Projekt SK5 - Analiza sieci

Sprawozdanie końcowe

15.01.2020

1 Treść zadania

Należy pobrać zrzut („dump") Wikipedii, utworzyć na jego podstawie graf niezorientowany, a następnie dopasować do tego grafu parametry wybranego modelu sieci (B-A, W-S, itp.).

2 Dane

Projekt opiera się na danych z ~~polskojezycznej~~ Wikipedii w jidysz z najnowszego dostępnego na dzień 01.12.2019 zrzutu [1]. Użyte zostaną pliki `.sql`, zawierające nazwy i identyfikatory poszczególnych stron i przestrzeni nazw oraz stron, z którymi posiadają odniesienia.

3 Narzędzia

- Python [2] - główny język programowania użyty do tworzenia listy sąsiedztwa grafu oraz komunikacji z zewnętrznymi bibliotekami pomocniczymi;
- ~~sqlparse [3] - moduł w pythonie do przetwarzania SQLowych formuł;~~
- ~~graph-tool [4] - moduł w pythonie do wysoko wydajnych operacji na grafach, zostanie użyty do konstrukcji i wyciągania danych statystycznych z rzeczywistego zrzutu linków Wikipedii;~~
- networkx [5] - bardzo duży moduł w pythonie do operacji na grafach, jednak nieco mniej wydajny od graph-tool; zostanie użyty do modelowania grafów (np. Barbassi-Albert).
- matplotlib [6] - użyta do wizualizacji danych statystycznych dla grafu

4 Porównywane parametry

Do opisu charakterystyki grafu użyliśmy następujących parametrów:

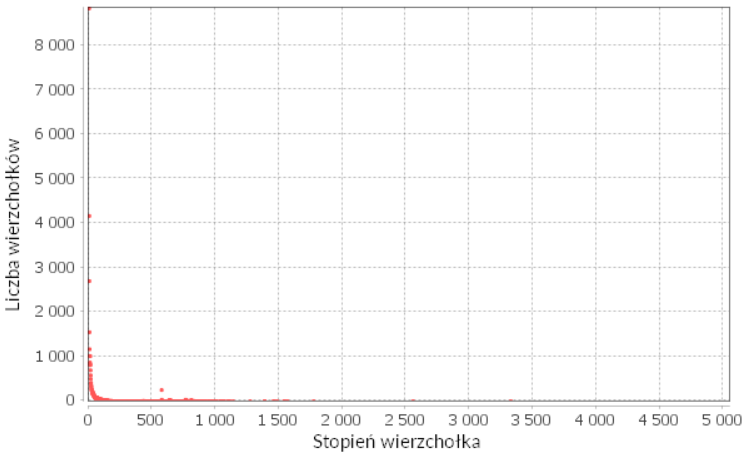
- średnia długość ścieżki - średnia ze ścieżek pomiędzy wszystkimi parami wierzchołków
- logarytmiczny rozkład stopni wierzchołków w grafie
- współczynnik klasteryzacji - stosunek liczby krawędzi sąsiadów danego wierzchołka do liczby wszystkich możliwych połączeń między nimi

5 Zebrane parametry grafu

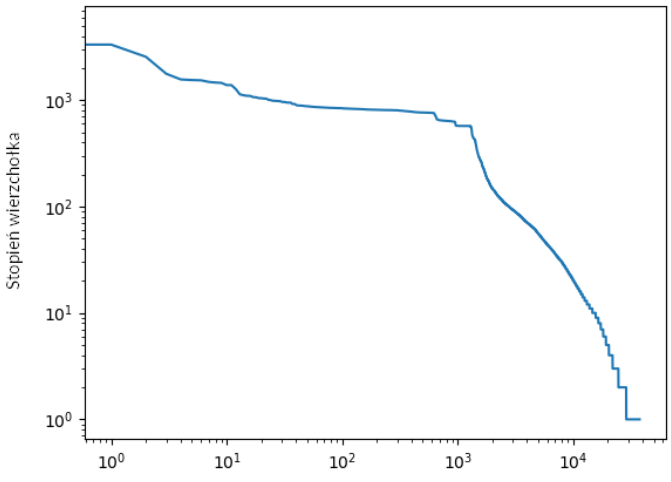
Wszystkie wykresy dotyczą grafu zbudowanego na podstawie grafu połączeń wikipedii w języku jidysz. Ogólne wartości parametrów, globalne dla całego grafu:

- średni stopień wierzchołków = 44.853

Średni stopień wierzchołków jest bardzo wysoki, ale wynika to najprawdopodobniej z jakiejś strony np. edycyjnej, do której referują prawie wszystkie inne strony. Widać to dobrze na wykresie na rysunku 1.



