Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Automatyki i Informatyki Stosowanej

Praca dyplomowa inżynierska

na kierunku Informatyka
w specjalności Systemy Informacyjno-Decyzyjne

Asystent wyboru OPP

Agnieszka Dunajska
Numer albumu 273687

promotor
dr inż. Mariusz Kamola

Warszawa 2018

Asystent wyboru OPP

Streszczenie

Celem pracy było utworzenie aplikacji, umożliwiającej porównanie wybranych cech organizacji pożytku publicznego, na podstawie treści składanych przez nie zeznań rocznych. W pracy przybliżono, czym są organizacje pożytku publicznego i publikowane przez nie sprawozdania merytoryczne oraz jaką mają budowę i jakie informacje są w nich zawarte (m.in.: opis działań organizacji, liczba wolontariuszy, liczba odbiorców, wysokość najwyższego i średniego wynagrodzenia członków zarządu). Opisano przeznaczenie aplikacji i zaprezentowano przykłady użycia. Przedstawiono etapy tworzenia pracy: automatyczne pobranie dokumentów PDF ze strony ministerstwa, użycie czytnika OCR do wydobycia tekstu z dokumentów PDF, wydobycie interesujących informacji z tekstów, zgromadzenie danych w bazie danych, utworzenie aplikacji. Opisano użyte narzędzia i technologie. Wskazano także problemy, jakie wystąpiły podczas realizacji. Zamieszczono wnioski z tworzenia pracy i uzyskanego efektu oraz perspektywy rozwoju.

Słowa kluczowe: organizacje pożytku publicznego, OCR, PDF, Python, Django, NLP, Selenium

Selection assistant of non-governmental organisations

Abstract

The main purpose of this thesis was to build an application to compare selected features of non-governmental organisations using their annual financial reports. Thesis includes description of non-governmental organizations and their financial reports, structure of financial reports and it contents (among others: descripton of organisation's actions, amount of volunteers, amount of recipients, the highest and average remuneration for organisation's management). Thesis contains an intended use of application and presentation of examples and results. Steps of creating thesis were presented: automatic downloading of PDF documents from departmental website, usage of OCR reader to extract text from PDF documents, extraction of interesting data from texts, creating database, development of application. Tools and technologies used in project were highlighted. Problems that occured during work were pointed out. At the end, there are conclusions and prospects for further development.

Keywords: non-governmental organizations,OCR, PDF, Python, Django, NLP, Selenium

załącznik do zarządzenia nr 28/2016 r.
Rektora PW

Politechnika Warszawska

„załącznik nr 3 do zarządzenia nr 24/2016 Rektora PW

.....
miejscowość i data

.....
imię i nazwisko studenta

.....
numer albumu

.....
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta"

Spis treści

Podziękowania

Składam serdeczne podziękowania dla Promotora pracy - Pana Doktora Inżyniera Mariusza Kamoli za inspirację, pomoc merytoryczną oraz okazaną cierpliwość i wyrozumiałość.

Agnieszka Dunajska

1. Wstęp

Organizacje Pożytku Publicznego

„Organizacje pożytku publicznego są to organizacje pozarządowe (między innymi stowarzyszenia i fundacje, niedziałające w celu osiągnięcia zysku), które na podstawie Ustawy o działalności pożytku publicznego i o wolontariacie (z 24 kwietnia 2003 r.) uzyskały w sądzie status pożytku publicznego."[1]

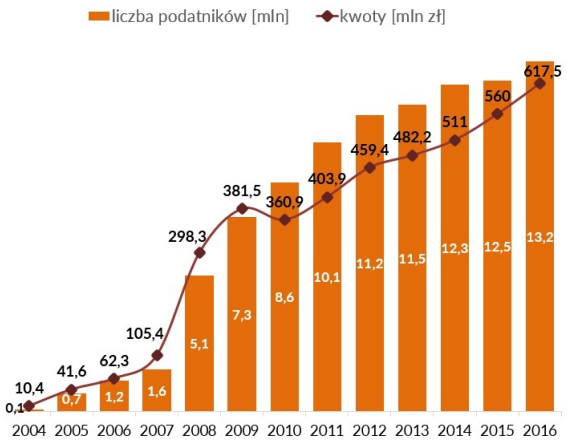
Jednym z ważnych przywilejów posiadanych przez OPP jest możliwość otrzymywania 1% podatku dochodowego od osób fizycznych. Podatnik, płacący podatek dochodowy, może przeznaczyć 1% swojego podatku na rzecz wybranej przez siebie organizacji pożytku publicznego, znajdującej się na oficjalnej liście tego rodzaju organizacji. Darczyńcy wybierają organizację, której chcą przekazać te środki za pomocą odpowiedniego wpisu w deklaracji PIT. „Darowizny przekazane przez osoby fizyczne organizacjom pożytku publicznego podlegają odliczeniu od dochodu, zmniejszając podstawę opodatkowania." [2]

Posiadanie przez organizację statusu organizacji pożytku publicznego wiąże się z obowiązkiem sprawozdawczości. Polega on na konieczności publikowania i składania w Departamencie Pożytku Publicznego Ministerstwa Pracy i Polityki Społecznej dorocznego sprawozdania finansowego i merytorycznego. Dzięki temu wszyscy zainteresowani mogą uzyskać informacje, w jaki sposób wykorzystane zostały pieniądze przekazane przez darczyńców i jak organizacja realizowała swoje cele. [2]

W Polsce zarejestrowanych jest 100 000 stowarzyszeń i 20 000 fundacji. Status OPP ma 8500. Najwięcej organizacji zajmuje się sportem (29%), edukacją (15%) oraz kulturą (13%). 18 % Polaków i Polek poświęca swój czas na pracę w organizacjach pozarządowych. 45% organizacji opiera się wyłącznie na pracy społecznej.[3]

W roku 2013 liczba osób, które przekazały swój 1% podatku dochodowego na organizacje pożytku publicznego, przekroczyła 12 milionów. W su-

mie przekazane kwoty wyniosły 506,6 milionów złotych. Źródło [3] podaje, że w roku 2016 z prawa do przekazania 1% podatku na rzecz wybranej organizacji pożytku publicznego skorzystało ponad 13 milionów Polaków. Stanowi to 49% wszystkich uprawnionych podatników. Liczba osób, które decydują się skorzystać z takiej formy pomocy rośnie od 2004 roku, kiedy została wprowadzona taka możliwość. „Początkowo korzystali z niej nieliczni: w 2004 roku zaledwie 80 tysięcy osób, rok później już 700 tysięcy, a po dwóch latach ponad 1 milion podatników."[3] Analogicznie rosły też kwoty przekazywane poprzez jednoprocentowy mechanizm. W pierwszym roku jego funkcjonowania organizacje pożytku publicznego zebrały w ten sposób nieco ponad 10 milionów złotych, a w kolejnym roku o 30 milionów więcej. W 2006 roku zebrano ponad 60 milionów złotych. „W 2016 roku, po ponad dziesięciu latach funkcjonowania, mechanizm przekazywania 1% podatku zapewnił organizacjom pożytku publicznego 617,5 milionów zł."[3] Wzrost obrazuje wykres na rysunku 1.1.



Rysunek 1.1: Wzrost kwot przekazywanych za pomocą 1%-towego mechanizmu w ciągu ostatnich lat. Źródło: [3]

Część darczyńców ma konkretnie sprecyzowane, komu chce pomóc – np. osobie bliskiej (poprzez subkonto) lub fundacji prowadzonej przez osobę znajomą. Część osób chce wspomóc organizację z konkretnej sfery. Motywację darczyńców przedstawia rysunek 1.2.

Rysunek 1.2: Motywacja darczyńców. Źródło: [3]

Cel

Darczyńcy mogą być zainteresowani tym, w jaki sposób wydatkowane są pieniądze organizacji, które rozważają wesprzeć, jak wysokie są wynagrodzenia zarządu, ilu było odbiorców organizacji i wolontariuszy. Takie dane są dostępne w sprawozdaniach merytorycznych umieszczanych przez organizacje w ministerialnej bazie sprawozdań. Każdy dokument liczy około kilkunastu stron i jest żmudny przy analizowaniu. Celem pracy jest zautomatyzowanie pobierania i przetwarzania danych organizacji oraz usprawnienie procesu ich wyświetlania, tak, aby użytkownik nie musiał pobierać wielu dokumentów PDF i każdego analizować oddzielnie. Szybko mógłby zorientować się jak duże kwoty poszczególne organizacje wypłacają członkom zarządu, ilu mają odbiorców, itd.. Mógłby także przefiltrować organizacje, tak, żeby wyświetlane wyniki reprezentowały interesujący go, ze względu na m.in. lokalizację i sferę działalności, podzbiór zbioru wszystkich OPP.

Istniejące rozwiązania

Obecnie nie istnieją rozwiązania, które umożliwiałyby porównywanie danych finansowych organizacji pożytku publicznego ani wyświetlanie ich na wykresie. Dostępne są strony internetowe, na których można wyświetlić listę organizacji i przeczytać ich opisy. Są także programy wspierające wypełnianie deklaracji PIT i przekazywanie 1%. Przykładem takiego rozwiązania jest: https://www.e-pity.pl/baza-opp-wykaz-organizacji-pozytku-publicznego/ oraz http://www.opp.aid.pl/.

2. Koncepcja

Wymagania

Aplikacja powinna umożliwić użytkownikowi wyszukanie organizacji na podstawie wypełnionych przez niego kryteriów. Powinna być też możliwość wyboru kolejności sortowania. Organizacje, będące wynikiem wyszukania, powinny być przedstawione na wykresie (na osiach umieszczone cechy, które zostały wybrane przez użytkownika przed posortowaniem jako dwie najważniejsze cechy). Ponadto rekordy, spełniające wymagania użytkownika, wyświetlone zostaną w postaci tabelarycznej, prezentując zgromadzone w utworzonej bazie danych cechy organizacji, pochodzące ze sprawozdań merytorycznych.

Scenariusze użycia

Jednym ze scenariuszy użycia jest wsparcie darczyńców w procesie wyboru OPP - wyszukanie przez użytkownika organizacji, które spełniają jego oczekiwania (m.in. lokalizacyjne i finansowe) oraz pomoc w znalezieniu organizacji, która najmniejsze kwoty przeznacza na wynagrodzenia, jednocześnie pomagając dużej liczbie odbiorców. Wyniki mogą pomóc użytkownikowi w wyborze organizacji do przekazania 1% podatku.

Innym zastosowaniem aplikacji jest wyszukanie organizacji, które przeznaczają na wynagrodzenia najwyższe kwoty lub w innych cechach znacząco odstają od pozostałych organizacji.

Aplikacja może też pomóc osobom zainteresowanym konkretnymi OPP w zorientowaniu się, jaki rząd wielkości stanowią kwoty przeznaczane na wynagrodzenia oraz porównać je z przychodami organizacji. Ciekawym dodatkiem jest możliwość zweryfikowania liczby wolontariuszy i odbiorców działań organizacji, które skorzystały z jej pomocy.

Kolejnym przykładem zastosowania aplikacji jest możliwość weryfikacji danych wprowadzonych w sprawozdaniach merytorycznych przez przedsta-

wicieli organizacji pożytku publicznego. Organizacje mogłyby znaleźć ewentualne błędy w swoich danych finansowych.

Struktura dokumentów PDF

Poniżej zamieszczone zostały przykładowe fragmenty sprawozdania, obrazujące jego charakter. Na rysunku 2.1 znajduje się opis głównych działań podjętych przez organizację.

1.1. Opis głównych działań podjętych przez organizację	Inicjatywy osób indywidualnych i instytucji realizowane przez Fundację: Fundusze: Fundusz Obywatelski, wspierać działania osób i instytucji na rzecz ochrony praw i wolności obywatelskich, wartości konstytucyjnych, utrzymania pokoju społecznego oraz tworzenia przestrzeni dla pluralistycznej debaty wokół istotnych spraw publicznych, w duchu troski o dobro wspólne; Fundusz im. E. Wende wspierający krzewienie i podnoszenie kultury prawnej w Polsce poprzez przyznawanie corocznej nagrody osobie szczególnie zasłużonej dla dobra publicznego; Fundusz Bądźmy solidarni, którego celem jest udzielanie pomocy agentom firmy Aviva oraz członkom ich najbliższych rodzin w przypadku długotrwałej choroby lub innych zdarzeń losowych, mających wpływ na zdrowie; Fundusz Złota Perła, którego celem jest wspieranie młodych, zdolnych osób oraz przedsięwzięć z dziedziny kultury, sztuki i nauki. Programy Fundacji: Program Filantropii Indywidualnej, którego celem jest promowanie postaw filantropijnych wśród osób indywidualnych, zwiększenie ich zaangażowania w działania społeczne oraz zachęcanie ich do prowadzenia własnej działalności filantropijnej; Program Transnational Giving Europe, dzięki któremu darczyńcy mogą wspierać przedsięwzięcia społeczne w wybranym przez siebie kraju z dowolnej dziedziny, np. kultury, sztuki, edukacji, sportu i skorzystać z ulg podatkowych w kraju zamieszkania; Program Młodzież i Filantropia wzmacniający postawy filantropijne i zaangażowanie na rzecz innych wśród młodzieży gimnazjalnej i licealnej.

