Na 30 października:

|  |
| --- |
| * ~~Każdy dokument z pliku documents.txt składa się z tytułu (pierwsza linia) oraz treści właściwej; tworząc jego reprezentację w przestrzeni TF-IDF potraktuj tytuł i treść jako "całość"; dokumenty oddzielone są pustą linią;~~ * ~~W pliku keywords.txt termy znajdują się w osobnych liniach;~~ * ~~Wszystkie termy (z keywords i documents) muszą być poddane działaniu algorytmu stemmingu (skorzystaj z gotowej implementacji z http://tartarus.org/~martin/PorterStemmer/, o samym algorytmie możesz poczytać w pliku potter-algorithm.txt); nie zapomnij o zamianie wszystkich liter na małe i wyeliminowaniu z dokumentów znaków interpunkcyjnych, itd.; (powinna istnieć możliwość wyświetlenia dokumentów i termów już w przetworzonej postaci);~~ * ~~Program musi umożliwiać wczytanie pliku z dokumentami i pliku ze zbiorem termów;~~ * Program musi umożliwiać zadanie dowolnego zapytania; * Wyniki powinny być prezentowane w formie listy uporządkowanej zgodnie z malejącą miarą podobieństwa do zapytania; pojedyncza pozycja na liście powinna mieć format (tytuł dokumentu; wartość miary podobieństwa); * Wymagany (prosty) interface użytkownika; * Weryfikacja poprawności działania odbywa się na podstawie zadania kilku zapytań i sprawdzeniu wynikowego uporządkowania dokumentów; * ~~Dowolny język programowania spośród tych, dla których dostępny jest PorterStemmer;~~ |