Rok akademicki 2014/2015

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Automatyki i Informatyki Stosowanej

Praca dyplomowa inżynierska

Analizator sieci społecznościowej w aspekcie kwalifikacji zawodowych

Monika Pawluczuk

Opiekun pracy
dr inż. Mariusz Kamola

Ocena: .....
.....
Podpis Przewodniczącego Komisji Egzaminu Dyplomowego

Specjalność: Systemy Informacyjno-Decyzyjne
Data urodzenia: 26.08.1992
Data rozpoczęcia studiów: październik 2011

Życiorys

Nazywam się Monika Pawluczuk i 26 sierpnia 1992 r. urodziłam się w Lublinie. W 2005 roku ukończyłam Szkołę Podstawową nr 4 w Lublinie, w 2008 roku Gimnazjum im. św. Stanisława Kostki w Lublinie. W latach 2008-2011 uczęszczałam do Liceum Ogólnokształcącego im. Stanisława Staszica w Lublinie.
Studia na Wydziale Elektroniki i Technik Informacyjnych rozpoczęłam w październiku 2011 roku. W pracy zawodowej miałam do czynienia z analizą i wizualizacją danych różnego rodzaju dla celów naukowych. Poza informatyką w wolnych chwilach interesuję się czytaniem książek, tańcem oraz fitnessem.

……
Podpis studenta

Egzamin dyplomowy

Złożyła egzamin dyplomowy w dniu ……20__ r z
z wynikiem ……
Ogólny wynik studiów: ……
Dodatkowe uwagi i wnioski komisji: …….
……
……

Streszczenie

Głównym celem niniejszej pracy jest:

- Wyłonienie kategorii zawodowych na podstawie danych publicznie udostępnionych na profilach serwisu GoldenLine
- Zamodelowanie sieci społecznej na podstawie informacji o kontaktach na profilach
- Stworzenie aplikacji działającej w chmurze pozwalającej, na podstawie uprzednio zgromadzonych przeze mnie informacji, na wyszukiwanie profili:
    - osób posiadających zadane kwalifikacje zawodowe,
    - charakteryzujące się wybranym parametrem sieci społecznej.

Motywacją dla jej stworzenia jest rozwiązanie problemu braku narzędzi specjalizujących się w wyszukiwaniu potencjalnych pracowników o zadanych kwalifikacjach zawodowych.

Praca rozpoczyna się od wprowadzenia do tematyki związanej z tworzeniem kategoryzacji oraz sieci społecznościowej, a także omówienia technik pozwalających na zebranie potrzebnych do analizy danych.

W głównej części pracy zostały wyjaśnione metody analizy danych, dzięki którym wyłoniona została dwupoziomowa kategoryzacja polskiego rynku pracy - zawody oraz ich specjalności. Ponadto opisano sposoby tworzenia sieci społecznościowej, metody liczenia parametrów dla jej wierzchołków oraz znaczenie tych parametrów. Końcowym efektem tej analizy jest omówiona w następnej kolejności aplikacja działająca w chmurze.

Całość zakończona jest opisem zastosowanej architektury (zastosowanych języków programowania, baz danych oraz użytej platformy chmurowej) oraz proponowanymi metodami weryfikacji poprawności stworzonej kategoryzacji.

W podsumowaniu dodatkowo zostały przedstawione dalsze możliwości rozwoju opracowanego projektu.

Abstract

The main purposes of this thesis are:

- Extraction of occupations' categories based on public data shared on Golden Line user profiles
- Modeling of social network based on contact information on user profiles
- Creation of cloud application that allows, based on previously gathered information, to search for profiles with:
  - required professional qualifications
  - required value of social network parameter

The motivation of creating this thesis is to solve the problem of lack of tools that specialize in finding potential employees with required professional qualifications.

At the beginning of thesis there is an introduction to topics related to categorization and social networks, as well as techniques used to gather data required for analysis.

The main part contains explanation of methods used to analyze data, that led to two-level categorization of polish job market - professions and its specializations. Furthermore, there is description of who to create a social network, calculate its parameters for nodes and what does this parameters indicate. The end result of this analysis is cloud application, described in the next section.

Thesis ends with the specification of used architecture (programming languages, database and cloud platform) and proposed methods verifying created categorization.

Additionally, summary contains further development possibilities of this project.

Spis treści

1. Wprowadzenie
   - 1.1. Motywacja i cel pracy
   - 1.2. Słownik pojęć i skrótów
2. Przegląd dziedziny wiedzy i rozwiązań
   - 2.1. Omówienie technik scrapingu
   - 2.2. Omówienie technik analizy sieci społecznościowej
     - 2.2.1. Macierz sąsiedztwa
     - 2.2.2. Parametry sieci społecznościowej
   - 2.3. Kategoryzacja
   - 2.4. Przegląd istniejących rozwiązań
     - 2.4.1. Wyszukiwanie po słowach kluczowych
     - 2.4.2. Ręczne przeglądanie profili
3. Projekt rozwiązania
   - 3.1. Zebranie publicznych profili serwisu
     - 3.1.1. Stworzenie mapy profili
     - 3.1.2. Zebranie danych z profili z wybranych sekcji
     - 3.1.3. Zebranie informacji o kontaktach
   - 3.2. Analiza danych z profili
     - 3.2.1. Analiza najczęściej występujących słów
     - 3.2.2. Analiza najczęściej występujących par słów
     - 3.2.3. Porównanie wyników z polskim słownikiem frekwencyjnym
     - 3.2.4. Stworzenie dwupoziomowej kategoryzacji zawodów
   - 3.3. Analiza sieci społecznościowej
     - 3.3.1. Stworzenie sieci społecznościowej z informacji o kontaktach
     - 3.3.2. Stworzenie podsieci dla profili należących do jednej kategorii zawodowej
     - 3.3.3. Analiza parametrów dla wierzchołków danej sieci
   - 3.4. Aplikacja działająca w chmurze
     - 3.4.1. Przeglądanie kategorii zawodowych i ich specjalności
     - 3.4.2. Wyszukiwanie profili z danych kategorii
     - 3.4.3. Filtracja profili dla zadanego parametru sieci
4. Opis architektury i wybranych elementów implementacji
   - 4.1. Skrypty node.js
   - 4.2. Bazy danych sqlite oraz MongoDB
   - 4.3. Platforma chmurowa Heroku
5. Opis metody i wyników weryfikacji opracowanego rozwiązania