Rysunek 2.1: Opis głównych działań podjętych przez organizację

Na podstawie *opisu głównych działań podjętych przez organizację* w ramach pracy inżynierskiej zostanie stworzona lista słów kluczowych odpowiadających danej organizacji.

W aplikacji zostanie wykorzystane pole *Łączna kwota przychodów organizacji ogółem (zgodnie z rachunkiem wyników / zysków i strat)*, przedstawione na rysunku: 2.2.

III. Przychody i koszty organizacji pożytku publicznego w okresie sprawozdawczym	
1. Informacja o przychodach organizacji	
1. Łączna kwota przychodów organizacji ogółem (zgodnie z rachunkiem wyników/zysków i strat)	110,127.60 zł
a) Przychody z działalności nieodpłatnej pożytku publicznego	86,557.60 zł
b) Przychody z działalności odpłatnej pożytku publicznego	23,570.00 zł
c) Przychody z działalności gospodarczej	0.00 zł
d) Przychody finansowe	0.00 zł
e) Pozostałe przychody	0.00 zł

Rysunek 2.2: Informacja o przychodach organizacji

Na rysunkach 2.3 i 2.4 znajdują się wynagrodzenia w okresie sprawozdawczym.

VI. Wynagrodzenia w okresie sprawozdawczym	
1.Łączna kwota wynagrodzeń (brutto) wypłaconych przez organizację w okresie sprawozdawczym	209,672.73 zł
a) z tytułu umów o pracę	177,270.60 zł
- wynagrodzenie zasadnicze	177,270.60 zł
- nagrody	0.00 zł
- premie	0.00 zł
- inne świadczenia	0.00 zł
b) z tytułu umów cywilnoprawnych	32,402.13 zł

Rysunek 2.3: Wynagrodzenia w okresie sprawozdawczym

4. Wysokość **przeciętnego** miesięcznego wynagrodzenia (brutto) wypłaconego członkom organu zarządzającego organizacji, wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia oraz umowy cywilnoprawne *Aby określić przeciętne miesięczne wynagrodzenie należy: 1.zsumować wszystkie kwoty wynagrodzeń wypłacone w ciągu roku sprawozdawczego (wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia oraz umowy cywilnoprawne); 2. podzielić zsumowaną kwotę przez 12 (miesięcy)*	6,622.60 zł
5. Wysokość **przeciętnego** miesięcznego wynagrodzenia (brutto) wypłaconego członkom organu kontroli lub nadzoru, wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia oraz umowy cywilnoprawne *(patrz komentarz do punktu 4)*	0.00 zł
6. Wysokość **przeciętnego** miesięcznego wynagrodzenia (brutto) wypłaconego członkom innych, niż organu zarządzającego, kontroli lub nadzoru, organów organizacji, wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia oraz umowy cywilnoprawne *(patrz komentarz do punktu 4)*	0.00 zł
7. Wysokość **przeciętnego** miesięcznego wynagrodzenia (brutto) wypłaconego pracownikom organizacji, wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia, oraz osobom świadczącym usługi na podstawie umowy cywilnoprawnej *(patrz komentarz do punktu 4)*	5,646.00 zł

Druk: MPiPS 10

8. Wysokość **najwyższego** miesięcznego wynagrodzenia (brutto) wypłaconego członkom organu zarządzającego, wliczając wynagrodzenie zasadnicze, nagrody, premie i inne świadczenia oraz umowy cywilnoprawne	6,622.60 zł

Rysunek 2.4: Wynagrodzenia w okresie sprawozdawczym

Istotnymi danymi są dane finansowe. W aplikacji zostaną wykorzystane pola *Łączna kwota wynagrodzeń (brutto) wypłaconych przez organizację w*

okresie sprawozdawczym oraz *Wysokość przeciętnego i najwyższego miesięcznego wynagrodzenia wypłaconego członkom organu zarządzającego.*

Inne dane, jakie zostaną wykorzystane to liczba wolontariuszy, odbiorców działań organizacji oraz sfera działalności.

Planowany wygląd aplikacji

Aplikacja wyświetli formularz, do którego użytkownik będzie mógł wpisać swoje kryteria.

Formularz

Przewidywane pola, które znajdą się w formularzu to:

1. *Nazwa organizacji zawiera* -możliwość wyszukania organizacji, w których nazwie zawarty jest fragment wprowadzony przez użytkownika
2. *Województwo*
3. *Miejscowość*
4. *Sfera* - możliwość wyboru sfery działaności organizacji z listy rozwijanej
5. *Wynagrodzenie średnie zarządu mniejsze niż*
6. *Wynagrodzenie średnie zarządu większe niż*
7. *Wynagrodzenie najwyższe zarządu mniejsze niż*
8. *Wynagrodzenie najwyższe zarządu większe niż*
9. *Wynagrodzenie łączne mniejsze niż*
10. *Wynagrodzenie łączne większe niż*
11. *Przychody mniejsze niż*
12. *Przychody większe niż*
13. *Liczba odbiorców działań organizacji mniejsza niż*
14. *Liczba odbiorców działań organizacji większa niż*
15. *Liczba wolontariuszy mniejsza niż*
16. *Liczba wolontariuszy większa niż*
17. *Słowa kluczowe zawierają* - możliwość wyszukania organizacji, które w zbiorze słów kluczowych, uzyskanych na podstawie opisu ich działalności, posiadają słowo wskazane przez użytkownika

Dodatkowo, użytkownik otrzyma możliwość wyboru trzech kryteriów, według których nastąpi sortowanie organizacji. Dla każdego z nich użytkownik będzie mógł wskazać, czy sortowanie ma być rosnące, czy malejące. Do wyboru będą następujące kryteria: wynagrodzenie najwyższe, wynagrodzenie średnie, wynagrodzenie łączne, przychody, odbiorcy, wolontariusze.

Wynik działania aplikacji

Odpowiedzią aplikacji na wprowadzone przez użytkownika kryteria będzie wyświetlenie wykresu bąbelkowego, pokazującego przefiltrowane, spełniające kryteria organizacje. Na osiach wykresu będą cechy wybrane jako dwie pierwsze spośród kryteriów sortowania. Pod wykresem, w formie tabelki, znajdą się przefiltrowane organizacje wraz z województwem, miejscowością, sferą, najwyższym wynagrodzeniem wypłaconym członkom zarządu, średnim wynagrodzeniem wypłaconym członkom zarządu, wynagrodzeniem łącznym, wysokością przychodów, ilością wolontariuszy i ilością odbiorców.

Przegląd technologii i narzędzi

Selenium

Pakiet Selenium Web Driver służy do automatyzacji interakcji z przeglądarkami internetowymi. Zostanie zastosowany do automatycznego pobrania dużej ilości dokumentów PDF ze strony ministerstwa. Selenium często wykorzystywane jest do przeprowadzania testów automatycznych. Umożliwia między innymi nawigację po stronie internetowej, pobieranie elementów, automatyczne wypełnianie formularzy zadanymi danymi i ich przesyłanie oraz zaznaczanie przycisków wyboru. Selenium potrafi współpracować m.in. z przeglądarkami, takimi jak: Firefox, Internet Explorer, Safari i Chrome. [4]
Dostępna jest przydatna funkcja
clickAndWait, która sprawia, że Selenium po kliknięciu w dany element poczeka, aż strona się w pełni załaduje, przed wykonaniem następnych kroków.[5]
Biblioteki Selenium dostępne są w różnych językach programowania, także w języku Python.

Czytnik OCR

OCR to zestaw technik lub oprogramowanie służące do rozpoznawania znaków na zdjęciach i skanach. Jest używany do rozpoznania tekstu w zeskanowanym dokumencie. W pracy inżynierskiej zostanie wykorzystany do przekonwertowania dokumentów PDF do postaci txt.

Na przestrzeni lat nastąpił znaczny rozwój optycznego rozpoznawania znaków. W latach 90-tych XX wieku stosowano kosztowne oprogramowanie, które pozwalało na uzyskanie zadowalających wyników przy skanach dobrej jakości. W 2013 roku możliwe było rozpoznawanie skanów kiepskiej jakości, z szumami na obrazkach, z tekstem napisanym pod nietypowymi kątami, w ponad 120 językach.[6]

Ciekawym zastosowaniem OCR jest wykorzystanie tej techniki przy dy-

gitalizacji zasobów bibliotek. Rozpoznawany jest tekst na stronach książek i tworzone są ich elektroniczne wersje, w których można wyszukiwać konkretne frazy. Chroni to oryginały książek przed zniszczeniem i zwiększa dostępność takich książek. [8] OCR przydaje się także przy odczytywaniu danych z formularzy wypełnianych pismem odręcznym. Zdarzają się jednak błędy przy rozpoznawaniu. W przypadku wątpliwości, gdy OCR ma trudność z odczytaniem fragmentu tekstu, niezbędne jest sprawdzenie wyniku OCR przez człowieka.[6]

Przy rozpoznawaniu pisma stosuje się metody rozpoznawania wzorców, należące do metod sztucznej inteligencji. „Oprogramowanie OCR wykorzystuje różne metody segmentacji obrazu, na przykład progowanie, aby wyodrębnić poszczególne znaki z obrazu, które następnie są najczęściej osobno klasyfikowane jako poszczególne litery. Zwykle w tym procesie wykorzystywane są sieci neuronowe. Zazwyczaj, by wyeliminować pomyłki, program sprawdza całość rozpoznanego tekstu lub poszczególne wyrazy pod kątem poprawności ortograficznej i gramatycznej danego języka." [7]

NLP - Natural Language Processing

„Przetwarzanie języka naturalnego (ang. natural language processing, NLP) jest interdyscyplinarną dziedziną, łączącą zagadnienia sztucznej inteligencji i językoznawstwa, zajmującą się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer."[9] Termin „język naturalny" został wprowadzony, by łatwo odróżnić język, używany do komunikacji międzyludzkiej, taki jak polski czy angielski, od języka komputerowego, takiego jak Java czy Python. NLP zostanie zastosowane w pracy inżynierskiej do wydobycia słów kluczowych z opisu działalności organizacji.

System badający język naturalny przekształca próbki języka naturalnego na formalne symbole, które następnie mogą zostać poddane dalszemu przetwarzaniu przez programy komputerowe. Tekst w języku naturalnym to dane bez struktury. Nie ma narzuconej organizacji informacji, jak w przypadku tabeli. Dlatego tekst poddawany jest przetwarzaniu wstępnemu, aby umożliwić analizę i wnioskowanie. „Wiele problemów NLP wiąże się zarówno z generacją, jak i rozumieniem języka, np. model morfologiczny zdania (struktura słów), który komputer powinien zbudować, jest potrzebny zarazem do tego, by zdanie było zrozumiałe, jak i gramatycznie poprawne."[9]

Morfeusz

Program Morfeusz jest analizatorem morfologicznym dla języka polskiego. Ciąg znaków w tekście wydzielony odstępami lub znakami interpunkcyjnymi

nazywany jest słowem. Leksem jest jednostką języka, zwaną wyrazem słownikowym. Jest traktowany jako wspólna część wszystkich form fleksyjnych. Formą wyrazową (fleksyjną) nazywamy słowo zinterpretowane, czyli powiązane z konkretnym leksemem. Forma wyrazowa jest rozpatrywana łącznie z jej cechami gramatycznymi i znaczeniem. [10]

„Analiza morfologiczna polega na określeniu dla danego słowa wszystkich form wszystkich leksemów, których może dotyczyć. W procesie tym nie uwzględnia się kontekstu, w którym wystąpiło dane słowo."[11]

Celem hasłowania jest wyróżnienie dla każdego słowa, występującego w tekście, odpowiadającego mu leksemu. Przybliżone hasłowanie, polegające na odcięciu ze słów części zmieniającej się przy odmianie, nazywane jest stemowaniem. Stemowanie sprawdza się dla języków o ograniczonej fleksji, ale dla polskiego daje niesatysfakcjonujące wyniki. W przypadku Morfeusza wykonywane jest prawdziwe hasłowanie, a nie stemowanie.[11]

Morfeusz zostanie użyty w pracy inżynierskiej do sprowadzenia słów w języku polskim do postaci hasłowej. Na schemacie 2.5 przedstawiono przykład działania Morfeusza.

