

Wyszukiwarka składa się zasadniczo z 3 części.

Klasa *DocumentParser* odpowiada za ściągnięcie strony, usunięcie z niej w całości tagów odpowiedzialnych za css oraz skrypty, wyciągnięcie zawartości pozostałych tagów i przetworzenie jej w listę słów. Następnie z tej listy odfiltrowywane są słowa nie mające znaczenia dla wyszukiwania (wczytywane z plików *stopwords_eng* i *stopwords_pl*).

Klasa *PageIndex* odpowiada za przechowywanie linków wraz z najczęściej występującymi w nich słowami (i ich liczebnościami). Tworzy słownik, którego kluczami są adresy stron a wartościami listy krotek (liczebność, słowo). Klasa automatycznie zapisuje swój stan do pliku pickle (domyślnie *index.pickle*).

Klasa *SearchEngine* to już właściwa klasa wyszukiwarki internetowej. Korzysta z obiektu *PageIndex*, w nim szuka słów. Jeśli znajdzie mniej niż 3 wyniki pyta użytkownika czy kontynuować wyszukiwanie dla podobnych słów (tzn. krótszych z każdą iteracją o 1 znak aż do 2/3 długości słowa w celu znalezienia tego samego słowa w innej formie).

Wyszukiwarka najlepiej działa na linkach do artykułów, wpisów na blogach itd. dlatego napisałam jeszcze generator linków do angielskiej wikipedii. Zapisuje on podaną jaką pierwszy argument liczbę wylosowanych (unikalnych) adresów do pliku podanego jako drugi argument. Plik *pages* zawiera już przygotowane w ten sposób adresy, klasa *PageIndex* domyślnie czyta z niego adresy.

W celu włączenia wyszukiwarki należy uruchomić skrypt *SearchEngine.py*

Wersja Pythona: 2.7.12

Pisałam pod Linuxem, ale raczej powinno działać wszędzie.

```
dominika@desktop ~/py/python-lab/homework $ python SearchEngine.py
Loaded index file

Type word to search for (or type q to exit): music

Looking for word 'music'
Found 'music' in: https://en.wikipedia.org/wiki/C._Curtis-Smith
Found 'music' in: https://en.wikipedia.org/wiki/Dead,_Hot_and_Ready
Found 'music' in: https://en.wikipedia.org/wiki/The_Big_Hurt_(song)
Found 'music' in: https://en.wikipedia.org/wiki/Vinod_Kumar_Dwivedi
Found 'music' in: https://en.wikipedia.org/wiki/Matilde_D%C3%ADaz
Found 'music' in: https://en.wikipedia.org/wiki/Music_for_Bondage_Performance_2
Found 'music' in: https://en.wikipedia.org/wiki/MTV_Video_Music_Award_for_Video_of_the_Year
Found 'music' in: https://en.wikipedia.org/wiki/Old_and_in_the_Gray
Found 'music' in: https://en.wikipedia.org/wiki/A_Moment%27s_Worth
Found 'music' in: https://en.wikipedia.org/wiki/List_of_University_of_North_Texas_College_of_Music_faculty
Found 'music' in: https://en.wikipedia.org/wiki/Rajesh_Payal_Rai
Found 'music' in: https://en.wikipedia.org/wiki/Raw_Stylus
Found 'music' in: https://en.wikipedia.org/wiki/George_Fox_(singer)
Found 'music' in: https://en.wikipedia.org/wiki/John_Branca
Found 'music' in: https://en.wikipedia.org/wiki/Wilber_Pan
Found 'music' in: https://en.wikipedia.org/wiki/Shopping_(Ryan_Bang_song)
Found 'music' in: https://en.wikipedia.org/wiki/Francisco_Gabilondo_Soler
Found 'music' in: https://en.wikipedia.org/wiki/Regard_Extra%27s_Aame
Found 'music' in: https://en.wikipedia.org/wiki/0%27Stravaganza_%E2%80%93_Vivaldi_in_Ireland
Found 'music' in: https://en.wikipedia.org/wiki/Kevin_Ceballo
Found 'music' in: https://en.wikipedia.org/wiki/Come_Closer_(Miles_Kane_song)
Found 'music' in: https://en.wikipedia.org/wiki/Delta_Music_Museum
Found 'music' in: https://en.wikipedia.org/wiki/Prayag_Sangeet_Samiti
Found 'music' in: https://en.wikipedia.org/wiki/Haylie_Ecker
Found 'music' in: https://en.wikipedia.org/wiki/Swedish_Rhapsody_No._1
Found 'music' in: https://en.wikipedia.org/wiki/Cesium_137_(band)
Found 'music' in: https://en.wikipedia.org/wiki/CFND-FM
Found 'music' in: https://en.wikipedia.org/wiki/Marcus_Wibberley

Type word to search for (or type q to exit):
```