6. Podsumowanie
    6.1. Zalety i wady rozwiązania
    6.2. Możliwości dalszego rozwoju
7. Bibliografia

1. Wprowadzenie

Tematem niniejszej pracy jest stworzenie dwupoziomowej kategoryzacji zawodowej dla polskiego rynku na podstawie zawartości dostępnych publicznie profili serwisu GoldenLine oraz stworzenie sieci społecznościowej na podstawie kontaktów serwisu GoldenLine i wyliczenie odpowiednich parametrów dla jej wierzchołków. W oparciu o te analizy, powstaje serwis pozwalający na wyszukiwanie kandydatów do pracy o zadanych kwalifikacjach.

Rozdział 2. przybliża nam techniki scrapingu w kontekście tego projektu, czyli automatyzacji zbierania danych z profili GoldenLine oraz podstaw teoretycznych dotyczących sieci społecznościowych - ich konstrukcją, zastosowaniem i miarami. Dodatkowo, pokazane zostały dotychczasowe dostępne rozwiązania pozwalające na wyszukiwanie i przeglądanie publicznych profili serwisu.

Rozdział 3. opisuje projekt rozwiązania. Przedstawia, jakie metody były stosowane, aby zebrać dane z profili oraz jak następnie te dane były przechowywane i przetwarzane. Następnie omawia rodzaje metod analizy danych, które miałyby wyłonić kategorie zawodowe i jakie były wyniki ich stosowania. Ostatni podrozdział przedstawia jak działa aplikacja, która wykorzystuje poprzednio zebrane i zanalizowane dane, aby pomagać znaleźć potencjalnych kandydatów do pracy.

Rozdział 4. przedstawia projekt od strony technicznej - kluczowe elementy architektury i implementacji, jak wykorzystane języki programowania, sposób przechowywania danych w bazach danych oraz krótki opis platformy pozwalającej na umieszczenie aplikacji w chmurze.

Rozdział 5. prezentuje metody pozwalające na weryfikację zastosowanych analiz oraz ich wyniki.

Rozdział 6. krótko podsumowuje całą pracę, wady i zalety rozwiązania oraz jego możliwości dalszego rozwoju.

Rozdział 7. to spis wszystkich źródeł na podstawie których niniejsza praca powstała. Są to zarówno źródła potrzebne do napisania pracy pisemnej, jak i te kluczowe do stworzenia projektu rozwiązania.

1.1 Motywacja i cel pracy

W dzisiejszych czasach dynamicznie rozwijają się firmy specjalizujące się w branży zarządzania zasobami ludzkimi (ang. *human resources, HR*). Dużą częścią ich pracy jest nabór pracowników, to znaczy przygotowanie puli kandydatów do pracy, posiadających odpowiednie kwalifikacje na zadane stanowisko.

Aby uprościć zadanie wyszukiwania takich osób, powstają serwisy społecznościowe, specjalizujące się w kontaktach zawodowo-biznesowych. Przykładem ogólnoświatowym takiego serwisu może być LinkedIn, mający ponad 300 milionów[1] użytkowników, natomiast jego polskim odpowiednikiem jest GoldenLine. Polski serwis ma prawie 2 miliony[2] użytkowników, z czego w przybliżeniu półtora miliona profili stanowią profile publicznie dostępne, niewymagające zalogowania.

Wadą polskiego serwisu jest jednak fakt, że nie kategoryzuje profili w sposób publicznie dostępny. Ponadto nie ma również zewnętrznych narzędzi które by na to pozwalały - wobec czego nie możemy łatwo znaleźć interesujących profili osób należących do wybranej branży.

Głównym zadaniem mojej pracy jest więc wyłonienie kategorii zawodowych i ich specjalizacji, poprzez analizę danych publicznie dostępnych na profilach serwisu GoldenLine. Jest to realizowane poprzez tworzenie skryptów, które są odpowiedzialne za zebranie listy profili publicznie dostępnych, danych z tych profili oraz informacji o kontaktach użytkownika. Dodatkowo, analizuję aspekty sieciowe, które mogą przyczynić się do znalezienia potencjalnych odbiorców ofert i reklam skierowanych do specjalistów różnych dziedzin.

Na podstawie wykonanej analizy, stworzona jest aplikacja działająca w chmurze, pozwalająca na wyszukiwanie w serwisie GoldenLine profili osób posiadających zadane kwalifikacje zawodowe i charakteryzujących się wybranym parametrem sieci społecznej.

[1] Źródło (13.08.2015): http://www.statista.com/statistics/274050/quarterly-numbers-of-linkedin-members/

[2] Źródło (11.02.2015): http://media.goldenline.pl/goldenline-pl-najwiekszym-serwisem-rekrutacyjnym-w-polsce/

1.2 Słownik pojęć

Crawler - robot internetowy, zbierający informacje o treściach znajdujących się w Internecie w różnych celach

HR - ang. *Human Resources*, czyli zasoby ludzkie. Firmy lub działy firm zajmujące się rekrutacją i zarządzaniem pracownikami

Korpus - zbiór tekstów służący badaniom lingwistycznym, np. określaniu częstości występowania form wyrazowych, konstrukcji składniowych, kontekstów w jakich pojawiają się dane wyrazy

Macierz sąsiedztwa - kwadratowa macierz grafu, w której wartość w kolumnie i oraz wierszu j oznacza liczbę krawędzi pomiędzy wierzchołkami i oraz j

Parser - analizator składniowy, przeznaczony najczęściej do analizy języków programowania

Scraping - technika komputerowa polegająca na zdobywaniu informacji

Sieć społecznościowa - struktura teoretyczna, przedstawiona za pomocą grafu, w którym wierzchołkami są ludzie, a krawędzie oznaczają dowolną zadaną relację pomiędzy osobami

Stemmer - narzędzie, którego zadaniem jest wydobycie z wybranego wyrazu tzw. rdzenia, czyli części, która jest odporna na odmiany przez przyimki, rodzaje itp.