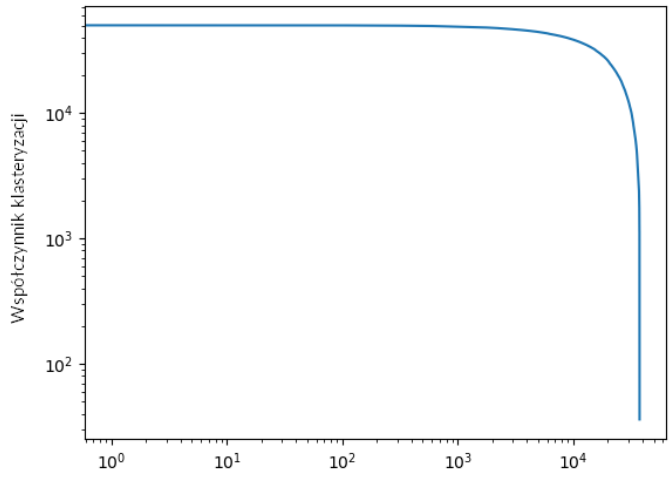
Rysunek 1: Rozkład stopni wierzchołków. Skala lin-lin



Rysunek 2: Rozkład stopni wierzchołków. Skala log-log

- średni współczynnik klasteryzacji = 0.230

Niski współczynnik klasteryzacji wskazuje na niewielką ilość klik, dzięki czemu mamy ciąg informacji, które odwołują się do kolejnych artykułów. Jest to pożądane zjawisko, bo widać, że połączenia wskazują źródło, od którego wywodzą się nowe informacje.



Rysunek 3: Rozkład współczynników klasteryzacji. Skala log-log

Rysunek 4: Rozkład współczynników klasteryzacji. Skala lin-lin

- średnia długość ścieżki = 3.7012

Średnia długość jest mała jak na graf takich rozmiarów jak nasz, co tak jak w przypadku średniego stopnia wierzchołków, najprawdopodobniej wynika z istnienia stron, do których odwołuje się duża ilość pozostałych wierzchołków w grafie nieskierowanym.

6 Dobór modelu sieci

Podsumowując, nasz graf posiada takie parametry:

- liczba wierzchołków = 37712,
- średni współczynnik klasteryzacji = 0.23,
- średnia długość ścieżki = 3.7012,
- średni stopień wierzchołków = 44.853,
- rozkład stopni wierzchołków mocno wykładniczy.

Takie też parametry będą brane pod uwagę przy doborze modelu. Pierwszym, najbardziej różnicującym parametrem jest rozkład stopni wierzchołków. W przypadku grafu Wikipedii jest on wyraźnie wykładniczy, co od razu odrzuca modele Erdős–Rényi, Wattsa–Strogatza (które reprezentują rozkład stopni wierzchołków jako gaussowski o niskim rozproszeniu dla większej liczby wierzchołków) oraz euklidesowski (zbyt duża średnia długość ścieżki). Ponadto, wysoki współczynnik klasteryzacji w porównaniu do średniej długości ścieżki, sugeruje, że ten graf spełnia właściwości *małego świata*.

Jedynie model Barabási–Alberta spełnia wszystkie wyżej wymienione wymagania. Rozkład stopni wierzchołków dla tego modelu wynosi $P(k) \simeq k^{-3}$, zaś średnia długość ścieżki to $\frac{lnN}{lnlnN} = 4.47$, a więc dość blisko względem modelowanej sieci.

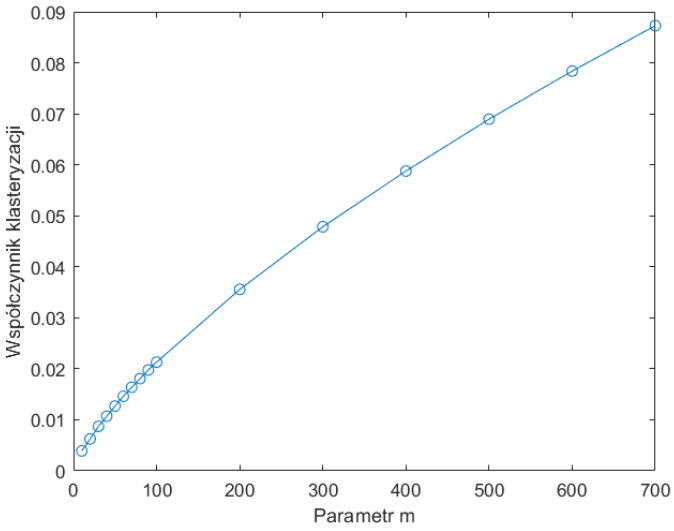
Sieć B-A cechuje właściwość, że średni stopień wierzchołków wynosi $\simeq 2m$, gdzie m to parametr sieci B-A, oznaczający ilość krawędzi dodawanych do grafu w każdym kroku. Jeśli kierować się tą cechą przy doborze parametrów sieci B-A i zachowując liczbę wierzchołków zgodną z grafem Wikipedii, otrzymamy sieć B-A o następujących parametrach:

- m = 22,
- średni współczynnik klasteryzacji = 0.007,
- średnia długość ścieżki = 2.91,
- średni stopień wierzchołków = 43.974,

Sugerując się średnim stopniem wierzchołków, uzyskana średnia długość ścieżki nie jest mocno odległa od grafu Wikipedii (różnica 21%). Jednakże, średni współczynnik klasteryzacji pozostaje na bardzo niskim poziomie i jest prawie 33 razy niższy niż modelowanego grafu.

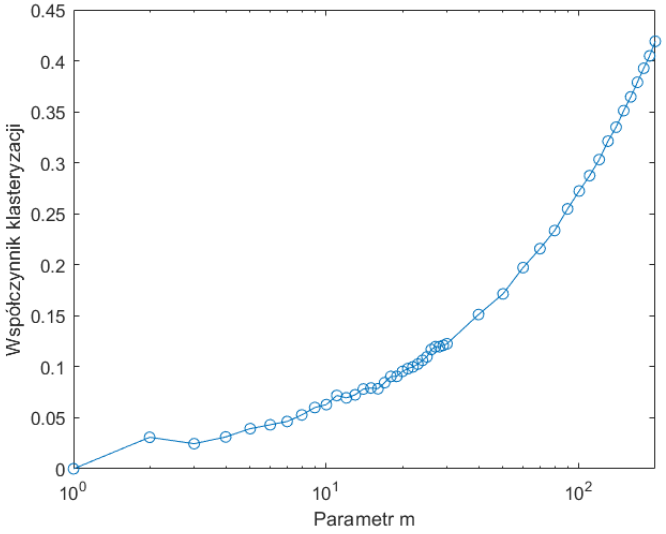
Jedynym sposobem na zwiększenie współczynnika klasteryzacji jest zwiększenie parametru m sieci B-A. Na rysunku 5 zbadaliśmy zależność średniego współczynnika klasteryzacji w zależności od parametru m dla grafu o wielkości identycznej jak Wikipedii. Nie udało się uzyskać wyników dla wartości m wyższych niż 700 ze względu na bardzo długi czas przetwarzania (wyliczenie średniego współczynnika klasteryzacji dla grafu o $m = 800$ nie zakończyło się po 12h).

Zależność średniego współczynnika klasteryzacji wykazuje charakter o przyroście logarytmicznym (jednak już funkcja liniowa przybliża ten rozkład całkiem dobrze). Jednakże, dla grafu wielkości Wikipedii (ok 37 tys. wierzchołków) sieć B-A osiąga już okolice 25 mln krawędzi przy $m = 700$ przy średnim współczynniku klasteryzacji $C = 0.087$. Taki wynik jest już niższy tylko 2.6 krotnie niż w modelowanym grafie Wikipedii.



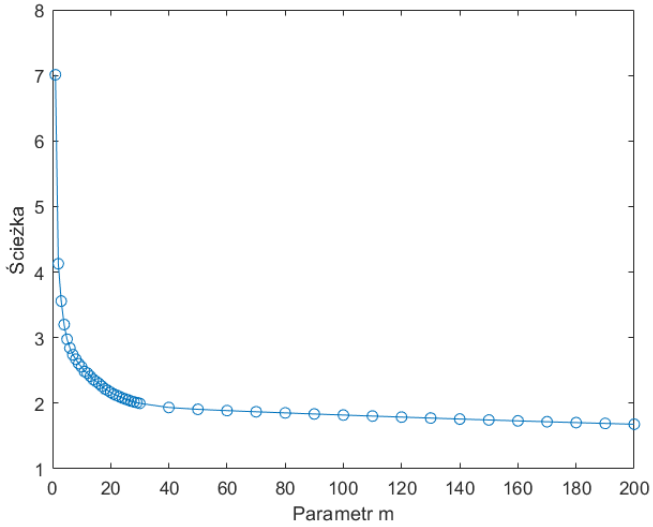
Rysunek 5: Rozkład średniego współczynnika klasteryzacji dla sieci B-A o rozmiarze 37712 wierzchołków. Skala lin-lin.