0	1	Mam	**mama** **mamić** **mieć**	subst:pl:gen:f impt:sg:sec:imperf fin:sg:pri:imperf
1	2	próbkę	**próbka**	subst:sg:acc:f
2	3	analizy	**analiza**	subst:sg:gen:f subst:pl:nom.acc.voc:f
3	4	morfologicznej	**morfologiczny**	adj:sg:gen.dat.loc:f:pos
4	5	.	**.**	interp

Rysunek 2.5: Przykład działania Morfeusza. Źródło: [11]

Baza danych MySQL

Dane organizacji będą przechowywane w bazie danych. Do tego celu zostanie wykorzystany poznany na studiach system zarządzania relacyjnymi bazami danych MySQL. Został wybrany, ponieważ serwer MySQL jest dostępny dla popularnych platform systemowych i różnych architektur procesorów. Również biblioteki klienckie MySQL, które umożliwiają korzystanie z serwera bazodanowego z poziomu aplikacji, dostępne są dla wielu języków programowania – m.in. dla C, C++ czy Python, użytego w pracy inżynierskiej. [12]

Django

Django to darmowy framework(szkielet) do tworzenia aplikacji webowych, napisany w Pythonie. Dostarcza gotowych do użycia komponentów, które organizują pracę programisty. W pracy inżynierskiej Django zostanie użyty do stworzenia aplikacji webowej.

Gdy użytkownik chce skorzystać z aplikacji, serwer otrzymuje żądanie użytkownika i przekazuje je dalej do Django, które odpowiada za ustalenie, czego dotyczy żądanie. Ta część jest wykonywana przez obecny w Django mechanizm rozpoznawania adresów (ang. Urlresolver). Urlresolver pobiera listę wzorców i próbuje dopasować adres URL. Jeśli uda się znaleźć pasującą regułę, przekazuje żądanie do odpowiedniej funkcji (zwanej widokiem). Aby przejść z URL-a do widoku, Django używa URLconfs, które mapuje wzorce URL na widoki. W funkcji widoku można połączyć się z bazą danych i wyszukać potrzebne informacje. Widok generuje odpowiedź, a Django wysyła ją do przeglądarki użytkownika.[13] Zaletą Django jest fakt posiadania ORM wysokiego poziomu. „ORM to mapowanie obiektowo-relacyjne (ang. Object-Relational Mapping). Jest to sposób odwzorowania obiektowej architektury systemu informatycznego na bazę danych (lub inny element systemu) o relacyjnym charakterze."[15] ORM umożliwia łatwe i bezpieczne operowanie na bazach danych bez użycia SQL.W pliku ustawień zdefiniowane są parametry połączenia z bazą danych. Django daje możliwość tworzenia formularzy i manipulacje na wpisanych danych.

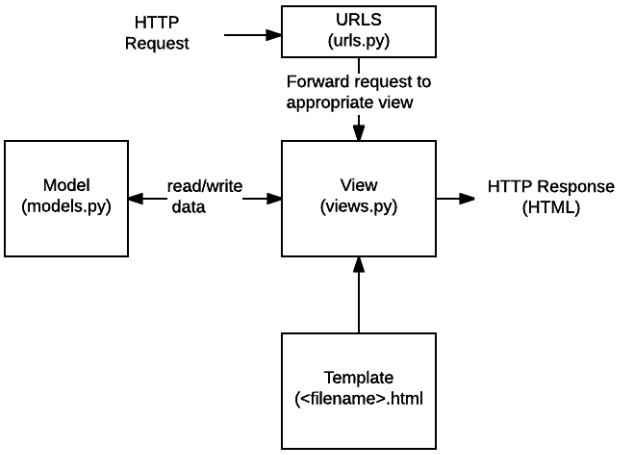
Model Szablon Widok

Django charakteryzuje się dobrą dokumentacją [14]. Opisano w niej zasadę działania wzorca architektonicznego Model Szablon Widok: Django udostępnia warstwę abstrakcji („modele") do strukturyzacji i manipulowania danymi aplikacji WWW. Model to pojedyncze, pełne źródło informacji o danych. Zawiera podstawowe pola i zachowania danych, które są przechowywane. Zazwyczaj każdy model odpowiada jednej tabeli w bazie danych. W przypadku niniejszej aplikacji tą tabelą jest tabela zawierająca dane organizacji pożytku publicznego. Dane reprezentowane przez model i informacje o nim są przechowywane w klasie modelu.

Django stosuje się do zasady DRY, opisanej w poniższej podsekcji Filozofie projektowe. Celem jest zdefiniowanie modelu danych w jednym miejscu i automatyczne czerpanie z niego informacji przez inne elementy.

Django korzysta z „widoków" do hermetyzacji logiki odpowiedzialnej za przetwarzanie zapytań użytkownika i zwracania odpowiedzi.[14] Warstwa szablonów udostępnia wygodną dla projektantów składnię renderowania in-

formacji prezentowanych użytkownikowi.[16] Na schemacie 2.6 przedstawiono komunikację między komponentami aplikacji napisanej w Django.



Rysunek 2.6: Schemat komunikacji Model Widok Szablon (Model View Template). Źródło: [16]

Serwer deweloperski

Django posiada własny, prosty serwer testowy, napisany w Pythonie. Serwer został zawarty w Django, aby możliwe było szybkie rozwijanie aplikacji bez konieczności zajmowania się konfiguracją produkcyjnego serwera, dopóki aplikacja nie będzie gotowa do wdrożenia. Serwer deweloperski automatycznie przeładowuje kod Pythona dla każdego żądania. W większości przypadków nie trzeba restartować serwera po dokonaniu zmian w kodzie. Niektóre działania, jak dodawanie plików, nie powodują automatycznego restartu i wtedy należy zrestartować serwer manualnie.[17]

Główne filozofie projektowe

Poniżej przedstawione zostały filozofie, którymi kierowali się projektanci Django podczas tworzenia frameworku. Filozofie zostały opisane w dokumentacji Django.[18]

Luźne powiązanie Różne warstwy frameworka nie powinny wiedzieć o innych więcej niż jest to niezbędne. Na przykład, szablon nie wie nic o żądaniu, warstwa bazy danych jest niezależna od wyświetlania danych, a widok nie ma informacji o szablonie.

Szybki rozwój Najważniejszym aspektem frameworku XXI wieku jest możliwość szybkiego wykonania powtarzalnych elementów tworzenia strony internetowej. Django pozwala na sprawne wytwarzanie oprogramowania.

Nie powtarzaj się DRY (Don't Repeat Yourself) Zaleca się unikanie powielania tych samych elementów kodu w różnych miejscach. Redundancja jest niekorzystna. Wiele z dynamicznych stron internetowych używa pewnego wspólnego sposobu projektowania - np. nagłówek i stopka. System szablonów Django ułatwia magazynowanie elementów w jednym miejscu, redukując możliwość duplikacji kodu.

Wprost jest lepiej niż nie wprost To jest główna zasada Pythona i oznacza, że w Django cel nie powinien być uzyskiwany w ukryty i trudny do wyjaśnienia sposób, ponieważ może to dezorientować programistów i być przyczyną błędów oraz trudnych do przewidzenia zachowań programu.

3. Wykonanie

Rozwiązanie - etapy tworzenia pracy

Poniżej przedstawiono etapy tworzenia pracy, które reprezentują poszczególne podzadania do wykonania. W każdym z etapów użyte zostało narzędzie, umożliwiające uzyskanie odpowiednich rezultatów.

1. Automatyczne pobranie dokumentów PDF z ministerialnej bazy sprawozdań i zapisanie ich pod nazwami organizacji, których dotyczą (domyślnie zapisywane były pod nazwami numerycznymi)

2. Przetworzenie dokumentów przez czytnik OCR do postaci txt

3. Przygotowanie dokumentów txt do ekstrakcji potrzebnych informacji

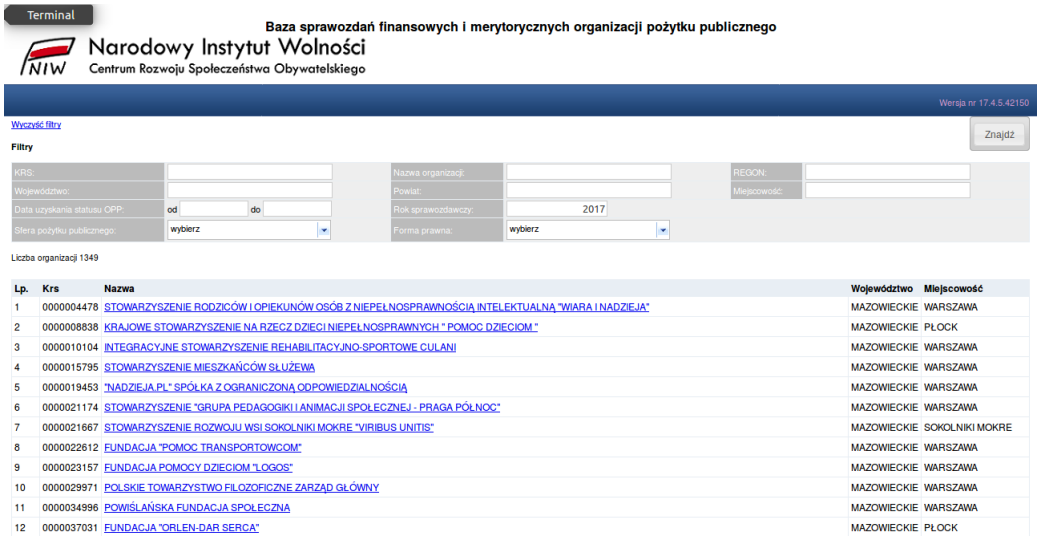
4. Ekstrakcja potrzebnych danych z dokumentów

5. Umieszczenie danych w bazie danych

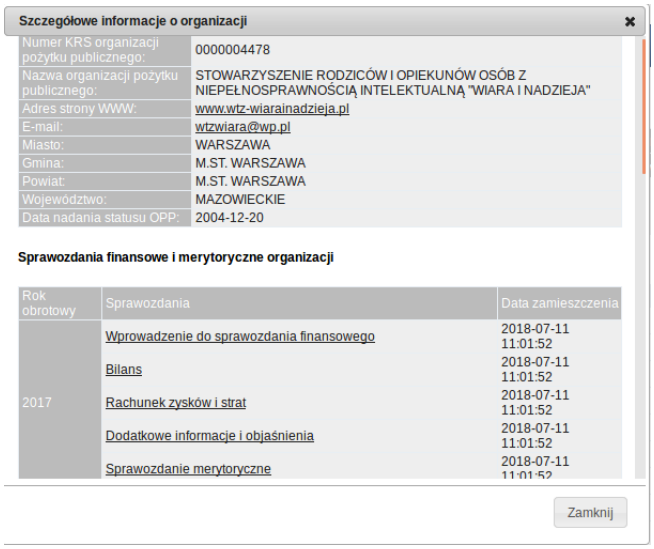
6. Utworzenie aplikacji, która korzystając z bazy danych, prezentuje dane w formie tabelarycznej oraz w formie wykresu, uwzględniając kryteria wprowadzone przez użytkownika

Pobieranie sprawozdań - Selenium

Sprawozdania zostały umieszczone w formie PDF w ministerialnej bazie sprawozdań dostępnej na stronie internetowej:
http://sprawozdaniaopp.mpips.gov.pl/. Na rysunkach 3.1 i 3.2 przedstawiono bazę sprawozdań merytorycznych i finansowych, z której pobrano dokumenty w ramach pracy inżynierskiej.