2. Przegląd dziedziny wiedzy i rozwiązań

2.1. Omówienie technik scrapingu

Scraping (w kontekście pracy *web scraping*) jest techniką komputerową, która polega w ogólności na zbieraniu danych z serwisów internetowych. Zbieranie danych, w zależności od ich ilości i struktury może się odbywać w różny sposób.

Czasami najlepszym rozwiązaniem jest ręczne kopiowanie i wklejanie w miejsce gdzie dane będziemy przechowywać. Jest to skuteczne, kiedy danych jest mało lub dane nie są w żaden sposób ustrukturyzowane, co sprawia że zautomatyzowanie tego działania jest prawie niemożliwe.

Pozostałe metody są już zautomatyzowane, to znaczy po ustaleniu jakie dane mają zostać zgromadzone, nie jest już potrzebna interakcja użytkownika. Jedną z takich metod jest dopasowywanie tekstu do wyrażeń regularnych - co jest bardzo wygodne w użyciu, gdyż każdy z języków programowania pozwala na przetwarzanie wyrażeń regularnych, a więc dobór narzędzi nie jest w żaden sposób ograniczeniem.

Metodą, z której skorzystałam jest tak zwany HTML *parsing*. Język HTML charakteryzuje się drzewiastą strukturą, a więc można przechodzić w jego hierarchii w różne strony, używając właśnie parserów, czyli analizatorów składniowych.

Często strony posiadają wspólny szablon według którego są dynamicznie tworzone, wypełniając odpowiednie sekcje danymi na przykład z bazy danych. Posiadając pełny kod HTML analizowanej strony internetowej, znając jej szablon i korzystając z parsera HTML, możemy z łatwością nawigować po dokumencie, wydobywając potrzebne nam dane.

2.2. Omówienie technik analizy sieci społecznościowej

2.2.1. SIEĆ SPOŁECZNOŚCIOWA

Sieć społecznościowa jest pojęciem interdyscyplinarnym, wykorzystywanym w dziedzinach psychologii społecznej, socjologii, statystyki i teorii grafów. Jest to teoretyczna, abstrakcyjna struktura społeczna złożona z węzłów, którymi są społeczni aktorzy (mogą to być osoby indywidualne bądź organizacje). Węzły są ze sobą powiązane w taki sposób, aby odzwierciedlać dowolne, różne relacje między węzłami - pokrewieństwo, hierarchię, znajomość czy wspólne zainteresowania.

W mojej pracy sieć społecznościową będę traktować jako graf nieskierowany, w którym węzłami są użytkownicy serwisu GoldenLine, a połączenia między nimi są odzwierciedleniem informacji o kontaktach (znajomych) w serwisie.

2.2.2. MACIERZ SĄSIEDZTWA

Macierz sąsiedztwa to pojęcie związane z teorią grafów. Chcąc zapisać strukturę grafu (zarówno węzły, jak i relacje między nimi), aby móc go później odtworzyć, należy stworzyć macierz sąsiedztwa.

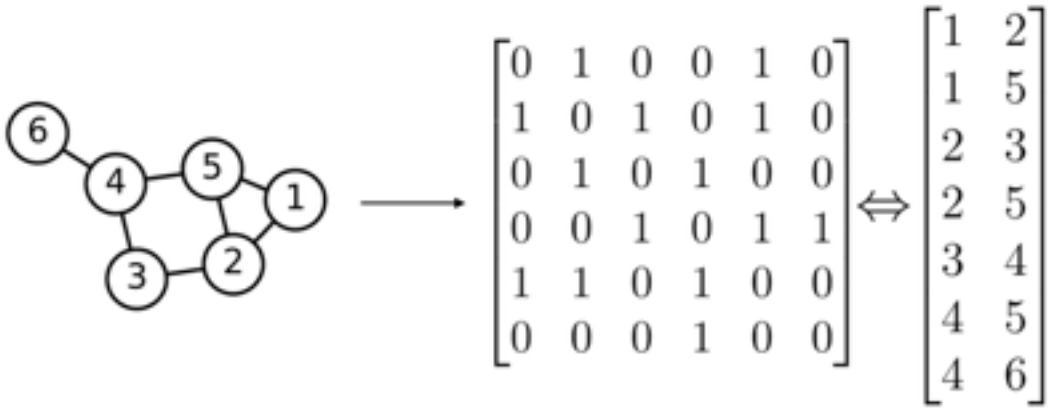
Jest to kwadratowa macierz, której kolumnami i wierszami są kolejne wierzchołki grafu. Wartość dla danej kolumny i wiersza oznacza ilość krawędzi pomiędzy odpowiadającymi im wierzchołkami. Gdy graf jest nieskierowany, tzn. relacja między dwoma wierzchołkami jest zawsze obustronna, macierz będzie symetryczna.

W moim przypadku, graf sieci jest nieskierowany, ponieważ jako relację przyjmuję znajomość dwóch osób w serwisie, a więc jest to zawsze relacja dwustronna. Profili, z których dane zbieram jest ponad milion - a więc istotną wartością jest tu oszczędne przechowywanie danych. Tworzenie tabeli o wymiarach milion na milion przekroczyłoby możliwości standardowych baz danych z których korzystam.

