania.

Zadanie 4. 1-4p ⋆ To jest zadanie bonusowe, które daje:

[1] Na SKOS powinna pojawić się informacja o różnych wektorach i korpusach. Wykład 5.05.2021 będzie zawierał więcej informacji o zanurzeniach i metodach przydatnych w tym zadaniu

- 1p, jeżeli Twój system odpowiadania na pytania odpowie na 130-149 pytań
- 2p, jeżeli Twój system odpowiadania na pytania odpowie na 150-169 pytań
- 3p, jeżeli Twój system odpowiadania na pytania odpowie na 170-199 pytań
- 4p, jeżeli Twój system odpowiadania na pytania odpowie na ponad 200 pytań