Michal Rynduch 131190

Robot - web crawler

1. Treść zadania

Napisz robota internetowego, który przegląda zasoby w obrębie podanej (rozsądnie dużej, min. 3000 podstron) domeny należącej do uczelni i zapisuje na dysku kopie dokumentów oraz analizuje graf połączeń
między nimi.

2. Opis

Robot został napisany w języku **Python 2.7 (CPython)** z użyciem bibliotek standardowych oraz **requests** do pobierania dokumentów.

Robot został zrównoległowny przy pomocy biblioteki **multiprocessing.** Robot działa na wszystkich procesach procesora (8 na moim komputerze).

Dla lepszego zorganizowania pracy robota pracy robota wykorzystano pliku sitemap.xml zawartego w http://www.english.paris-sorbonne.fr/robots.txt.

Każdy proces zaczynał pracę pobierając jeden dokument wylistowny w sitemap.xml.

Wszystkie strony z sitemap.xml zostały uznany za “popularne”, co oznacza, że robot nigdy nie pobierał ich ponownie.

Robot przestrzega warunków z robot.txt omijając podstrony:

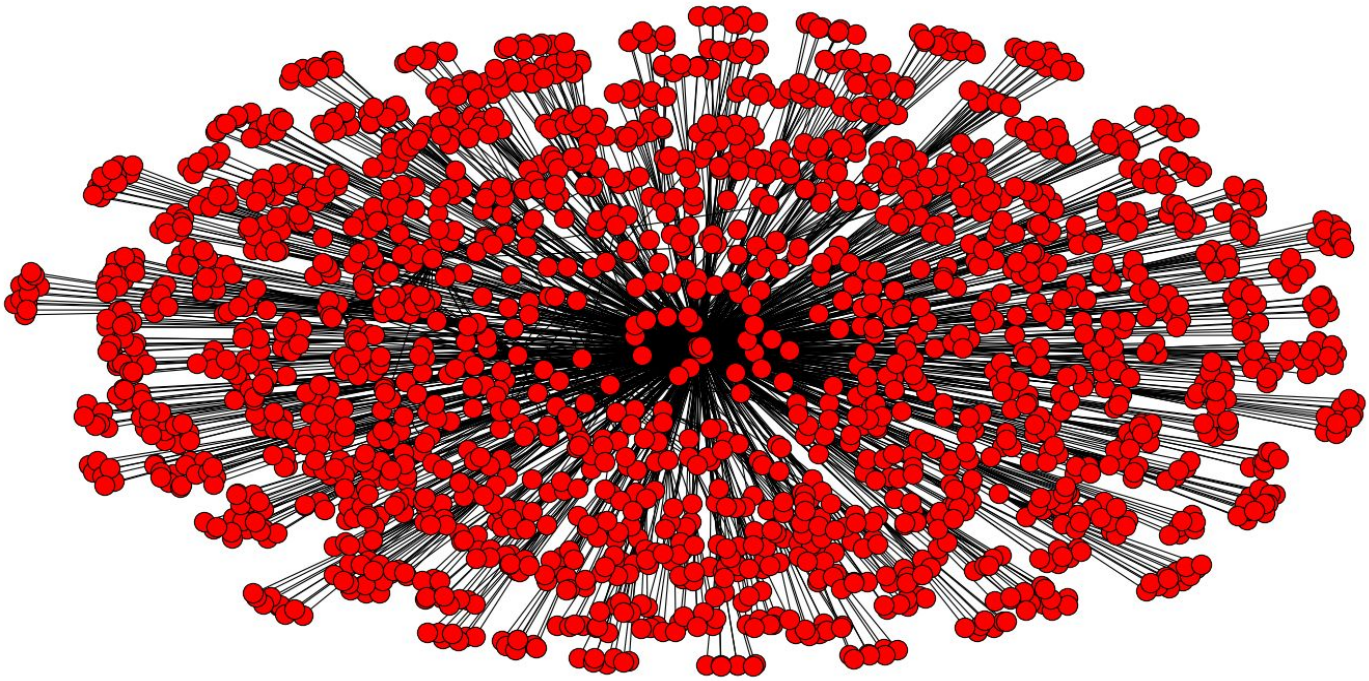
- /local/
- /ecrire/
- /extensions/
- /lib/
- /plugins/
- /prive/
- /squelettes-dist/
- /squelettes/

