Paweł Troka 132334

Zadanie 2. Robot internetowy i graf WWW.

Modelowanie Internetu

Sprawozdanie

Wybrano domenę htttp://pg.edu.pl, ze względu na to, że jako jedyna z badanych spełniała wymóg ponad 3000 poprawnie linkowanych I nadających się do analizy podstron. Dokładna ich liczba wyniosła około 4500, co oznacza również spory potencjał programu.

Poza główną domeną politechniki analizie podlegały również (zależnie od potrzeb, także w celach czysto testowych).

- http://cui.pg.edu.pl (45 podstron, równa 2 średnica)
- http://pg.edu.pl (4967 postron)
- http://www.sportowapolitechnika.pl (208 postron)
- http://csa.pg.edu.pl (89 postron)
- http://cas.pg.edu.pl (22 podstron)
- https://pomoc.pg.gda.pl (0 podstron, ponieważ cała domena ma disallow dla botów)

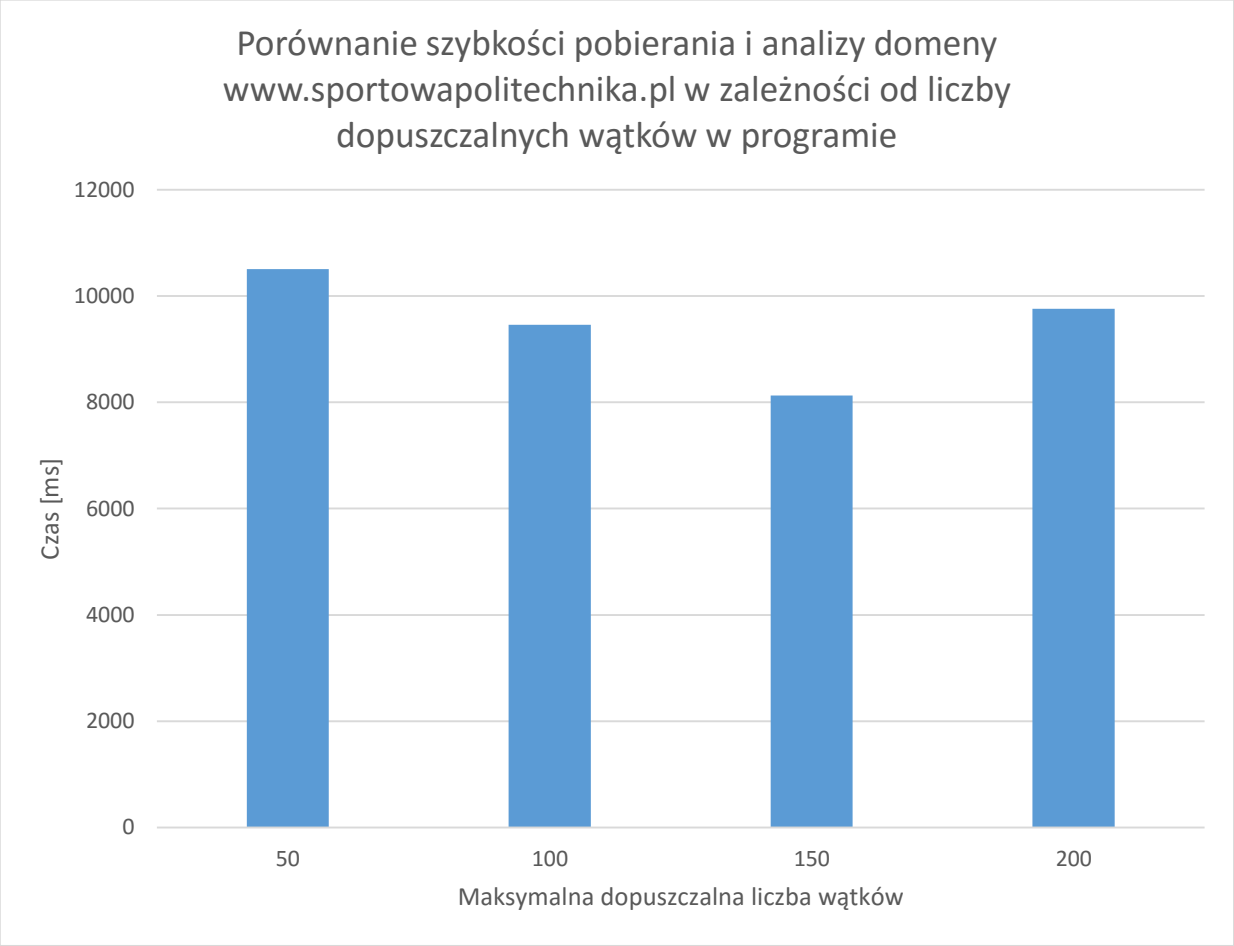
1. Analiza czasowa w wielowątkowym ściąganiu dokumentów