Dlatego też taką macierz zapisuję w inny sposób. Właściwości, które pozwalają mi na zredukowanie ilości trzymanych danych bez utraty informacji to:

- dwustronność relacji, tzn. graf jest nieskierowany,
- standardowa macierz sąsiedztwa byłaby bardzo rzadka (byłaby w większości wypełniona zerami)
- jedyne możliwe wartości to 0 lub 1,
- dane o wierzchołkach są trzymane w innym miejscu, aby powiązać je z danymi dotyczącymi samego profilu.

Dzięki tym właściwościom, macierz sąsiedztwa zapisuję w dwukolumnowej tabeli, gdzie każdy wiersz oznacza unikalne połączenie pomiędzy dwoma wierzchołkami.



Rys.2.1. Schemat przedstawiający sposób zapisywania informacji o strukturze grafu w macierzy sąsiedztwa i moje przekształcenie.

2.2.3. PARAMETRY SIECI SPOŁECZNOŚCIOWEJ

Posiadając strukturę sieci społecznościowej, można badać jej właściwości dotyczące wierzchołków lub krawędzi i na tej podstawie, w zależności od tego jakie relacje są odzwierciedlone, odkrywać nowe wnioski.

Centralność stopnia (ang. *degree centrality*) - w istocie jest stopień wierzchołka w grafie, tzn. liczba krawędzi które dany wierzchołek posiada. Może być interpretowana na przykład jako miara popularności danego wierzchołka.

Centralność bliskości (ang. *closeness centrality*) - miara ta dla wierzchołka w grafie jest odwrotnością średniej odległości najkrótszej ścieżki z tego wierzchołka do wszystkich pozostałych w grafie. Może być interpretowana jako miara wydajności danego wierzchołka w rozprzestrzenianiu informacji. Im większa centralność bliskości dla wierzchołka, tym krótsza jest średnia odległość z tego wierzchołka do wszystkich pozostałych, co oznacza lepszą pozycję wierzchołka w rozprzestrzenianiu informacji do pozostałych.

Betweenness centrality - miara ta dla wierzchołka w grafie jest frakcją najkrótszych ścieżek pomiędzy wszystkimi parami węzłów w grafie, które przechodzą przez ten węzeł. Może być interpretowana, w pewnym sensie, jako miara wpływu wierzchołka na rozprzestrzenianie się informacji w sieci przyjmując, że do przepływu są wykorzystywane najkrótsze ścieżki.

2.3. Kategoryzacja

Celem analizy tekstów z publicznych profili serwisu GoldenLine jest stworzenie kategoryzacji zawodów na polskim rynku.

Kategoryzacja ta jest dwupoziomowa: pierwszy poziom stanowi istniejące na polskim rynku zawody, drugi to specjalizacje zawodów z poziomu pierwszego. Kategoryzacja ta powinna pozwolić na przyporządkowanie jak największej liczby profili do zadanych kategorii i specjalizacji. Stuprocentowe przyporządkowanie nigdy nie będzie możliwe, ze względu na brak pełnych publicznie dostępnych danych na profilu (użytkownik może publicznie udostępnić jedynie wybrane sekcje) lub profile puste, bez żadnych informacji.

Kategoryzacja profili w serwisie GoldenLine miałaby działać w sposób analogiczny do popularnych serwisów aukcyjnych. To znaczy, możemy odgórnie wybrać kategorię z dostępnego zestawu kategorii lub ich specjalizacji, i dostać profile które do niej należą. Różnicą jest tutaj ważna właściwość, że jeden profil może należeć do więcej niż jednej kategorii lub specjalizacji - zakładamy, że osoba może być wykształcona w więcej niż jednym kierunku, i często jest to bardzo pożądane wśród pracodawców.

2.4. Przegląd istniejących rozwiązań

2.4.1. WYSZUKIWANIE PO SŁOWACH KLUCZOWYCH

Serwis udostępnia wyszukiwarkę pełnotekstową (ang. *full-text search*). Działanie takiej wyszukiwarki opiera się na tym, czy słowa z szukanej frazy występują w przeszukiwanym dokumencie - jeśli tak, to będzie on wyświetlony jako rezultat.

Niestety, takie wyszukiwanie nie będzie wygodne w dwóch przypadkach:

- gdy nie mamy sprecyzowanej szukanej grupy zawodowej - na przykład, gdy chcemy najpierw określić, które grupy mogą być przydatne lub lukratywne ,
- wyszukiwanie pełnotekstowe może dawać nieistotne wyniki wyszukiwania. POPRAWIĆ

DODAĆ WIĘCEJ

2.4.2. RĘCZNE PRZEGLĄDANIE PROFILI

Serwis udostępnia również mapę profili publicznie dostępnych. Jednak w przypadku gdy nie szukamy konkretnej osoby, której dane osobowe znamy, jest ono właściwie nieprzydatne ze względu na zbyt dużą ilość profili.

DODAĆ WIĘCEJ

3. Projekt rozwiązania

3.1. Zebranie publicznych profili serwisu

3.1.1. STWORZENIE MAPY PROFILI

Mapa profili to lista adresów www do wszystkich dostępnych publicznie profili, a więc takich z których chcemy pobrać dane. Serwis GoldenLine udostępnia taką mapę (pod adresem: http://www.goldenline.pl/profile/mapa/a), tzn. publikuje wszystkie linki publicznych profili, segregując je pod względem liter na które się zaczynają.

Dla każdej litery z alfabetu (A-Z) określiłam liczbę podstron, na której umieszczone są linki, poprzez parsowanie strony. Mając określoną literę alfabetu i podstronę, można wygenerować link do danej podstrony z profilami, tzn:
http://www.goldenline.pl/profile/mapa/**litera_alfabetu**/s/**podstrona**.