Rysunek 3.1: Lista organizacji publikujących sprawozdania na stronie http://sprawozdaniaopp.mpips.gov.pl/



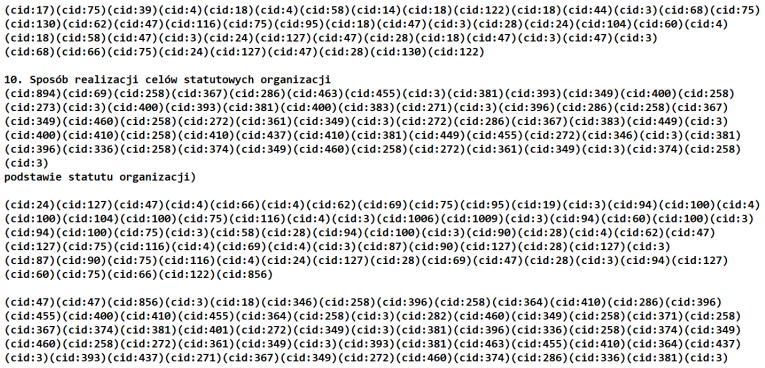
Rysunek 3.2: Zestawienie dokumentów możliwych do pobrania, opublikowanych przez jedną z organizacji. Dane ze strony http://sprawozdaniaopp.mpips.gov.pl/

W województwie mazowieckim było około 1300 organizacji. Aby zautomatyzować pobieranie dokumentów w ramach pracy został napisany skrypt w języku Python, który pobrał dokumenty PDF do folderu na dysku. Skorzy-

stano w tym celu z Selenium, pozwalającego na zautomatyzowanie interakcji z przeglądarką internetową.

Konwersja PDF do postaci txt

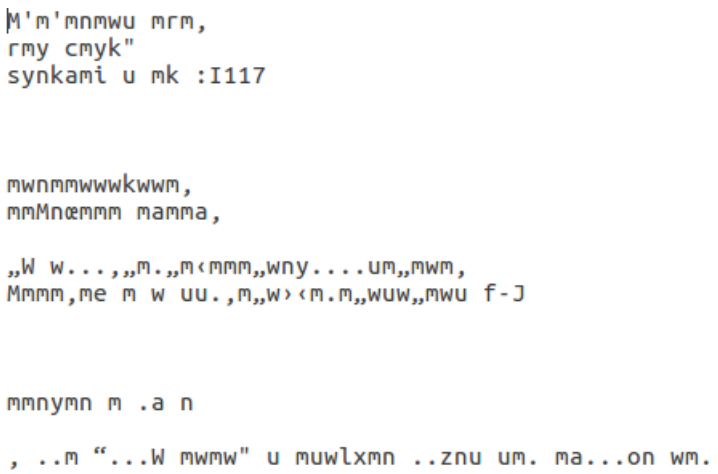
Mając pobrane dokumenty PDF, należało odczytać ich zawartość. Miały one specyficzną strukturę, zawierającą tabele i czcionki osadzone. Były także zabezpieczone przed kopiowaniem. Znaczącą trudność przysporzyło ich przetworzenie. Dostępna biblioteka w języku Java (Apache PDFBox) nie potrafiła prawidłowo poradzić sobie z dokumentami. Popularny PDFMiner, napisany w języku Python także nie potrafił przetworzyć dokumentów. Efekt jego działania został przedstawiony na rysunku 3.3.



Rysunek 3.3: Przykład niezadowalającego działania PDFMiner

Wiele czytników OCR nieprawidłowo przetwarzało znaki, dając w efekcie znaki, znacznie różniące się od występujących w dokumentach PDF. W przypadku innych dokumentów PDF narzędzia te wykazują się dobrą skutecznością. Czytniki, które zawiodły w kontekście rozpatrywanych sprawozdań merytorycznych to m.in.: PDFelement 6, Able2Extract, free-ocr.

Na rysunku 3.4 został pokazany efekt konwersji dokumentu PDF do postaci txt, uzyskany za pomocą narzędzia free-ocr. Wynik działania jest wysoce niezadowalający i dyskwalifikuje to narzędzie w przypadku pracy ze sprawozdaniami merytorycznymi.

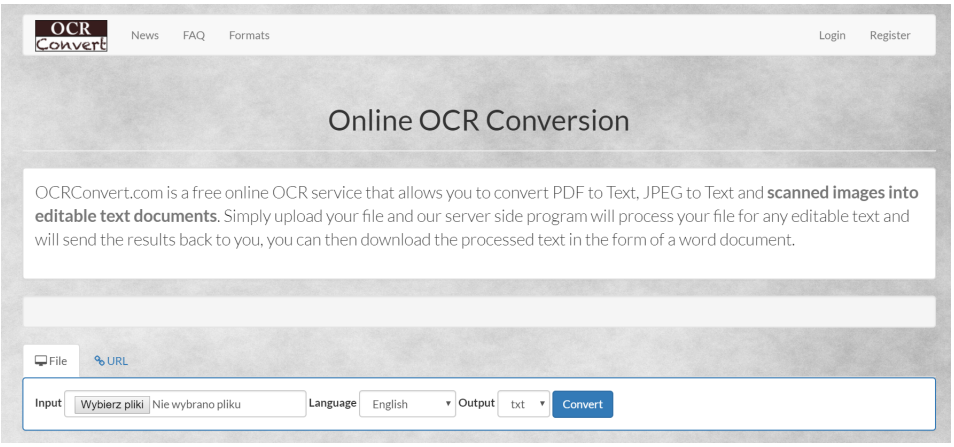


Rysunek 3.4: Przykład niezadowalającego działania free-ocr

Narzędzie dostępne pod adresem: https://lightpdf.com/pl/ocr również nie spełniło wymagań. Tekst w wyjściowym pliku był nieuporządkowany, niezgodny z kolejnością występującą w dokumencie PDF. Nie można było zidentyfikować, który opis dotyczy danej kwoty.

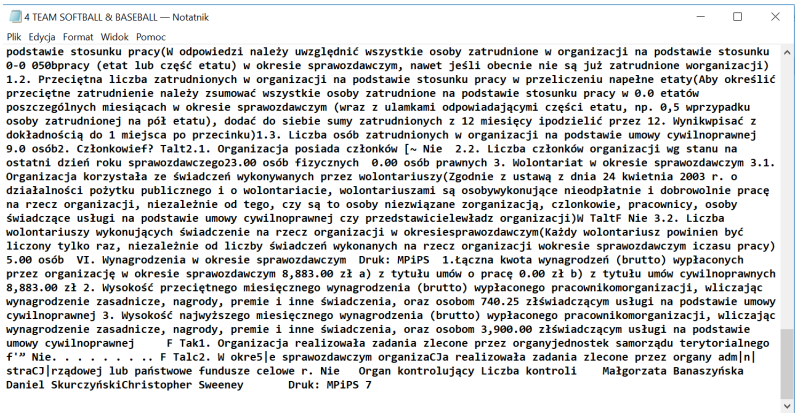
Narzędzie, które najlepiej poradziło sobie z dokumentami to https://www.onlineocr.net/ - jednak użycie go do przetwarzania większej ilości stron jest odpłatne. W przypadku sprawozdań merytorycznych zdarzały się dokumenty nawet kilkunastostronicowe. Wiele narzędzi do przetwarzania większej liczby dokumentów i stron wymagało dość wysokich opłat.

Narzędzie, które poradziło sobie z problemem i zostało wykorzystane w pracy to czytnik OCR:
https://www.ocrconvert.com/ , wspierający następujące formaty plików wejściowych: PDF, GIF, BMP, JPEG, PNG. Czytnik ten nie ma limitu na liczbę konwersji. Wygląd czytnika został przedstawiony na rysunku 3.5.

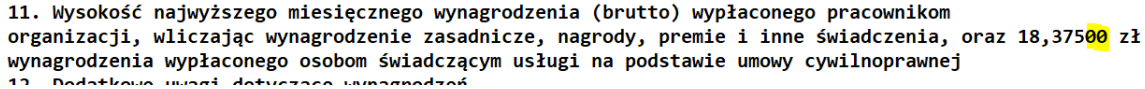


Rysunek 3.5: Wybrany czytnik OCR. Źródło: https://www.ocrconvert.com/

Przetworzył on dokumenty PDF do postaci plików tekstowych txt. Uzyskane zostały zadowalające rezultaty, przedstawione na rysunku 3.6. Jednak czas przetwarzania dokumentów był długi – wahał się od 1-4 minut (jeden dokument). Zdarzały się także błędy w plikach wynikowych. Na przykład czytnik gubił istotną kropkę w danych finansowych lub pojawiała się zamiana litery *l* na *I*. Przykład takiego błędu zamieszczono na rysunku 3.7. Za pomocą wyrażeń regularnych możliwe było poradzenie sobie z tym problemem. W większości przypadków czytnik działał jednak poprawnie. Sporadycznie pojawiały się błędy, leżące po stronie organizacji pożytku publicznego. Przykładowo, organizacja miała wynagrodzenia zarządu większe niż 0, a wynagrodzenia łączne równe 0. Wskazuje to najprawdopodobniej na błąd osoby uzupełniającej sprawozdanie w takiej organizacji. Zdarzały się też sprawozdania z niekompletnymi informacjami.



Rysunek 3.6: Przykład działania wybranego czytnika OCR: przetworzony dokument



Rysunek 3.7: Przykład niepoprawnego działania wybranego czytnika OCR: zgubiona kropka przez ostatnimi dwoma zerami

Aby przyspieszyć przetwarzanie dokonane zostało jego zrównoleglenie, tak, że jednocześnie przetwarzane było kilka dokumentów, każdy w osobnym wątku. Interakcja z czytnikiem przebiegała przy pomocy skryptu, który został napisany w ramach pracy w języku Python. Skorzystano z pakietu Selenium, tak jak w przypadku pobierania dokumentów PDF. Funkcja od-

powiedzialna za pobranie lokalizuje pole, w którym wgrywany jest plik do przetworzenia, ładuje go, wybiera język polski z listy rozwijanej, czeka na przetworzenie pliku przez konwerter a następnie go pobiera.

Wyrażenia regularne

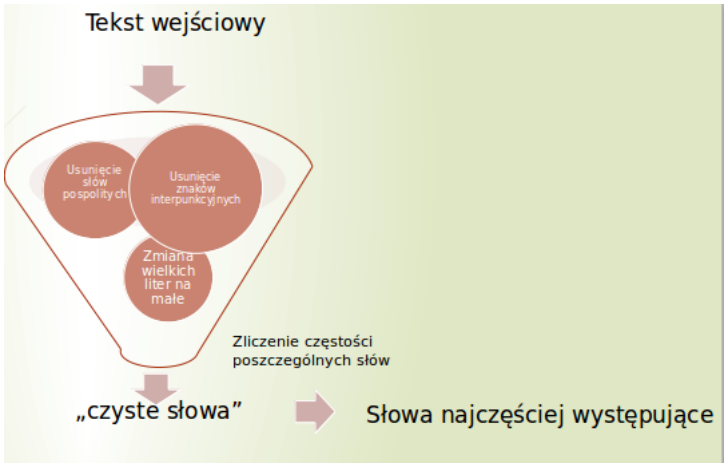
Do wydobycia interesujących danych z plików txt zostały użyte wyrażenia regularne. Wyrażenia regularne to pewne wzorce, za pomocą których można opisać ciąg znaków.[19] Istnieje wiele notacji, które służą do zapisywania wyrażeń regularnych . Jedną z nich jest notacja Pythona, użyta przy tworzeniu aplikacji. Aby skorzystać z możliwości oferowanych przez wyrażenia regularne, należy zaimportować moduł re. [20] W pracy inżynierskiej, dzięki wyrażeniom regularnym, możliwe było znalezienie w plikach txt odpowiednich danych finansowych. Program znajdował zdefiniowane wcześniej słowa poprzedzające poszukiwaną liczbę (np. wartość wynagrodzeń), a następnie samą liczbę i sprawdzał, czy ma ona odpowiedni format. Znaleziona liczba wpisywana była do bazy danych.

