AGH

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

Grafowy analizator notek prasowych

Karolina Mura
Paweł Nowak

Spis treści

1. Sformułowanie zadania projektowego

Celem projektu jest stworzenie oprogramowania pozwalającego na pobieranie notek prasowych na bieżąco, zapisywanie ich do bazy danych, tagowanie oraz złożoną analizę za pomocą sieci społecznych.

1.1 Pobieranie notek

Oprogramowanie powinno umożliwiać pobieranie notek prasowych z wybranych serwisów oraz zapisanie ich w bazie danych.

1.2 Tagowanie notek

Oprogramowanie powinno umożliwiać tagowanie notek w kilku kategoriach, co pozwoli na ich późniejszą analizę.

1.3 Analiza notek

Oprogramowanie powinno udostępniać przeprowadzenie analizy zgromadzonych notatek na podstawie przypisanych tagów przez stworzenie grafu zależności między tagami i przeprowadzenie analizy sieci społecznych.

2. Wizja rozwiązania

W naszym oprogramowaniu zostaną wykorzystane fragmenty kodu należącego do projektu (zwanego dalej “Pierwszym Projektem”) “Środowisko do pozyskiwania i analizy treści na przykładzie informacji prasowych” autorstwa Małgorzaty Olszewskiej, Julii Samól oraz Damiana Smutka.

2.1 Pobieranie notek

Do pobierania notek prasowych zostanie użyte narzędzie z Pierwszego Projektu. Po otagowaniu, wiadomości prasowe zostaną zapisane w bazie danych. Ściąganie następuje raz na dobę, może też być uruchomione komendą użytkownika.

2.2 Tagowanie notek

Tagi zostaną przypisane zaraz po ściągnięciu notek z internetu. Dodawane są na podstawie 3 kategorii: państwa, waluty, nazwy organizacji międzynarodowych. Algorytmy tagowania zostaną opisane w dalszej części dokumentacji.

2.3 Analiza notek

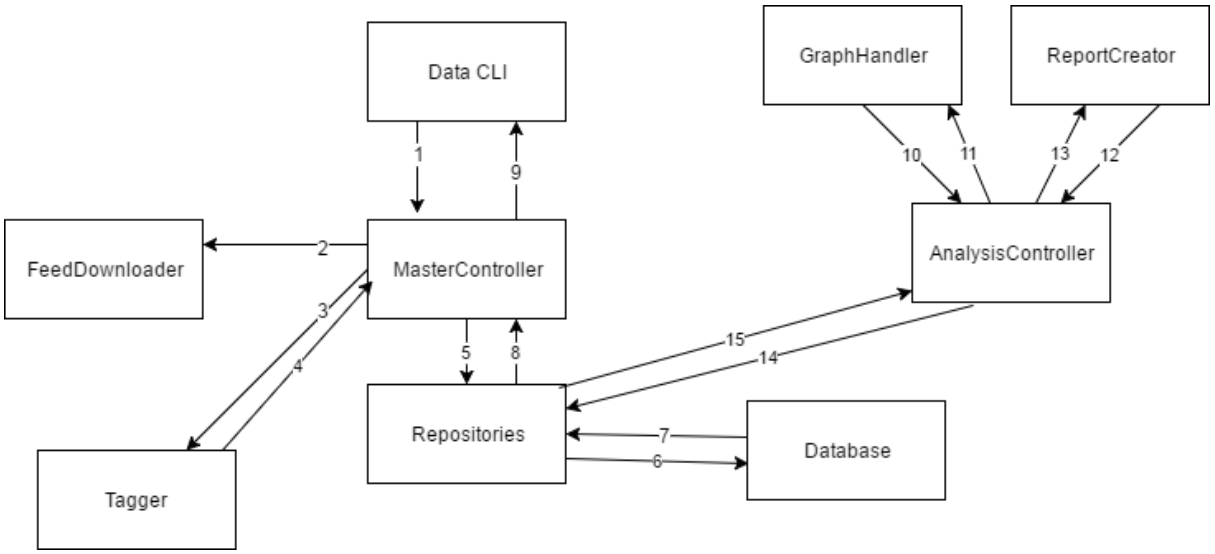
Na podstawie poleceń wprowadzonych w linii poleceń narzędzie pobiera z bazy danych potrzebne notatki, odczytuje przypisane do nich tagi, tworzy graf i przeprowadza na nim analizy, którego wynikiem są pliki .csv zawierające zarówno wyniki analizy, jak i strukturę otrzymanego grafu. Umożliwia również przeprowadzenie analiz na podzbiorze otrzymanych danych i porównanie wyników dla różnych czasopism dla wybranych ram czasowych. Algorytmy stosowane do analizy zostaną omówione w dalszej części dokumentacji.

2.4 Wykorzystane technologie

Całość aplikacji została napisana w języku Java (1.8) i jest zbudowana na bazie frameworka Spring. Do budowania grafu i przeprowadzania analizy sieci społecznych została wykorzystana biblioteka Gephi [1], do tworzenia i eksportowania wykresów biblioteka xChart [2], a do tworzenia raportów .pdf biblioteka iText [3]. Oprócz tego skorzystaliśmy z biblioteki opencsv [4] - wykorzystanej w Pierwszym Projekcie.

3. Architektura systemu

3.1 Moduły



Aplikacja posiada dwa główne interfejsy użytkownika:

a) Moduł Data CLI (klasa NewsAnalyzerMain), który udostępnia zarządzanie bazą danych oraz niegrafowe analizy
b) AnalysisController (klasa MainUI), który udostępnia analizę grafową.

1. Pobranie nowych notek prasowych

Moduł Data CLI wysyła request (1) do master kontrolera. Wtedy on wywołuje procedurę pobrania nowych notek prasowych oraz zapisania ich do plików csv(2). Następnie uruchamiany jest tagger(3), który zapisuje notki wraz z tagami do plików csv i zwraca status do modułu Data CLI(4). Następnie otagowane notki są zapisywane przez pośrednictwo Spring Crud repozytoriów (5) do bazy (6). Następnie status zostaje zwrócony do modułu Data CLI (7), (8), (9)

2. Analiza niegrafowa

Moduł Data CLI wysyła request(1) do master kontrolera, który wysyła request(5) do repozytoriów, które pobierają potrzebne do analizy dane z bazy (6), (7), a następnie zwracają mu je (8). Kontroler wykonuje potrzebne operacje, a następnie zapisuje wyniki w plikach lub zwraca je do Data CLI(9).

3. Analiza sieci społecznych

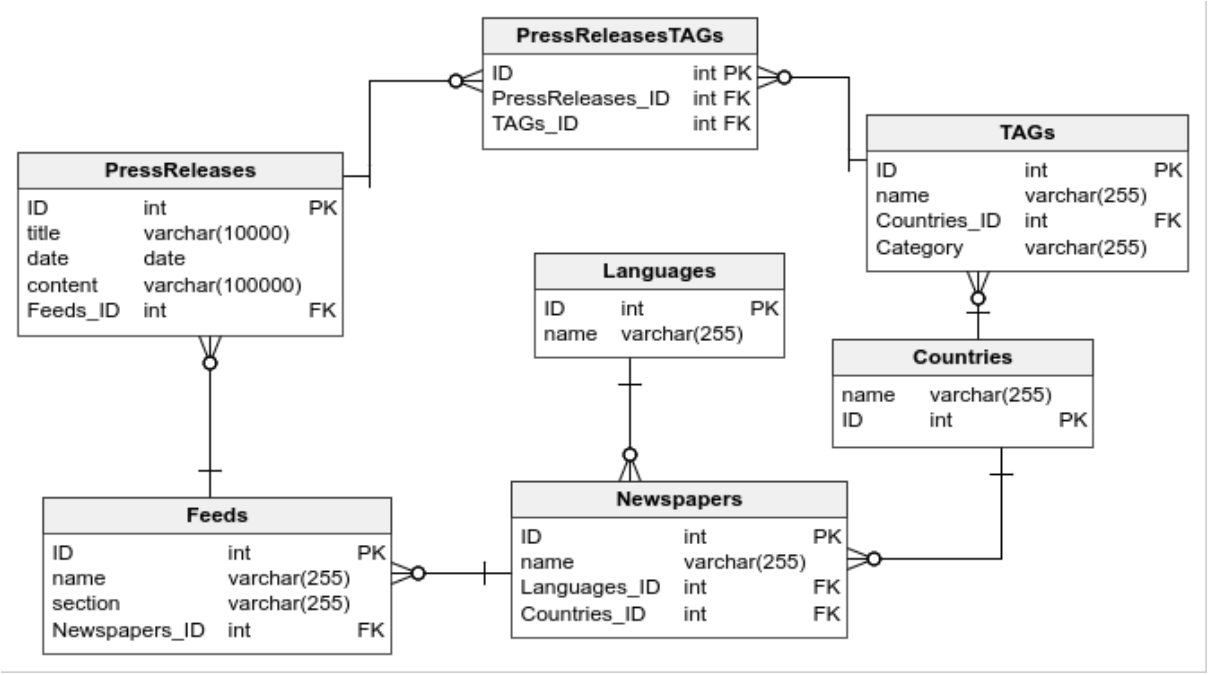
Analiza wstępna: moduł AnalysisController wysyła request za pośrednictwem repozytoriów (14) do bazy danych (6) w celu pobrania danych potrzebnych do analizy. Otrzymane dane(7)(15) przesyła do GraphHandler (11). GraphHandler tworzy graf dla tagów, oblicza wszystkie metryki dla otrzymanego grafu i zapisuje otrzymane wyniki w plikach .csv

Analiza szczegółowa: AnalysisController nie korzysta z bazy danych, lecz z plików .csv otrzymanych w analizie wstępnej. Przekazuje do ReportController (13) te pliki .csv, z których trzeba wyodrębnić dane dla wybranego przedziału czasowego. ReportController tworzy nowe pliki .csv, które są przesyłane(12) do GraphHandler (11) w celu przeprowadzenia analizy dla wybranego podzbioru danych. Pliki utworzone przez GraphHandler zostają na końcu przesłane(10) do ReportController(11) w celu ukończenia analizy szczegółowej.
W przypadku wybrania przez użytkownika opcji stworzenia raportu, AnalysisController zleci ReportController (13) stworzenie pustego pliku .pdf, który potem jest przekazywany wielokrotnie w celu dalszego uzupełnienia (12)(13).

3.2 Baza danych

Schemat bazy został w większości zaczerpnięty z Pierwszego Projektu. Dokonano w nim jedynie dwóch zmian:

a) W tabeli TAGs dodano pole “Category” przyjmujące jedną z trzech wartości: “Currency”, “Country Name”, “Organization”
b) Zmieniono relację między tabelami Countries i TAGs z 1..1 na 1..*.



Poniżej opis tabel zaczerpnięty z Pierwszego Projektu
Opis tabel:
• PressReleases: tabela notek prasowych zebranych do analizy treści, zawiera kanał RSS, z którego pochodzi notka, tytuł, datę opublikowania oraz zawartość,

• Feeds: tabela kanałów RSS, z których pochodzą notki prasowe zebrane do analizy treści, zawiera nazwę, gazetę, z której pochodzi oraz nazwę działu wiadomości, którego dotyczy,

• Neswpapers: tabela magazynów, gazet, z których pochodzą kanały RSS, zawiera nazwę, kraj, w którym są wydawane oraz język, w którym pisane są artykuły

• Languages: tabela języków, w których napisane zostały notki prasowe,

• Countries: tabela państw, z których pochodzą kanały RSS oraz o których wspomina się w notkach,

• TAGs: tabela tagów (znaczników) państw występujących w treści notek,

• PressReleasesTAGs: tabela łącznikowa dla relacji wiele-do-wielu między tabelami PressReleases i TAGs.

4. Analiza danych

4.1 Format wyników analizy wstępnej

Analiza wstępna może być przeprowadzona w jeden z następujących sposobów:

1. dla wszystkich gazet jednocześnie, z podziałem na daty (miesięczne)
2. dla wszystkich dat jednocześnie, z podziałem na gazety
3. z podziałem na gazety, a wewnątrz nich z podziałem na miesiące
4. z podziałem na gazety, a wewnątrz nich z podziałem na tygodnie
5. z podziałem na gazety, a wewnątrz nich z podziałem na dni

Uzyskane w ten sposób wyniki mogą zostać wykorzystane w analizie szczegółowej, która wykorzystuje wyniki czwartej opcji.

Pliki otrzymane dla analiz:

- *tytul*.csv
- *tytul*_nodes.csv
- *tytul*_edges.csv

gdzie *tytul* dla odpowiednich analiz przyjmuje następującą postać:

1. Month
2. Newspaper
3. *tytuł gazety*(months)
4. *tytuł gazety*(weeks)
5. *tytuł gazety*(days)

tytul.csv - zawiera parametry dla całości grafu (np. *density*) dla grafu dla całego przedziału czasowego
tytul_nodes.csv - zawiera wszystkie tagi i wartości ich parametrów (np. *degree*) dla całego przedziału czasowego
tytul_edges.csv - zawiera schemat grafu - listę wszystkich wierzchołków oraz ich sąsiadów (*Source* - wszystkie występujące wierzchołki, *Occurrences* - liczba wystąpień tagu w notatkach, kolejne kolumny - sąsiedzi danego wierzchołka (*N*i - i-ty sąsiad) na przemian z wagami (*W*i - waga krawędzi do i-tego sąsiada))

Pliki tytul.csv i tytul_nodes.csv mają wspólny format danych:
Date - data przedziału czasowego, którego dot. wynik
Newspaper - nazwa czasopisma
Param name - nazwa parametru (grafu lub wierzchołka)
Param value (albo *nazwa_tagu*) - wartość parametru (dla całego grafu lub dla danego tagu)
We wszystkich następnych kolumnach może występować tylko *nazwa_tagu* (tylko w pliku dot. parametrów wierzchołków).
Z kolei w tytul_edges.csv pierwsze kolumny to *Date* i *Newspaper* (analogiczne do tych wyżej opisanych), kolejne to:
Source -nazwa tagu określającego wierzchołek grafu

Occurrences - liczba wystąpień *Source* w notkach w danym przedziale czasu
*N*i - i-ty sąsiad wierzchołka *Source*
*W*i - waga krawędzi prowadzącej do i-tego sąsiada

4.2 Format wyników analizy szczegółowej

Analiza szczegółowa jest przeprowadzana dla dwóch czasopism lub jednego. Pliki otrzymane w wyniku analizy (w katalogu NewsAnalyzer/App/src/main/resources):

1. w katalogu csv:
    a. rdzen.csv
    b. rdzen_days.csv
    c. rdzen_nodes.csv
    d. rdzen_TOP.csv
    e. rdzen_days_TOP.csv
    f. rdzen_edges.csv
2. w katalogu charts:
    a. rdzen*.png użyte później w raporcie .pdf (nie wszystkie, które zostały wygenerowane, są tam zapisane - część występuje tylko w .pdf)
3. w katalogu reports:
    a. rdzen.pdf

gdzie:
rdzen - tytul1_tytul2_(data1_data2)
tytul1 - tytuł pierwszej gazety
tytul2 - tytuł drugiej gazety
data1 - pierwsza data brana pod uwagę
data2 - ostatnia data brana pod uwagę

Możliwe formaty daty:

- miesięczna: yyyy-MM
- tygodniowa: yyyy-MMwX (gdzie X - nr tygodnia, np. dla dni miesiąca z [1,7] X=1, dla dni [8,14] X = 2, itd.)
- dzienna: yyyy-MM-dd

rdzen.csv - zawiera parametry dla całości grafu (np. *density*) dla grafu dla całego przedziału czasowego
rdzen_days.csv - j.w., ale parametry są rozpisane dla każdego dnia z wybranego przedziału czasowego
rdzen_nodes.csv - zawiera wszystkie tagi i wartości ich parametrów (np. *degree*) dla całego przedziału czasowego
rdzen_TOP.csv - zawiera tagi, dla których określone parametry miały największą wartość oraz ich wartości (np. dla *degree* może być tag USA, obok którego będzie wartość stopnia wierzchołka grafu dla tego tagu) - dla całego przedziału czasowego (podzbiór danych z pliku rdzen_nodes.csv)

rdzen_days_TOP.csv - j.w., ale rozpisane na poszczególne dni przedziału czasowego
rdzen_edges.csv - zawiera schemat grafu - listę wszystkich wierzchołków oraz ich sąsiadów (*Source* - wszystkie występujące wierzchołki, *Occurences* - liczba wystąpień tagu w notatkach, kolejne kolumny - sąsiedzi danego wierzchołka (*N*i - i-ty sąsiad) na przemian z wagami (*W*i - waga krawędzi do i-tego sąsiada))

Pliki rdzen.csv, rdzen_nodes.csv, rdzen_days.csv i rdzen_days_TOP.csv mają wspólny format danych, analogiczny do formatu plików tytul.csv i tytul_nodes.csv. Format rdzen_edges.csv jest taki sam, jak dla tytul_edges.csv

W raporcie .pdf są umieszczone wykresy dla wyniku porównania czasopism dla danego przedziału czasowego. W tytułach wykresów jest nazwa parametru (i ew. nazwa tagu), etykietami na osi X są daty (dla parametrów grafu lub parametrów najważniejszych wierzchołków) lub nazwy tagów (tylko dla parametrów wierzchołków). W tytułach został umieszczony również współczynnik korelacji obliczony wg. korelacji rangowej Spearmana. Wzór:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

gdzie:
$d_i = \mathrm{R}\, x_i - \mathrm{R}\, y_i$ - różnica między rangami zmiennych X i Y dla i-tego wyniku
n - liczba wyników

4.3 Miary do podstawowej analizy grafu i analizy sieci społecznych

Dostarczone przez samo Gephi:

- **liczba wierzchołków** (graph.getNodeCount())
- **liczba krawędzi** (graph.getEdgeCount())
- **ClusteringCoefficient**
  - *dla wierzchołka:* miara jak kompletne jest sąsiedztwo wierzchołka (jak dużo brakuje jego sąsiedztwu do stworzenia kliki) - wzór z wikipedii:

    $$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

    gdzie:
    N_i - sąsiedztwo wierzchołka v_i, $N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}.$
    E - zbiór krawędzi
    k_i = $|N_i|$
  - *dla całej sieci:* średnia z wartości tej miary dla wszystkich wierzchołków - wzór z wikipedii

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i.$$

gdzie:
n - liczba wszystkich wierzchołków

- opis na wiki gephi [5]

- **ConnectedComponents** – na grafach skierowanych wykrywa silne i słabe spójne składowe, na grafach nieskierowanych tylko (słabe) spójne składowe
  - opis na gephi wiki [6] i odniesienie do algorytmu
- **Degree** - stopień wierzchołka (inne możliwości: stopień wejściowy, wyjściowy, średni stopień wierzchołków sieci)
- **EigenvectorCentrality** – miara ważności wierzchołka w sieci na podstawie połączeń między wierzchołkami
  - wzór do otrzymania miary (wikipedia)

    Dla grafu $G := (V, E)$ z liczbą wierzchołków $|V|$ niech $A = (a_{v,t})$ będzie macierzą sąsiedztwa, (tj. $a_{v,t} = 1$ jeśli wierzchołek v jest połączony z wierzchołkiem t, a $a_{v,t} = 0$ w przeciwnym wypadku). Wzór:

    $$x_v = \frac{1}{\lambda}\sum_{t \in M(v)} x_t = \frac{1}{\lambda}\sum_{t \in G} a_{v,t} x_t$$

    gdzie:

    $M(v)$ - zbiór sąsiadów v

    λ - stała

    Po przekształceniu może być zapisane w notacji wektorowej jako równanie:

    $$\mathbf{Ax} = \lambda \mathbf{x}$$
  - opis na wiki gephi [7]
- **Hits** – Hyperlink-Induced Topic Search (znane również jako hubs and authorities - koncentratory i autorytety) - miara określa ważności wierzchołków w sieci, autorytet definiuje wartość danego wierzchołka, a koncentrator szacuje wartość połączeń wychodzących z wierzchołka
  - opis na wiki gephi [8]
- **GraphDensity** - oblicza gęstość grafu - jak blisko grafu pełnego jest dana sieć. Graf pełny (ze wszystkimi możliwymi krawędziami) ma gęstość = 1. calculates graph density
  - wzór z wikipedii:

    $$D = \frac{|E|}{|V|\left(|V| - 1\right)}$$

    gdzie
    V - liczba wierzchołków
    E - liczba krawędzi
  - opis na wiki gephi [9]
- **GraphDistance** - gromadzi kilka rodzajów miar
  - **betweenness centrality** - określa ważność wierzchołka w sieci - jak często wierzchołek pojawia się na najkrótszych ścieżkach między wierzchołkami

        - opcjonalna normalizacja: podzielenie wyniku przez liczbę par wierzchołków (bez uwzględniania tego wierzchołka, dla którego liczymy wynik) - dla grafów skierowanych to (n-1)(n-2)
          gdzie n - liczba wierzchołków
        - opis na wiki gephi [10]
    - **closeness centrality** - średnia odległość od danego wierzchołka do pozostałych wierzchołków w sieci
        - opcjonalna normalizacja: przemnożenie wyniku przez (n-1)
          gdzie n - liczba wierzchołków w grafie
          (normalizacja ułatwia porównywanie węzłów w grafach różnych rozmiarów)
        - opis na wiki gephi [11]
    - **eccentricity** - największa odległość z danego wierzchołka do jakiegokolwiek innego wierzchołka (jak daleko jest dany wierzchołek od najdalszego wierzchołka w grafie)
    - **harmonic closeness centrality** - rola podobna do closeness centrality, ale daje lepsze rezultaty dla grafów mających dwie lub więcej spójnych składowych
        - wzór z wikipedii:

          $$H(x) = \sum_{y \neq x} \frac{1}{d(y,x)}.$$

          gdzie: d(y,x) - odległość między wierzchołkami y i x
        - opcjonalna normalizacja: : podzielenie wyniku przez (n-1)
          gdzie n - liczba wierzchołków w grafie
    - **getPathLength**() -zwraca długość najkrótszej ścieżki w grafie
    - **getDiameter**() - zwraca średnicę grafu - maksymalną wartość *eccentricity* spośród wszystkich wierzchołków
        - wyjaśnienie na wiki gephi [12]
    - **getRadius**() -zwraca promień grafu - najmniejszą wartość *eccentricity* spośród wszystkich wierzchołków
- **Modularity**
    - *dla wierzchołków*: przypisuje podspołeczność, do której należą
    - *dla całej sieci:* mierzy jak dzieli się sieć na podspołeczności (dla całej sieci) - wysoka wartość wskazuje na wyrafinowaną strukturę wewnętrzną
    - opis na wiki gephi [13]
- **PageRank** – określa ważność wierzchołka w sieci - przypisuje wierzchołkom prawdopodobieństwo dotarcia do danej strony
    - opis na wiki gephi [14]
- **WeightedDegree** -podobny do Degree, ale uwględnia wagi krawędzi (możliwości: wdgegree, wingree, woutdegree)

Dostarczone przez pluginy do gephi:

- **Prestige Plugin** -oblicza kilka metryk dla grafów skierowanych:
    - opis pluginu [15]
    - **Indegree prestige:** liczba innych węzłów, które łączą się bezpośrednio z wierzchołkiem *v*

        - opcjonalna normalizacja: wynik zostaje podzielony przez (n-1) gdzie n - liczba wierzchołków
          (0 = żadne inne wierzchołki nie łączą się z bezpośrednio *v*,
          1 = wszystkie inne wierzchołki łączą się bezpośrednio z *v*)
    - **Domain prestige:** jaka część sieci jest osiągalna z wierzchołka *v*
    - **Proximity prestige:** uwzględnia wagi krawędzi - krótsze ścieżki są bardziej wartościowe. Parametry i wzór:
      v - węzeł, dla którego liczymy miarę
      I - zbiór węzłów osiągalnych z *v*
      Sum(d) - suma długości najkrótszych ścieżek z każdego wierzchołka należącego do *I* do *v*
      n - liczba wszystkich wierzchołków
      above = |I| / (n-1)
      below = SUM(d) / |I|
      ProximityPrestige (v) = above / below = domain_prestige_normalized/below
    - **Rank prestige:** w każdym wierzchołku *v* sumuje wartość wybranego parametru wierzchołka (np. betweenness centrality) dla wszystkich sąsiadów *v*. Parametr (tzw. prominence), musi być wartością logiczną lub liczbą.
        - opcjonalna normalizacja: Min-Max-Normalized Rank-Value dla wierzchołka

4.4 Format wyników analizy walutowej

W ramach analizy walutowej możliwe jest stworzenie statystyk odnośnie następujących parametrów:

1. Liczba wystąpień danego tagu we wszystkich notkach
2. Liczba wystąpień danego tagu dla gazety (iteracja po wszystkich gazetach)
3. Liczba wystąpień pary tagów dla gazety (iteracja po wszystkich gazetach)
4. Liczba wystąpień pary tagów dla konkretnej gazety, rozłożona w czasie (iteracja po miesiącach).

Wyniki analiz znajdują się w plikach *.csv w następujących miejscach:

Typ analizy	Ścieżka do wyników
1	analysis/currencyTagsStats.csv
2	analysis/currencyTagStatsForNewspapers.csv
3	analysis/tagPairStatsForNewspaper/<nazwa_gazety>.csv
4	analysis/tagPairStatsForNewspaperForMonths/<data>.csv

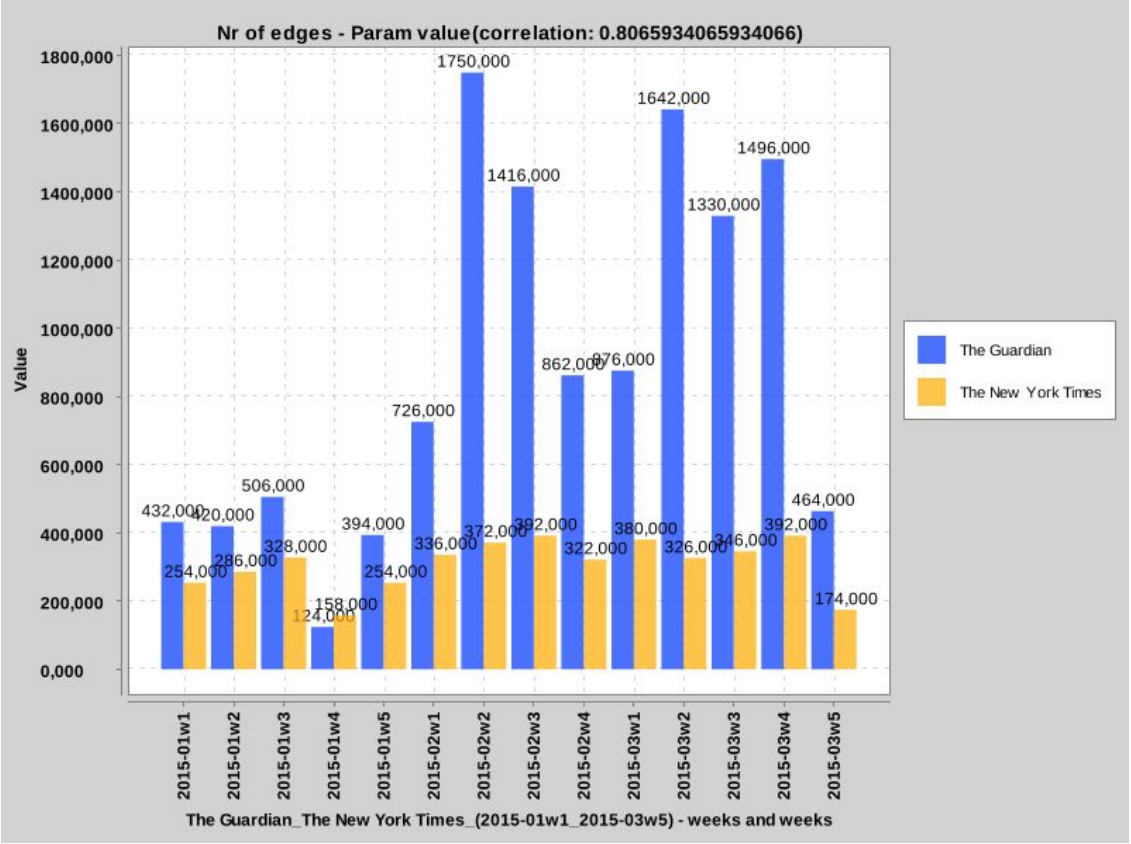
5. Wyniki

5.1 Wyniki porównania różnych gazet

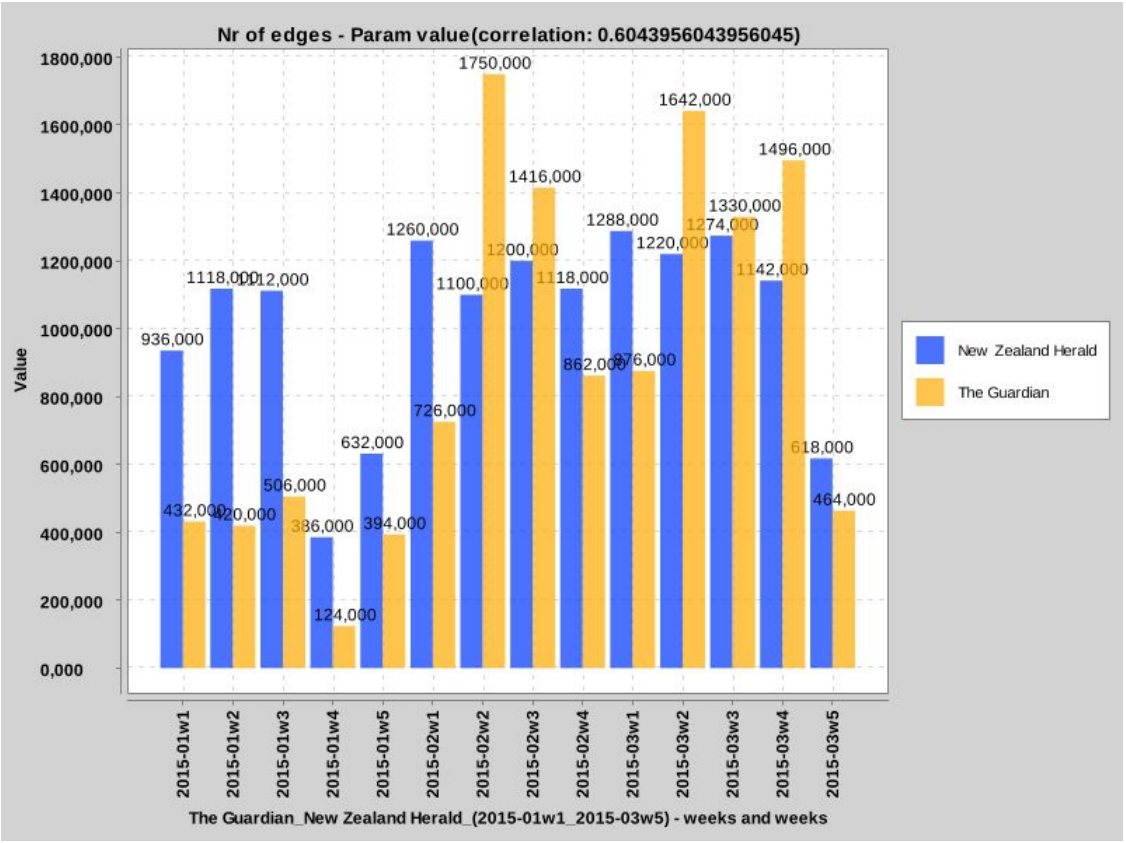
Wyniki powstały na podstawie porównania gazet *New Zealand Herald* i *The New York Times* z T*he Guardian* dla tygodni od 01.01.2015 do 31.03.2015 (losowo wybrany przedział czasowy). *New Zealand Herald* i *The Guardian* mają największą liczbę notatek prasowych, a *The New York Times* jest jedną z najważniejszych amerykańskich gazet (trzecią pod względem nakładu gazetą w Stanach Zjednoczonych).