Na każdej podstronie bezpośrednie linki do profili są umieszczone w postaci nieuporządkowanych list, które pobrałam i zapisałam w bazie danych.

3.1.2. ZEBRANIE DANYCH Z PROFILI Z WYBRANYCH SEKCJI

Mając już poprzednio przygotowaną listę linków do profili, mogłam zebrać dane z profili. Podstawowe dane: nazwa i ID profilu możemy zawsze pobrać, ponieważ są umieszczone zawsze w tym samym miejscu:

```
<article>
        <header>
                <section class="basicInfo" data-id="264254"
                        data-urlname="jakub-kowalski6">
                        <a href="https://..." class="fancybox thumb-150">
                <img src="https://..." alt="Jakub Kowalski"/>
                </a>
                <section>
                <div class="nameSurname">
                        <h1>Jakub Kowalski</h1>
                </div>
 /*
...
*/
</article>
```

Listing 1. Fragment kodu HTML publicznego profilu GoldenLine, gdzie mamy informacje o nazwie i ID użytkownika

Profile w serwisie mogą mieć udostępnioną różną zawartość - niektórzy użytkownicy nie udostępniają żadnych informacji publicznie, niektórzy wszystkie dostępne sekcje (plus na przykład swoje własne) i kontakty. Jeśli jednak jakaś sekcja została udostępniona przez użytkownika, to zawsze jest ona generowana w identyczny sposób.

W mojej pracy pobieram, jeśli są udostępnione, z profili cztery sekcje: Edukacja, Wykształcenie, Podsumowanie oraz Tagi.

3.1.3. ZEBRANIE INFORMACJI O KONTAKTACH

Zbieranie informacji o kontaktach działa na podobnej zasadzie jak przy mapie profili. Po wczytaniu liczby podstron kontaktów, mając adres profilu oraz podstronę, można wygenerować link do danej podstrony z kontaktami, tzn:
http://www.goldenline.pl/**nazwa_profilu**/kontakty/s/**podstrona**.

Na każdej podstronie, podobnie jak wcześniej bezpośrednie linki do profili są umieszczone w postaci nieuporządkowanych list, które pobrałam i zapisałam w bazie danych.

Różnica polega na tym, że potrzebujemy z kontaktu wydobyć informacje na temat ID użytkownika, tak aby móc stworzyć listę sąsiedztwa. O ile profil został wcześniej przetworzony, informacja o identyfikatorze będzie już w bazie danych. Jednak w przypadku, gdy:

- profil nie został przetworzony przez poprzednie skrypty: nie wszystkie profile zostały z mapy profili zostały przetworzone lub profil został utworzony po pobraniu mapy profili (brak aktualizacji),
- profil jest niepubliczny.

zamiast identyfikatora zachowujemy link do profilu, tak aby nie stracić informacji o danym kontakcie i móc wiernie odtworzyć sieć społecznościową.

Jeżeli znajomy został już wcześniej przetworzony, to informacji o kontakcie nie dodajemy, aby nie duplikować połączeń (relacja znajomości jest zawsze obustronna, więc jeśli mamy w bazie parę kontaktów (a,b) to wiemy, że istnieje również kontakt (b,a)).

3.1.4. PRZYJAZNY WEBSCRAPING

Opisane w poprzednich podpunktach skrypty opierają się na masowym pobieraniu danych z internetu. Aby nie przeciążać serwerów wystawiających aplikację oraz uniknąć wykrycia przez systemy anty-scrapingowe, co może skutować znalezieniem się na czarnej liście adresów IP, zostały zastosowane następujące metody:

- powolne pobieranie danych - nielimitowanie połączeń wysyłanych do serwera ze skryptów byłoby przykładem przeciążania serwerów poprzez atak DoS[3], dlatego wysyłane zapytania do serwera przy pobieraniu dowolnej strony są poprzedzane losową przerwą o czasie z zakresu od 1 do 10 sekund (losowość zmniejsza prawdopodobieństwo wykrycia zautomatyzowanego procesu),
- zmiana adresów IP, z których wysyłane są zapytania poprzez korzystanie z różnych sieci oraz korzystanie z serwerów Proxy,
- zmiana User-Agenta - w nagłówku zapytania do serwera umieszczane są losowe (ale poprawne formalnie) informacje o programach klienckich, które imitują używanie popularnych przeglądarek internetowych lub Googlebota[4].

3.2. Analiza danych z profili

Dane, które zostały zebrane do analizy składają się z pełnego tekstu 741198 profili z czterech sekcji: Edukacja, Wykształcenie, Podsumowanie oraz Tagi.

---

[3] Atak DoS (ang.*Denial of Service*) to „zalewanie" sieci (ang. *flooding*) nadmiarową ilością danych mających na celu wysycenie dostępnego pasma, którym dysponuje atakowany host. Niemożliwe staje się wtedy osiągnięcie go, mimo że usługi pracujące na nim są gotowe do przyjmowania połączeń.

[4] Googlebot jest robotem Google indeksującym sieć (czasami jest również nazywany „pająkiem"). Indeksowanie to proces, podczas którego Googlebot wykrywa nowe i zaktualizowane strony, aby dodać je do indeksu Google. (źródło: (http://www.google.com/bot.html)

