

”Automatyczna analiza rynku pracy.”

Bartłomiej Sielicki

Łukasz Skarżyński

Praca inżynierska
wersja 0.1

Promotor:
Barbara Rychalska



Wydział Matematyki i Nauk Informatycznych
Politechniki Warszawskiej

23.11.2017

Streszczenie

AUTOMATYCZNA ANALIZA RYNKU PRACY IT

Rynek pracy w branży IT jest bardzo dynamiczny. Szczególnie jest to zauważalne w ostatnich latach, kiedy to nieustannie zauważamy wzajemne wypieranie się technologii. Na zmiany reagują wszyscy - począwszy od firm, które aby dorównać konkurencji stale wdrażają nowe rozwiązania, przez pracowników, chcących nieustannie poszerzać własne kompetencje, po uczelnie wyższe, które starają się przekazać swoim przyszłym absolwentom jak najbardziej aktualną wiedzę i umiejętności. Istotną staje się zatem możliwość wglądu w trendy i oparta na niej analiza zachodzących zmian.

Aplikacja której poświęcona jest ta praca jest próbą udowodnienia, że proces taki da się zautomatyzować i bez wymogu nadmiernej ingerencji użytkownika zbierać, analizować, wizualizować oraz udostępniać najbardziej istotne (z punktu widzenia autorów) informacje. W ramach pracy wykonano system informatyczny składający się z dwóch komponentów. Pierwszym z nich jest zaplecze analityczne (back end), zajmujące się automatycznym pobieraniem zamieszczanych na jednym z największych polskich serwisów rekrutacyjnych (Pracuj.pl) ofert z branży IT oraz ich późniejszą analizą pod kątem umieszczanych w nich technologii. Drugim wykonanym komponentem jest aplikacja WWW prezentująca wyniki użytkownikowi oraz umożliwiająca eksport zebranych danych do samodzielnej analizy.

Słowa kluczowe: *Rynek pracy IT, Automatyzacja, Web Scraping*

Abstract

AUTOMATED JOB MARKET ANALYSIS

Streszczenie po angielsku.

Spis treści

Wstęp	5
1 Specyfikacja projektu	6
1.1 Słownik pojęć	6
1.2 Opis biznesowy	7
1.3 Wymagania funkcjonalne	8
1.3.1 Wyszukiwanie ofert wg klucza	8
1.3.2 Wyświetlenie oferty	8
1.3.3 Wyświetlenie statystyk serwisu	9
1.3.4 Wyświetlenie statystyk klucza	9
1.4 Wymaganie нефункционалне	11
1.5 Schemat architektury	11
1.5.1 Moduł zbierający dane	12
1.5.2 Pipeline	12
1.5.3 Backend	13
1.5.4 Aplikacja WWW	13
1.6 Model wytwórczy	15
1.7 Harmonogram	16
2 Specyfikacja komponentów systemu	17
2.1 Moduł zbierający dane	17
2.1.1 Serwis zewnętrzny	17
2.1.2 Napotkane problemy	22
2.1.2.1 Aktualizacja ofert w serwisie	22
2.1.2.2 Utrzymywanie stanu scrapera	22
2.1.2.3 Zabezpieczenia serwisu przed zbieraniem danych	22
2.1.3 Scrapy	23
2.1.4 Uruchomienie i użytkowanie	23
2.1.5 Struktura kodu źródłowego	26
2.1.6 Główne klasy modułu	26
2.1.7 Testy	27
2.1.7.1 Testy jednostkowe	27

	2.1.7.2	Testy akceptacyjne	28
2.2		Moduł przechowujący dane	31
	2.2.1	Wymagania	31
	2.2.2	Interfejs	32
	2.2.3	Baza danych	32
	2.2.4	Integracja z kolejnym modulem	33
	2.2.5	Uruchomienie	33
	2.2.6	Struktura kodu	33
	2.2.7	Główne klasy modułu	34
	2.2.8	Testy	34
	2.2.8.1	Testy jednostkowe	34
	2.2.8.2	Testy akceptacyjne	35
2.3		Moduł ekstrakcji kluczy	38
	2.3.1	Przeznaczenie modułu	38
	2.3.2	Architektura i algorytm	38
	2.3.2.1	Rozpoznawanie umiejętności i technologii	39
	2.3.2.2	Uzasadnienie wyboru powyżej opisanego algorytmu	40
	2.3.2.3	Znajdowanie podobnych kluczy za pomocą modelu Word2Vec	42
	2.3.3	Celery	43
	2.3.4	Struktura kodu	44
	2.3.5	Główne klasy modułu	45
	2.3.6	Testy	45
	2.3.6.1	Testy jednostkowe	45
	2.3.6.2	Testy akceptacyjne	46
2.4		Moduł statystyk i wyszukiwarki	49
	2.4.1	Wymagania	49
	2.4.2	Interfejs	50
	2.4.3	Baza danych	52
	2.4.3.1	Indeksy	52
	2.4.4	Uruchomienie	52
	2.4.5	Struktura kodu	53
	2.4.6	Główne klasy modułu	53
	2.4.7	Algorytm generowania statystyk	54
	2.4.8	Testy	55
	2.4.8.1	Testy jednostkowe	55
	2.4.8.2	Testy akceptacyjne	55
2.5		Aplikacja webowa	58
	2.5.1	Wymagania	58
	2.5.2	VueJS	58

2.5.2.1	Dodatkowe biblioteki	59
2.5.3	Uruchomienie	59
2.5.3.1	Wersja developerska	59
2.5.4	Struktura kodu	59
2.5.5	Główne komponenty aplikacji	60
2.5.6	Instrukcja użytkowania oraz testy akceptacyjne	61
2.5.6.1	Zakładka statystyk	61
2.5.6.2	Zakładka wyszukiwania	61
2.5.6.3	Zakładka informacji	62
Historia zmian dokumentu		65
Bibliografia		66

Wstep

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam et turpis gravida, lacinia ante sit amet, sollicitudin erat. Aliquam efficitur vehicula leo sed condimentum. Phasellus lobortis eros vitae rutrum egestas. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec at urna imperdiet, vulputate orci eu, sollicitudin leo. Donec nec dui sagittis, malesuada erat eget, vulputate tellus. Nam ullamcorper efficitur iaculis. Mauris eu vehicula nibh. In lectus turpis, tempor at felis a, egestas fermentum massa.