5.1.1 Liczba krawędzi (tj. liczba wystąpień par tagów)

5.1.1.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)

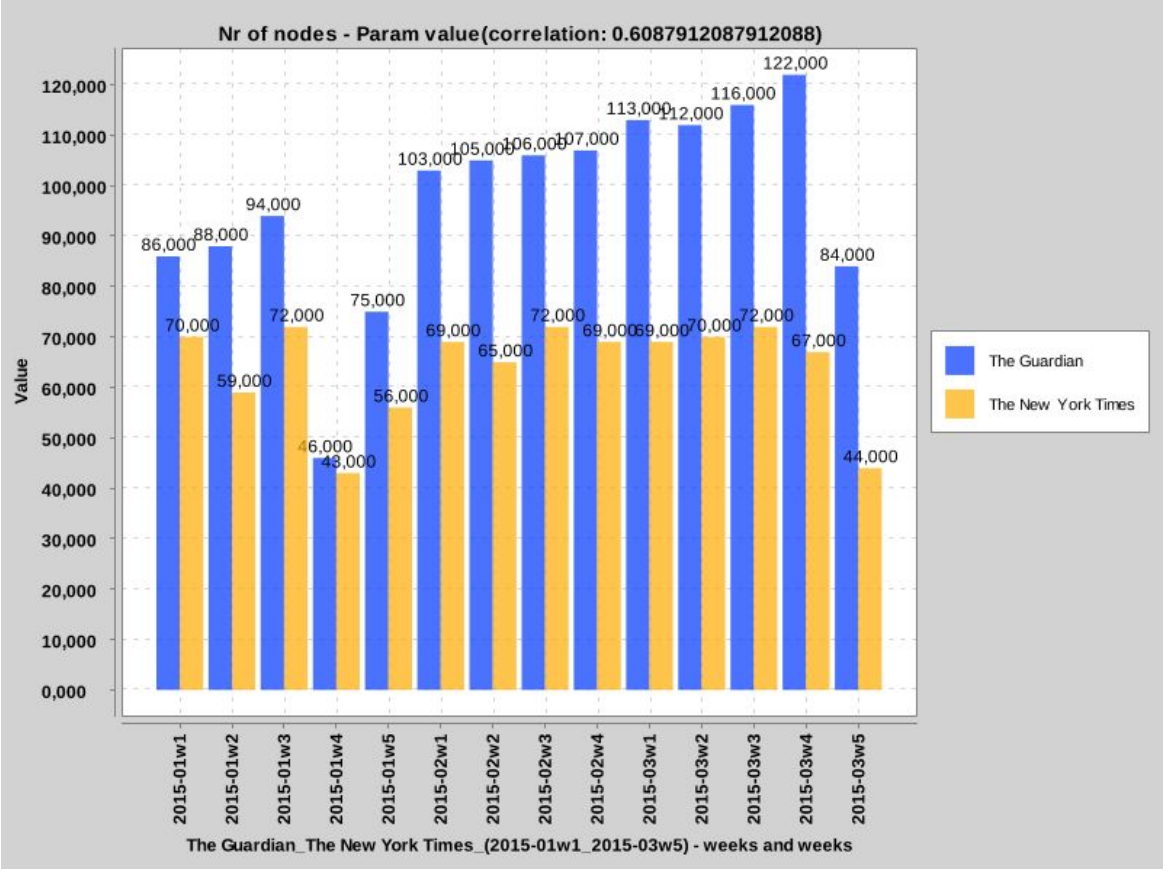


5.1.1.b The Guardian i New Zealand Herald (2015-01w1 - 2015-03w5)

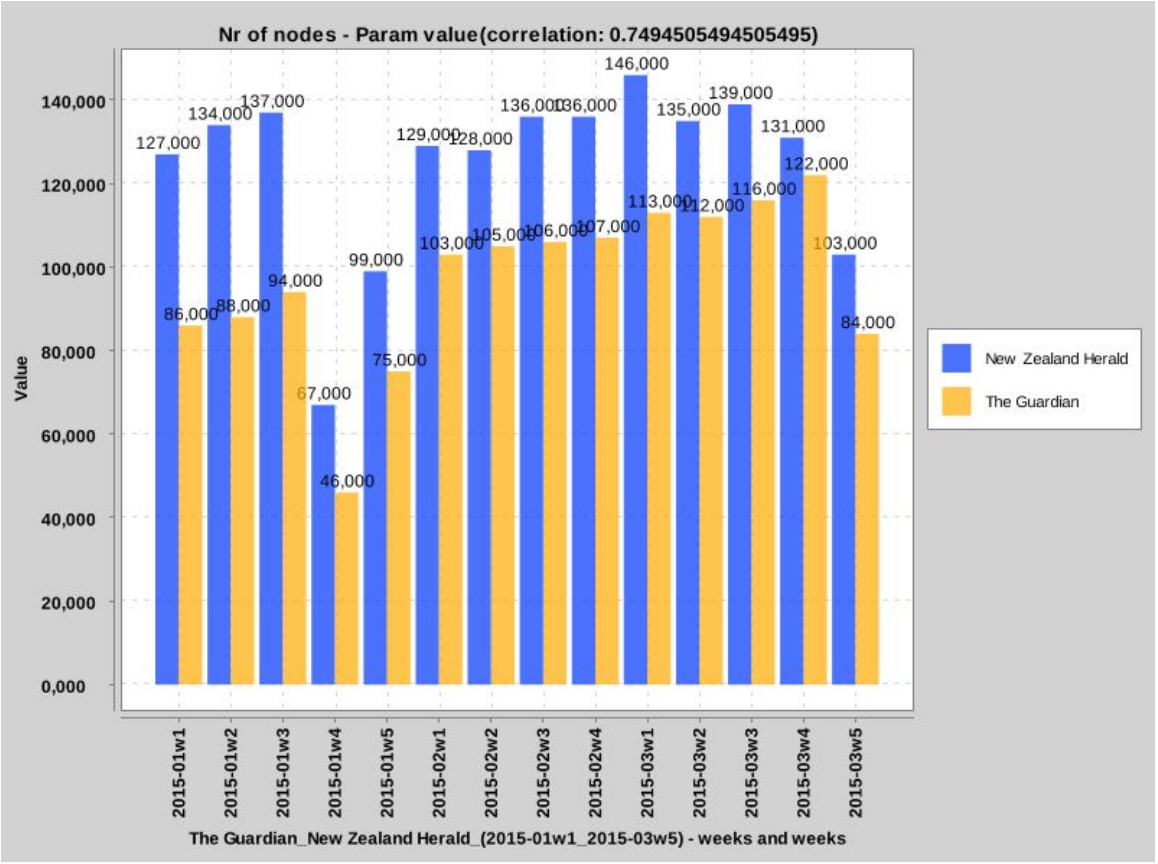


5.1.2 Liczba wierzchołków

5.1.2.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)

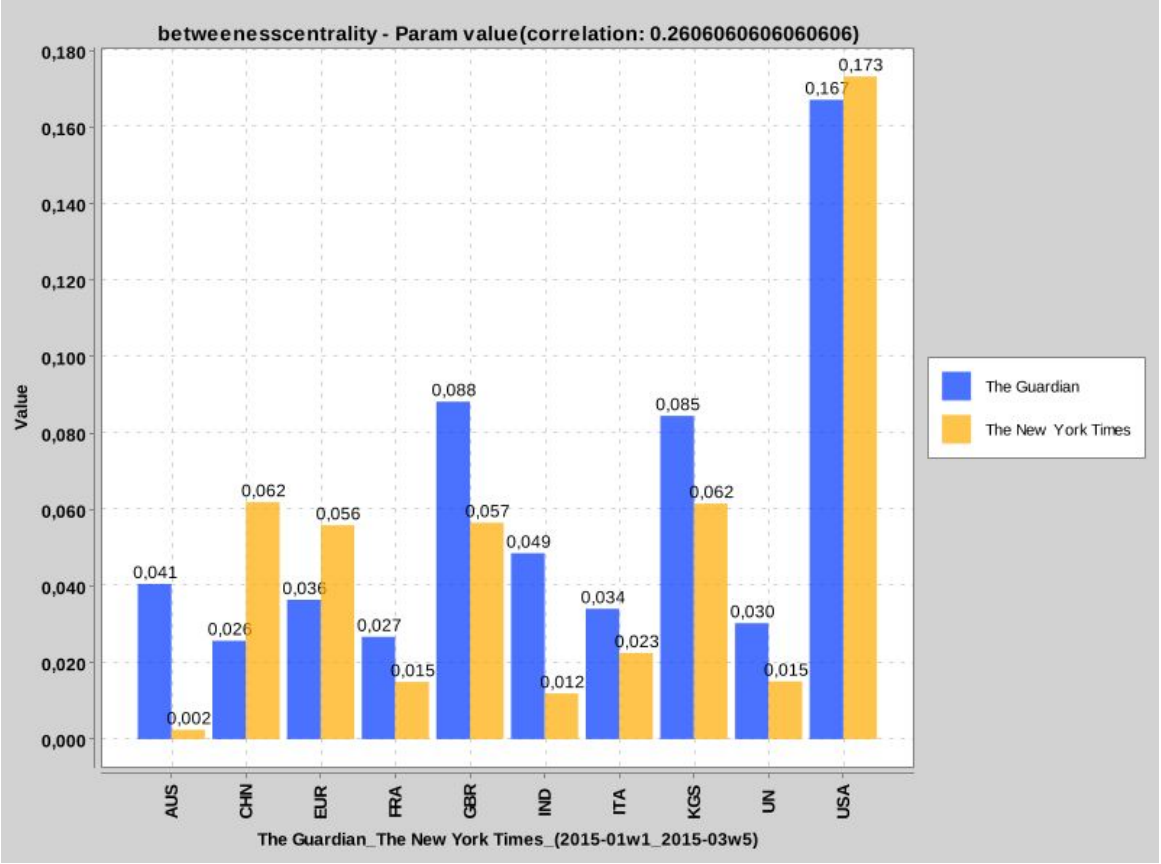


5.1.2.b The Guardian i New Zealand Herald (2015-01w1 - 2015-03w5)

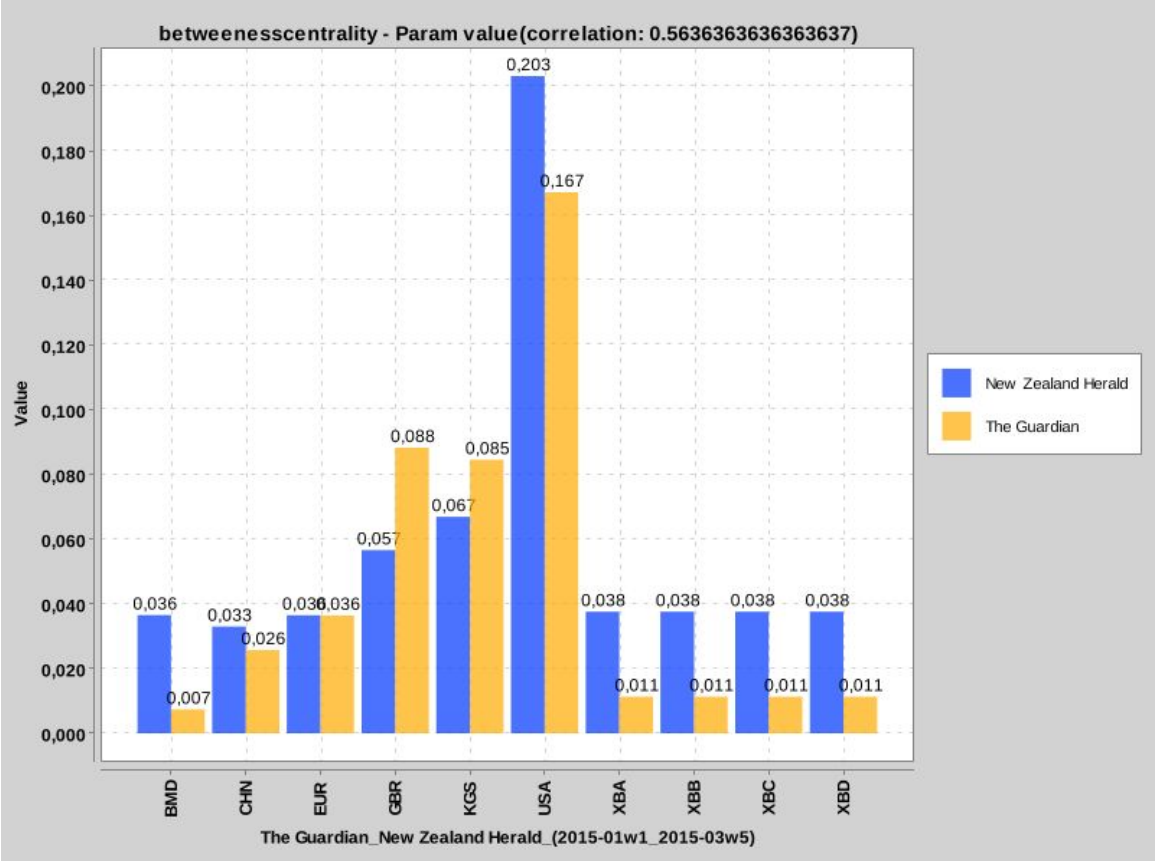


5.1.3 Betweenness centrality tagów

5.1.3.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)

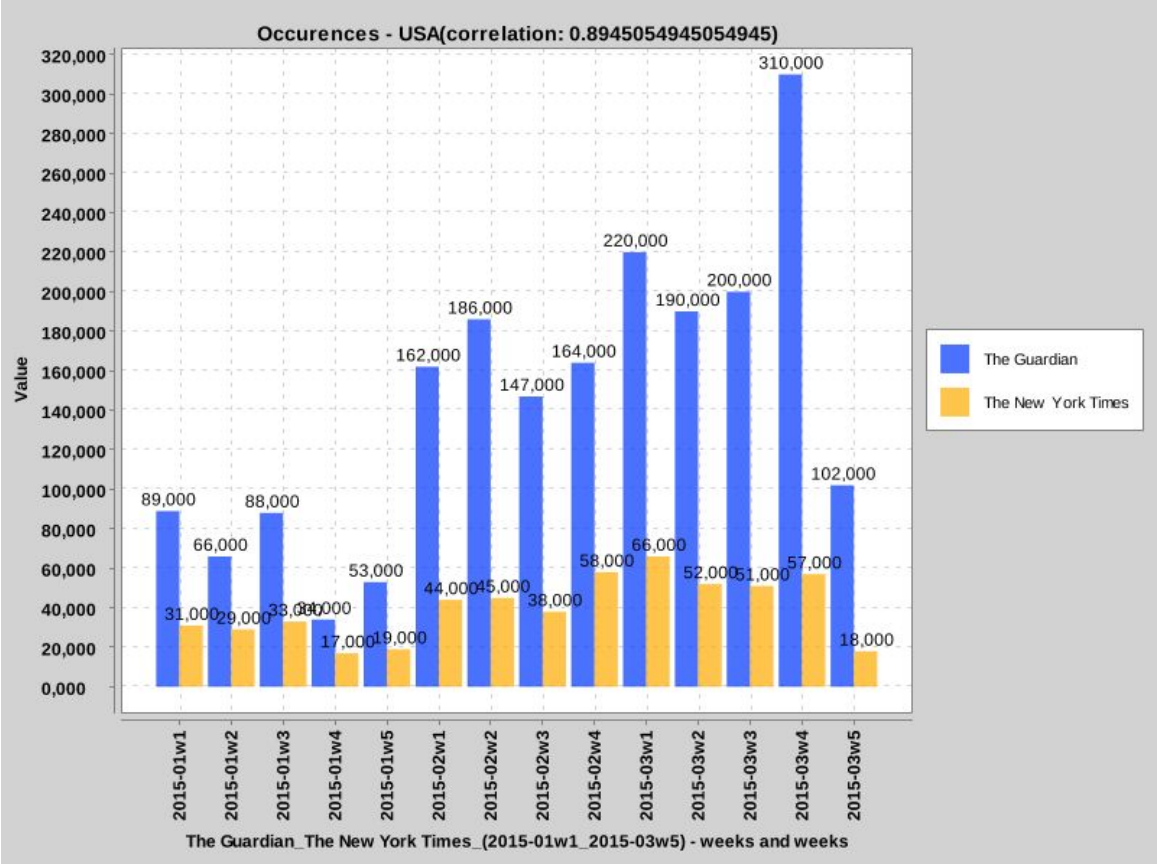


5.1.3.b The Guardian i New Zealand Herald (2015-01w1 - 2015-03w5)