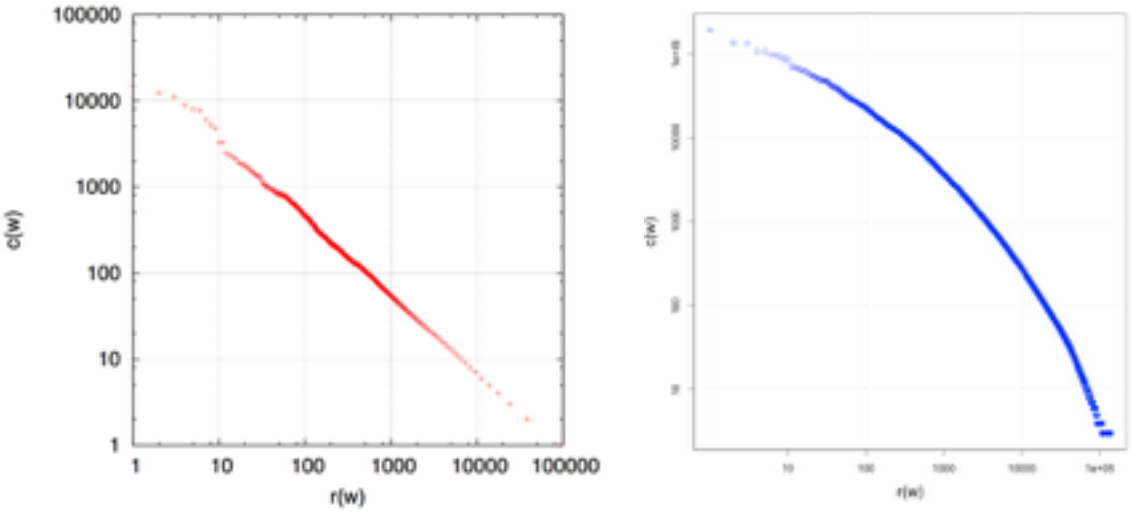
W pierwszej sekcji, Edukacja, użytkownik może uzupełnić swoje dane dotyczące szkół do których uczęszczał oraz wykształcenia które uzyskał. Kolejna przeznaczona jest do uzupełnienia historii zatrudnienia, pracodawcy oraz zajmowanego stanowiska. Ostatnie dwie sekcje nie są usystematyzowane, użytkownik może dowolnie opisać swoją osobę w Podsumowaniu i oznaczyć profil dowolnymi, według niego odpowiednimi, tagami.

Dane te posłużyły za korpus do analizy danych - teksty z profili są traktowane jako podstawa wiedzy, za pomocą której zostanie wyłoniona dwupoziomowa kategoryzacja. Przed analizą danych, ważnym elementem było zweryfikowanie czy korpus jest prawidłowym, lecz okrojonym odzwierciedleniem języka polskiego - czy nie występują żadne anomalie.

Prawo Zipfa określa, że mając częstość wystąpień w tekście **c**(**w**) dla słowa **w** oraz jego rangę **r**(**w**), tzn. pozycję w liście posortowanej malejąco, pod względem liczby wystąpień w tekście:

Częstość słowa c(w) jest odwrotnie proporcjonalna do jego rangi r(w). Na przykład, jeśli słowo **w1** ma rangę 10 razy większą niż słowo **w2**, to słowo **w1** ma częstość 10 razy mniejszą niż słowo **w2**.

Analiza częstości występowania słów okazała się zgodna z Prawem Zipfa dla korpusów.



Ryc. 4.1 Po lewej, wykres zależności częstości wystąpień słów od ich rangi z korpusu Słownika frekwencyjnego Polszczyzny Współczesnej - źródło: Łukasz Dębowski, *Prawo Zipfa próba objaśnień (*http://www.ipipan.waw.pl/~ldebowsk/docs/seminaria/zipf3.pdf*).* Po prawej wykres dla korpusu stworzonego na podstawie danych z GoldenLine.
Legenda: c(w) - częstość słowa, r(w) ranga słowa

Po upewnieniu się, że korpus jest poprawny, teksty z profili zostały przygotowane przed dalszą analizą. Wszystkie słowa zostały sprowadzone do małych liter, usunięto znaki interpunkcyjne, liczby oraz zbiór tzw. „stop words". „Stop words" to najczęściej występujące słowa języka, które zwykle nie niosą ze sobą żadnych istotnych treści. Ich usunięcie zmniejsza obciążenie serwera poprzez zredukowanie ilości trzymanych danych. Przykładem takich wyrazów są np. aczkolwiek, albo, ktokolwiek, mój, przy, przez. Jako listę takich słów użyłam słowa ignorowane w wyszukiwarce polskiej wersji Wikipedii, dostępną pod adresem: https://pl.wikipedia.org/wiki/Wikipedia:Stopwords.

Następnie, po wstępnej analizie częstotliwości występowania słów zostały usunięte te słowa, które zostały uznane za mało istotne przy szukaniu kategorii zawodowych. Były to informacje związane z:

- czasem (np.lata uczęszczania do szkoły, pracy),
- miejscem (np. miejsca pracy),
- wykształceniem (uniwersytety, szkoły),
- niektóre ręcznie wybrane słowa (np. „kat. b").

Ostatnim etapem, po analizie ilości słów w korpusie i częstości ich występowania, zostały usunięte bardzo rzadko występujące słowa, tzn. takie, które pojawiły się na profilach maksymalnie 3 razy. Redukuje to ilość trzymanych danych i wspomaga wydajność obliczeń na nich wykonywanych, które musiały być zoptymalizowane ze względu na ograniczoną ilość dostępnej pamięci RAM.

3.2.1.ANALIZA NAJCZĘŚCIEJ WYSTĘPUJĄCYCH SŁÓW

Pierwszym etapem analizy danych z profili było policzenie frekwencji wszystkich słów występujących w korpusie. Analiza tych słów wykazała jednak, że słowa najczęściej pojawiające się na profilach zazwyczaj nie są związane z kategoriami zawodowymi.

Aby wspomóc filtrowanie, zostały wykorzystane dodatkowo wartości z polskiego słownika frekwencyjnego. Pozwoliło to na filtrowanie słów, które są rzadko używane w języku polskim, ale występują często w serwisie GoldenLine - czyli potencjalnych kategorii.

