Temat 2
Dla wskazanego kandydata w wyborach w USA, przedstaw w sposób czytelny na mapie finansowanie jego kampanii wyborczej, z uwzględnieniem darowizn bezpośrednich i poprzez komitety wyborcze.

Onaszkiewicz Przemysław, Gadawski Łukasz

19 grudnia 2015

1 Funkcjonalności

1. Wybór rocznika wyborów prezydenckich Użytkownik będzie miał możliwość wybrania dane z których wyborów chce oglądać. Dostępne będą tylko dane z wyborów prezydenckich.

2. Wybór kandydata z listy. W systemie będzie dostępna lista kandydatów którzy brali udział w wyborach. Użytkownik będzie miał możliwość wybrania konkretnego kandydata i wyświetlenia danych o dotacjach dotyczących tylko jego kandydatury. Będzie też możliwość wyświetlenia zagregowanych danych dotyczących wszystkich kandydatów.

3. Wybór prezentowanych danych na mapce:

   (a) wielkość wszystkich środków uzyskanych z danego regionu

   (b) wielkość darowizn bezpośrednich

   (c) wielkość dotacji od komitetu wyborczego

   Będzie to dodatkowa możliwość. Po jej wybraniu nastąpi dodatkowe filtrowanie danych pod kątem regionu w którym wpłynęły, jak również rodzaju darowizn (bezpośrednie i od komitetów "POC").

4. Wybór prezentacji danych na mapie:

   (a) dla całego kraju

   (b) poszczególnych stanów

   (c) ewentualnie rejon danego kodu pocztowego

   W początkowej fazie dane będą pokazane na obszarze całego kraju. Po wejściu w konkretny stan zostaną wyświetlone bardziej szczegółowae dane na jego temat (podział na mniejsze okręgi).

2 Pobranie danych dotyczących wyborów

Do pobrania danych dotyczących wyborów rozpatrywaliśmy API jakie udostępnia New York Times na swoich stronach. Dane można uzyskać w formacie xml w formacie json. Rozpatrywano dwa sposoby pobierania danych. Pierwszym z nich było korzystanie z API wystawionego przez New York Times do pobierania danych w formacie json. Kolejnym było pobranie na dysk plików z danymi dotyczącymi wyborów bezpośrednio na dysk twardy i przechowywanie ich w bazie danych.

2.1 New York Times API- pobieranie potrzebnych fragmentów danych w formacie json

Indywidualne wpłaty na danego kandydata:

Zapytanie [13] pozwala na pobranie danych o indywidualnych darczyńcach. Z uwzględnieniem ich miejsca zamieszkania. Filtrując dane można uzyskać dane o wysokości darowizn od darczyńców indywidualnych z danego regionu geograficznego, jak również o ilości wpłat na danego kandydata w danym regionie.

Finansowanie kandydata przez komitet z danego dystryktu stanu:

Zapytanie [12] pozwala uzyskać wysokość wpłat przekazanych na rzecz danego kandydata w danym dystrykcie wyborczym. Pozwala to uzyskać dane o wysokości wpłat komitetów na rzecz kandydata(i ich procentowego udziału w totalnej kwocie zebranej przez niego).

2.2 Pobranie plików zawierających dane o finansach bezpośrednio ze stron FEC

Drugim rozpatrywanym sposobem uzyskania danych na temat finansowania kampanii kandydatów w wyborach prezydenckich jest pobieranie danych do kampanii bezpośrednio ze stron FEC i załadowanie ich do bazy danych. Poniżej przedstawiono opis zawartości poszczególnych plików w celu przedstawienia jakie dane można za ich pomocą uzyskać.