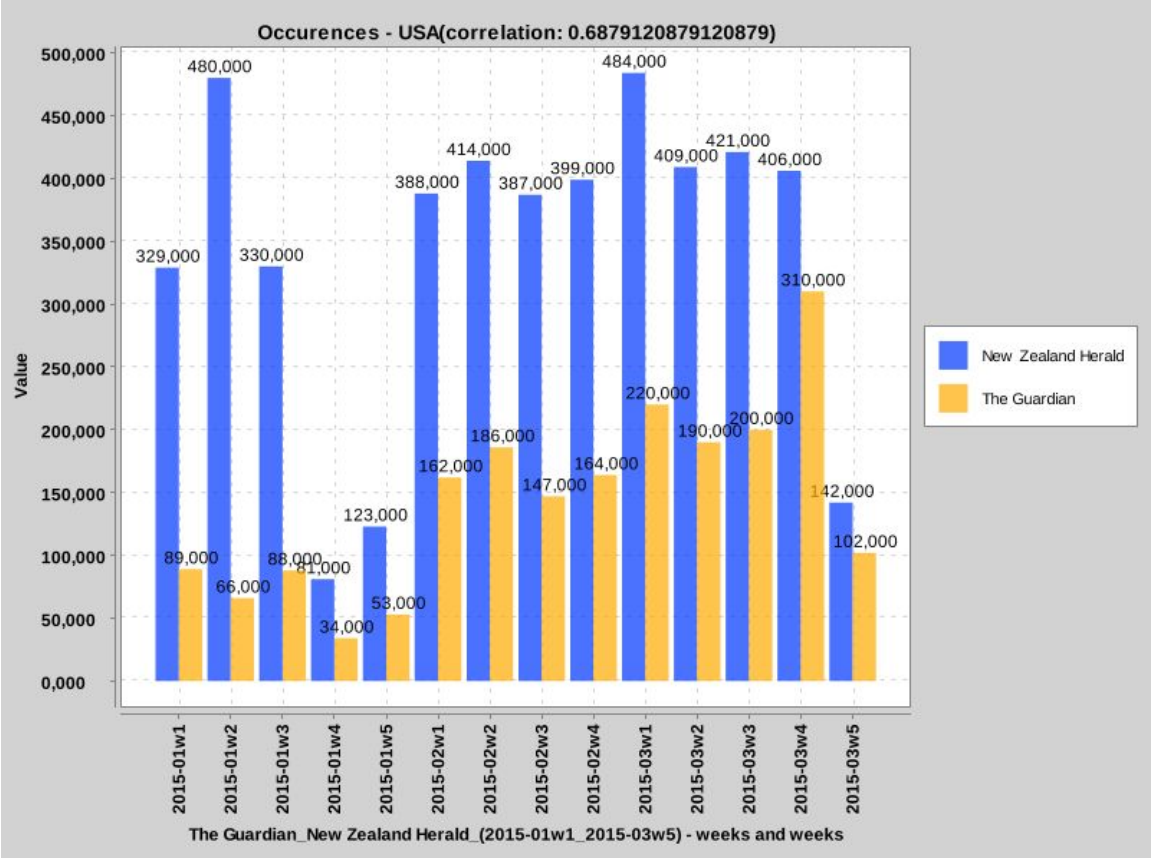


5.1.4 Liczba wystąpień tagu USA

5.1.4.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)

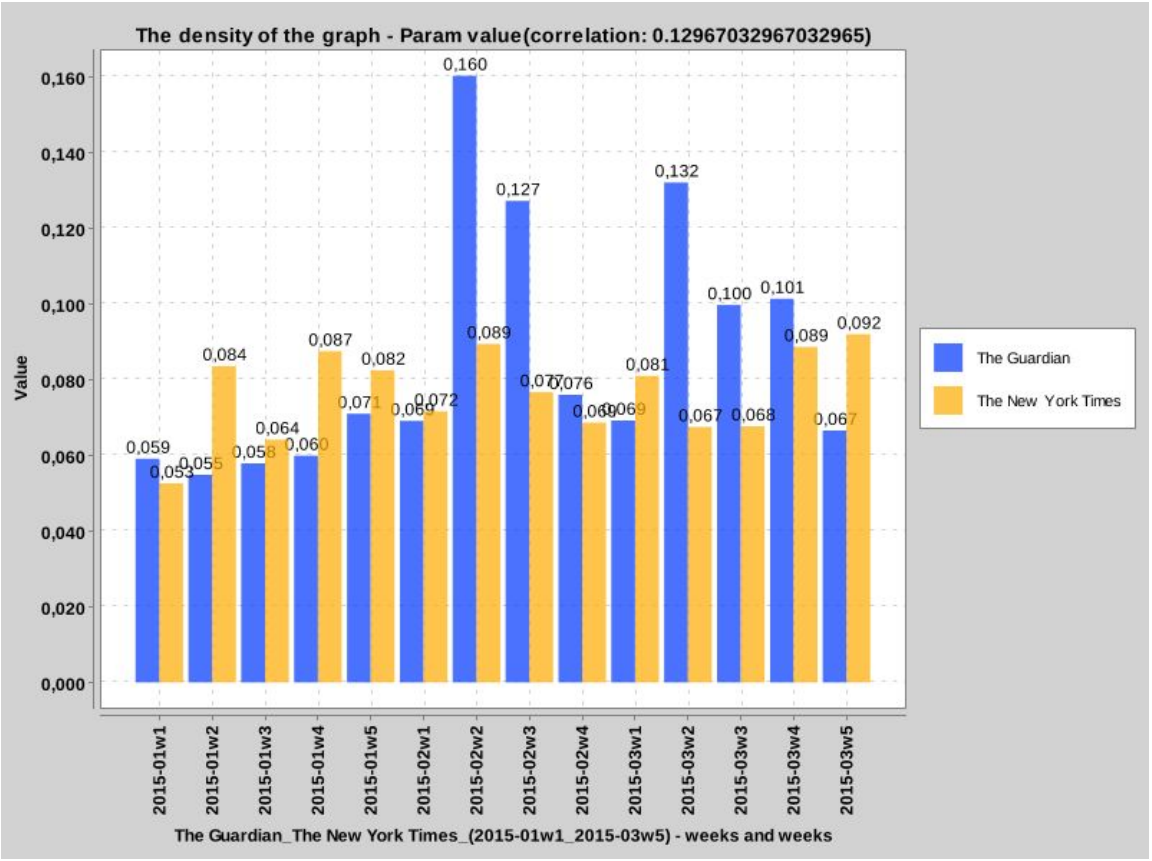


5.1.4.b The Guardian i New Zealand Herald (2015-01w1 - 2015-03w5)

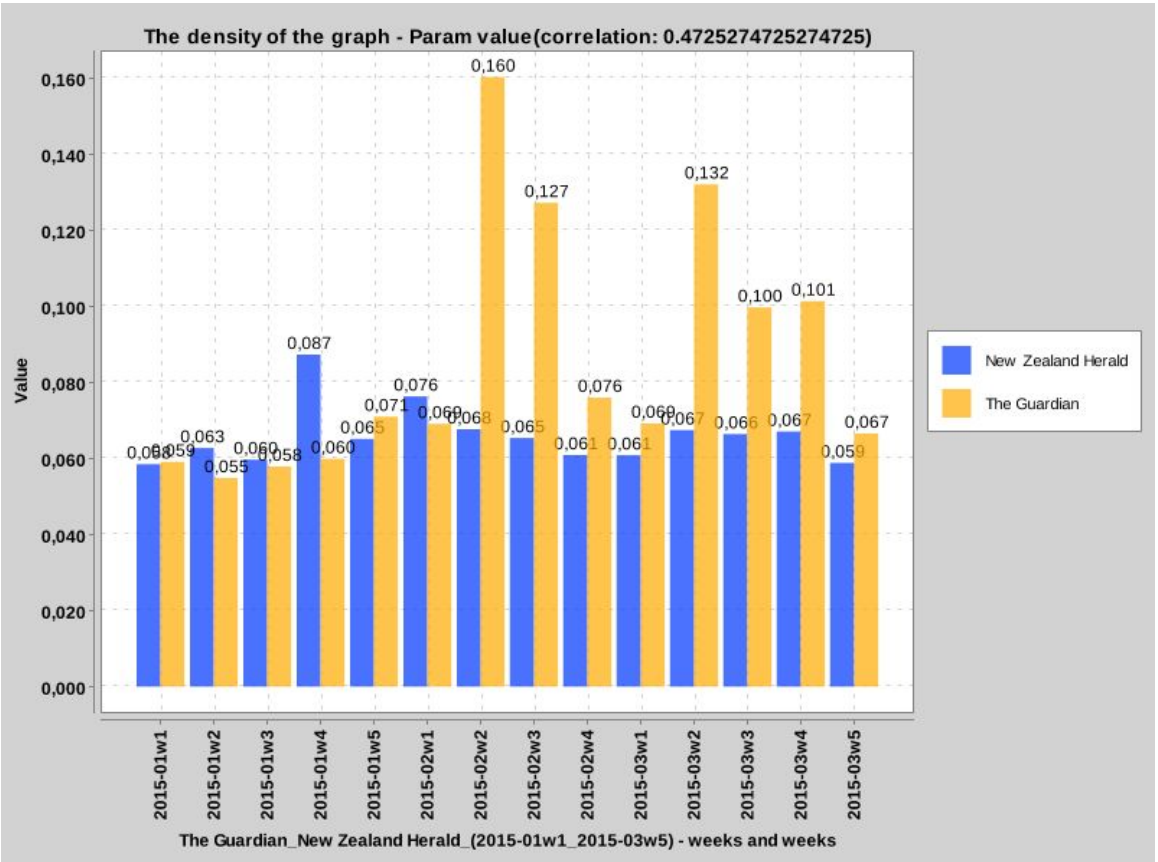


5.1.5 Gęstość grafu

5.1.5.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)

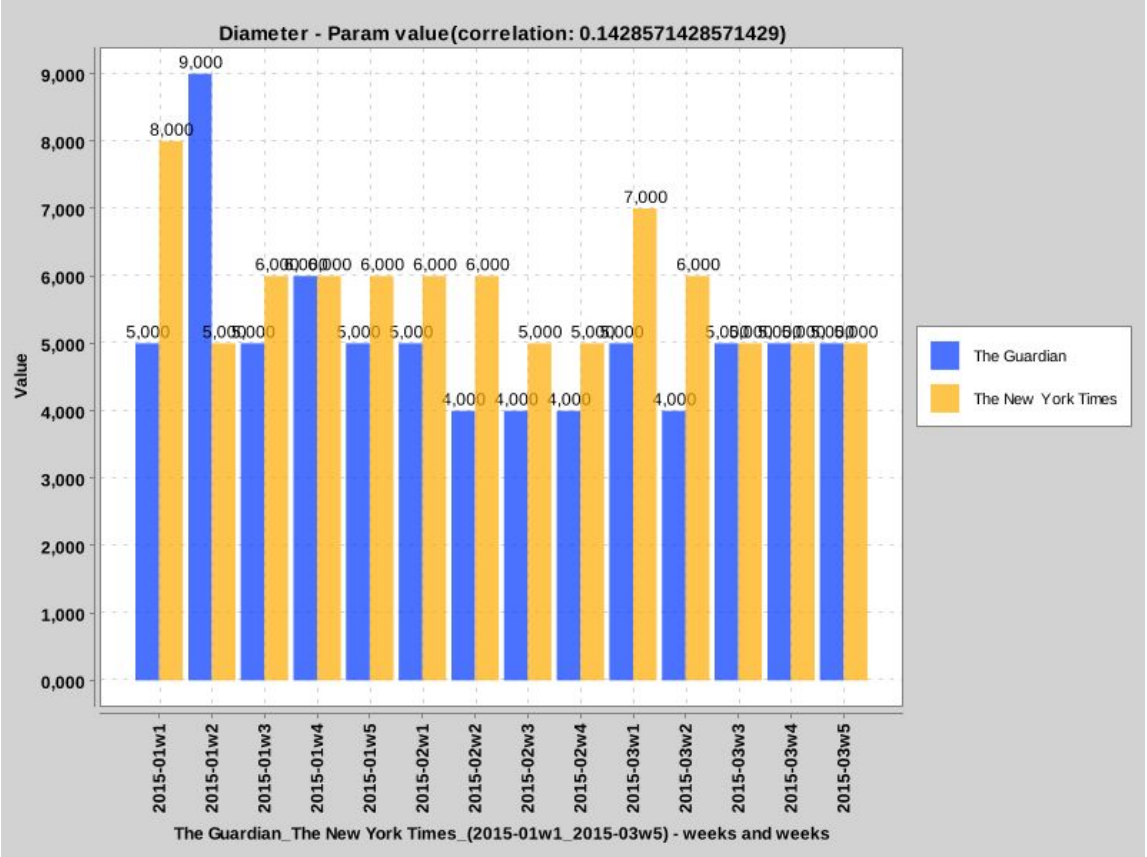


5.1.5.b The Guardian i New Zealand Herald(2015-01w1 - 2015-03w5)

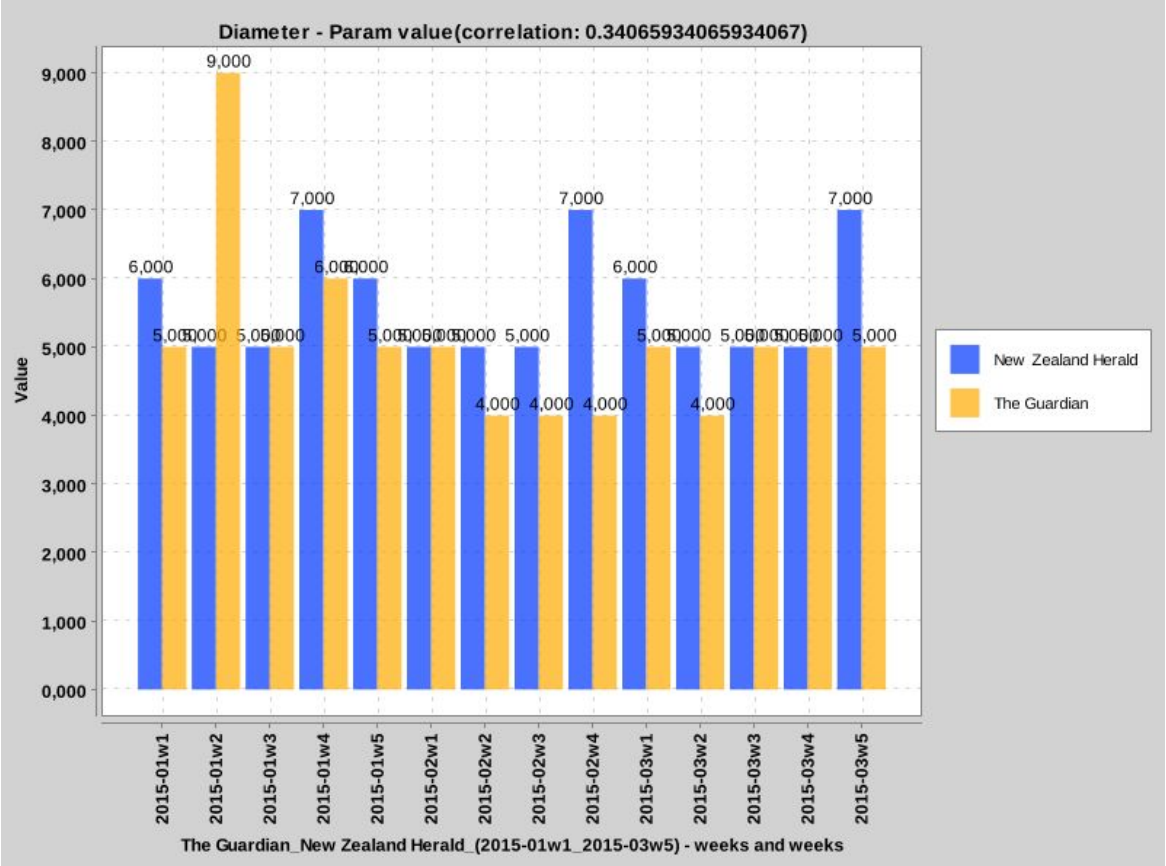


5.1.6 Średnica grafu

5.1.6.a The Guardian i The New York Times (2015-01w1 - 2015-03w5)