Ryc.4.2 Mapa słów: często występujące słowa w serwisie GoldenLine (powyżej 1000 wystąpień) i jednocześnie rzadko w polskim słowniku frekwencyjnym (poniżej 2500 wystąpień)

Lista wszystkich występujących na profilach słów, umożliwiła zbadanie korelacji pomiędzy występowaniem słów na profilach: jeżeli słowo **w1** pojawiło się na profilu, to z jakim prawdopodobieństwem pojawiło się na nim również słowo **w2**?

Szukanie słów o wysokiej korelacji może dawać dobre rezultaty, gdy znamy kategorię zawodową i szukamy jedynie dla niej specjalizacji. Jednak w przypadku, gdy kategorie nie są z góry znane metoda ta nie wnosi istotnych informacji.

Słowo	Korelacja
php	44%
javascript	35%
java	32%
mysql	30%

Tab 1. Przykładowe słowa z najwyższym stopniem korelacji dla wyrazu „programista"

3.2.2. ANALIZA NAJCZĘŚCIEJ WYSTĘPUJĄCYCH PAR SŁÓW

Drugim etapem analizy danych z profili było stworzenie rankingu wszystkich par słów, które pojawiły się w korpusie. Analiza par była podyktowana chęcią znalezienia par słów, które tworzyłyby dwupoziomową hierarchię: kategoria zawodowa + specjalizacja, jak na przykład *zarządzanie jakością* czy *inżynieria lądowa*.

3.2.3. ANALIZA KORELACJI CZĘSTO I RZADKO WYSTĘPUJĄCYCH SŁÓW

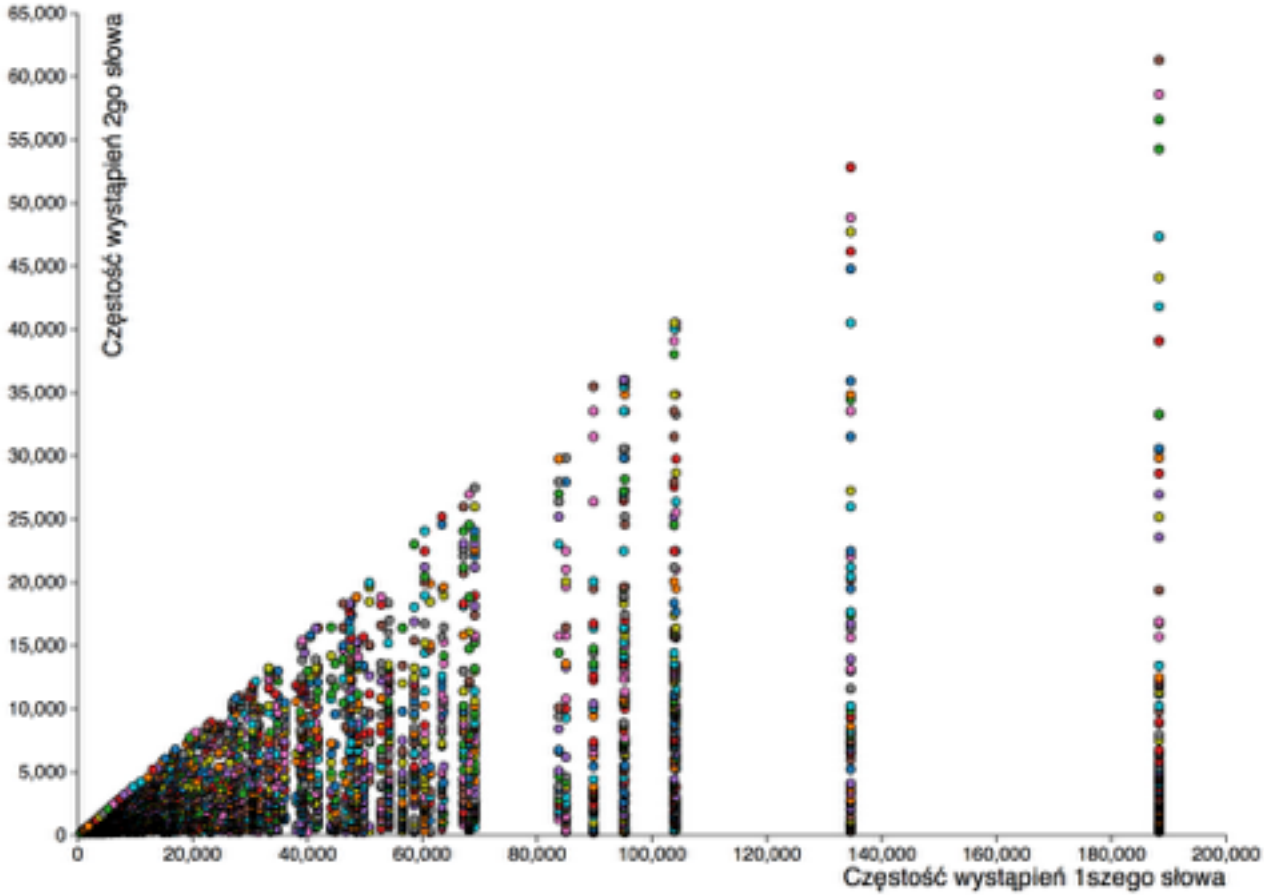
Korzystając z poprzednich analiz, użyjemy trzech rodzajów danych:

- zbiór wszystkich słów występujących na profilach i ich częstość występowania,
- zbiór wszystkich par słów występujących na profilach i ich częstość występowania,
- ranking 10000 najczęściej występujących par słów.

Wybieramy takie pary słów, które spełniają powyższe warunki:

- są jednymi z 10000 najczęściej występujących par słów,
- pierwsze słowo w parze ma wysoką częstość wystąpień, natomiast drugie niską - tzn. różnica między ich częstością wystąpień wynosi przynajmniej 60%.