Ze względu na ograniczony czas jak i ograniczenia sprzętowo-wydajnościowe do realizacji tego zadania wybrano domenę sportowapolitechnika.pl. Również stosunkowo wysoka prędkość pobierania z domeny ułatwiała zadanie.

W analizie postanowiono przetestować dwie rzeczy:

- Szybkość pobierania w zależności od maksymalnej dopuszczalnej liczby wątków

  W środowisku .NET całe Task Parallel Library – a więc cała obszerna biblioteka dotycząca wątków, równoległości i zadań korzysta z obiektu statycznego znanego pod nazwą ThreadPool. Modyfikując więc jedną z propercji tego obiektu wpływamy na zachowanie wszystkich mechanizmów całej biblioteki. Postanowiono więc potestować wydajność aplikacji dla róznej liczby wątków.

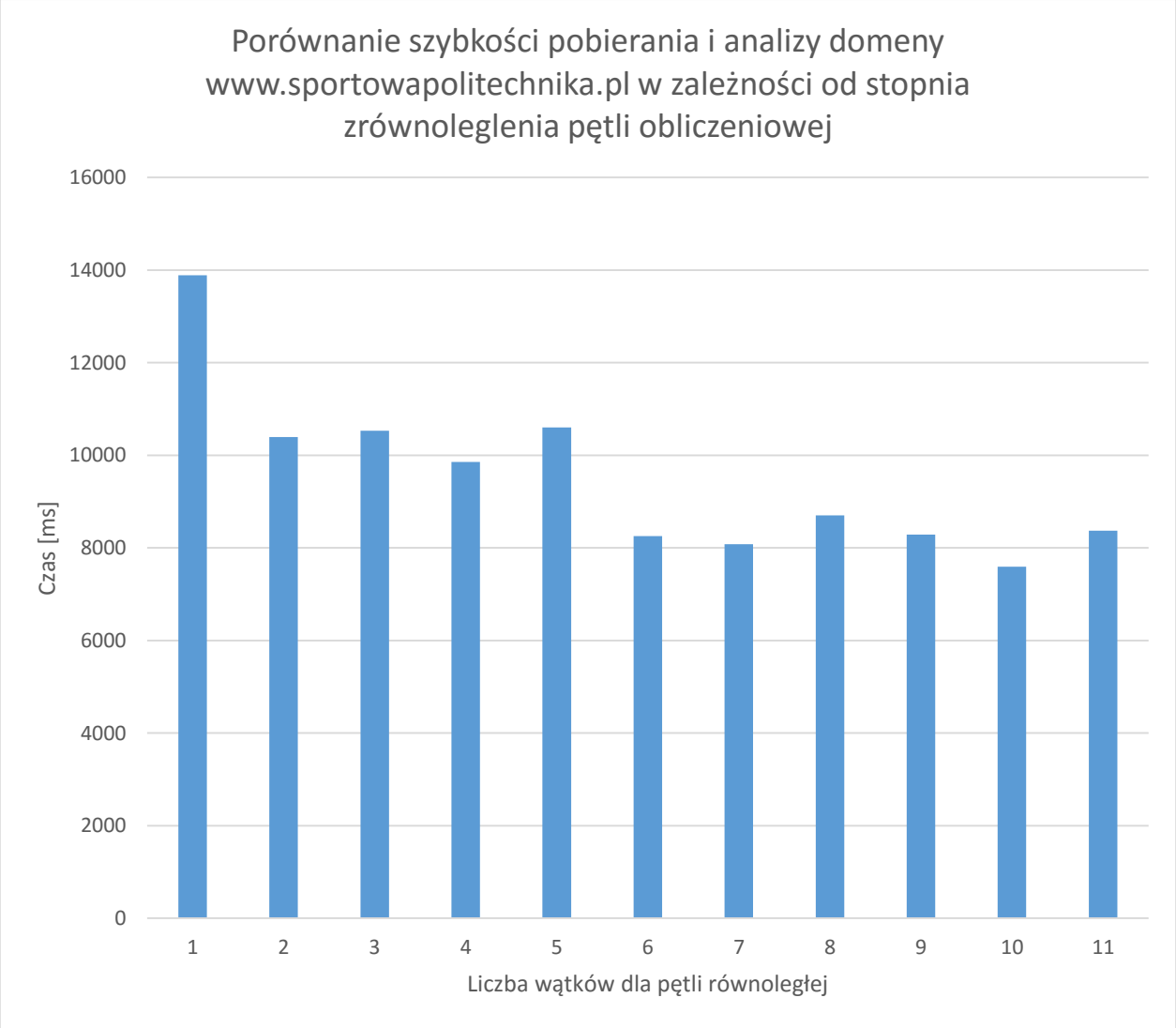


Po wielokrotnych pomiarach okazało się, że optimum dla pisanego programu oscyluje gdzieś w granicach 150 wątków, dalsze ich zwiększanie raczej zmniejszało wydajność.

- Szybkość pobierania w zależności od stopnia zrównoleglenia pętli obliczeniowej

W środowisku .NET dosyć łatwo można ustawić dla każdej równoległej pętli oliczeniowej opcje równoległości, w tym stopień zrównoleglenia. O ile ilość wątków ustawiana w poprzednim punkcie ma

charakter bardziej „globalny” i może źle wpływać na działanie niektórych modułów aplikacji na których działanie wcale nie chcieliśmy wpłynąć o tyle stopień zrównoleglenia pętli możemy ustawiać „bezkarnie” wpływając jedynie na wykonanie tej pętli.



Tutaj z kolei widać bardzo duży wzrost wydajności pomiędzy jednym dopusczalnym wątkiem (czyli tak naprawdę pętla niezrównoleglona) a dwoma. Dalsze zyski wydajności są już znacznie mniejsze ale i tak są znaczne.

Optimum wydaje się być w okolicach 6-10, co jest prawdopodobnie mocno związane z parametrami środowiska testującego (Core i7 720QM, 4 rdzenie, 8 wątków). Zwiększanie powyżej wartości 11 generalnie zmniejszało wydajność – narzut związany z dodatkowymi wątkami nie dawał poprawy wydajności.