5.1.6.b The Guardian i New Zealand Herald (2015-01w1 - 2015-03w5)



Wnioski

Dość wysoka korelacja dla liczby wierzchołków i krawędzi może wskazywać, że wydarzenia światowe miały wpływ na to, że odpowiednio więcej lub mniej rodzajów informacji było jednocześnie zawartych w notkach.
W obu porównaniach betweenness centrality dla USA jest najwyższe ze wszystkich - świadczy to o ogromnej roli, jaką odgrywa USA w polityce światowej.
Dość wysoka jest korelacja dla liczby wystąpień tagu USA w notkach - to także świadczy o tym, że jeśli jedna gazeta wspomniała o jakimś istotnym wydarzeniu w USA, to inna prawdopodobnie też. Jednak dla porównania New Zealand Herald i The Guardian ta korelacja jest niższa - jest to prawdopodobnie związane z tym, że jednak i The Guardian, i The New York Times są gazetami amerykańskimi, a New Zealand Herald nie.
Gęstości grafu są ze sobą trochę skorelowane, ale nie tak bardzo, jak inne parametry - może to świadczyć np. o rozbieżności tematyki różnych gazet. Podobna sytuacja jest ze średnicą grafu - nie są one tak bardzo skorelowane, jak pozostałe parametry. Może to wynikać z tego, że gazety te mają różne profile (piszą na trochę inne tematy) - wówczas w jednej gazecie mogą pojawiać się tagi związane z inną tematyką (i przez to inaczej ze sobą powiązane, co z kolei ma wpływ na średnicę) niż w innej gazecie.

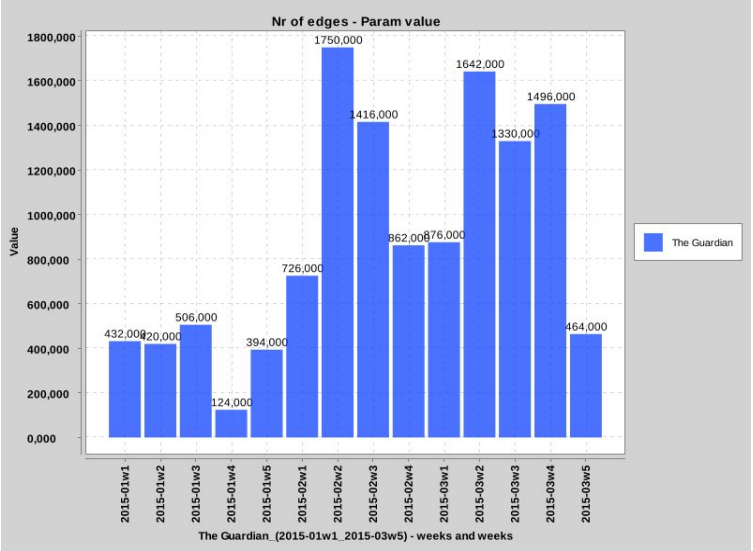
Z wybranych parametrów największa korelacja występuje w przypadku liczby wystąpień USA w notkach dla gazet “The Guardian” i “The New York Times” - wynosi ok. 0.895. To jest spowodowane najprawdopodobniej - jak już wcześniej zostało wspomniane - tym, że obie gazety są amerykańskie. Drugą w kolejności największą korelacją jest liczba krawędzi dla “The Guardian” i “The New York Times” - może to się wiązać z tym, że pewne wydarzenia na świecie wpływały na to, że więcej tagów występowało w notatkach parami, a nie pojedynczo.

5.2 Wyniki dla parametrów jednej gazety (ze wzgl. na tygodnie)

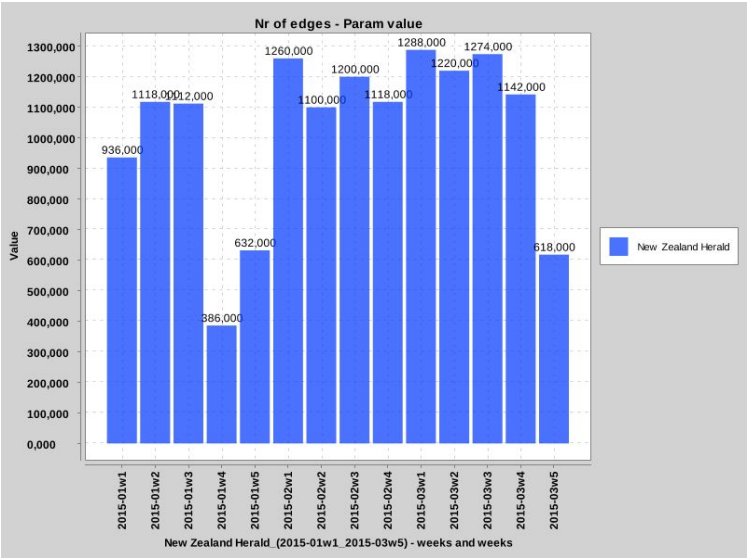
Analiza bazuje na wynikach z analizy wstępnej dla poszczególnych tygodni. Została przeprowadzona dla tego samego przedziału czasu i tych samych gazet, co powyższa. Złożenia tych wykresów są równoważne wykresom powyższej analizy.

5.2.1 Liczba krawędzi

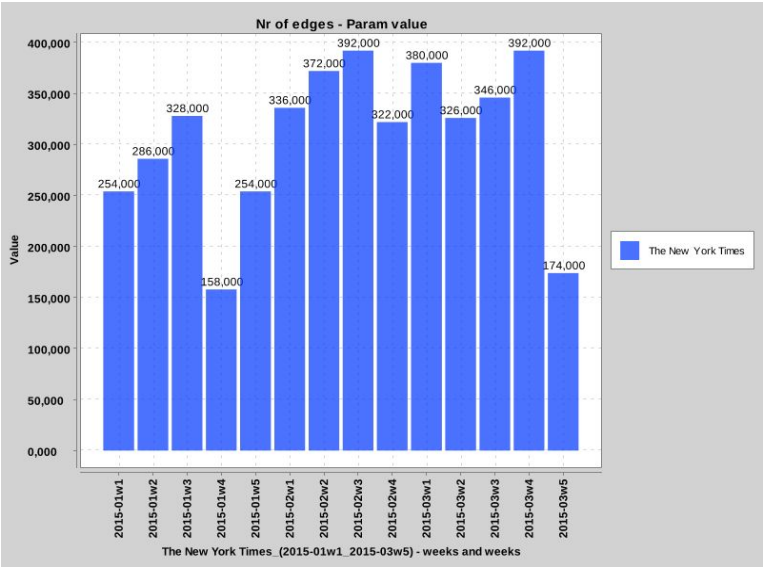
5.2.1.a The Guardian (2015-01w1 - 2015-03w5)



5.2.1.b New Zealand Herald (2015-01w1 - 2015-03w5)

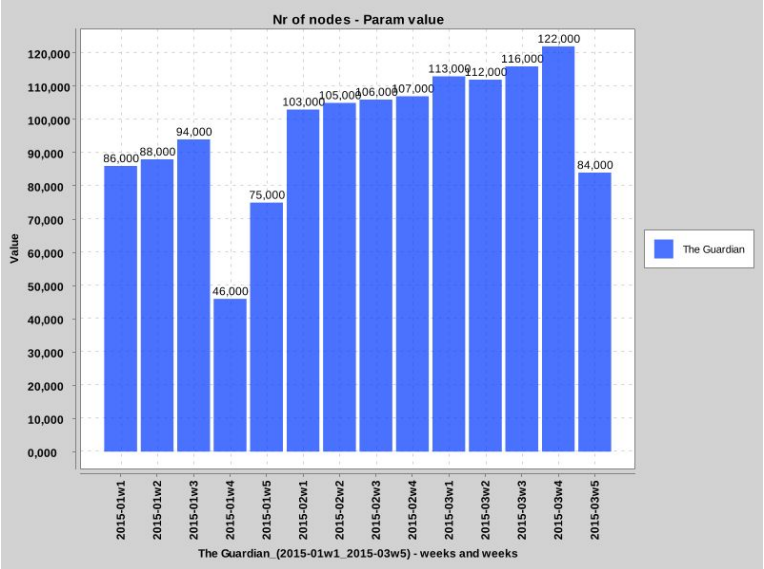


5.2.1.c The New York Times (2015-01w1 - 2015-03w5)

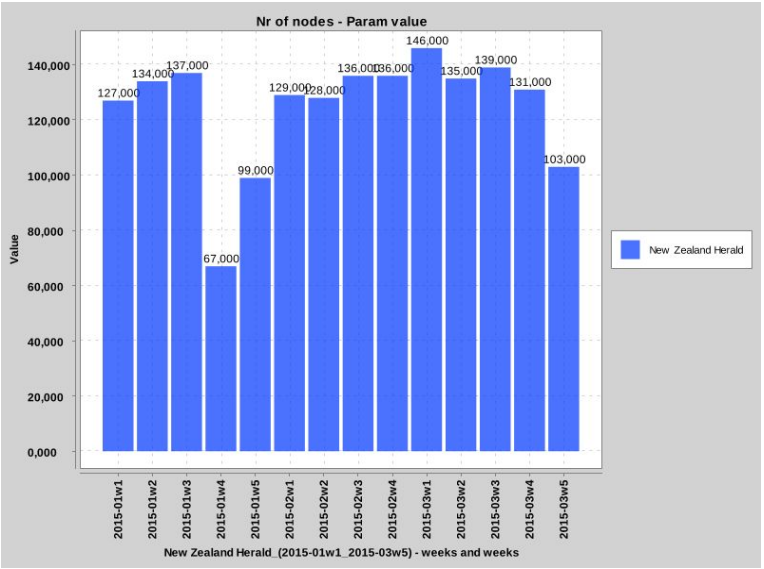


5.2.2 Liczba wierzchołków

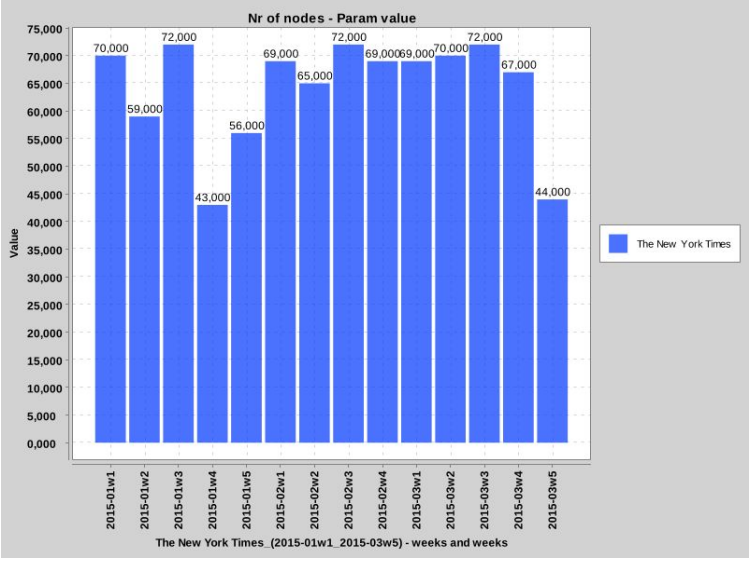
5.2.2.a The Guardian (2015-01w1 - 2015-03w5)



5.2.2.b New Zealand Herald (2015-01w1 - 2015-03w5)

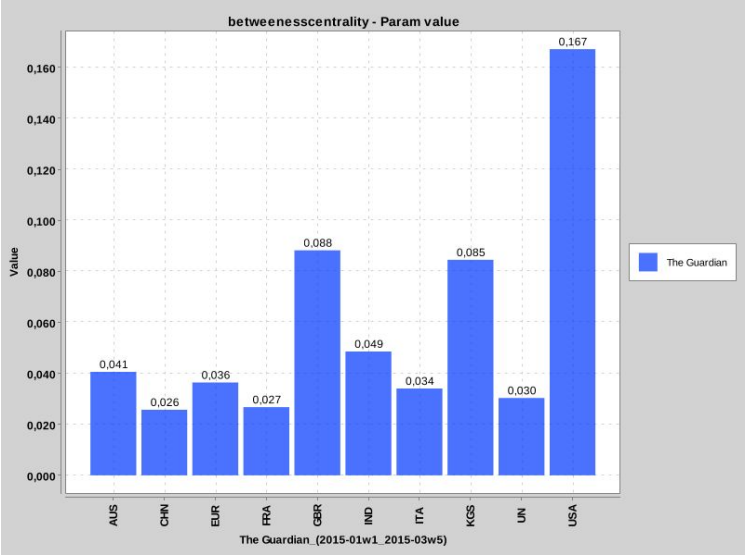


5.2.2.c The New York Times (2015-01w1 - 2015-03w5)

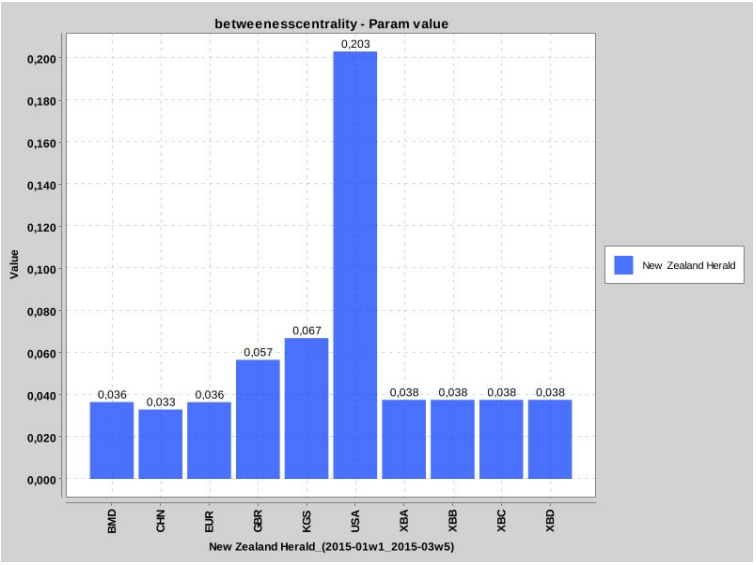


5.2.3 Betweenness centrality wierzchołków

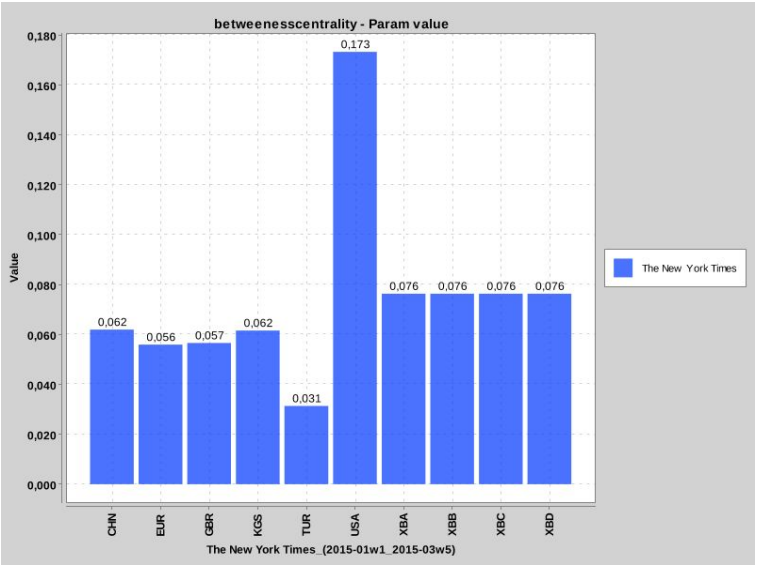
5.2.3.a The Guardian (2015-01w1 - 2015-03w5)