Rozdział 1

Specyfikacja projektu

1.1 Słownik pojęć

- **back-end** - Odpowiada za operacje w tle, których przebiegu użytkownik nie widzi bezpośrednio. Zajmuje się przetwarzaniem, wykonywaniem zadań na podstawie otrzymanych danych. W projekcie słowem back-end określamy wszystkie moduły za wyjątkiem aplikacji internetowej z której korzysta użytkownik.
- **front-end** - Warstwa wizualna systemu - interfejs użytkownika. Głównym jego zadaniem jest pobieranie danych od użytkownika oraz przekazywanie ich do back-endu oraz ewentualne pokazanie odebranej odpowiedzi. W systemie którego dotyczy dana dokumentacja front-end jest stroną internetową.
- **web scraper** - program, którego głównym zadaniem jest zbierać określone dane ze stron internetowych. Gdy w dokumentacji używamy słowa “scraper” zawsze odnosi się ono właśnie do “web scraper’a”.
- **pipeline** - łańcuch przetwarzania danych (funkcji), w którym wejściem kolejnego etapu jest wyjście poprzedniego.
- **framework** - szkielet do budowy różnego rodzaju aplikacji. Definiuje on strukturę aplikacji oraz ogólny mechanizm jej działania, oraz dostarcza zestaw komponentów i bibliotek ogólnego przeznaczenia do wykonywania określonych zadań.
- **API** - (ang. Application Programming Interface) interfejs programowania aplikacji, jest to ściśle określony zestaw reguł i ich opisów, w jaki programy komputerowe komunikują się między sobą. Najczęściej używamy tego słowa w odniesieniu do WEB API - interfejsu komunikacji korzystającego z HTTP oraz formatu JSON do przesyłania danych.

1.2 Opis biznesowy

Głównym zadaniem aplikacji ma być automatyczna analiza rynku pracy. System zajmuje się agregacją oraz przetwarzaniem (analizą) ofert zebranych z serwisu zewnętrznego (Pracuj.pl) i przedstawianiem wyników użytkownikowi.

W szczególności proces działania systemu sprowadza się do:

1. Pobierania ofert pracy z serwisu zewnętrznego wyłuskując kluczowe elementy: tytuł ogłoszenia, opis, datę dodania, itp.
2. Zapisania tak zebranych ogłoszeń do własnej bazy danych
3. Przeprowadzenia analizy na treści ogłoszenia uzyskując zeń listę *kluczy*, tj:
 - obszaru branży IT którego dotyczy ogłoszenie
 - stanowiska którego dotyczy
 - wymaganych od pracownika opanowanych technologii / umiejętności
4. Udostępnienia użytkownikowi interfejsu do wyszukiwania zebranych w systemie ofert oraz wyświetlania statystyk przy użyciu wyżej wspomnianych kluczy.

Wymienione działania realizują odrębne komponenty systemu. Bardziej szczegółowy podział i opis znajduje się w sekcji wymagań funkcjonalnych oraz rozdziale drugim będącym dokumentacją każdego z modułów.

Użytkownik bezpośrednio będzie korzystał tylko z ostatniego modułu - aplikacji WWW będącej interfejsem dla zebranych i przetworzonych przez system danych.

Aplikacja będzie niosła korzyść dużej grupie odbiorców takich jak:

- studenci - dzięki aplikacji będą mogli znaleźć świetnie dopasowane stanowisko do ich umiejętności lub obserwując dane stanowisko określić jakie umiejętności są na nim elementarne, a także jakiej technologii powinni się nauczyć w najbliższym czasie.
- osoby szukające pracy - łatwo na podstawie swoich umiejętności znajdą stanowiska dla siebie.
- pracodawcy - na podstawie trendów wśród używanych technologii będą mogli podjąć decyzje dotyczące przyszłych projektów.
- pozostali użytkownicy zainteresowani analizą rynku pracy.

1.3 Wymagania funkcjonalne

Podstawą interakcji użytkownika jest strona WWW - użytkownik powinien być w stanie otworzyć ją na dowolnym komputerze z dostępem do internetu i wyposażonym w aktualną wersję jednej z wiodących na rynku przeglądarek.

- Google Chrome w wersji 49 lub wyższej
- Mozilla Firefox w wersji 52 lub wyższej
- Safari w wersji 10.1 lub wyższej

Posiadanie przeglądarki innej niż wymienione, lub w starszej wersji nie oznacza że strona nie będzie działać, jednak jako twórcy nie możemy zagwarantować że będzie to działanie w pełni poprawne.

Na stronie nie przewidujemy kont użytkowników, nawet administracyjnego. Każdy wchodzący na stronę będzie miał dostęp do tych samych danych oraz takie same możliwości.

Funkcjonalność strony rozbita jest na dwa moduły. Moduł wyszukiwarki oraz moduł statystyk. Możliwości użytkownika prezentuje załączony na końcu sekcji diagram przypadków użycia.

1.3.1 WYSZUKIWANIE OFERT WG KLUCZA

Jednym z podstawowych zastosowań strony jest wyszukiwanie ofert zebranych i umieszczonych w bazie przez nasz system. Jako kryterium wyszukiwania (przez mechanizm filtrowania) może zostać użyty tzw. *klucz*, czyli wyłuskana z opisu ogłoszenia jego cecha. Klucze dzielimy na trzy kategorie:

- **Obszary** (np. Mobile development, Helpdesk)
- **Stanowiska** (np. Software Developer, Data Scientist)
- **Technologie i umiejętności** (np. Java, Docker, AWS)

Pozwoli to na kompleksowe wyszukiwanie ofert ze względu na branżę czy pozycję którą interesuje się użytkownik oraz posiadane przez niego umiejętności. Możliwe jest podanie wielu kluczy jako kryterium.