Jednak, z uwagi na ogromną liczbę krawędzi w grafach o $m > 50$ nie byliśmy już w stanie wyznaczyć średniej długości ścieżki na grafie B-A o wielkości analogicznej do grafu Wikipedii. W celu analizy posłużyliśmy się dużo mniejszym grafem, o wielkości 1000 wierzchołków. Na rysunku 6 została zaprezentowana relacja pomiędzy średnim współczynnikiem klasteryzacji a parametrem m. Na rysunku 7 został zaprezentowany stosunek pomiędzy średnią długością ścieżki w grafie w zależności od m.



Rysunek 6: Zależność średniego współczynnika klasteryzacji od parametru m dla grafu z 1000 wierzchołków.

Z rysunku 7 można zauważyć, że średnia długość ścieżki maleje już bardzo wolno przy parametrze m stanowiącym 3% z liczby wierzchołków. Dla grafu wielkości Wikipedii, takie załamanie miałoby miejsce w granicach $m = 1100$.



Rysunek 7: Zależność średniej ścieżki w grafie od parametru m dla grafu z 1000 wierzchołków.

Korzystając z wykładniczej aproksymacji średniego współczynnika klasteryzacji $C = 0.0007m^{0.7371}$ wyznaczonego na podstawie wykresu na rysunku 5, w celu osiągnięcia średniego współczynnika klasteryzacji na poziomie 0.23 dla modelu Wikipedii, m powinno wynosić **m=2596** dla modelu B-A o rozmiarze 37712 wierzchołków.

7 Podsumowanie

Graf Wikipedii w języku Jidysz najlepiej modeluje sieć Barabási–Alberta o 37712 wierzchołkach. W zależności od oczekiwań stawianych modelowi, możliwe są 2 rozwiązania:

1. jeśli głównie interesuje nas zachowanie średniego stopnia wierzchołków (kosztem średniej klasteryzacji niższej 33 krotnie), $m = 22$

2. jeśli głównie zależy nam na średnim współczynniku klasteryzacji (kosztem 1.5-krotnie niższej wartości średniej długości ścieżki i ponad 115-krotnie większym średnim stopniu wierzchołków), $m = 2596$

Wartość średniej długości ścieżki dla m=2596 została oszacowana na podstawie wykresu z rysunku 7. Wiedząc, że dla m stanowiącego ok. 7% liczby wierzchołków, średnia długość ścieżki wyniosła $\simeq 2$ dla przypadku modelu sieci B-A o rozmiarze 1000 wierzchołków.

8 Errata

- 2: niestety nie udało się przetworzyć polskojezycznej wikipedii. Przy wykorzystaniu biblioteki networkx 14GB pamięci ram zajmowało już 64 miliony krawędzi, a w polskojezycznej wikipedii jest ich ok. 158 milionów. Użyliśmy wielokrotnie mniejszej (ok. 91 razy) wikipedii w języku jidysz [1]. Dla np. węgierskojęzycznej czy macedońskiej policzenie rozkładu stopni było jeszcze możliwe, ale już klasteryzacja była poza zasięgiem.
- 3:
  - biblioteka sqlparse okazała się całkowicie zbyteczna, czysty regex okazał się wydajniejszy i prostszy,
  - do wyciągania danych z plików SQLowych użyliśmy biblioteki re [7] do zwykłych wyrażeń regularnych
  - biblioteka graph-tool okazała się nie współpracować z Windowsem (wymaga pełnej kompilacji, która jest wspierana tylko na linuxie).
  - dodano matplotlib

Referencje

[1] *YI Wikipedia Dumps*. URL: `https://dumps.wikimedia.org/yiwiki/latest/`.

[2] *Python*. URL: `https://www.python.org/`.

[3] *sqlparse*. URL: `https://pypi.org/project/sqlparse/`.

[4] *graph-tool*. URL: `https://graph-tool.skewed.de/`.

[5] *networkx*. URL: `https://networkx.github.io/`.

[6] *matplotlib*. URL: `https://matplotlib.org/`.

[7] *re - biblioteka do wyrażeń regularnych*. URL: `https://docs.python.org/3/library/re.html`.