5.2.3.b New Zealand Herald (2015-01w1 - 2015-03w5)

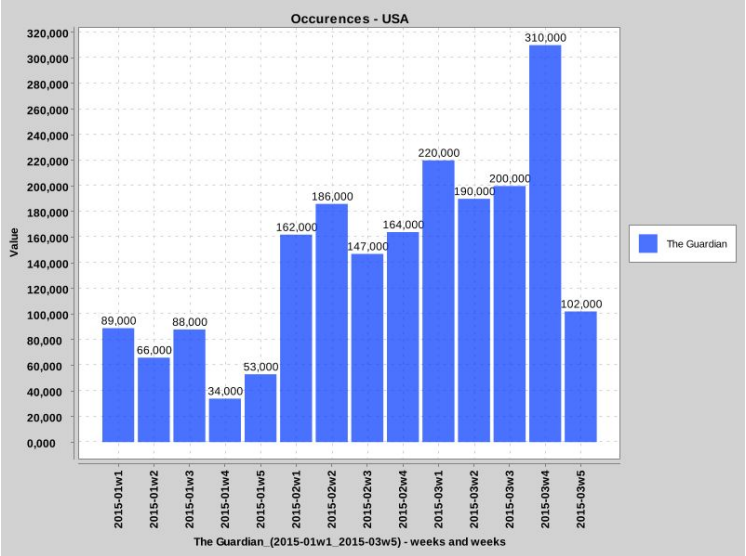


5.2.3.c The New York Times (2015-01w1 - 2015-03w5)

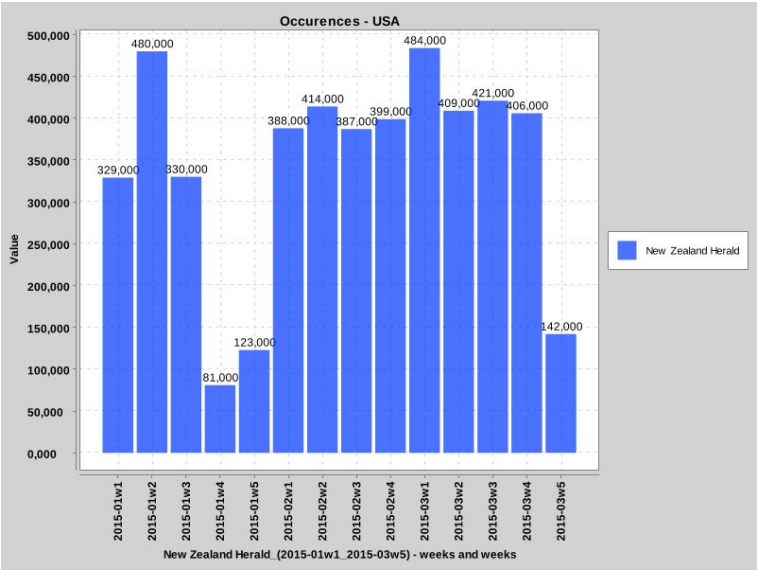


5.2.4 Liczba wystąpień tagu USA

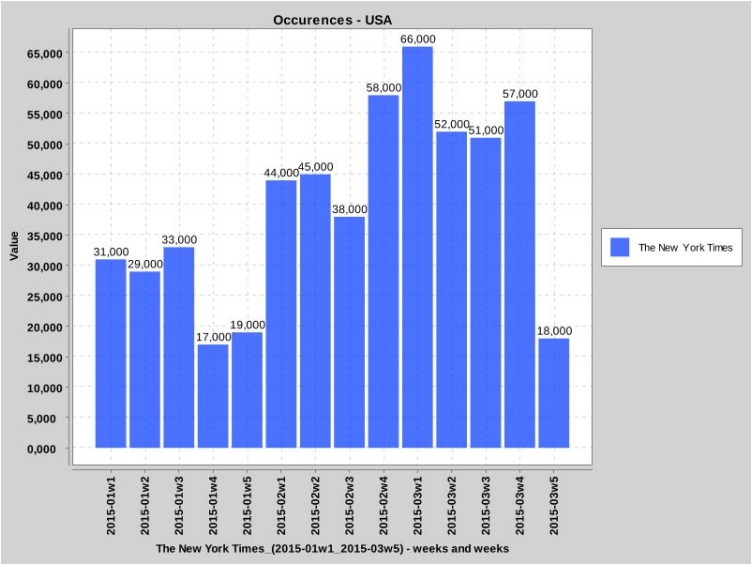
5.2.4.a The Guardian (2015-01w1 - 2015-03w5)



5.2.4.b New Zealand Herald (2015-01w1 - 2015-03w5)

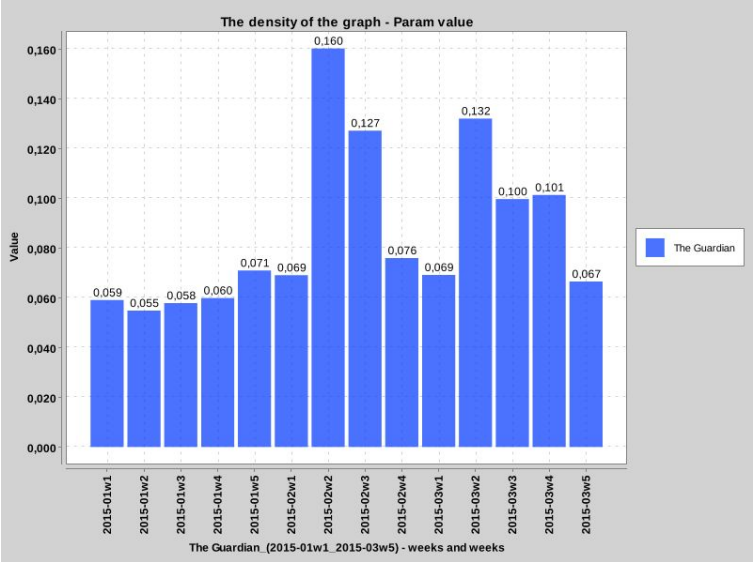


5.2.4.c The New York Times (2015-01w1 - 2015-03w5)

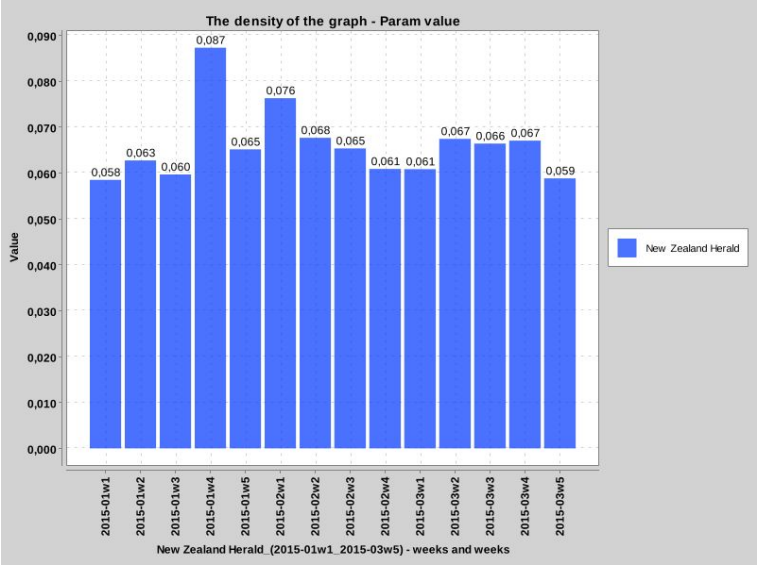


5.2.5 Gęstość grafu

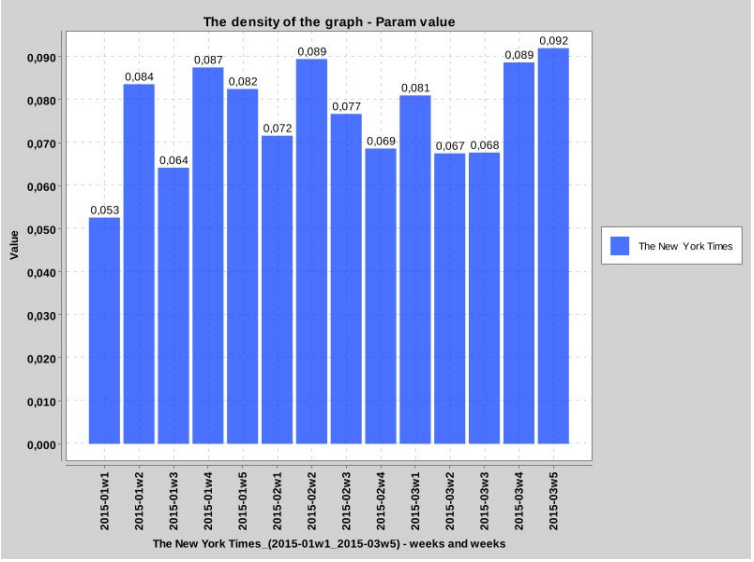
5.2.5.a The Guardian (2015-01w1 - 2015-03w5)



5.2.5.b New Zealand Herald (2015-01w1 - 2015-03w5)

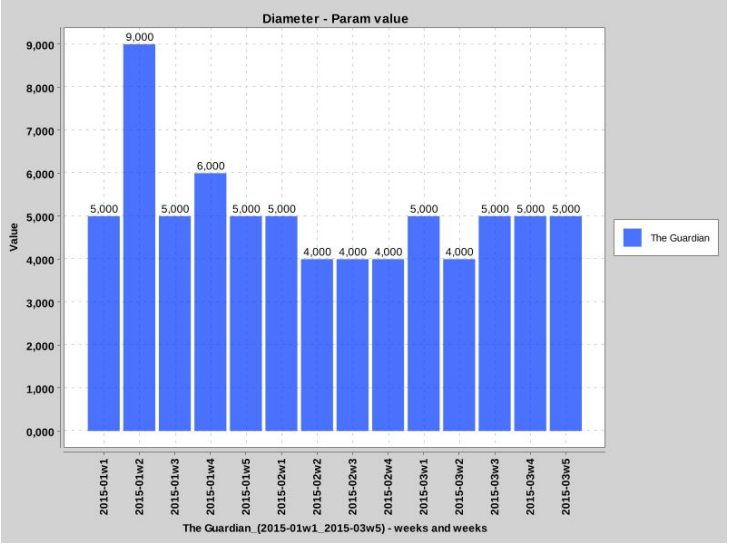


5.2.5.c The New York Times (2015-01w1 - 2015-03w5)

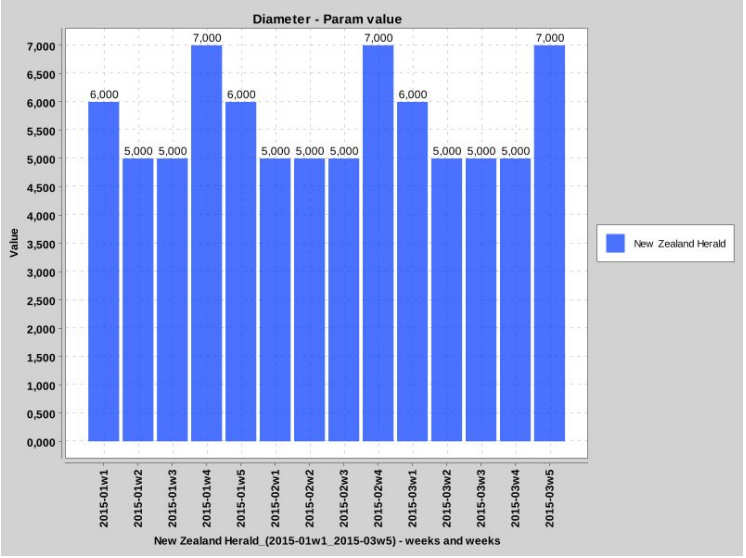


5.2.6 Średnica grafu

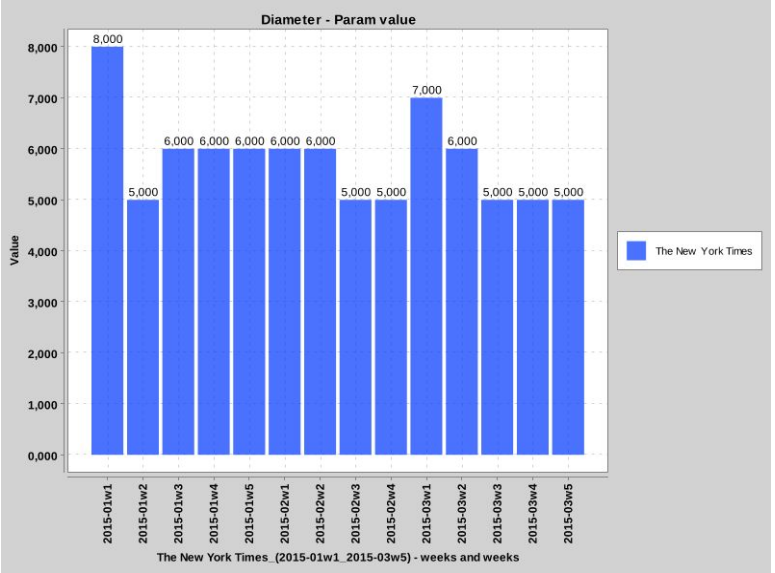
5.2.6.a The Guardian (2015-01w1 - 2015-03w5)



5.2.6.b New Zealand Herald (2015-01w1 - 2015-03w5)



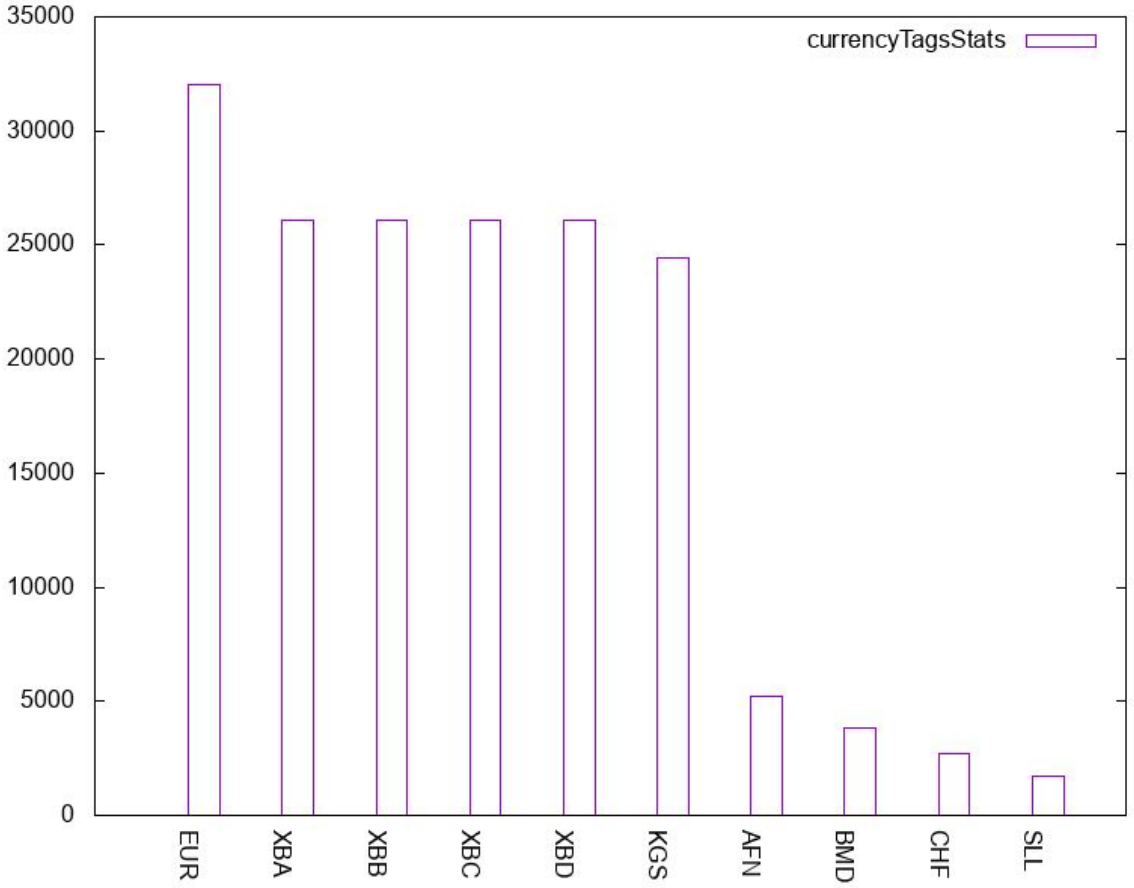
5.2.6.c The New York Times (2015-01w1 - 2015-03w5)



5.3 Wyniki analizy walutowej

Sprawdzono, które tagi oznaczające waluty występują w notkach najczęściej. Wyniki dla 10 najpopularniejszych tagów przedstawiono poniżej.