Uzyskanie słów kluczowych

W pracy inżynierskiej, do uzyskania słów kluczowych na podstawie opisu działań organizacji został użyty NLTK Natural Language Toolkit, będący darmowym zestawem bibliotek i programów do pracy z językiem naturalnym i jego statystycznego przetwarzania. Umożliwia m.in. tokenizację, analizę składniową i klasyfikację tekstu. [21]

Tekst wejściowy (opis działań organizacji) został poddany tokenizacji. Tokenizacja służy do podziału tekstu na pojedyncze słowa. Znakiem podziału jest spacja i inne znaki przestankowe.Następnie zbiór tokenów poddano lematyzacji, czyli sprowadzeniu słów do ich podstawowej formy (lematów).[22] Np. formy: *dziecku* i *dziecka* dotyczą tego samego słowa - *dziecko*. Więc do takiej formy powinny zostać sprowadzone. W pracy inżynierskiej do sprowadzenia słów w języku polskim do postaci hasłowej został wykorzystany Morfeusz. W następnym etapie identyfikowane były słowa powszechne, które często występują w danym języku. W języku polskim są to na przykład: *w, do, który, gdzie, przed, na*. Nie identyfikują one tematu przetwarzanego tekstu. Podczas obliczeń statystycznych wykonywanych na przetworzonym tekście takie słowa występowałyby bardzo często, częściej niż inne, istotne słowa. Dlatego ważne jest odfiltrowanie popularnych słów przed wykonaniem analizy statystycznej. W przypadku asystenta wyboru OPP, na potrzeby poprawnego działania aplikacji, istniejący zbiór słów popularnych został powiększony o słowa dziedzinowe, takie jak *organizacja, stowarzyszenie, związek*. Pojawiają się one w opisach działań większości organizacji, lecz nie przyczyniają się do

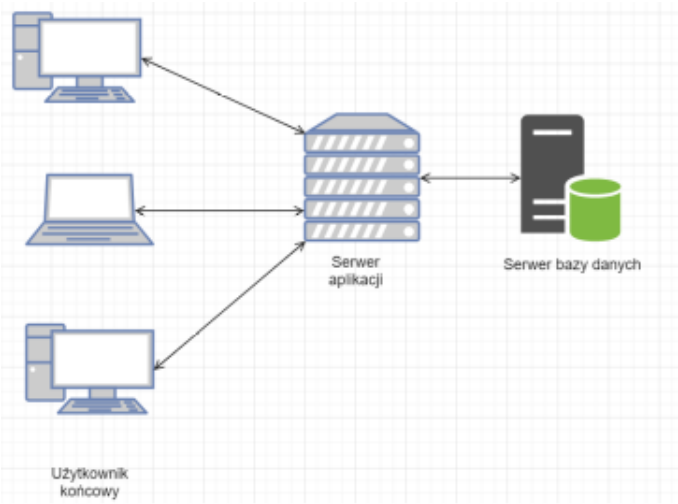
możliwości rozróżnienia organizacji między sobą pod względem przedmiotu działalności. Dlatego zostały dodane do listy słów popularnych, aby wykluczyć je z oznaczenia jako słowa kluczowe. Zliczenie częstotliwości występowania słów poddanych ostatecznemu przetworzeniu pozwala na wyodrębnienie słów najczęściej występujących - kluczowych. Schemat 3.8 przedstawia sposób uzyskania słów kluczowych.



Rysunek 3.8: Etapy tworzenia słów kluczowych. Schemat wykonany przez autora pracy.

Aplikacja webowa

Aplikacja została stworzona przy użyciu języka Python i frameworku Django. Do tworzenia wykresu bąbelkowego wykorzystano bibliotekę Chart.js, umożliwiającą tworzenie różnych typów wykresów, m.in.: liniowego, słupkowego, kołowego.[23] Aplikacja korzysta z bazy danych MySQL, w której zostały umieszczone dane organizacji uzyskane w poprzednich etapach pracy. Architektura aplikacji została przedstawiona na rysunku 3.9.

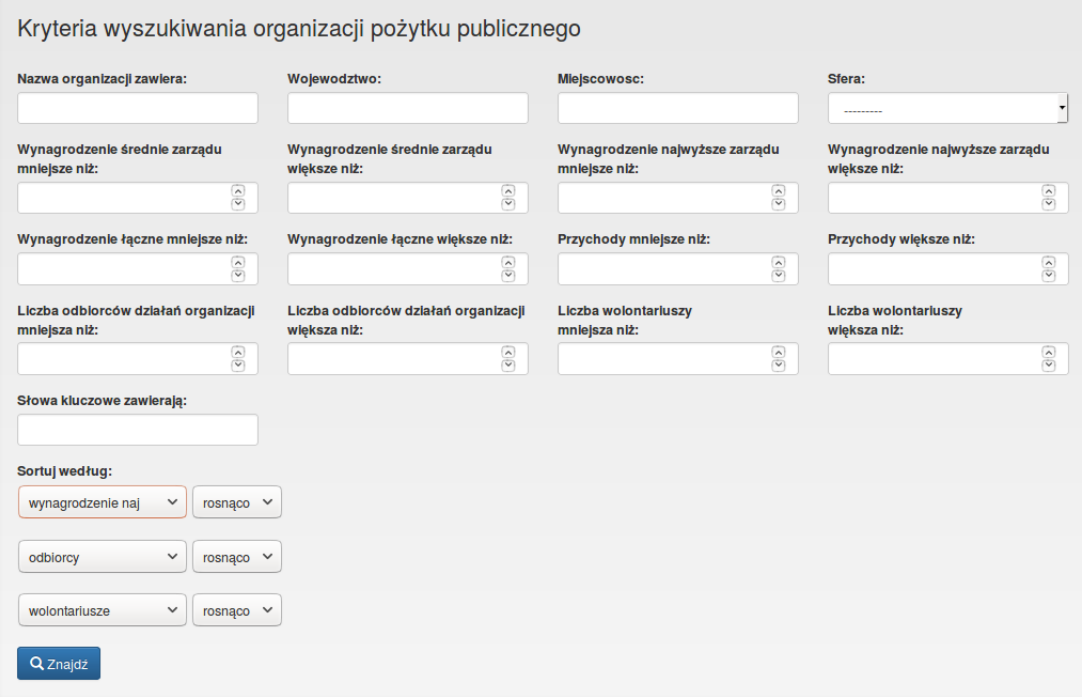


Rysunek 3.9: Architektura rozwiązania

Poniżej zaprezentowane zostało działanie aplikacji przy różnych danych wprowadzonych przez użytkownika. Przy użytkowaniu, po najechaniu na wybrany bąbelek na wykresie, wyświetlana jest nazwa organizacji reprezentowanej przez bąbelek oraz jej współrzędne.

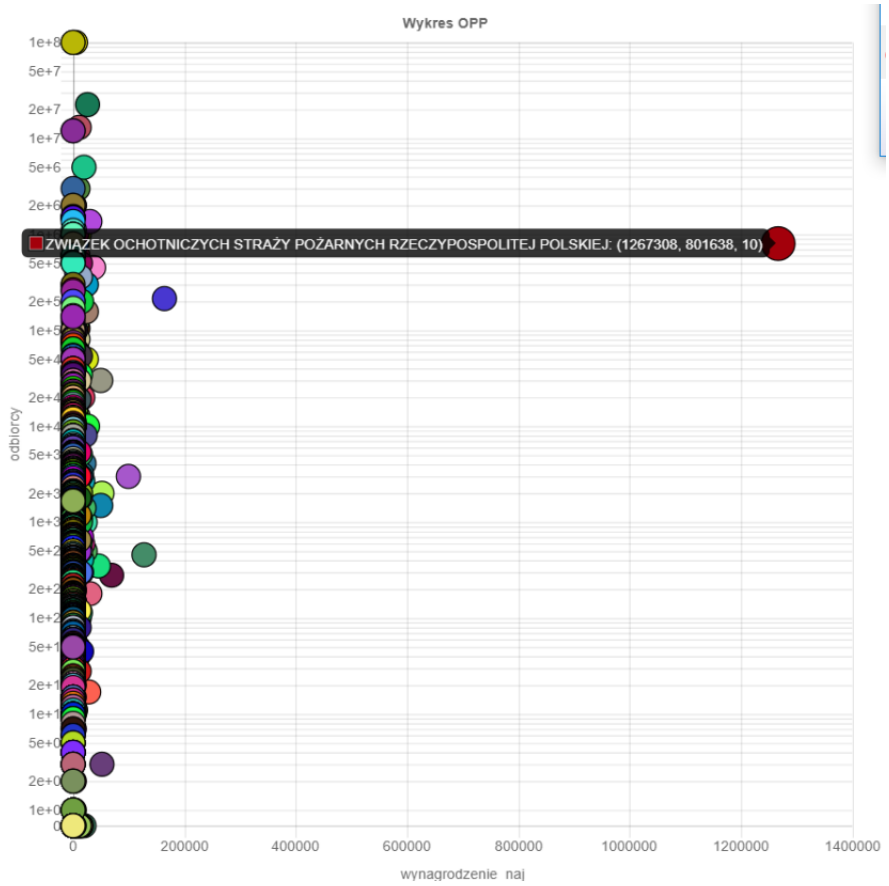
Przykład 1

Po uruchomieniu aplikacji użytkownik widzi formularz przedstawiony na rysunku 3.10.



Rysunek 3.10: Formularz z kryteriami wyszukiwania OPP

Użytkownik nie wprowadza żadnych parametrów wyszukiwania. Domyślnie dane sortowane są według: *wynagrodzenia najwyższego (rosnąco), odbiorców (malejąco), wynagrodzenia łącznego (rosnąco)*. Na rysunku 3.11 przedstawiono wyświetlający się wykres organizacji bez wpisanych kryteriów użytkownika.



Rysunek 3.11: Wykres zawierający wszystkie organizacje

Na rysunku 3.12 przedstawiono listę pojawiających się organizacji bez wpisanych kryteriów użytkownika.

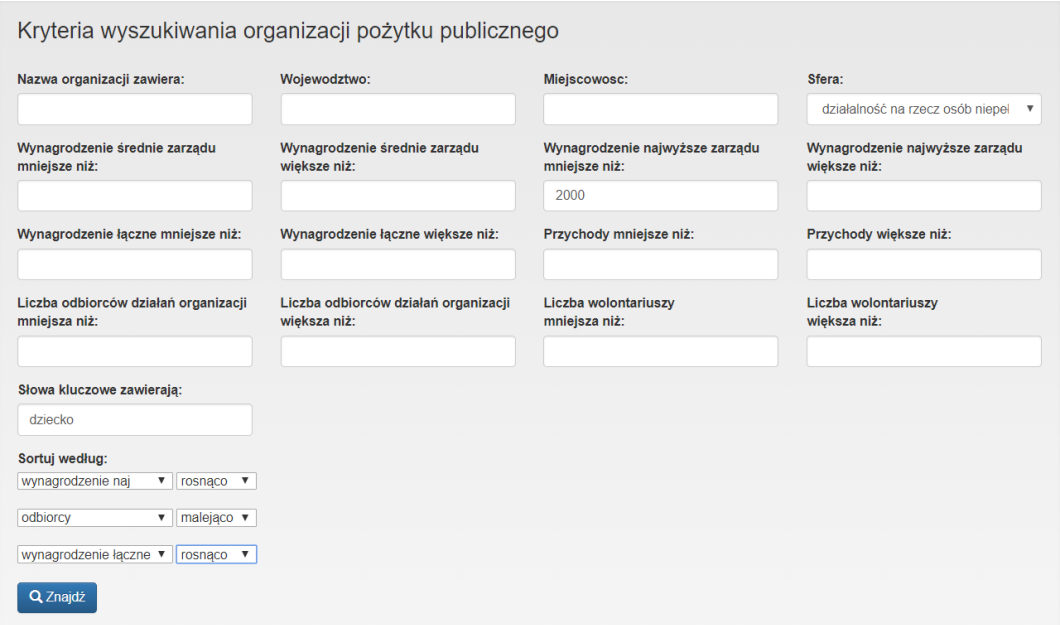
Nazwa	Województwo	Miejscowość	Sfera	Najwyższe wynagrodzenie wypłacone członkom zarządu	Średnie wynagrodzenie wypłacone członkom zarządu	Łączne wynagrodzenie wypłacone w roku sprawozdawczym	Przychody	Wolontariusze	Odbiorcy działań organizacji
STOWARZYSZENIE MIŁOSIERDZIA ŚW. WINCENTEGO A PAULO PRZY PARAFII ŚW. KRZYŻA	MAZOWIECKIE	WARSZAWA	działalność charytatywnej	0	0.0	48737	157751.0	3000	99999999
FUNDACJA HUTNIK 1957	MAZOWIECKIE	WARSZAWA	działalność na rzecz dzieci i młodzieży, w tym wypoczynek dzieci i młodzieży	0	0.0	5499	10434.0	7	12000000
STOWARZYSZENIE PRZYJACIÓŁ INTEGRACJI	MAZOWIECKIE	WARSZAWA	działalność na rzecz osób niepełnosprawnych	0	0.0	1568954	3105630.0	50	3000000
STOŁECZNE WODNE OCHOTNICZE POGOTOWIE RATUNKOWE	MAZOWIECKIE	WARSZAWA	ratownictwo i ochrona ludności	0	0.0	1706272	2661090.0	320	2000000
STOWARZYSZENIE MIŁOŚĆ NIE WYKLUCZA	MAZOWIECKIE	WARSZAWA	działalność charytatywnej	0	0.0	79134	16559.0	20	2000000
POLSKA AKCJA HUMANITARNA	MAZOWIECKIE	WARSZAWA	działalność wspomagająca rozwój wspólnot i społeczności lokalnych	0	0.0	4820643	71206300.0	73	1537165
ZWIĄZEK STOWARZYSZEŃ POLSKA ZIELONA SIEĆ	MAZOWIECKIE	WARSZAWA	upowszechnianie i ochrona praw konsumentów	0	0.0		1461100.0	0	1500000

Rysunek 3.12: Część listy organizacji

Na wykresie 3.11 widać, że organizacja *Związek Ochotniczych Straży Pożarnych Rzeczypostpolitej Polskiej* ma najwyższe wynagrodzenia miesięczne dla członków zarządu, znacząco wyższe od pozostałych organizacji.