1.3.2 WYŚWIETLENIE OFERTY

Wynikiem wyszukiwania jest lista ofert. Każdą z nich użytkownik może wyświetlić uzyskując dostęp do takich informacji jak:

- Tytuł oferty
- Data dodania i wygaśnięcia oferty w macierzystym serwisie
- Nazwa pracodawcy i miejsce pracy
- Wszystkie wyłuskane przez system klucze
- Odnośnik do oryginalnego ogłoszenia w macierzystym serwisie

1.3.3 WYŚWIETLENIE STATYSTYK SERWISU

W osobnej sekcji użytkownik ma dostęp do wyświetlenia zbiorczych statystyk dotyczących serwisu i dostępnych w nim danych. Zbiór dostępnych statystyk planowo zostanie rozwinięty podczas pracy nad ostatnimi dwoma modułami systemu. Bazowe, przewidziane już teraz to:

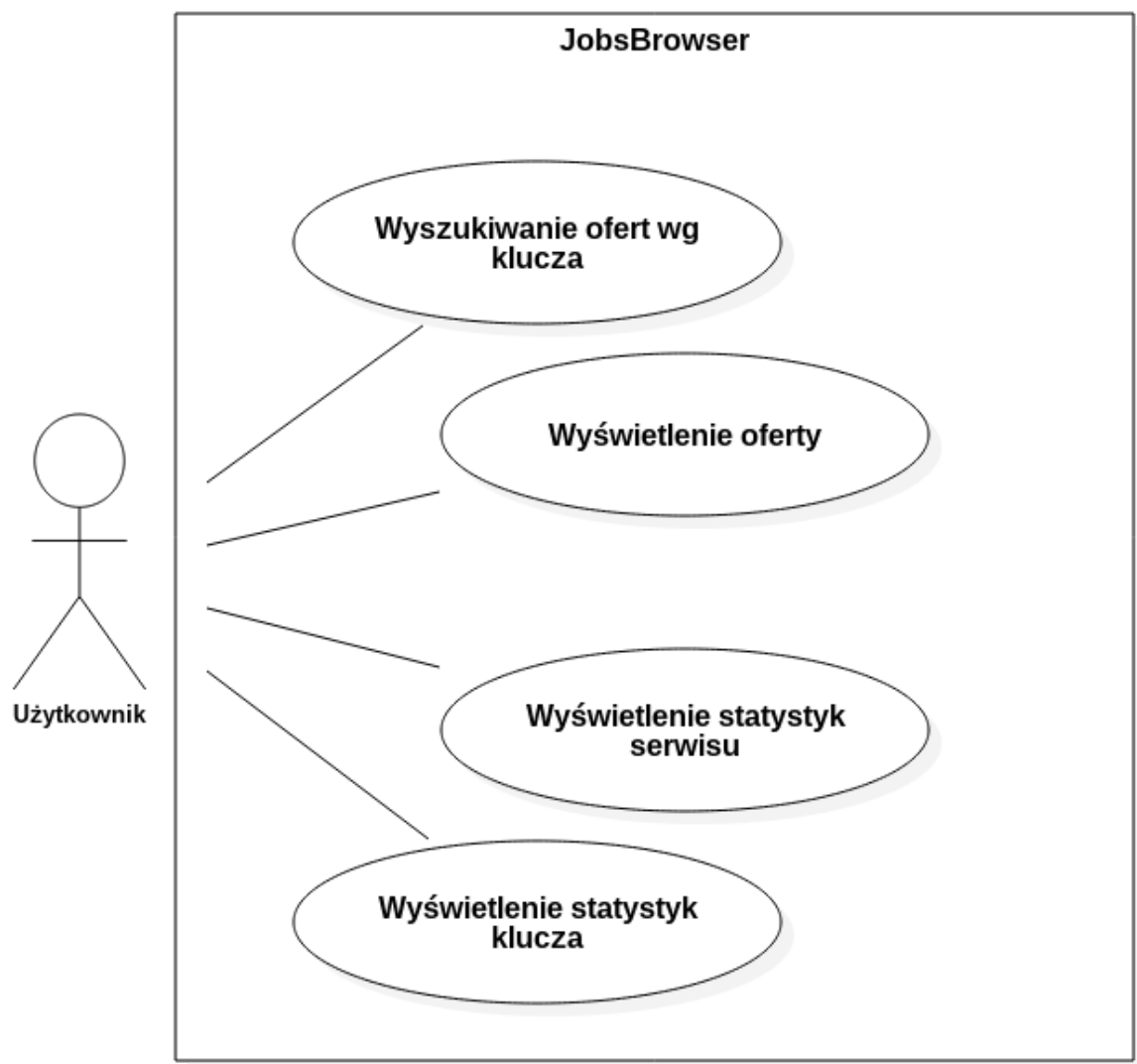
- Liczba wszystkich ofert w bazie systemu
- Wykres powyższej wartości względem czasu
- Datę ostatniego zebrania danych z serwisu macierzystego
- Listę wszystkich istniejących w systemie kluczy

1.3.4 WYŚWIETLENIE STATYSTYK KLUCZA

Jest to druga obok wyszukiwania podstawowa funkcjonalność strony. Pozwala ona użytkownikowi na wybór klucza oraz wyświetlenie:

- Wykresu ilości ofert zawierających ten klucz względem czasu
- Wykresu procentowej ilości ofert z systemu zawierających ten klucz
- Pogrupowanych w kategorie kluczy które występują z tym kluczem najczęściej w jednym ogłoszeniu

Możliwe jest podanie wielu kluczy. Wtedy pod uwagę przy generowaniu powyższych statystyk będą brane tylko ogłoszenia które zawierają każdy z nich.



Rysunek 1.1: Przypadki użycia.

1.4 Wymaganie niefunkcjonalne

W poniższej tabeli przedstawione zostały wymagania niefunkcjonalne tworzonej aplikacji.

Tabela 1.1: Wymagania niefunkcjonalne

Obszar wymagań	Opis
Używalność (Usability)	Wymagany dostęp do internetu w celu skorzystania z aplikacji. Aplikacja WWW jest intuicyjna w obsłudze dla użytkownika. Aplikacja WWW powinna być dostępna dla użytkownika w każdym wybranym dla niego momencie.
Niezawodność (Reliability)	Dostęp do aplikacji WWW powinien być możliwy przez 24 godziny, 7 dni w tygodniu. Za wyjątkiem prac serwisowych nie dłuższych niż 2 h w tygodniu przy założeniu stabilnego połączenia internetowego. Pozostałe moduły powinny działać bez problemów na oddzielnych serwerach.
Wydajność (Performance)	Aplikacja WWW powinna działać płynnie na każdym komputerze z dowolnym systemem operacyjnym na 1 z wcześniej wymienionych przeglądarek. Pozostałe moduły powinny uruchamiać się same co określony odstęp czasu. Początkowo planowane jest uruchamianie ich raz dziennie.
Wsparcie (Supportability)	W razie jakichkolwiek problemów w aplikacji WWW dostępny jest formularz kontaktowy.

1.5 Schemat architektury

Przewidziana architektura ma strukturę modułową. Schemat komponentów oraz ich połączenie przedstawia załączony na końcu sekcji diagram.