- Plik zawierający dane na temat wpływów do kandydatów od komitetów. [10]
  - Candidate Identification Number
  - Filer Identification Number (CMTE_ID)
  - Zip Code
  - State
  - City/Town
  - Transaction Amount
  - Transaction ID
  - Entity Type
    - CAN = Candidate
    - CCM = Candidate Committee
    - COM = Committee

        * IND = Individual (a person)
        * ORG = Organization (not a committee and not a person)
        * PAC = Political Action Committee
        * PTY = Party Organization)

- Plik zawierający dane na temat wpływów do komitetów z indywidualnych dotacji. [11]
  - Filer Identification Number (CMTE_ID)
  - Zip Code
  - State
  - City/Town
  - Transaction Amount
  - Transaction ID
  - Entity Type

- Plik zawierający złączenie kandydat - komitet którego bardziej szczegółowy opis znajduje się na stronie [7]
  - Committee Identification (CMTE_ID)
  - Committee Name
  - Zip Code
  - State
  - City or Town

- Plik zawierający dane na temat kandydatów i ich szczegółowy opis [8].
  - Candidate Identification
  - Candidate Name
  - Candidate Office
    * H = House
    * P = President
    * S = Seante
  - Year of Election
  - Candidate State
  - Candidate District

- Plik zawierający komitety którego opis znajduje się pod adresem [9].
  - Committee Identification (CMTE_ID)
  - Committee Name
  - Zip Code
  - State
  - City or Town

Dane pochodzące ze strony FEC dane dotyczące finansowania są przechowywane w plikach tekstowych o przedstawionej powyżej strukturze. Na potrzeby naszej aplikacji powstanie narzędzie parsujące te pliki tekstowe oraz przeprowadzające ich zapis do bazy danych o analogicznej strukturze tabel.

3 pobranie danych geograficznych

Dane geograficzne można pobrać w dwóch formach. Jedn z nich jes formą punktową a druga obszarową.

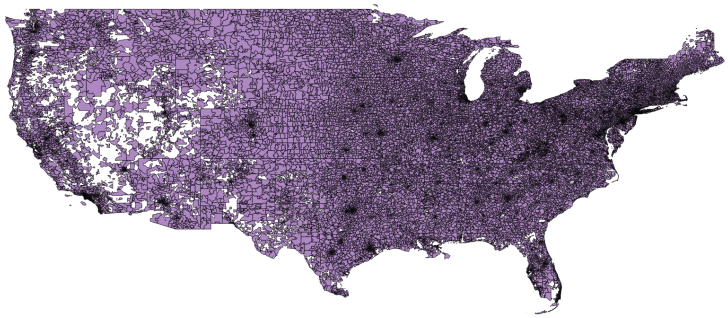
3.1 Punktowe dane o ZIP kodach

Dane o ZIP kodach można pobrtać w postaci plików z rozszerzeniem CSV. Zawierają one dane takie jak ZIP, latitude(długość) i longtitude(szerokość).

Sprawia to że zip kody są na umiejscowione na mapie, ale ich występowanie jast punktowe. Być może na potrzeby projektu taka reprezentacja była by wystarczająca. dane dotyczące finansowania wyborów na danym obszarze przedstawiano by za pomocą słupków w tych punktach.

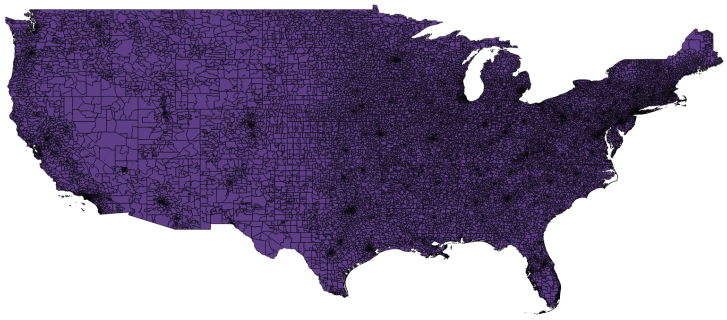
3.2 Obszarowe dane o ZIP kodach

Dane geograficzne można odnaleźć w kilku miejscach. Między innymi jest to otwarty plik Catrographic Shapefile pochodzący ze strony [6, US Census Bureau] i zawierający około 33143 kodów.



Rysunek 1: Rysunek przedstawiający pokrycie kodów pocztowych z bazy USCB na obszarze USA

Kolejną bazą ZIP kodów jest baza kodów ESRI dostępna pod adresem [1, Arcgis] na stronie istnieje również opis jak można skonwertować dane do Shapefile'a. Kodów dostępnych w tej bazie jest około 30541.