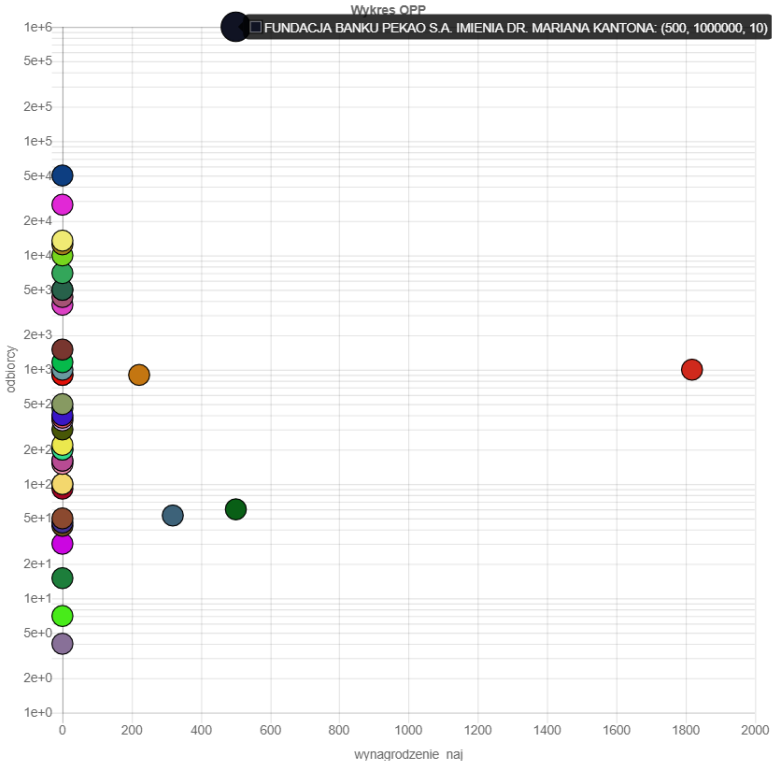
Przykład 2

Użytkownik wybiera sferę działalności organizacji: *działalność na rzecz niepełnosprawnych*, słowo kluczowe: *dziecko* oraz wynagrodzenie najwyższe zarządu mniejsze niż: *2000*. Sortowanie pozostaje domyślne. Wypełnione kryteria użytkownika zostały przedstawione na rysunku 3.13.

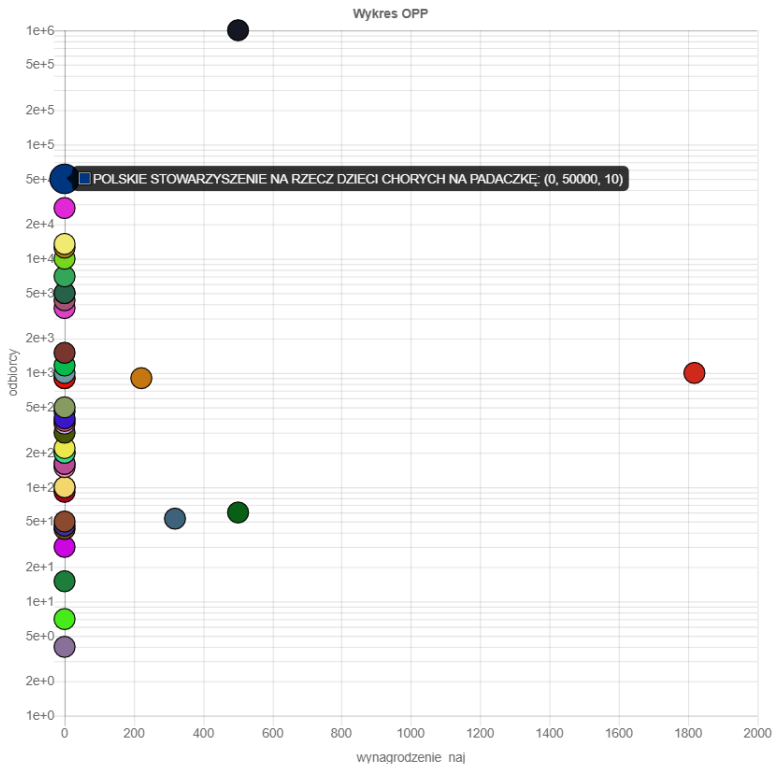


Rysunek 3.13: Kryteria użytkownika

Po naprowadzeniu kursora myszy na bąbelek znajdujący się najwyżej można zobaczyć nazwę organizacji, która ma najwięcej odbiorców spośród wyszukanych organizacji, spełniających kryteria użytkownika. Po naprowadzeniu kursora myszy na bąbelek znajdujący się najbardziej z prawej strony można zobaczyć nazwę organizacji, która ma największe najwyższe wynagrodzenia zarządu. Na rysunku 3.14 i 3.15 przedstawiono wyświetlający się wykres organizacji po uzupełnieniu kryteriów użytkownika.



Rysunek 3.14: Wykres zawierający przefiltrowane organizacje



Rysunek 3.15: Wykres zawierający przefiltrowane organizacje

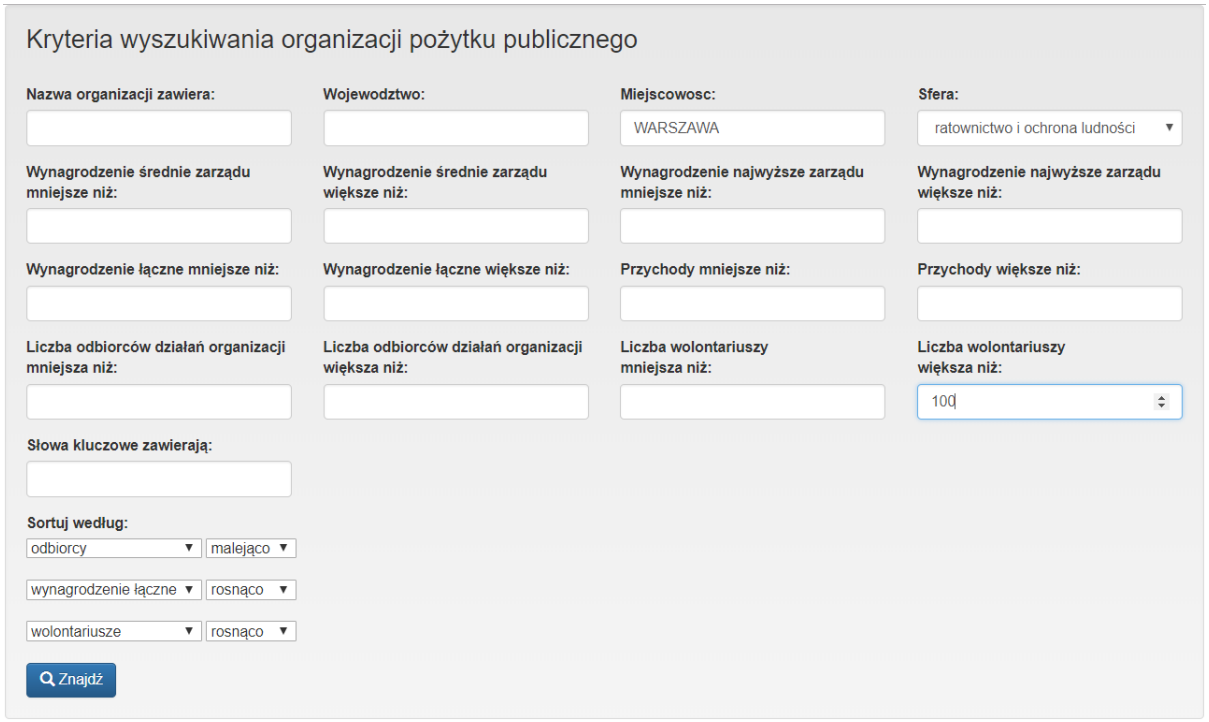
Na rysunku 3.16 przedstawiono listę przefiltrowanych organizacji.

Nazwa	Województwo	Miejscowość	Sfera	Najwyższe wynagrodzenie wypłacone członkom zarządu	Średnie wynagrodzenie wypłacone członkom zarządu	Łączne wynagrodzenie wypłacone w roku sprawozdawczym	Przychody	Wolontariusze	Odbiorcy działań organizacji
POLSKIE STOWARZYSZENIE NA RZECZ DZIECI CHORYCH NA PADACZKĘ	MAZOWIECKIE	WARSZAWA	działalność na rzecz osób niepełnosprawnych	0	0.0	4400	80861.0	0	50000
POLSKIE STOWARZYSZENIE POMOCY CHORYM NA FENYLOKETONURIĘ I CHOROBY RZADKIE ARS VIVENDI, SKRÓCONA NAZWA ARS VIVENDI	MAZOWIECKIE	RASZYN	działalność na rzecz osób niepełnosprawnych	0	0.0	99775	1160090.0	20	27769
TOWARZYSTWO PRZYJACIÓŁ DZIECI ZARZĄD MAZOWIECKIEGO ODDZIAŁU WOJEWÓDZKIEGO	MAZOWIECKIE	WARSZAWA	działalność na rzecz osób niepełnosprawnych	0	0.0	6371784	13271200.0	89	13463
FUNDACJA SYNAPSIS	MAZOWIECKIE	WARSZAWA	działalność na rzecz osób niepełnosprawnych	0	0.0	4882896	8006650.0	11	12500
STOWARZYSZENIE WOLONTARIUSZY NA RZECZ POMOCY DZIECIOM I MŁODZIEŻY SERCE-SERCU	MAZOWIECKIE	DOBACZEWO	działalność na rzecz osób niepełnosprawnych	0	0.0	430388	8697.0	0	10000
STOWARZYSZENIE DOMU DZIECKA-POMNIKA IM. DZIECI ZAMOJSZCZYZNY	MAZOWIECKIE	SIEDLCE	działalność na rzecz osób niepełnosprawnych	0	0.0	135422	381222.0	0	7000

Rysunek 3.16: Lista przefiltrowanych organizacji

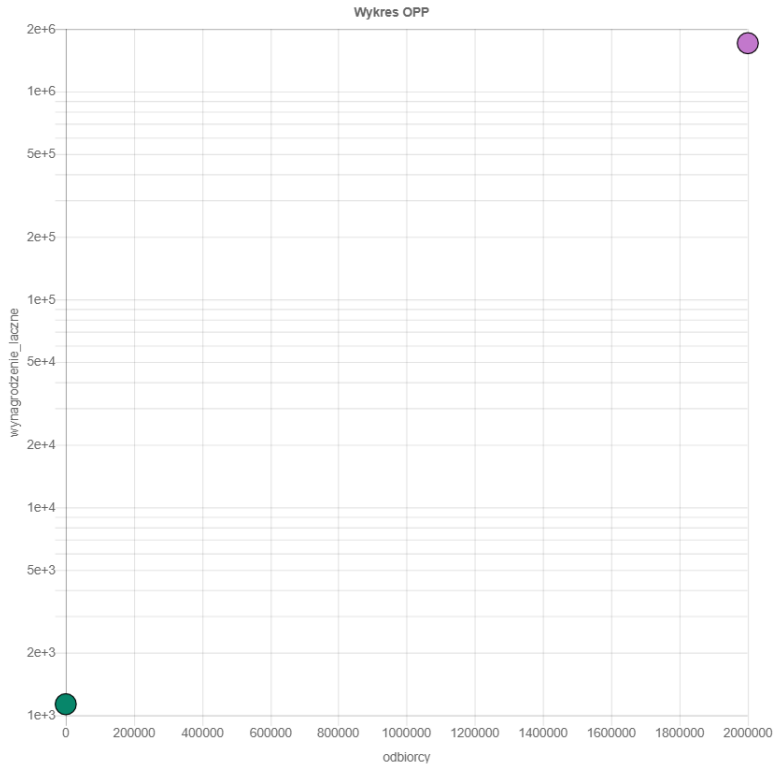
Przykład 3

Użytkownik wybiera sferę działalności organizacji: *ratownictwo i ochrona ludności*, miejscowość: *Warszawa*, liczbę wolontariuszy większą niż: *100*. Sortuje według: *odbiorcy (malejąco), wynagrodzenie łączne (rosnąco), wolontariusze (rosnąco)*. Wypełnione kryteria użytkownika zostały przedstawione na rysunku 3.17.