Łatwo policzyć, że maksymalna uzyskana tutaj wydajność to nawet 25 stron/sekundę – co jest uznawane w tematyce crawler’ów generalnie za wydajność przynajmniej dobrą [13]. Niestety dalsze próby zwiększania wydajności, także te nie związane z liczbą wątków, nie przyniosły poprawy.

2. Analiza grafu połączeń między dokumentami dla domeny http://pg.edu.pl i domeny http://www.sportowapolitechnika.pl

Podstawowe oznaczenia i pojęcia:

- V – zbiór wierzchołków w grafie
- stopień wchodzący wierzchołka v (indegree) – ilość krawędzi kończących się w v
- stopień wychodzący wierzchołka v (outdegree) – ilość krawędzi wychodzących z v
- E – zbiór łuków w grafie
- |Q| - moc zbioru Q, tj liczba elementów należących do zbioru Q
- Odległość / odległość geodezyjna – odległość w grafie między dwoma wierzołkami w sensie długości najkrótszej ścieżki między nimi – wartość wyznaczna w algorytmie Floyda-Warshalla
- ε(v) – „mimośród" (eccentricity) wierzchołka v w grafie - największa odległość geodezyjna między v a dowolnym innym wierzchołkiem

Niektóre z wyznaczanych parametrów analizowanego grafu:

- liczba wierzołków - |V|
- liczba łuków: - |E|
- rozkłady stopni (in, out) – ukazane na wykresie w punkcie 3
- najkrótsze ścieżki (wszystkie pary) – wyznaczone z algorytmu Floyda-Warshalla wypisywane tylko do pliku z powodów wydajnościowych
- średnia odległość

$$\frac{1}{|V|}\sum_{u,v\in V} dist[u,v]$$

  gdzie V – zbiór wierzchołków, dist – macierz odległości z algorytmu Floyda-Warshalla
- średnica grafu
  największa odleglość między dwoma dowolnymi wierzchołkami w grafie

$$d = \max_{v\in V} \epsilon(v)$$

Poza tym wyznaczono również wartości nie wymienione w poleceniu zadania takie jak:

- promień grafu:

$$r = \min_{v\in V} \epsilon(v).$$

- średni stopień wchodzący wierzchołka:
- średni stopień wychodzący wierzchołka:

- średnia wartość PageRank dla dokumentów:

Udało się również wyznaczyć wartość PageRank dla każdej podstrony. Przebadano zbieżność algorytmu iteracyjnego i najlepsze wyniki uzyskano dla wartości współczynnika tłumienia około 0.85.

$$PR_x = \frac{1-d}{N} + d\left(\frac{PR_y}{L_y} + \frac{PR_z}{L_z}...\right)$$

Gdzie:

- PR – PageRank danej strony
- d – współczynnik tłumienia, liczba pomiędzy 0 i 1. Dla obliczeń przyjmuje się zazwyczaj wartość 0,85
- N – liczba stron internetowych
- L – liczba linków do których odsyła dana strona internetowa

Nawet bez tłumienia (wartość współczynnika 1) algorytm był dość szybkozbieżny.

a) http://pg.edu.pl

Pomimo stosunkowo dużego rozmiaru domeny (prawie 5 tysięcy podstron) udało ją się pobrać I zanalizować w stosunkowo niedużym czasie 5520 s i przy stosunkowo niskiej wydajności 0.9 stron/sekundę. Niska wydajność wynika prawdopodobnie z faktu wielu wadliwych łączy na podstronach domeny.

Niestety z powodu dużej liczby podstron i złożoności algorytmu Floyda-Warshalla – $O(n^3)$ nie udało się w zadowalającym czasie wyznaczyć najkrótszych ścieżek dla wszystkich par (tutaj akurat sama wydajność generowania/wypisywania tak dużej liczby par), średnicy grafu, promienia grafu i średniej odległości.
Liczba stron: 4965

Czas pobierania i analizy dokumentów: 5519864.513ms

Prędkość pobierania i analizy dokumentów: 0.89947859921322 stron/sekundę

Liczba wierzchołków: 4965

Liczba łuków: 470969

Średni stopień wchodzący wierzchołka: 94.857804632427

Średni stopień wychodzący wierzchołka: 94.9111782477341

Średnia wartość PageRank: 0.000134989991043876

b) http://www.sportowapolitechnika.pl

Tutaj jest znacznie lepiej jeżeli chodzi o wydajność. Znacznie mniejsza liczba podstron oznacza również bezproblemowe wykonanie w zadowalającym czasie wszystkich algorytmów.