W systemie wyróżnić możemy cztery główne, niezależne od siebie (na tyle że mogą, a nawet powinny, być uruchamiane na różnych maszynach) moduły. Ta sekcja zawiera ich ogólny, wysokopoziomowy opis. Szczegóły dotyczące architektury i implementacji znajdują się w następnym rozdziale.

1.5.1 MODUŁ ZBIERAJĄCY DANE

Komponent ten odpowiada za automatyczne pobieranie ogłoszeń z serwisu zewnętrznego. Jest to program, który cyklicznie łączy się z udostępniającym ogłoszenia serwisem i automatycznie pobiera ich treść. Z pobranych ogłoszeń w najprostszy sposób (poprzez analizę kodu HTML) wyłuskuje podstawowe elementy, takie jak:

- Tytuł ogłoszenia
- Treść
- Data dodania
- Pracodawca
- Miejsce pracy

W kolejnym kroku tak “rozbite” ogłoszenie przesyła do kolejnego komponentu systemu.

Opisany proces nie obejmuje zapisu do żadnej bazy danych. W tej kwestii moduł zbierający dane polega całkowicie na komponencie, do którego przekazuje pobrane ogłoszenia. To samo dotyczy kwestii rozpoznawania duplikatów - program przed każdym rozpoczęciem skanowania prosi swojego “odbiorcę” o listę już zeskanowanych ogłoszeń aby uniknąć przetwarzania ich ponownie.

1.5.2 PIPELINE

Zebrane ogłoszenia trafiają do komponentu *Pipeline* (Łańcucha przetwarzania). Nazwa komponentu odnosi się do zasady jego działania. Odbierane ogłoszenia przekazywane są przez kolejne podmoduły, które wykonują na nich stosowne operacje.

Pierwszym elementem jest moduł przechowujący dane. Odbiera on nowe ogłoszenia od modułu zbierającego i zapisuje je w bazie danych. Jednocześnie oferuje usługę odczytania listy kluczy (adresów URL) dodanych już ogłoszeń, z czego moduł zbierający korzysta przed rozpoczęciem skanowania serwisu zewnętrznego.

Następnie ogłoszenia trafiają do modułu ekstrakcji kluczy. Tutaj zgodnie z wymaganiami funkcjonalnymi systemu z ogłoszenia wyłuskane są elementy z trzech kategorii:

- Obszar branży IT którego dotyczy ogłoszenie
- Stanowisko
- Technologie i umiejętności

Po ukończeniu procesu model obiekt ogłoszenia powiększa się o zestaw kluczy, a następnie zostaje wysłany do kolejnego komponentu.

1.5.3 BACKEND

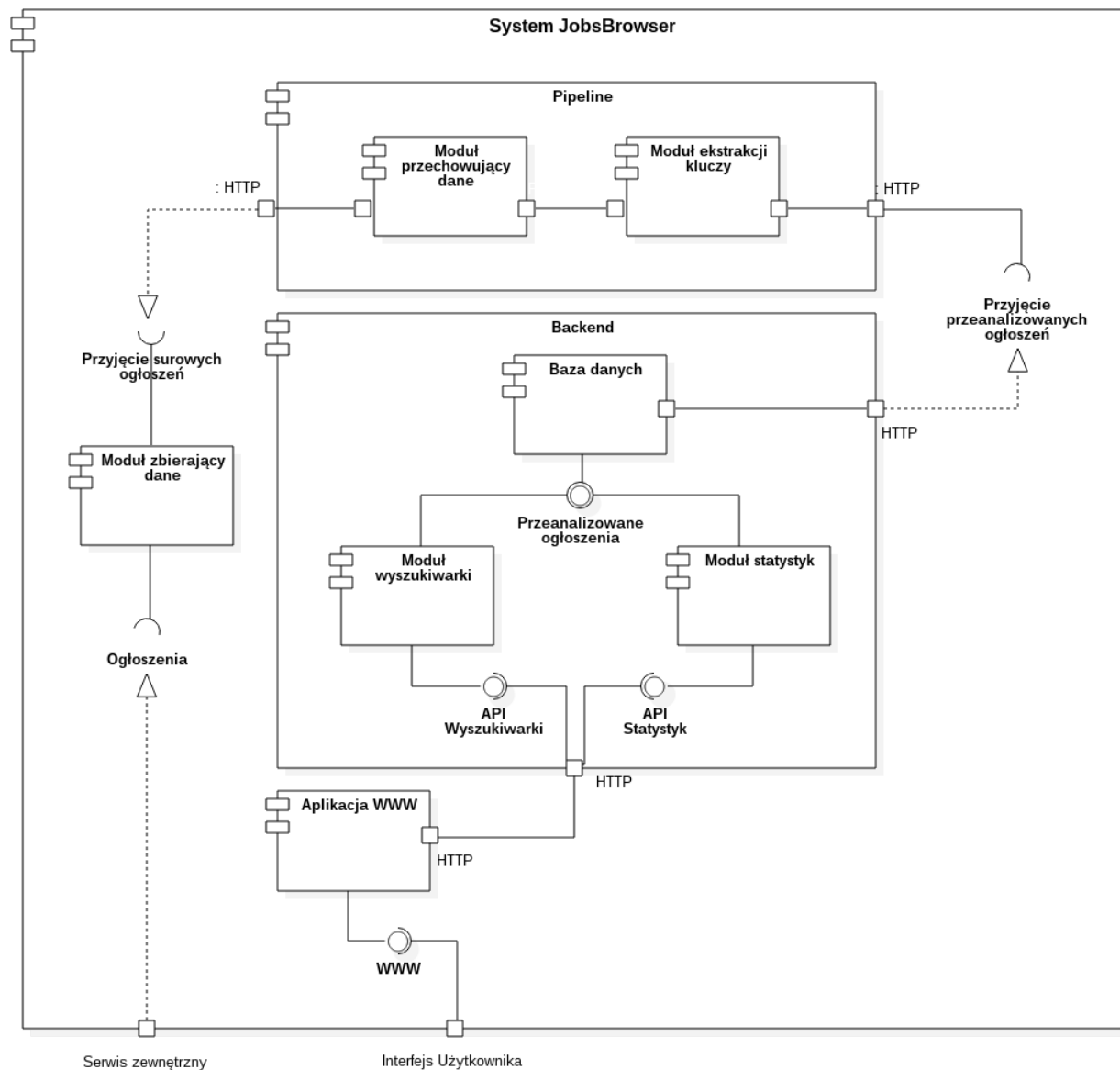
Moduł ten zajmuje się dostarczaniem użytkownikowi (a konkretnie aplikacji WWW) informacji na podstawie przeanalizowanych ofert. Do bazy danych zapisywane są ogłoszenia wraz z kluczami, a podmoduły pozwalają na jej odpytywanie.