Rysunek 3.17: Kryteria użytkownika

Na rysunku 3.18 przedstawiono wyświetlający się wykres organizacji po uzupełnieniu kryteriów użytkownika.



Rysunek 3.18: Wykres zawierający przefiltrowane organizacje

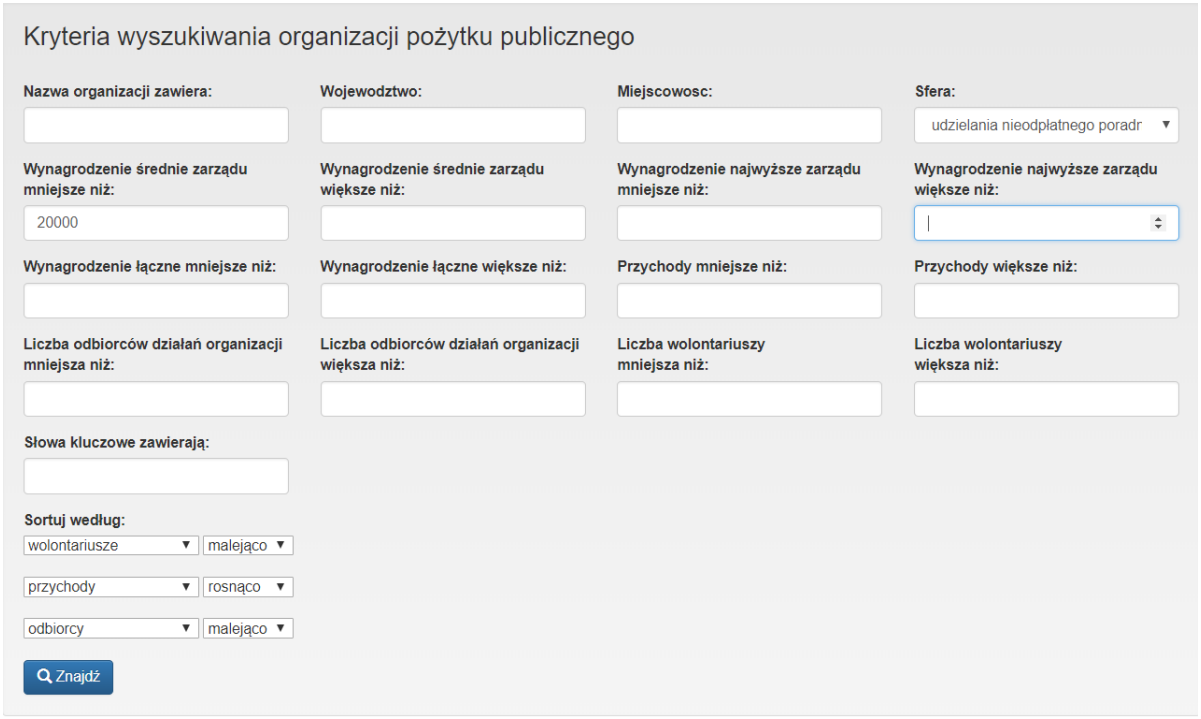
Na rysunku 3.19 przedstawiono listę przefiltrowanych organizacji.

Nazwa	Województwo	Miejscowość	Sfera	Najwyższe wynagrodzenie wypłacone członkom zarządu	Średnie wynagrodzenie wypłacone członkom zarządu	Łączne wynagrodzenie wypłacone w roku sprawozdawczym	Przychody	Wolontariusze	Odbiorcy działań organizacji
STOŁECZNE WODNE OCHOTNICZE POGOTOWIE RATUNKOWE	MAZOWIECKIE	WARSZAWA	ratownictwo i ochrona ludności	0	0.0	1706272	2661090.0	320	2000000
OCHOTNICZA STRAŻ POŻARNA URSUS	MAZOWIECKIE	WARSZAWA	ratownictwo i ochrona ludności	0	0.0	1130	141510.0	3000	0

Rysunek 3.19: Lista przefiltrowanych organizacji

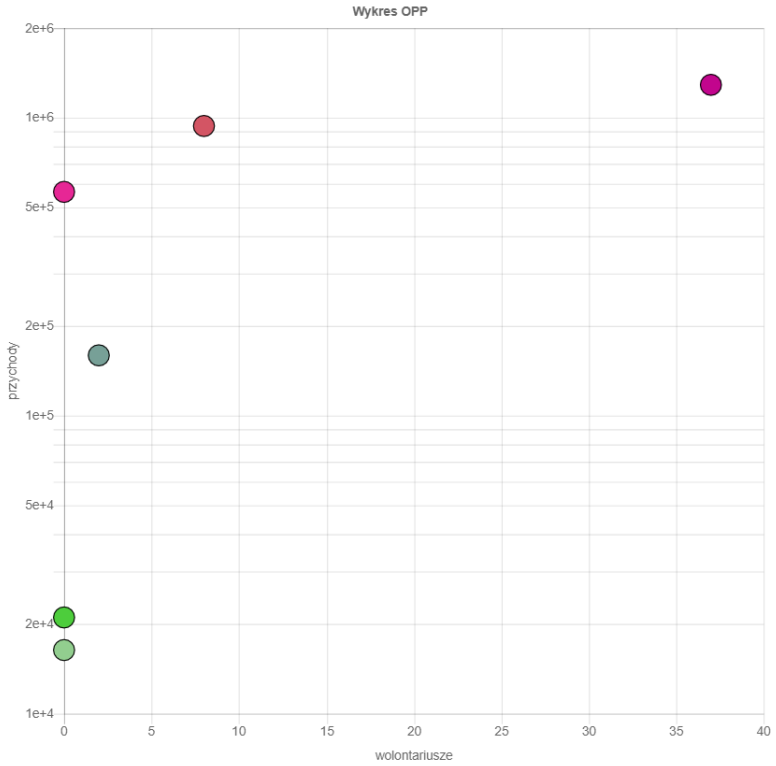
Przykład 4

Użytkownik wybiera sferę działalności organizacji: *udzielania nieodpłatnego poradnictwa obywatelskiego*, wynagrodzenie średnie zarządu mniejsze niż: *20000*. Sortuje według: *wolontariusze(malejąco), przychody(rosnąco), odbiorcy(malejąco)*. Wypełnione kryteria użytkownika zostały przedstawione na rysunku 3.20.



Rysunek 3.20: Kryteria użytkownika

Na rysunku 3.21 przedstawiono wyświetlający się wykres organizacji po uzupełnieniu kryteriów użytkownika.



Rysunek 3.21: Wykres zawierający przefiltrowane organizacje

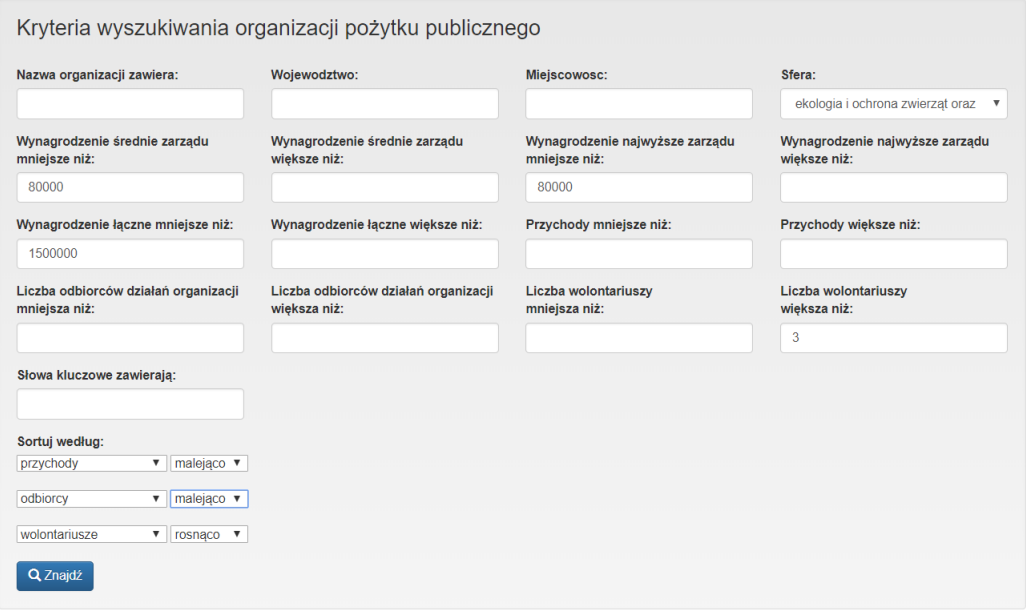
Na rysunku 3.22 przedstawiono listę przefiltrowanych organizacji.

Nazwa	Województwo	Miejscowość	Sfera	Najwyższe wynagrodzenie wypłacone członkom zarządu	Średnie wynagrodzenie wypłacone członkom zarządu	Łączne wynagrodzenie wypłacone w roku sprawozdawczym	Przychody	Wolontariusze	Odbiorcy działań organizacji
FUNDACJA FORANI	MAZOWIECKIE	WARSZAWA	udzielania nieodpłatnego poradnictwa obywatelskiego	953	0.0	170012	1288910.0	37	0
STOWARZYSZENIE WOLNEGO SŁOWA	MAZOWIECKIE	WARSZAWA	udzielania nieodpłatnego poradnictwa obywatelskiego	0	0.0	430388	937903.0	8	300000
STOWARZYSZENIE POMOCNA DŁOŃ	MAZOWIECKIE	GARWOLIN	udzielania nieodpłatnego poradnictwa obywatelskiego	0	0.0	134910	159382.0	2	1432
SISKOM - STOWARZYSZENIE INTEGRACJI STOŁECZNEJ KOMUNIKACJI	MAZOWIECKIE	WARSZAWA	udzielania nieodpłatnego poradnictwa obywatelskiego	0	0.0	0	16342.0	0	1500
WYRÓWNYWANIE SZANS - FUNDACJA	MAZOWIECKIE	WARSZAWA	udzielania nieodpłatnego poradnictwa obywatelskiego	0	0.0	0	21027.0	0	180
WSPÓLNOTA ROBOCZA ZWIĄZKÓW ORGANIZACJI SOCJALNYCH	MAZOWIECKIE	WARSZAWA	udzielania nieodpłatnego poradnictwa obywatelskiego	0	0.0	314538	563687.0	0	1000

