Eksploracja tekstów
ćwiczenia 4
Zajęcia 30. maja i 1. czerwca

Zadanie 1. Wiemy, że N-gramy są a) użyteczne, b) zajmują sporo miejsca w pamięci. Zaproponuj efektywny sposób obliczania częstości N-gramów z dużego tekstu w następujących scenariuszach :

a) mamy dowolnie wiele pamięci (albo tekst jest wystarczająco krótki)

b) wydaje się, że wszystkie n-gramy powinny się zmieścić w pamięci, ale (najwyraźniej) standardowe struktury danych są zbyt rozrzutne (bo się nie mieszczą). Jakie modyfikacje algorytmu z punktu a) można zaproponować?

c) Pamięci operacyjnej jest za mało na wszystkie N-gramy i żadna kompresja tego najprawdopodobniej nie zmieni. Mamy za to wystarczającą ilość pamięci dyskowej.

Zadanie 2. Dokumenty Wikipedii mają swoje kategorie. Te z kolei mają „nadkategorie", które z kolei ... (itd). Co można powiedzieć o tym grafie (jest drzewem? dagiem?). Zaproponuj metodę wykorzystania grafu kategorii w znajdywaniu dokumentów podobnych.

Zadanie 3. Kategorie mają różną użyteczność w określaniu podobieństwa artykułów. Jedną z miar użyteczności jest liczba dokumentów w kategorii (małe kategorie są generalnie bardziej *zwarte*. Ale nie powinna to być miara jedyna. Przykładowo kategorie: *Pochowani na cmentarzu X*, *urodzeni w roku N* są mniej użyteczne od np. *Drzewa (informatyka)*, nawet wówczas, kiedy mają mniej elementów. Jak rozwiązać ten problem?

Zadanie 4. Zaproponuj metodę wyznaczania zanurzeń dla tytułów Wikipedii. Istotne jest, by były one umieszczone we wspólnej przestrzeni ze słowami, tak, żeby na przykład w otoczeniu słowa *wino*, mogły się znaleść tytuły *piwo_pszeniczne*, czy *wiśniówka_(wódka)*.

Zadanie 5. Zaproponuj metodę wyznaczania zanurzeń dla tytułów Wikipedii **oraz** dla kategorii (znajdujących się w tej samej przestrzeni). Jak wykorzystać te zanurzenia, żeby zmienić strukturę kategorii Wikipedii na drzewo (ew. na las)?

Zadanie 6. Pewne informacje o słowie można znaleźć w jego budowie. Przykładem są słowa: biologia, dendrologia, biolog, dendrolog, minimalizm, czołgista, sprzedawczyni, itd. Zaproponuj scenariusz użycia word2vec, w którym wyznaczane są zanurzenia dla słów, i zarazem wykorzystywana jest (w pewnym stopniu) budowa słowa.[1]

Zadanie 7. Załóżmy, że zanurzenia, które otrzymałeś pozwalają wyznaczać różne analogie, typu: mężczyzna do kobiety ma się tak jak król do królowej, albo Paryż do Francji ma się tak, jak Moskwa do Rosji. Zaproponuj przynajmniej dwa sposoby wykorzystania tego zjawiska.

Zadanie 8. Na potrzeby tej listy zdefiniujemy *drzewo rozbioru* zdania jako sposób postawienia w tym zdaniu (zagnieżdżonych) nawiasów, w taki sposób, by każdy nawias zawierał dokładnie dwa elementy (wyrazy, albo frazy w nawiasie), a ponadto, by te nawiasy oddawały jakoś strukturę gramatyczną zdania. Przykładowo, dla zdania *Młody krokodyl szybko pożerał kąpiące się różowiutkie prosięta* drzewem rozbioru mogłoby być:

((Młody krokodyl) ((szybko pożerał) ((kąpiące się) (różowiutkie prosięta))))

Zaproponuj prosty, zachłanny algorytm, który wyznacza drzewo rozbioru, korzystając z zanurzeń słów.

Zadanie 9. Jeżeli wykonamy zanurzenia dla korpusów w dwóch językach, to wektory dla *women* oraz *kobieta* nie będą miały ze sobą związku. Zaproponuj jakiś sposób liczenia zanurzeń, w których słowa będące swoimi tłumaczeniami będą otrzymywać podobne wektory.

Powinieneś wykorzystywać korpusy w obu językach i (być może) jakieś inne dane.

---

[1] Oczywiście (zob. W12) podobną rzecz robi FastText. Ale w tym zadaniu powinieneś wykorzystać word2vec, traktując go jako czarną skrzynkę i nie modyfikując jego kodu