Moduł wyszukiwarki wystawia interfejs pozwalający przeglądarce na wykonanie zapytania o ogłoszenia zawierające jednej lub więcej kluczy.

Interfejs modułu statystyk pozwala na dostęp do statystyk systemu jako całości oraz statystyk poszczególnych kluczy. Bardziej szczegółowe wymagania tego interfejsu znajdują się w sekcji wymagań funkcjonalnych.

1.5.4 APLIKACJA WWW

Ostatnim elementem systemu, jedynym dostępnym bezpośrednio dla użytkownika jest aplikacja WWW. Nie posiada ona własnej bazy, a co za idzie kont użytkowników. Korzysta wyłącznie z interfejsu wyszukiwania oraz statystyk pozwalając użytkownikowi na przejrzysty i wygodny dostęp do informacji.



Rysunek 1.2: Diagram komponentów.

1.6 Model wytwórczy

Do wytworzenia opisywanego systemu wybraliśmy metodykę zwinną - Agile. Jest ona jedną z najszybszych metod wytwarzania oprogramowania i idealnie wpasowuje się w modułową strukturę architektury systemu. Wybór tej metodyki oznacza prowadzenie prac w modelu iteracyjnym - komponent po komponencie. W naszym przypadku oznacza to, że z każdą kolejną iteracją gotowy projekt będzie powiększał się o nowy, w pełni działający fragment. Po pierwszym etapie będzie to moduł zbierający dane z zewnętrznych serwisów, następnie moduł przechowujący i udostępniający zebrane dane, i tak dalej.

Wymagania funkcjonalne projektu zostały opracowane w pierwszej kolejności. Pozwoliło to na nakreślenie zarysu systemu i tego na co musi pozwalać, bez wdawania się w szczegóły implementacyjne. Po wyodrębnieniu komponentów systemu, uznaliśmy że podejście kaskadowe byłoby zbyt ryzykowne. Zaprojektowanie dokładnej struktury i szczegółowych wymagań implementacji każdego z nich już na wstępie, niosłoby za sobą ryzyko niedopasowania się do przewidzianego na implementację czasu. Zachodziłoby również niebezpieczeństwo przeoczenia błędów w planowaniu, które wyszłyby na powierzchnię dopiero w momencie implementacji bądź testów, kiedy metoda kaskadowa nie przewiduje już zmiany wymagań ani planu. Jakość końcowego produktu zdecydowanie bardzo by na tym ucierpiała.

Komponenty systemu są na tyle niezależne i wyizolowane, że mając opracowane ogólne ich założenia i przeznaczenie, podczas każdej z iteracji jesteśmy w stanie skupić się na nich indywidualnie, bez ingerowania w resztę projektu. Pozwoli to na dokładne ich dopracowanie, co jest szczególnie istotne biorąc pod uwagę, że część z nich opiera się w znacznym stopniu na zewnętrznych projektach i technologiach, które do udanej integracji będą wymagały głębszego poznania.

1.7 Harmonogram

Załączona tabela 1.2 przedstawia planowany harmonogram postępów w pracy nad projektem - przewidywane terminy ukończenia każdego z modułów. Wytłuszczone daty to terminy laboratoriów na których przewidziana jest kontrola postępów.

Tabela 1.2: Harmonogram

Rozpoczęcie iteracji	Ukończenie iteracji	Opis	Moduły
26.10.2017	02.11.2017	Przygotowanie harmonogramu oraz wstępnych wymagań	
02.11.2017	09.11.2017	Moduł zbierający dane Tydzień 1	
09.11.2017	16.11.2017	Moduł zbierający dane Tydzień 2	1/6
16.11.2017	23.11.2017	Moduł przechowujący dane	2/6
23.11.2017	30.11.2017	Moduł ekstrakcji kluczy Tydzień 1	
30.11.2017	07.12.2017	Moduł ekstrakcji kluczy Tydzień 2	
07.12.2017	14.12.2017	Moduł ekstrakcji kluczy Tydzień 3	3/6
14.12.2017	21.12.2017	Moduł statystyk Tydzień 1	
21.12.2017	28.12.2017	Moduł statystyk Tydzień 2	4/6
28.12.2017	04.01.2018	Moduł wyszukiwarki	5/6
04.01.2018	11.01.2018	Strona WWW	6/6

Rozdział 2

Specyfikacja komponentów systemu

2.1 Moduł zbierający dane

Kod źródłowy znajduje się pod adresem: github.com/jobsbrowser/scrapper.

Komponent zajmuje się automatycznym pobieraniem ofert z serwisu zewnętrznego, ekstrakcją podstawowych informacji oraz przesłaniem tak przetworzonych ofert do kolejnego modułu. Zdecydowaliśmy się na serwis Pracuj.pl jako źródło ofert oraz na framework **Scrapy**(Scrapy n.d.) jako bazę naszego modułu.