Liczba stron: 208

Czas pobierania i analizy dokumentów: 19236.8758ms

Prędkość pobierania i analizy dokumentów: 10.8125665603143 stron/sekundę

Liczba wierzchołków: 208

Liczba łuków: 11675

Średni stopień wchodzący wierzchołka: 56.1298076923077

Średni stopień wychodzący wierzchołka: 56.1298076923077

Średnia odległość: 5.76370654585799

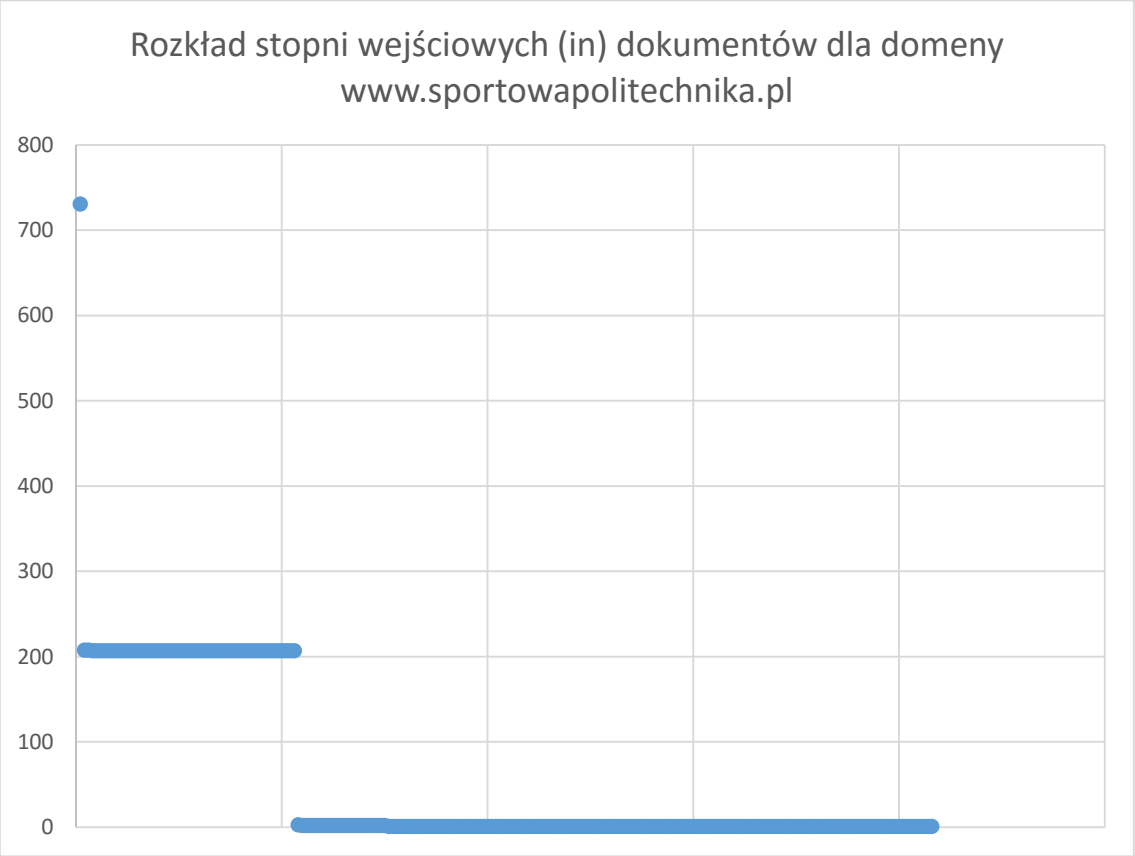
Średnica grafu: 15

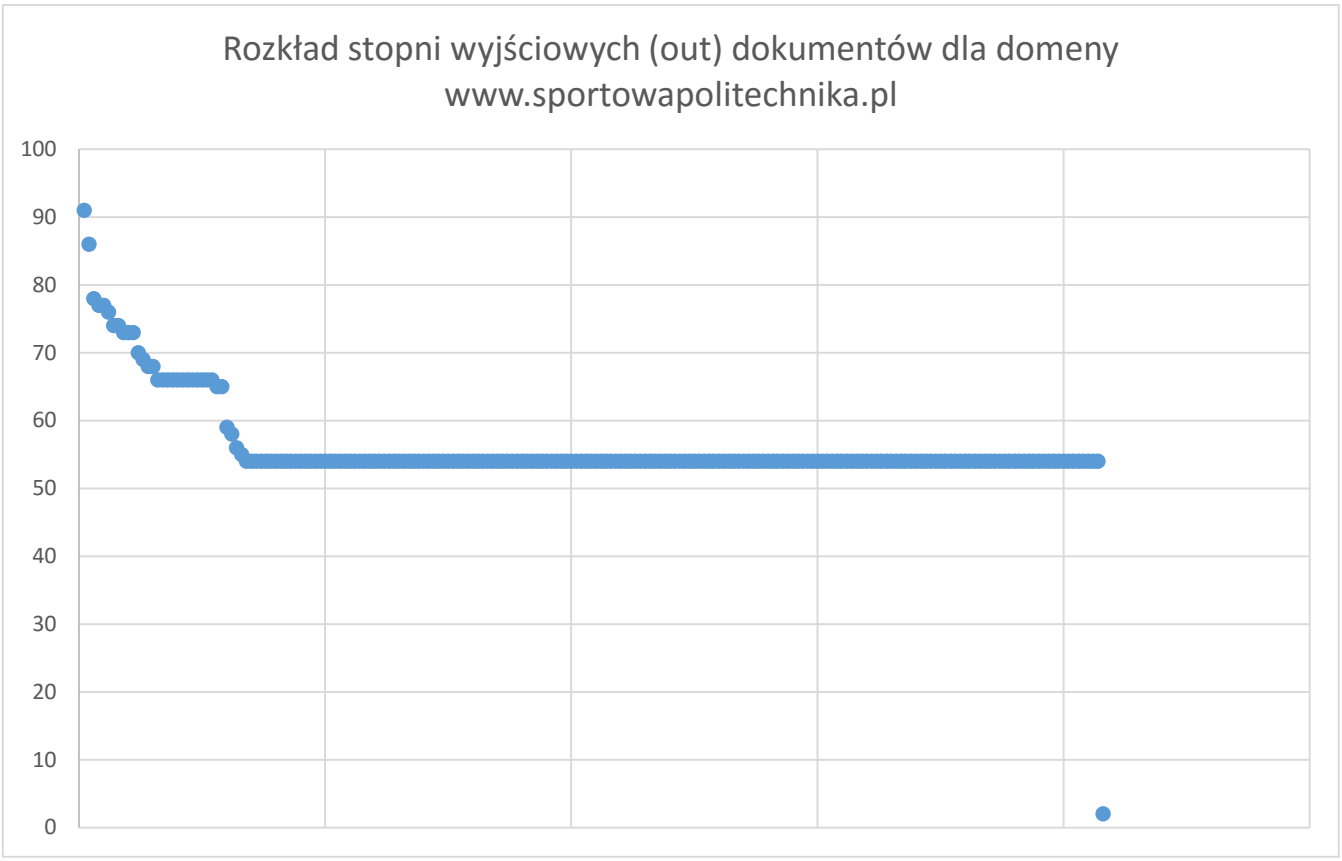
Promień grafu: 6

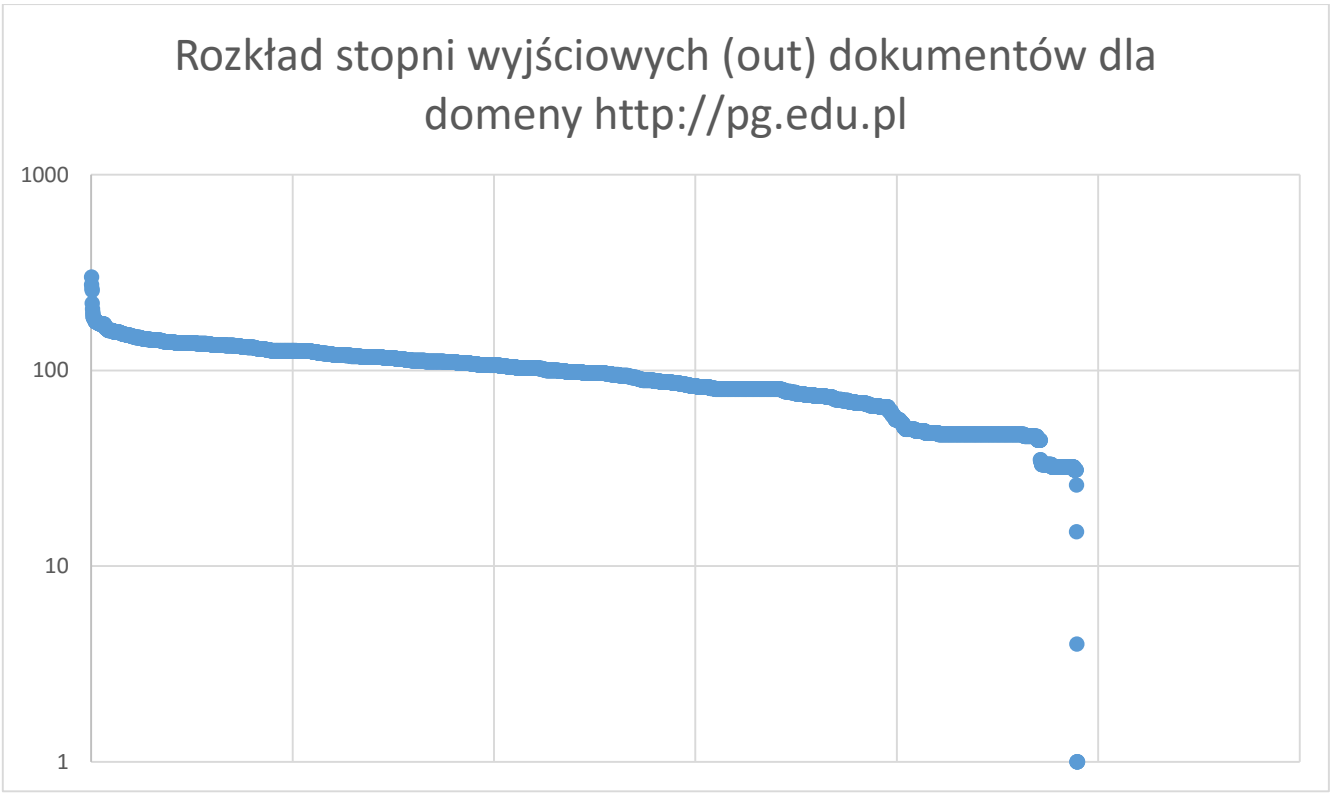
Średnia wartość PageRank: 0.00378411334342211