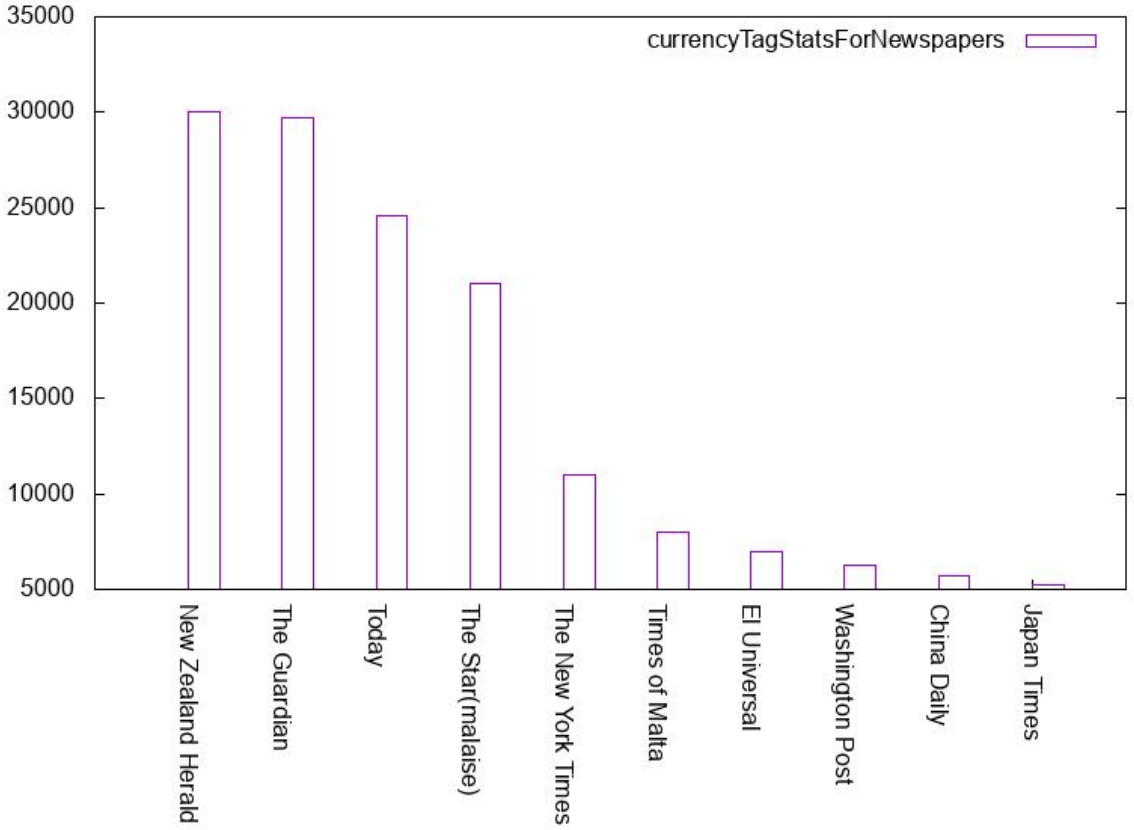
5.3.1 częstotliwość występowanie poszczeglnych tagów walutowych



Najczęściej opisywaną walutą jest euro, zaraz za nim plasują się jednostki rynków obligacji. Popularność europejskiej waluty może być spowodowana jej istotną rolą na świecie: jest używana przez największą liczbę krajów na świecie. Warto zauważyć również, że wśród 10 najpopularniejszych walut znajduje się frank szwajcarski, brakuje natomiast dolara amerykańskiego. Ten fakt mógłby być przedmiotem dalszej analizy w przypadku rozbudowywania projektu.

Sprawdzono, które czasopisma najczęściej piszą nt. walut. Poniżej diagram przedstawiający 10 najczęściej piszących na ten temat gazet.

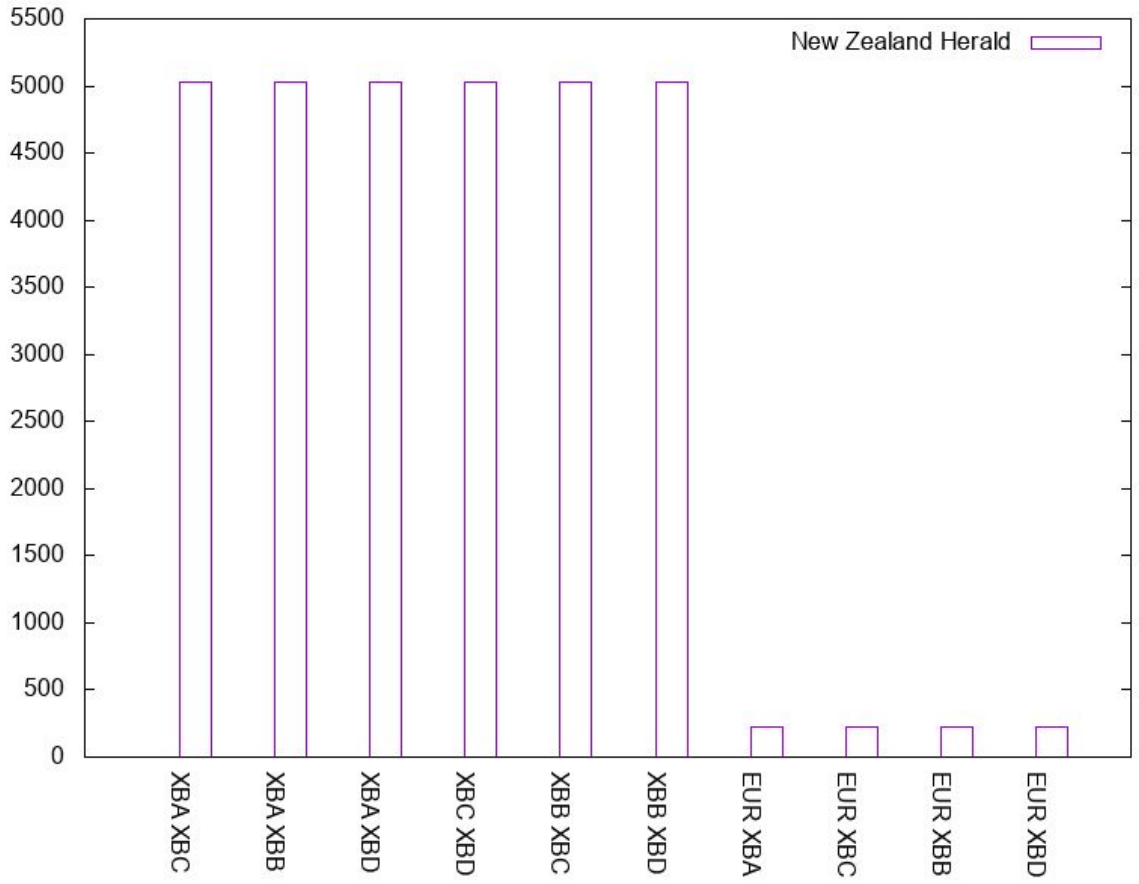
5.3.2 Liczba tagów walutowych dla poszczególnych gazet



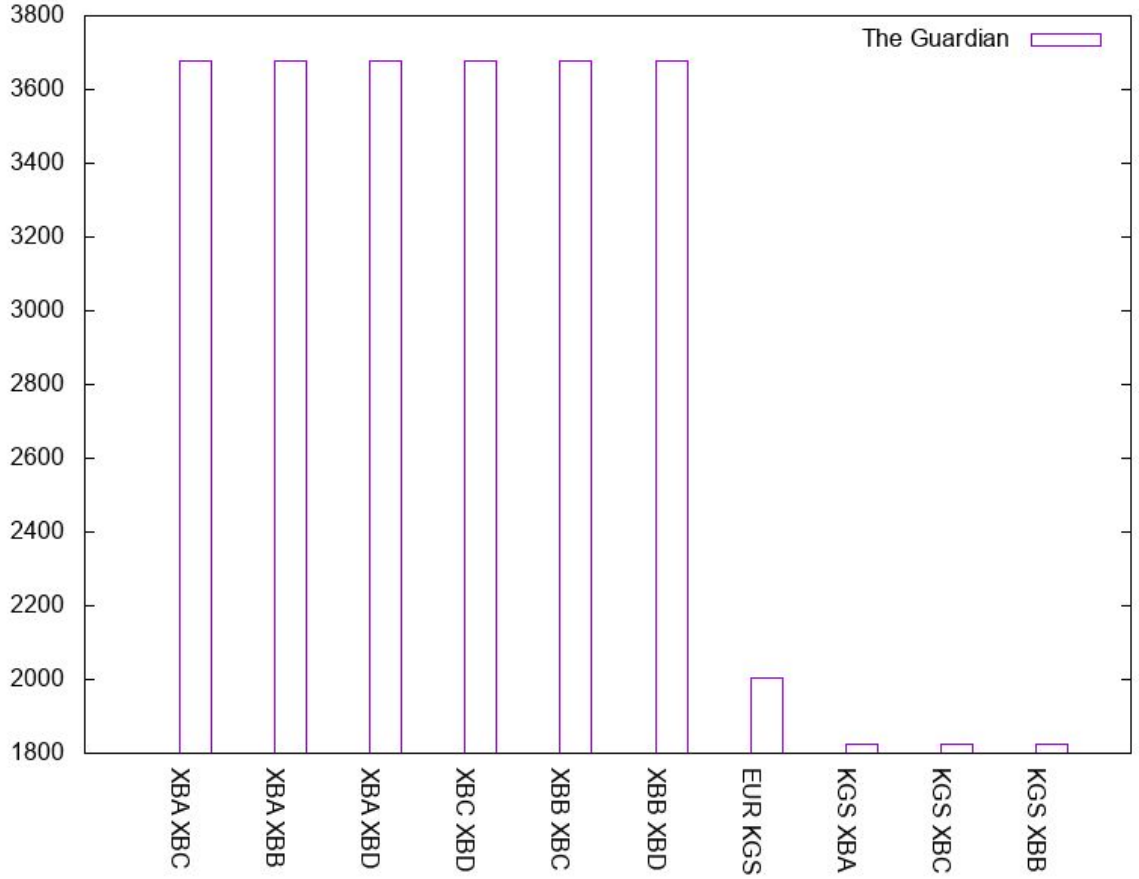
Najwięcej o walutach piszą gazety z Nowej Zelandii oraz Wielkiej Brytanii. Zauważalnym faktem jest duża rozbieżność między tytułami z tych krajów, a gazetami z tak wielkiego państwa, jak Stany Zjednoczone.

Przeprowadzono również analizę dla częstotliwości występowania par tagów dla poszczególnych gazet. Poniżej wyniki dla najczęściej występujących par dla dwóch największych gazet.

5.3.3 Najczęściej występujące pary tagów dla gazety *New Zealand Herald*



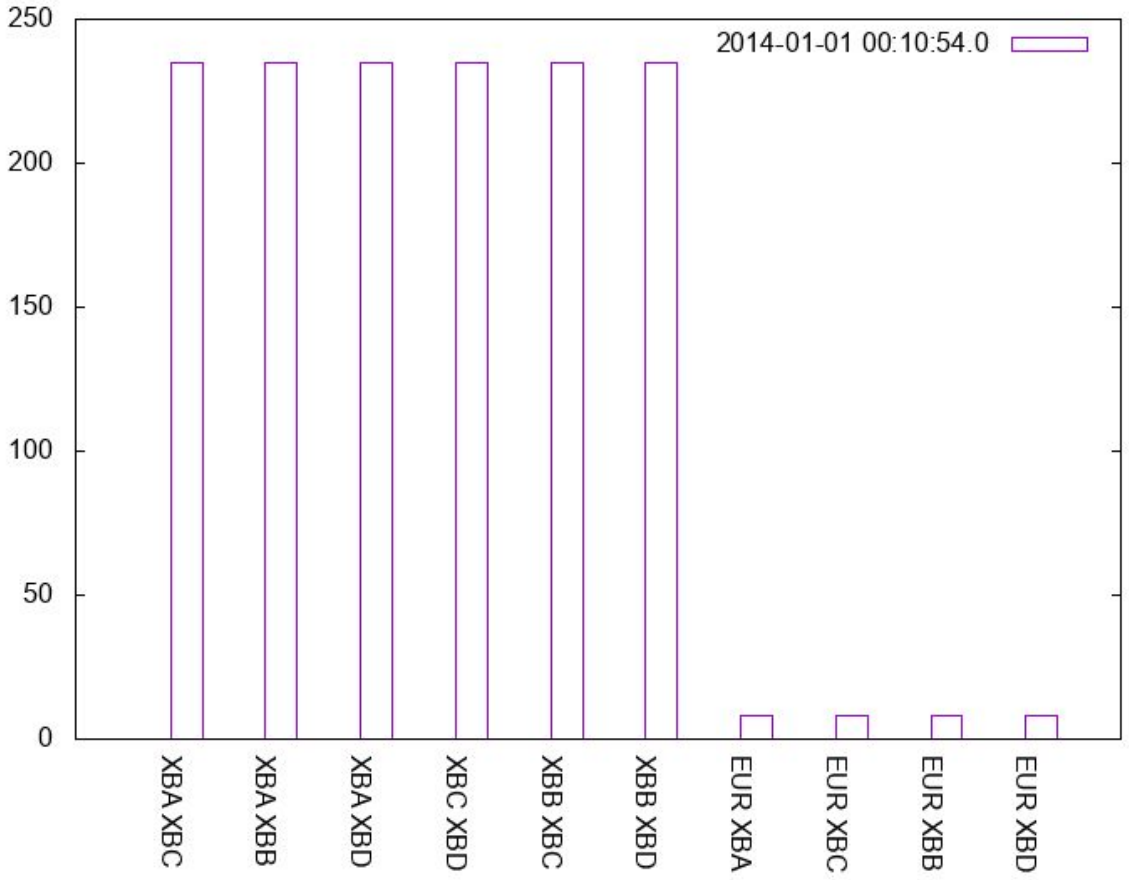
5.3.4 Najczęściej występujące pary tagów dla gazety *The Guardian*