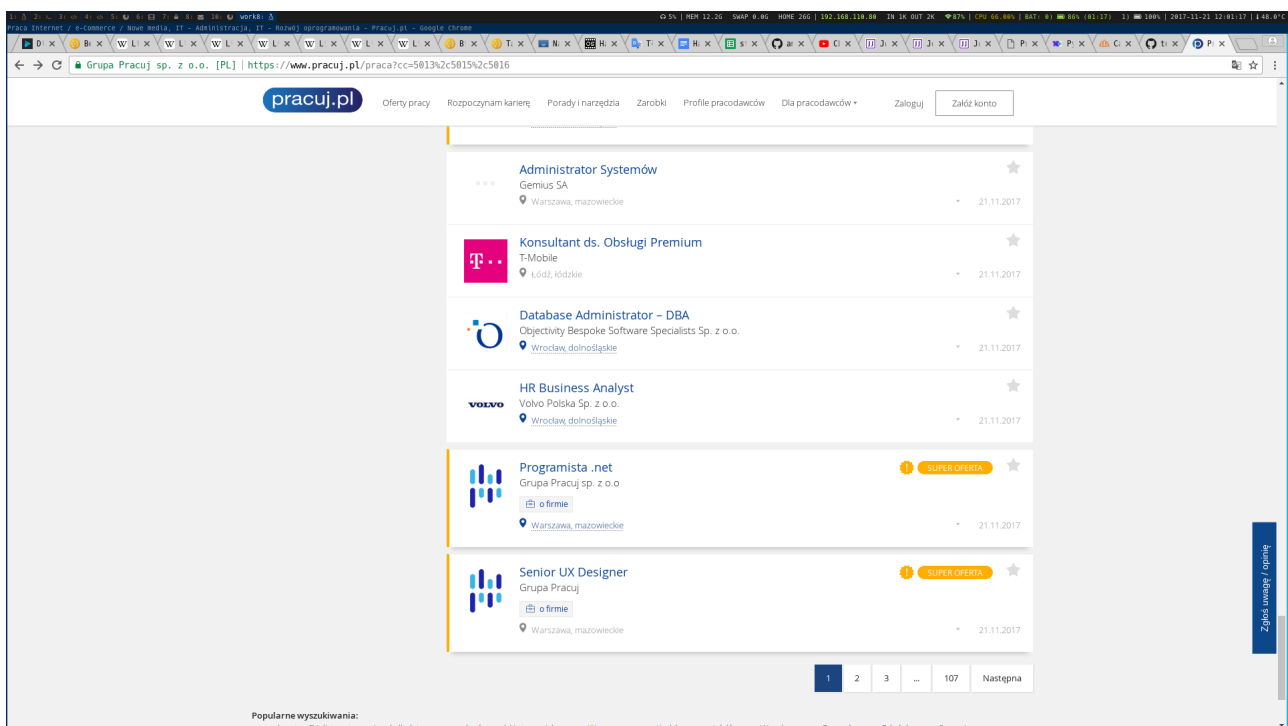
2.1.1 SERWIS ZEWNĘTRZNY

Pracuj.pl dzieli zamieszczone w nim oferty na kategorie. My, z racji tematyki pracy skupiamy się wyłącznie na trzech z nich:

- Internet / e-Commerce / Nowe media
 - E-marketing / SEM / SEO
 - Media społecznościowe
 - Projektowanie
 - Sprzedaż / e-Commerce
 - Tworzenie stron WWW / Technologie internetowe
- IT - Administracja
 - Administrowanie bazami danych i storage
 - Administrowanie sieciami
 - Administrowanie systemami

- Bezpieczeństwo / Audyt
 - Wdrożenia ERP
 - Wsparcie techniczne / Helpdesk
 - Zarządzanie usługami
- IT - Rozwój oprogramowania
 - Analiza biznesowa
 - Architektura
 - Programowanie
 - Testowanie
 - Zarządzanie projektem

Widok listy ogłoszeń w serwisie umożliwia wybór kategorii, z których ogłoszenia chcemy zobaczyć. Skorzystaliśmy z tej możliwości, aby uzyskać bazowy link od którego zaczniemy pobieranie ofert.



Rysunek 2.1: Widok paginacji ogłoszeń w witrynie pracuj.pl.

Pracuj.pl przy zbiorczym wyświetlaniu ofert używa paginacji. Oznacza to, że scraper musi poradzić sobie nie tylko z pobieraniem podstron poszczególnych ofert, ale też z poruszaniem się pomiędzy ponumerowanymi stronami listy.

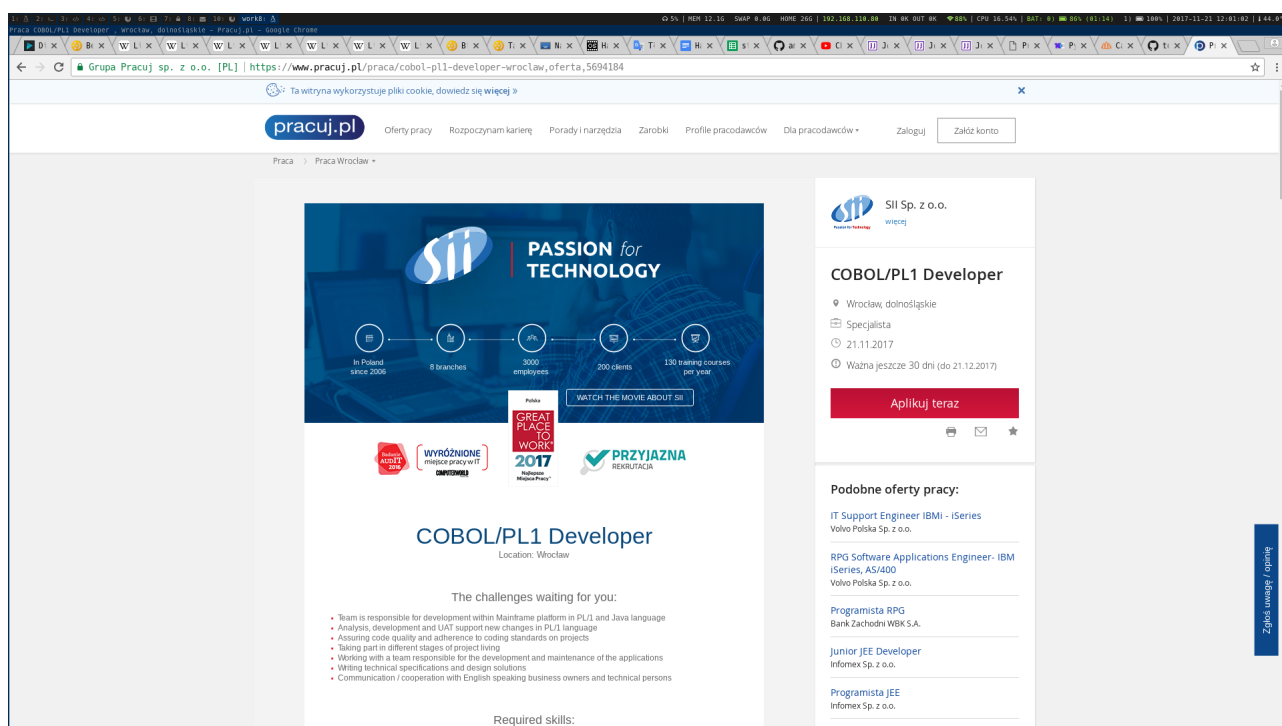
Do każdej z ofert na stronie (których znajduje się ok. 50) prowadzi bezpośredni link, który można wyłuskać z kodu HTML listy. Podstrona pojedynczej oferty zawiera wszystkie interesujące nas informacje również możliwe do wyłuskania z kodu HTML przy użyciu odpowiednich selektorów CSS. Dostępne w przystępny sposób informacje to:

- Data dodania oferty
- Data wygaśnięcia
- Nazwa dodającego (pracodawca)
- Lokalizacja (miasto i województwo)
- Tytuł oferty
- Treść oferty
- Kategorie w których znajduje się dana oferta

Ponadto, w łatwy sposób można uzyskać również dwie wartości jednoznacznie identyfikujące ofertę:

- Adres URL oferty
- ID (będące częścią adresu URL)

Te dwie wartości z powodzeniem mogą służyć za klucz pozwalający np. szybko sprawdzać czy oferta jest już w bazie systemu - czyli czy została już kiedyś przetworzona.