Czas analizy grafu: 6076ms

3. Rozkłady stopni IN, OUT




Rozkład stopni wyjściowych (out) dokumentów dla domeny www.sportowapolitechnika.pl
100
90
80
70
60
50
40
30
20
10
0


Rozkład stopni wyjściowych (out) dokumentów dla domeny http://pg.edu.pl
1000
100
10
1

Wykresy dla małej pod względem liczby podstron domeny www.sportowapolitechnika.pl są zwykłymi wykresami punktowymi z liniową skalą na osiach x i y.

Natomiast żeby uzyskać większą czytelność wykresy rozkładów stopni in,out dla domeny http://pg.edu.pl o dużej iczbie podstron wykreślono na wykresie punktowym ale z logarytmiczną skalą na osi y i liniową na osi x.

4. Odporność na ataki i awarie

Program został „uodporniony” na wszelkie ataki i awarie za pomocą następujących mechanizmów:

- Każdy analizowany link jest najpierw analizowany pod kątem poprawności składniowej
- Dopuszczamy tylko linki które nie są wykluczone przez plik robots.txt zgodnie z „robots exclusion protocol”
- Każdy link który jest poprawny składniowo i niewykluczony przez robots.txt przed dodaniem do grafu dokumentu jest najpierw analizowany jako dokument, sprawdzane jest m in.
    - Czy istnieje, czy witryna internetowa w odpowiedzi na request o zasob reprezentowany przez link odpowiada kodem sukcesu
    - Czy jest właściwego typu, czy nadaje się do dalszej analizy np. zdjęcia się nie nadają
- Jeżeli dokument spełnia powyższe kryteria jest dodawany do grafu dokumentu
- Przed rozpoczęciem analizy cały graf dokumentu jest ponownie sprawdzany pod kątem spójności, poprawności
- Graf jest dodatkowo „czyszczony” ze wszystkich błędnych wpisów, błędnych dokumentów, linków które z jakichś powodów zostały wykluczone w trakcie analizy ale znalazły się w grafie dokumentów
- Usunięcie losowego wierzchołka lub nawet wierzchołka o najwyższym stopniu wywołałoby po prostu ponowne „czyszczenie” grafu, co przywróciłoby spójność danych ale graf mógłby przestać być grafem połączonym (spójnym) – co jest oczywiście naturalne w tym przypadku

5. Bibliografia

1. http://kaims.pl/~mima/mi2016/projekt/zadanie2.txt
2. http://kaims.pl/~mima/mi2016/W2/wyklad1-2015.pdf
3. http://kaims.pl/~mima/mi2016/W2/wyklad2-2015.pdf
4. https://en.wikipedia.org/wiki/Distance_(graph_theory)
5. https://pl.wikipedia.org/wiki/Odleg%C5%82o%C5%9B%C4%87_(teoria_graf%C3%B3w)
6. https://pl.wikipedia.org/wiki/PageRank
7. https://en.wikipedia.org/wiki/PageRank
8. https://github.com/jeffersonhwang/pagerank/blob/master/PageRank.cs
9. https://msdn.microsoft.com/en-us/library/system.threading.tasks.paralleloptions.maxdegreeofparallelism(v=vs.110).aspx
10. https://msdn.microsoft.com/en-us/library/system.threading.threadpool(v=vs.110).aspx
11. https://msdn.microsoft.com/en-us/library/0ka9477y(v=vs.110).aspx
12. https://en.wikipedia.org/wiki/Robots_exclusion_standard
13. Wykład z przedmiotu „Inteligentne wyszukiwanie informacji” – dr Julian Szymański, wykład nr 6, rok akademicki 2015/2016
14. https://en.wikipedia.org/wiki/Floyd–Warshall_algorithm