Robot nie pobiera również dokumentów i plików z podstron:

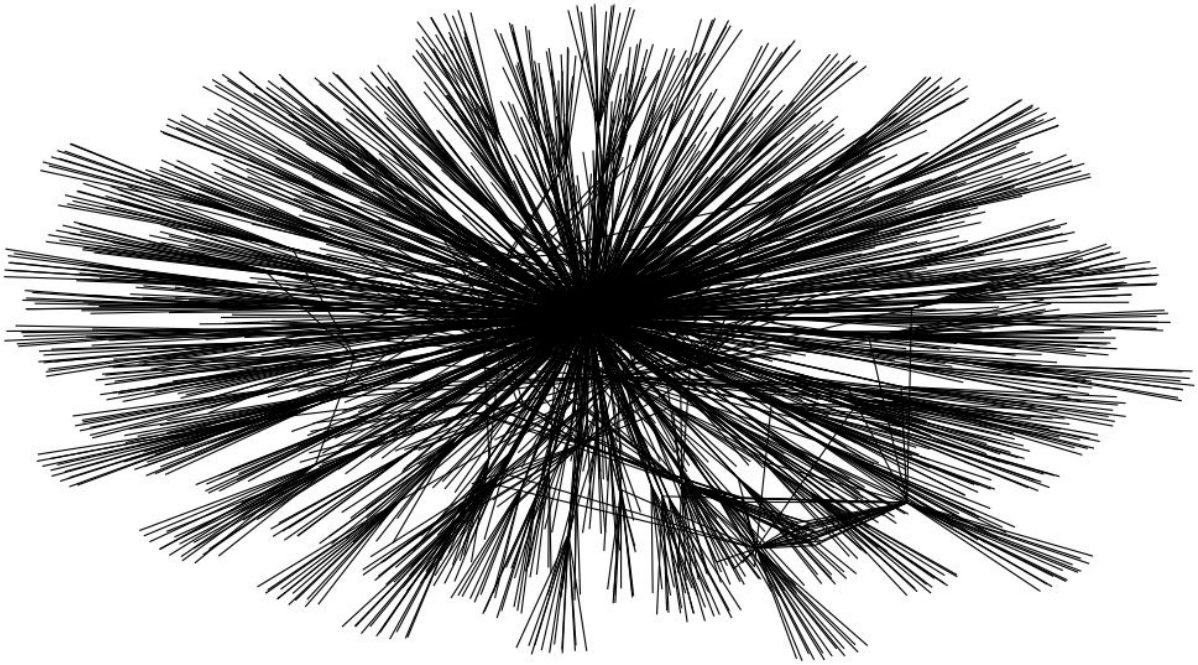
- /Files/
- /IMG/
- /libraries/
- /menu-footer/
- /menu-cache/

Po wszystkich wykluczonych podstronach z pobranych dokumentów utworzono graf:

- 1879 wierzchołków
- 2715 krawędzi
- 277 - najwyższy stopień wierzchołka - http://www.english.paris-sorbonne.fr/courses/french-proficiency-tests-59/
- 268 najwyższy stopień wierzchołka (in) http://www.english.paris-sorbonne.fr/courses/french-proficiency-tests-59/'
- 35 najwyższy stopień wierzchołka (out) http://www.paris-sorbonne.fr/l-international/diplomes-de-francais-pour-les/presentation-generale/presentation-4859/
- 1.44 - średni stopień
- 5 - średnica
- 3 - promień



Graf uniwersytetu



Krawędzie grafu