Rysunek 2.2: Widok oferty w witrynie pracuj.pl.

2.1.2 NAPOTKANE PROBLEMY

Zadaniem stojącym przed scraperem jest automatyczne przejście po wszystkich dostępnych stronach listy, wyłuskanie zeń adresów poszczególnych ofert, a stamtąd wszystkich potrzebnych informacji. Pobrane ogłoszenie z wydzielonymi fragmentami powinno trafić do innego komponentu systemu, który zajmie się jego zapisem czy dalszym przetwarzaniem. Podczas prac nad modułem natrafiliśmy na kilka kwestii które wymagały opracowania konkretnego rozwiązania.

2.1.2.1 Aktualizacja ofert w serwisie

Ofert na stronie wraz z upływem czasu będzie przybywać - i dla naszego systemu jest to bardzo istotne. Aby zapewnić stały przypływ ogłoszeń z serwisu Pracuj.pl scraper został tak przygotowany, aby mógł być uruchamiany cyklicznie. Przewidzianym interwałem jest 12 godzin, choć nie jest to wartość zdefiniowana w programie. To od uruchamiającego program na komputerze zależy jak ją ustawi.

2.1.2.2 Utrzymywanie stanu scrapera

Uruchomienie cykliczne wraz z brakiem stanu (bo przecież wszystkie przetworzone ogłoszenia trafiają do osobnego modułu) wiąże się z możliwością niechcianego przetwarzania jednej oferty wielokrotnie - przy każdym kolejnym uruchomieniu programu. Zdecydowaliśmy się na rozwiązanie tego programu przez zaimplementowanie w scraperze możliwości pobrania z modułu do którego trafiają dane kluczy (adresów URL) tych danych które już tam są. Oznacza to, że scraper przy każdym uruchomieniu przetworzy wszystkie dostępne strony, ale nie będzie pobierał ani przetwarzał podstron ofert które już ma na liście. Próba pobrania listy wykonywana jest po uruchomieniu programu, przed rozpoczęciem skanowania. Jeżeli nie uda się jej otrzymać (serwer nie odpowie, lub odpowie z błędem), program wypisze w konsoli stosowny komunikat ostrzegający i przejdzie do przetwarzania wszystkich ofert.

2.1.2.3 Zabezpieczenia serwisu przed zbieraniem danych

Wartym wspomnienia jest fakt stosowania przez serwis Pracuj.pl zabezpieczeń utrudniających automatyczne zbieranie danych. Podczas pierwszych testów naszego modułu wszystkie wychodzące żądania HTTP miały nagłówek `User-Agent` ustawiony na nazwę naszego projektu - `jobsbrowser`. Nie zauważyliśmy wtedy żadnych utrudnień ani trudności związanych z uzyskaniem przez scrapera dostępu do zasobów serwisu. Po kilku dniach jednak, sytuacja się

zmieniała. Okazało się, że wszystkie nasze żądania (nawet z innych adresów IP) podpisane jako `jobbrowser` były odrzucane przez serwer. Konieczne było wprowadzenie poprawki w kodzie modułu, tak żeby nagłówkiem `User-Agent` imitował przeglądarkę internetową. Wybór padł na Google Chrome w wersji 41 i to rozwiązało problem.

2.1.3 SCRAPY

Do napisania scrapera zdecydowaliśmy się na framework Scrapy. Jest to framework napisany w języku Python, przy wykorzystaniu biblioteki Twisted. Dzięki temu działa całkowicie asynchronicznie, co czyni go niezwykle wydajnym przy relatywnej łatwości pisania własnych scraperów. Wydany jest na licencji BSD3.

2.1.4 URUCHOMIENIE I UŻYTKOWANIE

W katalogu `jobbrowser` repozytorium znajduje się wykonywalny skrypt `run.sh`. Nie przyjmuje on żadnych parametrów, a jego uruchomienie powoduje uruchomienie procesu scraper'a. W celu cyklicznego uruchamiania, zalecamy skorzystanie z menadżera *cron*.

Za pomocą parametru `--log-level=LOG_LEVEL`, gdzie `LOG_LEVEL` może przyjąć jedną z poniższych wartości (wartości wymienione są od najmniej restrykcyjnej do najbardziej):

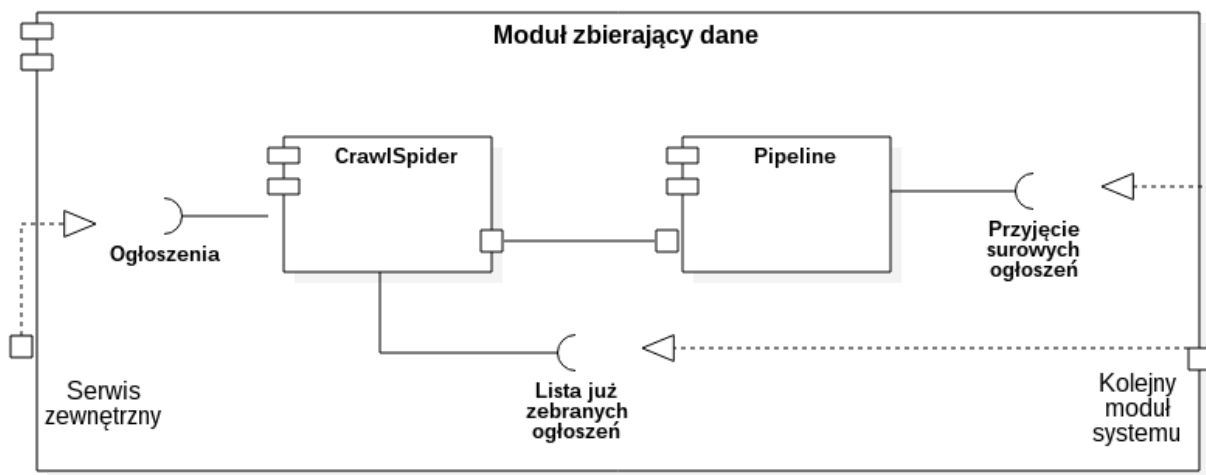
- `DEBUG`
- `INFO`
- `WARNING`
- `ERROR`
- `CRITICAL`