Rysunek 3.22: Lista przefiltrowanych organizacji

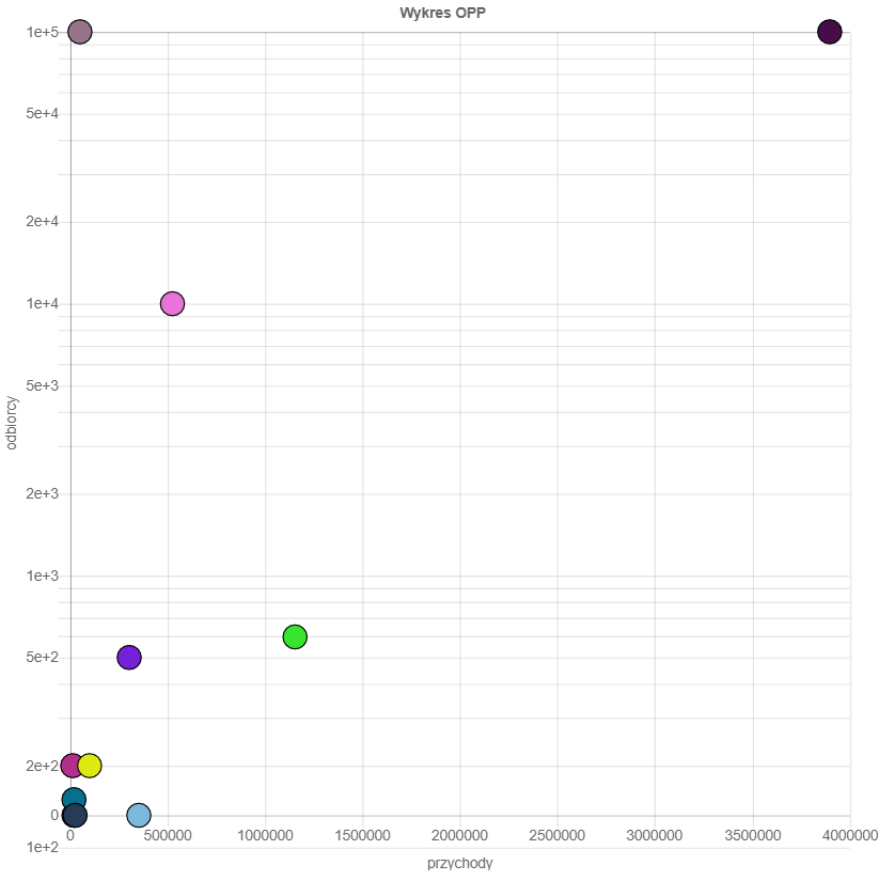
Przykład 5

Użytkownik wybiera sferę działalności organizacji: *ekologia i ochrona zwierząt oraz dziedzictwa przyrodniczego*, wynagrodzenie średnie zarządu mniejsze niż: *80000*, wynagrodzenie najwyższe zarządu mniejsze niż: *80000*, wynagrodzenie łączne mniejsze niż: *1500000*, liczba wolontariuszy większa niż: *3*. Sortuje według: *przychody(malejąco), odbiorcy(malejąco), wolontariusze(rosnąco)*. Wypełnione kryteria użytkownika zostały przedstawione na rysunku 3.23.



Rysunek 3.23: Kryteria użytkownika

Na rysunku 3.24 przedstawiono wyświetlający się wykres organizacji po uzupełnieniu kryteriów użytkownika.



Rysunek 3.24: Wykres zawierający przefiltrowane organizacje

Na rysunku 3.25 przedstawiono listę przefiltrowanych organizacji.

Nazwa	Województwo	Miejscowość	Sfera	Najwyższe wynagrodzenie wypłacone członkom zarządu	Średnie wynagrodzenie wypłacone członkom zarządu	Łączne wynagrodzenie wypłacone w roku sprawozdawczym	Przychody	Wolontariusze	Odbiorcy działań organizacji
LIGA OCHRONY PRZYRODY	MAZOWIECKIE	WARSZAWA	ekologia i ochrona zwierząt oraz ochrona dziedzictwa przyrodniczego	0	0.0	1198092	3900290.0	435	100000
RADOMSKIE TOWARZYSTWO OPIEKI NAD ZWIERZĘTAMI	MAZOWIECKIE	RADOM	ekologia i ochrona zwierząt oraz ochrona dziedzictwa przyrodniczego	0	0.0	1052	1155290.0	37	595
FUNDACJA DLA RATOWANIA ZWIERZĄT BEZDOMNYCH EMIR	MAZOWIECKIE	ODDZIAŁ	ekologia i ochrona zwierząt oraz ochrona dziedzictwa przyrodniczego	0	0.0	0	525273.0	38	10000
TOWARZYSTWO OPIEKI NAD ZWIERZĘTAMI W POLSCE ODDZIAŁ W SOCHACZEWIE	MAZOWIECKIE	SOCHACZEW	ekologia i ochrona zwierząt oraz ochrona dziedzictwa przyrodniczego	0	0.0	130382	353169.0	5	0
JOKOT	MAZOWIECKIE	WARSZAWA	ekologia i ochrona zwierząt oraz ochrona dziedzictwa	0	0.0	0	303785.0	102	500

Rysunek 3.25: Lista przefiltrowanych organizacji

Testy

Praca była testowana na kilku poziomach. Na etapie pobierania dokumentów PDF ze strony ministerialnej sprawdzono manualnie, czy liczba dokumentów pobranych do folderu zgadza się z liczbą organizacji dostępną na stronie dla województwa mazowieckiego. Sprawdzono także, czy dokumenty zostały pobrane w całości i zapisane pod nazwami organizacji, których dotyczą.

Na etapie konwersji przez czytnik OCR sprawdzono, czy wszystkie dokumenty PDF z folderu zostały przetworzone w całości oraz czy prawidłowo przetworzyły potrzebne dane finansowe. Wykryto sporadyczne błędy w danych finansowych, takie jak pominięcie kropki lub przecinka.

W aplikacji webowej napisano testy, sprawdzające poprawność działania filtrów. Skorzystano z biblioteki TestCase. Na potrzeby testów operujących na bazie danych tworzona jest tymczasowa testowa baza, która nie wprowadza zmian w prawdziwej. Po zakończeniu testów baza tymczasowa jest niszczona.

Zweryfikowano także, czy na wykresie pojawiają się wszystkie przefiltrowane organizacje, które są dostępne w tabelce oraz czy osie wykresu odpowiadają wyborowi użytkownika.

W aplikacji użytkownik wprowadza swoje kryteria do formularza. W takiej sytuacji aplikacja może być podatna na atak SQL injection, polegający na wykonaniu niepożądanych operacji na bazie danych, innych niż przewidziane podczas tworzenia aplikacji (na przykład usunięcie rekordów lub dostęp do informacji, które nie miały być udostępnione). Django dba o bezpieczeństwo. Użycie ORM pozwala na zabezpieczenie przed możliwością SQL injection. Zapytania są konstruowane z użyciem parametryzacji. Kod SQL zapytania jest zdefiniowany oddzielnie i odseparowany od danych użytkownika. [24]

4. Zakończenie

Wnioski

Aplikacja spełnia założenia. Wykres okazał się korzystną formą prezentowania danych i pokazywania różnic między organizacjami. Najsłabszym ogniwem pracy okazał się czytnik OCR (ze względu na sporadyczne błędy i długi czas przetwarzania). Rozwiązaniem tego problemu mógłby być wybór innego, płatnego czytnika. Możliwe także, że w przyszłości wprowadzone zostaną ulepszenia w czytniku zastosowanym w ramach pracy inżynierskiej, które pozwolą na uniknięcie błędów, utrudniających pracę z danymi finansowymi. Jeśli w przyszłości zostanie zmieniona forma dokumentów PDF, może okazać się, że z pomocą bibliotek dostępnych w językach programowania będzie możliwe odczytanie danych finansowych, bez konieczności użycia czytnika OCR.

Framework Django dobrze sprawdził się przy tworzeniu aplikacji. Podczas używania aplikacji można zauważyć, że organizacje znacząco różnią się pod względem wybranych cech. Dodatkowo, znaleziono błędy w sprawozdaniach merytorycznych, leżące po stronie organizacji. Na przykład wynagrodzenie średnie nie zostało podzielone przez liczbę 12 (odpowiadającą liczbie miesięcy), przez co kwota wpisana przez organizację w tym polu była znacznie wyższa od poprawnej. Aplikacja wykonana w ramach pracy inżynierskiej mogłaby okazać się pomocnym narzędziem dla organizacji pożytku publicznego do weryfikacji, czy poprawnie uzupełniły sprawozdania. Na błąd organizacji może wskazywać wyższa kwota wynagrodzenia średniego członków zarządu od kwoty wynagrodzenia najwyższego członków zarządu.

Możliwości rozwoju

Aplikacja może zostać rozszerzona o ilość informacji finansowych, dotyczących organizacji, na przykład liczbę pracowników, wynagrodzenia pracowników i konsekwentnie zwiększyć ilość kryteriów wyszukiwania dla użytkownika. Można także zwiększyć zakres aplikacji do organizacji z całej Polski, nie tylko województwa mazowieckiego. Na wykresie bąbelkowym jest także

możliwość prezentowania trzeciego wymiaru (wielkość bąbelka), odpowiadającego trzeciej z wybranych przez użytkownika cech. W pracy inżynierskiej trzeci wymiar nie został wykorzystany z powodu zaciemnienia wykresu w przypadku dużej ilości organizacji, które się na nim wyświetlały. W takiej sytuacji ciężej było odczytywać dane z wykresu.

Bibliografia

[1] Co to są organizacje pożytku publicznego?
http://poradnik.ngo.pl/co-to-sa-organizacje-pozytku-publicznego
(data dostępu: 4.09.2018r.)

[2] Organizacja pożytku publicznego (OPP)
https://pl.wikipedia.org/wiki/Organizacja_pożytku_publicznego
(data dostępu: 4.09.2018r.)

[3] Fakty o NGO, Organizacje pożytku publicznego (OPP)
http://fakty.ngo.pl/opp
(data dostępu: 4.09.2018r.)

[4] Selenium
https://www.techbeamers.com/selenium-webdriver-python-tutorial/
(data dostępu: 4.09.2018r.)

[5] Pierwsze kroki z Selenium – Selenium IDE
https://www.kainos.pl/blog/pierwsze-kroki-z-selenium-selenium-ide/
(data dostępu: 4.09.2018r.)

[6] Optyczne rozpoznawanie znaków
https://pl.wikipedia.org/wiki/Optyczne_rozpoznawanie_znaków
(data dostępu: 4.09.2018r.)

[7] Optyczne rozpoznawanie znaków, Zasada działania
https://pl.wikipedia.org/wiki/Optyczne_rozpoznawanie_znaków
(data dostępu: 4.09.2018r.)

[8] Dygitalizacja
https://pl.wikipedia.org/wiki/Digitalizacja_(bibliotekarstwo)
(data dostępu: 4.09.2018r.)

[9] Przetwarzanie języka naturalnego
https://pl.wikipedia.org/wiki/Przetwarzanie_języka_

naturalnego
(data dostępu: 4.09.2018r.)

[10] Forma wyrazowa, Wyraz słownikowy
http://hamlet.edu.pl/shi/uczen/?id=wyrazy
(data dostępu: 4.09.2018r.)

[11] Morfeusz
http://sgjp.pl/morfeusz/morfeusz.html
(data dostępu: 4.09.2018r.)

[12] MySQL
https://pl.wikipedia.org/wiki/MySQL
(data dostępu: 4.09.2018r.)

[13] Czym jest Django?, Co się dzieje, gdy ktoś chce otworzyć stronę z Twojego serwera?
https://tutorial.djangogirls.org/pl/django/
(data dostępu: 4.09.2018r.)

[14] Dokumentacja Django
https://docs.djangoproject.com/pl/2.1/
(data dostępu: 4.09.2018r.)

[15] Mapowanie obiektowo-relacyjne
https://pl.wikipedia.org/wiki/Mapowanie_obiektowo-relacyjne
(data dostępu: 4.09.2018r.)

[16] Django Tutorial Part 5: Creating our home page
https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Home_page
(data dostępu: 4.09.2018r.)

[17] Serwer deweloperski
https://docs.djangoproject.com/pl/2.1/intro/tutorial01/
(data dostępu: 4.09.2018r.)

[18] Filozofie projektowe
https://docs.djangoproject.com/pl/2.1/misc/design-philosophies/
(data dostępu: 4.09.2018r.)

[19] Regular expressions
https://www.regular-expressions.info/
(data dostępu: 4.09.2018r.)

[20] Wyrazenia regularne
https://www.learnpython.org/pl/Wyrazenia_regularne
(data dostępu: 4.09.2018r.)

[21] Natural Language Toolkit
https://www.nltk.org/
(data dostępu: 4.09.2018r.)

[22] Natural Language Pricessing is fun
https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e
(data dostępu: 4.09.2018r.)

[23] Chart.js
https://www.chartjs.org/docs/latest/
(data dostępu: 4.09.2018r.)

[24] SQL injection protection
https://docs.djangoproject.com/en/2.1/topics/security/
(data dostępu: 4.09.2018r.)