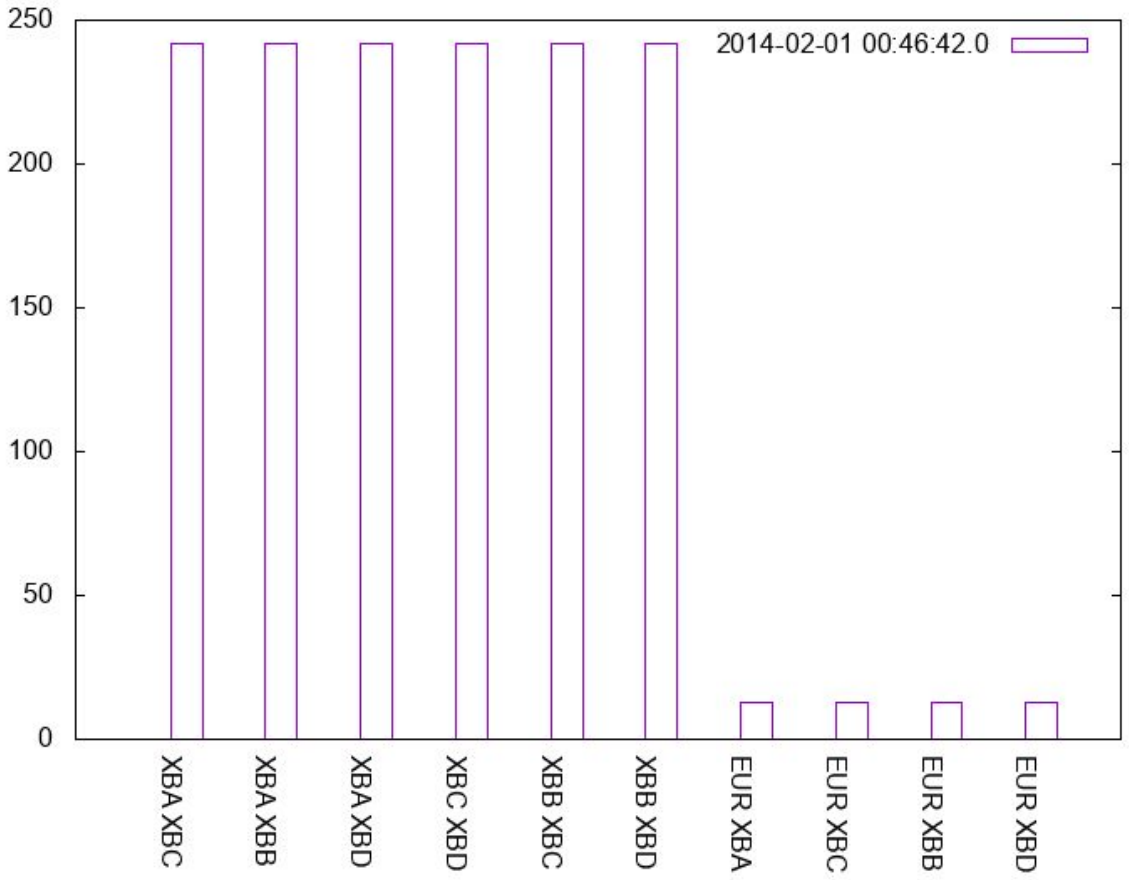
Zdecydowanie w notkach prasowych najczęściej występują informacje nt. rynków obligacji. W kolejności następne są informacje nt. euro oraz rynków obligacji, przy czym ww. par jest kilka razy mniej. Gdyby połączyć ten fakt z największą popularnością waluty euro, można by dojść do ciekawych wniosków: chociaż czasopisma często piszą o euro, rzadko porównują ją z innymi walutami. Prawdopodobną przyczyną tej rozbieżności jest duże prawdopodobieństwo występowania wszystkich jednostek obligacji w jednej notce dotyczącej rynku obligacji, co znacząco podbija statystyki tych par.

Poniżej przedstawiono występowanie najczęstszych par tagów walutowych dla gazety *New Zealand Herald* na przestrzeni 5 miesięcy.

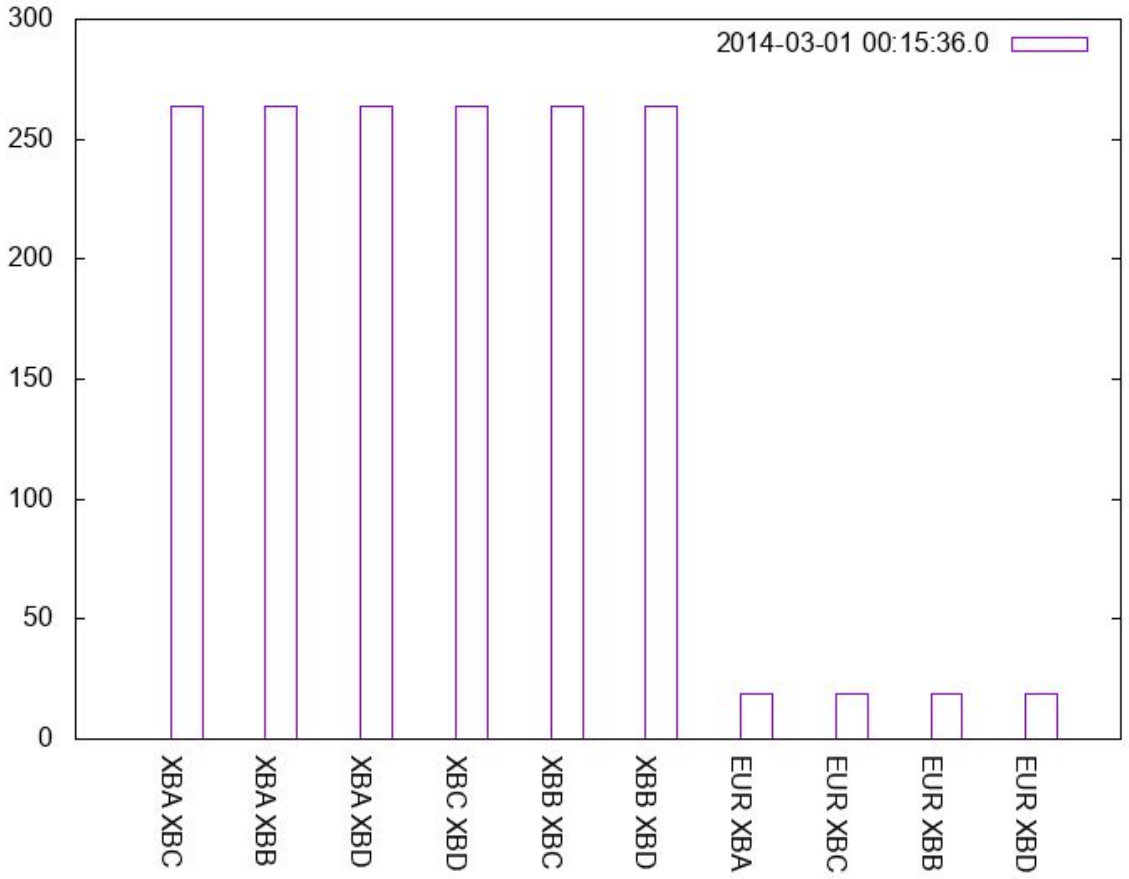
5.3.5 styczeń 2014



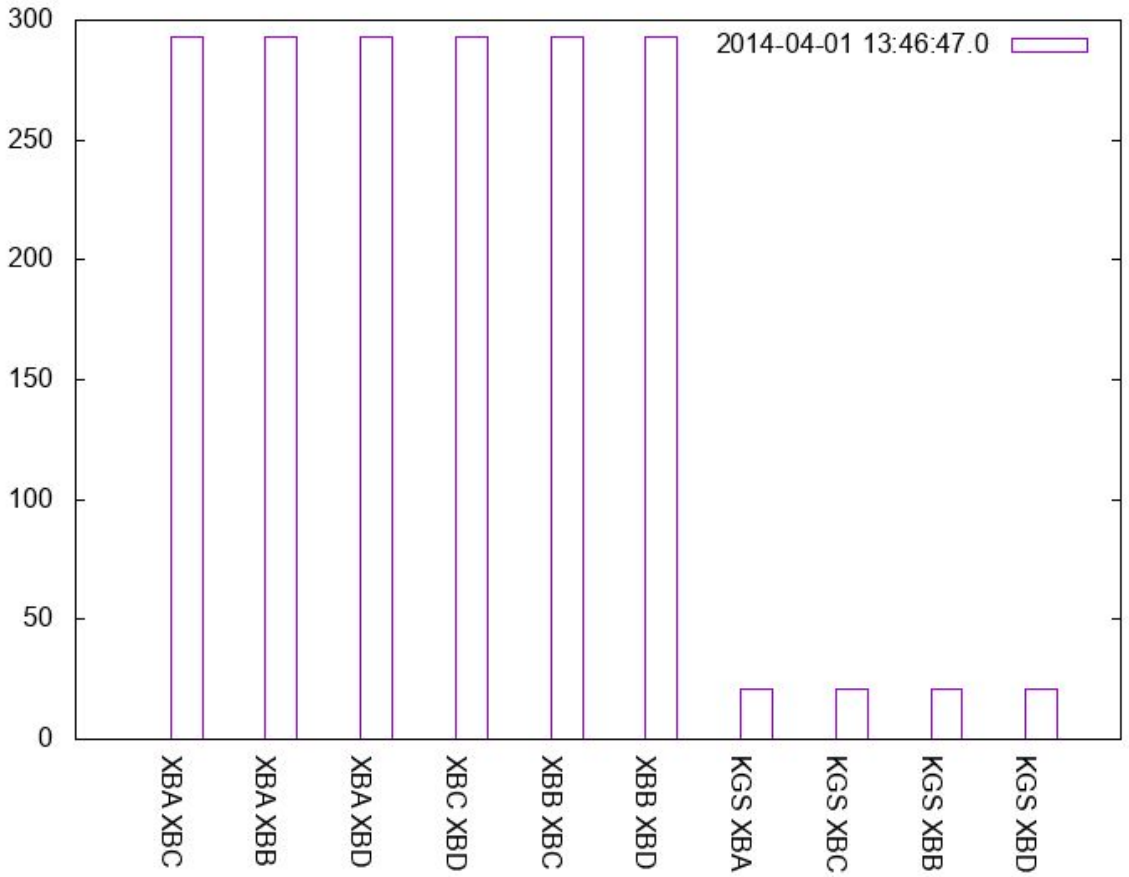
5.3.6 luty 2014



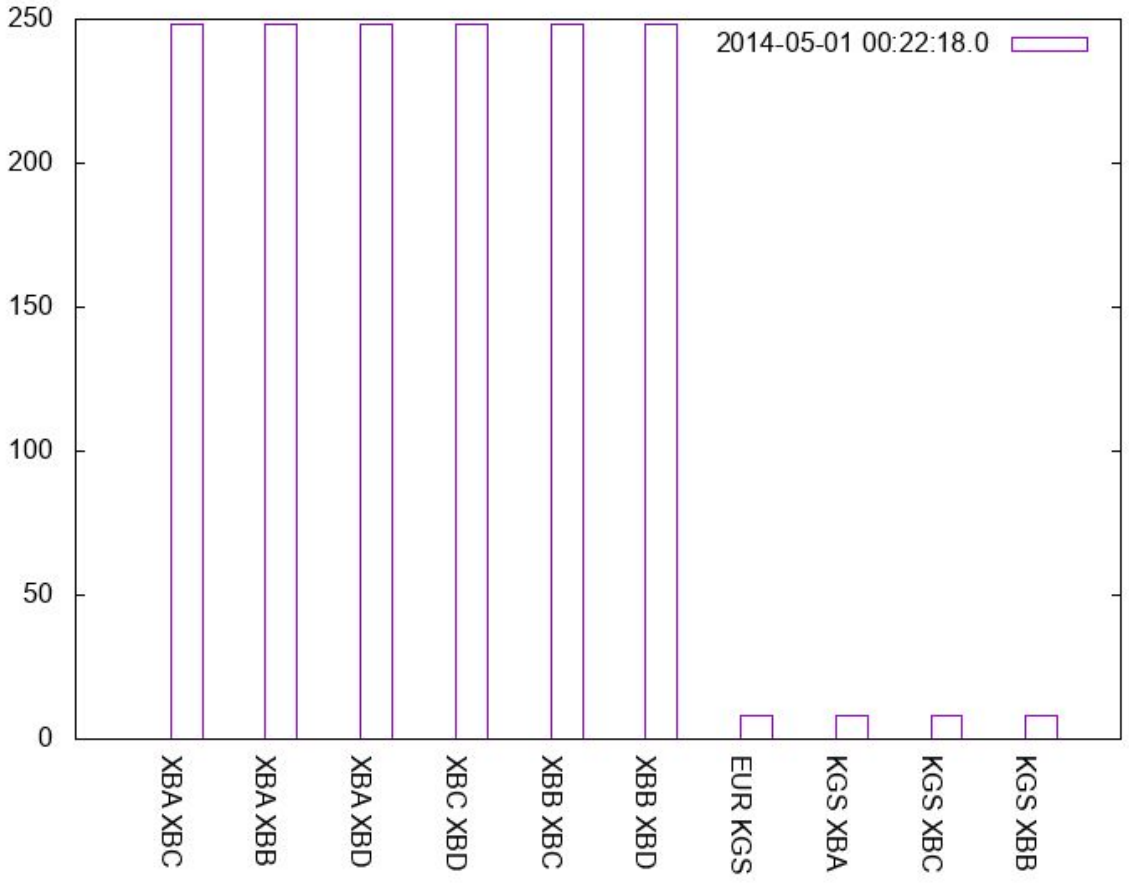
5.3.7 marzec 2014



5.3.8 kwiecień 2014



5.3.9 maj 2014



Widać, że choć wyniki w kolejnych miesiącach trochę się różnią, ogólna tendencja jest zachowana. Pokrywa się ona z ogólną analizą par tagów – liczba tych dotyczących obligacji zdecydowanie przewyższa wszystkie pozostałe.

6. Bibliografia

[1] Gephi toolkit - https://gephi.org/toolkit/

[2] xChart - http://knowm.org/open-source/xchart/

[3] iText - http://itextpdf.com/

[4] opencsv - http://opencsv.sourceforge.net/

Miary do analizy sieci społecznych:

[5] Average Clustering Coefficient
https://github.com/gephi/gephi/wiki/Average-Clustering-Coefficient

[6] Connected Components
https://github.com/gephi/gephi/wiki/Connected-Components

[7] Eigenvector Centrality
https://github.com/gephi/gephi/wiki/Eigenvector-Centrality

[8] HITS https://github.com/gephi/gephi/wiki/HITS

[9] Density https://github.com/gephi/gephi/wiki/Graph-Density

[10] Betweenness Centrality
https://github.com/gephi/gephi/wiki/Betweenness-Centrality

[11] Closeness Centrality
https://github.com/gephi/gephi/wiki/Closeness-Centrality

[12] Diameter
https://github.com/gephi/gephi/wiki/Diameter

[13] Modularity https://github.com/gephi/gephi/wiki/Modularity

[14] PageRank https://github.com/gephi/gephi/wiki/PageRank

[15] Prestige Plugin https://gephi.org/plugins/#/plugin/prestige-plugin

[16] draw.io https://www.draw.io/

[17] vertabelo https://www.vertabelo.com/