Text mining
Pracownia 2
Zajęcia 4 i 5

Uwaga: Jedno zadanie z tej listy można przełożyć na pracownię 3.

Zadanie 1. (6p) Napisz wyszukiwarkę dla Wikicytatów, która uwzględnia lematyzację. Tzn. dla zapytania $w_1w_2w_3$ zwraca te cytaty, które zawierają jakiś lemat wspólny z każdym ze słów w_1, w_2 oraz w_3.

Zadanie 2. (8p) Napisz program MCR (Mistrz Ciętej Riposty), który umożliwia przeprowadzenia dialogu człowieka z komputerem, w którym wypowiedzi komputera są

a) wzięte z wikicytatów,

b) pasujące w pewnym stopniu do wypowiedzi przedmówcy.

Należy zaproponować podział dłuższych cytatów na „minicytaty", które będą potencjalnymi odpowiedziami. Program dodatkowo powinien:

1. tworzyć sobie wewnętrzny ranking pasujących wypowiedzi do danego pytania,
2. losować odpowiedź w ten sposób, by odpowiedzi na szczycie rankingu miały dużo większą szansę być wylosowanymi (losowanie powinno być możliwe do wyłączenia, w tej wersji odpowiedź jest po prostu liderem rankingu)
3. korzystać z prostego stemmera lub lematyzacji,
4. (generalnie) nie powtarzać wypowiedzi w ramach jednego dialogu,
5. pilnować, by wypowiedź nie była nigdy kopią (lub prawie kopią) wypowiedzi poprzedzającej
6. reagować jakimś ogólnym cytatem w przypadku nieznalezienia niczego pasującego
7. unikać takich słów jak *on*, *go*, *ich*, których znaczenie zależy od kontekstu.

Przykładowe dialogi hipotetycznego MCR-a znajdują się na SKOS-ie. Do tworzenia Mistrza Ciętej Riposty jeszcze wrócimy na kolejnych listach.

Zadanie 3. (8p) W zadaniu tym powinieneś napisać wyszukiwarkę dla Wikipedii. Wymagamy tu oddzielenia procesu indeksowania treści od wyszukiwania, czyli podczas indeksowania powinny powstać listy postingowe, które następnie należy zapisać do bazy danych (lub do pliku). Proces wyszukiwania z kolei powinien startować szybko i wczytywać listy postingowe do pamięci w sposób leniwy (czyli tylko wtedy, gdy któraś jest potrzebna).

Wyszukiwarka powinna prezentować wyniki w kolejności uwzględniającej następujące rzeczy:

a) Trafienia w tytule są cenniejsze od trafienia poza tytułem

b) Trafienia dokładne są cenniejsze od trafienia niedokładnego (czyli zgodności lematu)

c) Lekko preferowane są dokumenty o mniejszych identyfikatorach (czyli występujące wcześniej w pliku z Wikipedyjką)

Do prezentacji wyników wykorzystaj kolory, w celu zwiększenia czytelności. Wyświetlaj zarówno tytuł artykułu, jak i fragment jego treści, zawierający termy z zapytania. Wystarczy, że Twój program zadziała dla zmniejszonej Wikipedii (patrz SKOS).

Zadanie 4. 7 Wygeneruj dla słów formy superbazowe (patrz lista ćwiczeniowa 1). Zdefiniujemy formę superbazową dla frazy jako połączenie form superbazowych dla składników tej frazy. Czyli dla frazy:

piłem wino na barce

forma superbazowa to na przykład

pić|picie|piła wino|wina na barka|bark|barek

Napisz program, który przegląda 2 i 3-gramy z Narodowego Korpusu Języka Polskiego (linki znajdziesz na SKOS) i znajduje sytuacje, w których tę samą formę superbazową mają frazy, w których na pewnej pozycji są wyrazy nie posiadające wspólnego lematu. Przykładowo

piłą win na **bark**
piła wino na **barce**

Jak procentowo jest to częste dla 2 i 3-gramów? Czy wydaje Ci się to groźne dla wyszukiwarki?