Rysunek 2: Rysunek przedstawiający pokrycie kodów pocztowych z bazy ESRI na obszarze USA

Wszystkich kodów w USA jest około 42000, najwięcej zawiera baza *U.S. Census Bureau* przytoczona powyżej (33143) i podaje przybliżone regiony, natomiast dane pobrane ze strony firmy ESRI zawierają 30541 rekordów. Pomimo to przy wizualizacji obu warstw okazuje się, że dane od firmy ESRI mają lepszą dokładność i większe pokrycie USA. Może to być spowodowane tym że dane z US CB posiadają zduplikowane dane bądź regiony na siebie zachodzą i występuje redundancja danych.

3.3 Wybranie klucza złączenia

Na podstawie danych dostarczonych przez FEC kluczem złączenia danych dotyczących finansowania kampanii kandydatów mogą być następujące atrybuty zawierające dane geograficzne:

- Zip Code
- State
- City or Town

Dana dotycząca stanu może być na nasze potrzeby zbyt ogólna, podobnie informacja o mieście lub wsi, taki klucz może nie być wystarczająco precyzyjny, co więcej dane geograficzne zawierające takie dane mogą zawierać tylko wycinki obszarów. Najbardziej dokładnym atrybutem jest kod pocztowy. Należy mieć jednak na uwadze, że taka dana również nie jest idealną reprezentacją, ponieważ kod pocztowy w głównym zamyśle powstał dla ułatwienia organizacja przesyłek pocztowych, dlatego też często zdarza się, że wieżowiec lub wielkie przedsiębiorstwo zajmujące spory obszar geograficzny może posiadać swój indywidualny kod pocztowy.

Kody pocztowe mogą zostać przedstawione w dwojakiej formie. Pierwszą jest wizualizacja danych na podstawie punktowych danych dotyczących kodów pocztowych. Tzn. każdy kod pocztowy posiada współrzędne geograficzne do niego przypisane. Znalezione dane punktowe są najdokładniejsze, ponieważ zawierają współrzędne dla 42522 kodów i można wnioskować, że są to wszystkie występujące dane. Problemem w przypadku ewentualnej wizualizacji naszego problemu finansowania kampanii wyborczej może być sposób prezentacji na mapie takich

danych. Można byłoby tego dokonać za pomocą różnej wielkości słupków, bądź okręgów przypisanych do danej długości i szerokości geograficznej. Niewątpliwą zaletą tej metody jest dokładność, ponieważ najprawdopodobniej wszystkie dane geograficzne uzyskałyby złączenie z danymi dotyczącymi finansowania.

Drugim sposobem są przybliżone obszary w postaci wielokątów na mapie obrazujących dany kod pocztowy. Pobrane dane geograficzne ze strony firmy ESRI po dokładnym przyjrzeniu się mapom wydają się być najbardziej precyzyjne i obejmować niemalże cały obszar USA, pomimo tego, że zawierają 30541 rekordów w stosunku do około 42000 wszystkich kodów pocztowych występujących w USA. Zaletą tego sposobu przedstawienia danych byłaby możliwość przedstawienia danych dotyczących finansowania poprzez gradient konkretnych obszarów. Wadą jest w tym przypadku potencjalna niedokładność prezentacji danych, lub wręcz brak prezentacji niektórych danych finansowych poprzez brak złączenia z danymi geograficznymi.

4 Technologie wykonania

Na potrzeby tworzonej aplikacji zostanie stworzona struktura przechowująca pobrane dane dotyczące finanosowania kandatów. Zostanie stworzony skrypt, który zapisze potrzebne dane z plików CSV do bazy danych ***PostgreSQL***. Dane geograficzne również będą przechowywane w bazie danych, dodatkowo będzie zainstalowana nakładka ***PostGIS*** umożliwiająca wykonywanie zapytań przestrzennych.

Pobieranie danych o ZIP kodach było rozpatrywane w dwóch wariantach: punktowym (3.1) i obszarowym (3.2). Zdecydowano się na pobranie ich w formie obszarowej. Najprawdopodobniej będzie to lepiej służyło potrzebom aplikacji. Spośród dwóch rozpatrywanych baz (ESRI i USCB) lepsza wydaje się baza ***ESRI***. Ponieważ pomimo mniejszej ilości występujących w niej kodów, na rysunkach 2 i 1 widać że baza ***ESRI*** ma większe pokrycie na terenie całego USA. To pokrycie obszaru USA w połączeniu z danymi adresowymi darczyńcy będzie mogło posłużyć do złączenia danych nawet w przypadku nie występowania ZIP kodu darczyńcy w bazie.