Wybieramy więc zatem pary, które często występują na profilach (pierwszych 10 000 z 419 388 par - około 2,4%), ale jednocześnie jedno ze słów występuje dużo częściej w porównaniu do drugiego, co może wskazywać na hierarchię: słowo częściej występujące jest ogólnym pojęciem, a rzadsze jego specjalizacją.



Ryc.4.3 Wykres punktowy dla 10 tysięcy najczęściej występujących par słów w serwisie. Każdy punkt oznacza parę słów, a częstość wystąpień pierwszego słowa jest określona na osi poziomej, drugiego na osi pionowej. Punkty znajdujące się dokładnie na jednej lini pionowej to pary o wspólnym pierwszym słowie.

Taka klasyfikacja może jednak doprowadzić do sytuacji, gdy przeważająca część wyników będzie brana pod uwagę tylko ze względu na fakt popularności pierwszego słowa.

Aby się od tego uniezależnić, należałoby normalizować w pewien sposób wyniki. W tym celu wybieram 10000 takich par słów, które mają najwyższy stosunek częstości występowania pary do częstości występowania pierwszego słowa (im większy stosunek, tym większa niezależność od popularności pierwszego słowa).

//TODO sprawdzić z wagą częstości wystąpień pierwszego słowa (żeby nie było tak, że jest dużo zarządzan czyms tylko dlatego ze zarzadzanie jest popularne)

3.2.4.WYKORZYSTANIE SŁOWOSIECI

3.2.5.STWORZENIE DWUPOZIOMOWEJ KATEGORYZACJI ZAWODÓW

3.3. Analiza sieci społecznościowej

3.3.1.STWORZENIE SIECI SPOŁECZNOŚCIOWEJ Z INFORMACJI O KONTAKTACH

3.3.2.STWORZENIE PODSIECI DLA PROFILI NALEŻĄCYCH DO JEDNEJ KATEGORII ZAWODOWEJ

3.3.3.ANALIZA PARAMETRÓW DLA WIERZCHOŁKÓW DANEJ SIECI

3.4. Aplikacja działająca w chmurze

3.4.1.PRZEGLĄDANIE KATEGORII ZAWODOWYCH I ICH SPECJALNOŚCI

3.4.2.WYSZUKIWANIE PROFILI Z DANYCH KATEGORII

3.4.3.FILTRACJA PROFILI DLA ZADANEGO PARAMETRU SIECI

4. Opis architektury i wybranych elementów implementacji

4.1. Skrypty node.js

4.2. Bazy danych sqlite oraz MongoDB

Baza danych sqlite została utworzona w celu zgromadzenia pobieranych informacji z serwisu GoldenLine. Ten rodzaj bazy danych został wybrany ze względu na brak wymogu uruchamiania serwera bazy danych. Wszystkie dane są zawarte w jednym pliku, co umożliwia łatwe przenoszenie plików i uruchamianie skryptów na różnych serwerach, jeśli zaszłaby taka potrzeba (skrypty były uruchamiane zarówno na prywatnym komputerze jak i serwerze uczelnianym).

W bazie zostały utworzone następujące tabele:

- *hyperlinks* - tabela zawierająca podstawowe informacje o profilach: ID profilu (z serwisu), link do strony oraz dwa znaczniki informujące, czy tekst z profilu został pobrany oraz czy zostały zapisane kontakty dla tego profilu
- *users* - tabela zawierająca teksty z profili dla każdej istotnej dla pracy sekcji (Edukacja, Praca, Podsumowanie i Tagi)
- *adjacency* - tabela zawierająca informacje o kontaktach pomiędzy użytkownikami

4.3. Platforma chmurowa Heroku

5. Opis metody i wyników weryfikacji opracowanego rozwiązania

5.1 Porównanie kategorii z dostępnymi przy rejestracji profilu specjalizacjami

Przy rejestracji profilu, użytkownik jest proszony o udostępnienie informacji o branży, w której pracuje.

Dostępne opcje do wyboru to: Administracja biurowa, Badania i rozwój, Bankowość, BHP, Ochrona środowiska, Budownictwo, Edukacja, Szkolenia, Finanse, Ekonomia, Hotelarstwo, Gastronomia, Turystyka, Human Resources, Zasoby ludzkie, e-Commerce, Nowe media, Łańcuch dostaw, Media, Sztuka, Rozrywka, Nieruchomości, Praca fizyczna, Prawo, Produkcja, Public Relations, Reklama, Grafika, Kreacja, Fotografia, Sektor publiczny, Sprzedaż, FMCG, Sieci handlowe, Transport, Spedycja, Ubezpieczenia, Zakupy, Zarządzanie jakością, Zdrowie, Uroda, Rekreacja, Call Center, Inżynieria, IT - Administracja, IT - Rozwój oprogramowania, Obsługa klienta, Marketing.

5.2 Stopień pokrycia profili

Rodzaj danych	Liczba bezwględna	Procentowo
Zebranie linków z mapy publicznych profili	1431440/1431440 (stan na październik 2014)	100%
Zebranie danych z profili	741198/1431440	52%
Zebranie kontaktów		
Dopasowanie profili do kategorii		

6. Podsumowanie

6.1. Zalety i wady rozwiązania

6.2. Możliwości dalszego rozwoju

- Wielokryterialna analiza danych pod względem parametrów sieci - degree, betweenness, closeness, centrality,…
- Wieloetapowe filtrowanie danych najpierw po parametrze sieciowym, następnie po kategorii
- Analiza danych pod względem geograficznym

7. Bibliografia

1. John Scott: *Social Network Analysis*
2. Kazuya Okamoto, Wei Chen, and Xiang-Yang Li: *Ranking of Closeness Centrality for Large-Scale Social Networks*
   URL: http://research.microsoft.com/en-us/people/weic/faw08_centrality.pdf
3. M. E. J. Newman: *A measure of betweenness centrality based on random walks*
   URL: http://arxiv.org/pdf/cond-mat/0309045.pdf