Z uwagi na chęć zapewnienia największej elastyczności w pobieraniu oraz prezentacji danych zostanie stworzona aplikacja internetowa umożliwiająca serwowania danych w przeglądarce. Do tego celu zostanie wykorzystany ***Geoserver***, czyli serwer aplikacyjny, który będzie odbierał przy pomocy REST API zapytań HTTP o zawartość do wyświetlenia. ***Geoserver*** jako popularny kontener serwletów umożliwia w prosty sposób połączenia z bazą danych PostgreSQL oraz prezentację danych geograficznych.

Warstwa prezentacji, czyli stworzenie komponentów HTML umożliwiających opisane w pierwszym punkcie funkcjonalności (np. wybór wyborów oraz kandydatów), zostanie wykonana przy pomocy biblioteki języka JavaScript ***AngularJS***. Umożliwia on w prosty sposób konstruowanie żądań HTTP do danego API, a także pozwala jego zastosowanie pozwoli odseparować warstwę widoku od serwera aplikacji i danych.

5 Pomocne linki

W niniejszej sekcji znajdują się linki do źródeł, na które autorzy natrafili podczas prób zdiagnozowania problemu.

- [4, Atrykuł przedstawiający ładowanie ShapeFile do bazy danych PostGIS]
- [5, Atrykuł pokazujący obsługę GeoSerwera]
- [3, Wątek opisujący załadowanie pliku z bazą danych eESRI i konwersje fo Shapefile'a]
- [2, Strona Wiki opisująca sposób ekstrakcji danych geograficznych do bazy postGisowej]

Literatura

[1] arcgisdatabaseesri. http://www.arcgis.com/home/item.html?id=8d2012a2016e484dafaac0451f9aea24.

[2] Extract osm data to postgis. http://gis.stackexchange.com/questions/108006/how-to-convert-data-from-a-gdb-into-a-shapefile-without-arcmap.

[3] How to load esri data and convert it to shapefile. http://gis.stackexchange.com/questions/108006/how-to-convert-data-from-a-gdb-into-a-shapefile-without-arcmap.

[4] Loading shapefile data into postgis from the command line. http://suite.opengeo.org/opengeo-docs/dataadmin/pgGettingStarted/shp2pgsql.html.

[5] Serve postgis data. http://docs.geoserver.org/stable/en/user/gettingstarted/index.html.

[6] Unites states census bureau zip code database. https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html.

[7] NW Washington DC 20463 (800) 424-9530 In Washington (202) 694-1000 Federal Election Commission, 999 E Street. Fec candidate - comtiee linkage file site. http://www.fec.gov/finance/disclosure/metadata/DataDictionaryCandCmteLinkage.shtml.

[8] NW Washington DC 20463 (800) 424-9530 In Washington (202) 694-1000 Federal Election Commission, 999 E Street. Fec candidates master file site. http://www.fec.gov/finance/disclosure/metadata/DataDictionaryCandidateMaster.shtml.

[9] NW Washington DC 20463 (800) 424-9530 In Washington (202) 694-1000 Federal Election Commission, 999 E Street. Fec comitees master file site. http://www.fec.gov/finance/disclosure/metadata/DataDictionaryCommitteeMaster.shtml.

[10] NW Washington DC 20463 (800) 424-9530 In Washington (202) 694-1000 Federal Election Commission, 999 E Street. Fec contributions to candidates from comitees site. http://www.fec.gov/finance/disclosure/metadata/DataDictionaryContributionstoCandidates.shtml, dec.

[11] NW Washington DC 20463 (800) 424-9530 In Washington (202) 694-1000 Federal Election Commission, 999 E Street. Fec individual contributions master file description site. http://www.fec.gov/finance/disclosure/metadata/DataDictionaryContributionsbyIndividuals.shtml.

[12] Derek Willis. commitees donnations ny times api. http://developer.nytimes.com/docs/campaign_finance_api/campaign_finance_api_examples#committee-contributions-to-candidate-json, may 2015.

[13] Derek Willis. individual donnations ny times api. http://developer.nytimes.com/docs/campaign_finance_api/campaign_finance_api_examples#pres-state-zip-json, may 2015.