Innym użytecznym parametrem podczas weryfikowania działania scraper'a jest `-o/--output filename.EXTENSION`, gdzie `EXTENSION` może przyjąć jedną z niżej wymienionych wartości:

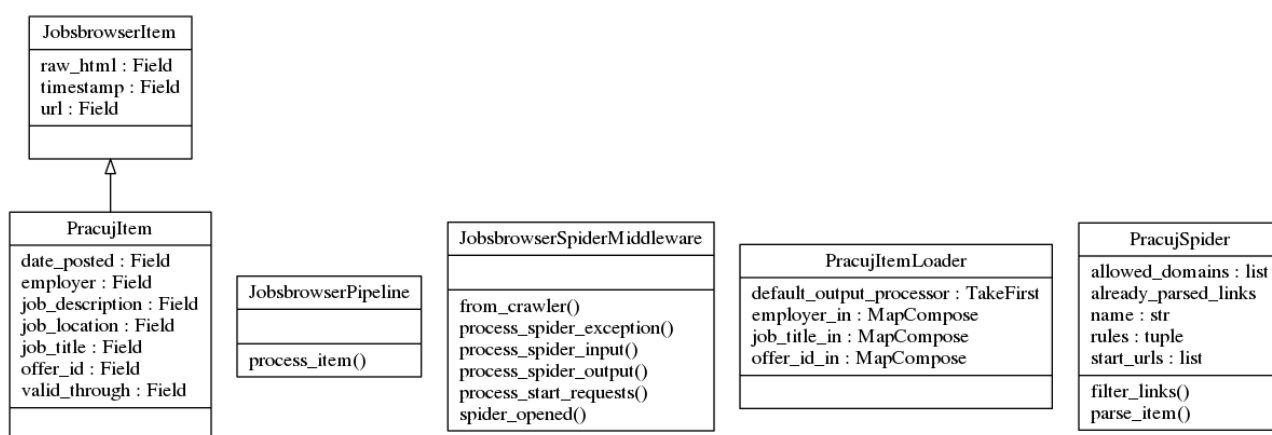
- `json`
- `jl` (json lines) każda linia jest oddzielnym obiektem JSON
- `csv`

Po podaniu tego parametru wszystkie dane zebrane przez scraper zostaną zapisane do podanego pliku w wybranym na podstawie rozszerzenia formacie.

W pliku `jobsbrowser/jobsbrowser/settings.py` znajdują się ustawienia programu. Większość z nich to ustawienia framework'a Scrapy. Ich opis znaleźć można w jego dokumentacji(Scrapy n.d.). Ustawieniami dotyczącymi konkretnie tego projektu są `STORAGE_SERVICE_ADD_URL` oraz `STORAGE_SERVICE_RETRIEVE_URL`, które oznaczają adresy URL pod który program może wykonać żądania w celu odpowiednio, dodania nowo przetworzonej strony oraz uzyskania listy adresów URL ofert których nie powinien przetwarzać.



Rysunek 2.3: Schemat architektury modułu Scraper'a.



Rysunek 2.4: Diagram klas modułu Scraper'a.

2.1.5 STRUKTURA KODU ŹRÓDŁOWEGO

Poniżej prezentujemy drzewo katalogów oraz plików projektu wraz z krótkim omówieniem:

```
jobsbrowser
  jobsbrowser
    __init__.py
    extractors.py    # definicje klas ekstraktujących linki z ofertami
    items.py         # definicje klas przechowujących zebrane dane
    loaders.py       # definicje klas ładujących zebrane dane
    loaders.py       # definicje klas ładujących zebrane dane
    middlewares.py
    pipelines.py     # definicje kolejnych etapów przetwarzania danych
    settings.py      # ustawienia pajków
    spiders          # katalog z definicjami pajków zbierających dane
    __init__.py
    pracuj.py        # definicja pajka zbierającego dane ze strony pracuj.pl
  tests/            # katalog z testami projektu
  run.sh            # plik wykonywalny uruchamiający proces zbierania danych
  scrapy.cfg        # konfiguracja całego projektu
  requirements.txt  # plik z wymaganiami projektu
```

2.1.6 GŁÓWNE KLASY MODUŁU

- **JobBrowserItem** - klasa bazowa definiująca pola które powinny być dostarczane przez wszystkie spider'y wykorzystywane w projekcie.
- **PracujItem** - klasa przechowująca pola które muszą zostać wypełnione przez scraper'y zbierające dane ze strony pracuj.pl.
- **JobsBrowserPipeline** - klasa która po zebraniu danych ze stron wysyła dane do modułu zapisy do bazy danych.
 - `process_item` - metoda w której za pomocą zapytania HTTP wysyłany jest aktualnie przetwarzany element(Item) do modułu bazy danych.
- **JobsBrowserSpiderMiddleware** - klasa wygenerowana automatycznie podczas tworzenia projektu za pomocą Scrapy'ego.
- **PracujLinkExtractor** - klasa wyciągająca linki do ofert pracy w serwisie pracuj.pl.

- **PracujItemLoader** - klasa przetwarzająca w prosty sposób zebrane dane.
 - `load_item` - wykonuje wszystkie operacje zdefiniowane w deskryptorach na aktualnie przetwarzanym elemencie
- **PracujSpider** - klasa wykonująca zapytania do serwisu pracuj.pl oraz zbierająca dane.
 - `start_urls` - adresy URL od których zaczynane jest zbieranie danych.
 - `rules` - zasady definiujące jakie elementy są ofertami oraz jak przejść do następnej strony z ofertami.
 - `already_parsed_links` - linki do ofert znajdujących się już w bazie.
 - `filter_link` - metoda sprawdzająca czy dane ogłoszenie nie jest już w bazie.
 - `parse_item` - zajmuje się wyciągnięciem potrzebnych danych z aktualnie przetwarzanej strony za pomocą `PracujItemLoader`'a.
 - `start_requests` - generuje zapytania HTTP do każdej strony z daną kategorią.

2.1.7 TESTY

W kodzie źródłowym modułu znajdują się testy jednostkowe pozwalające na przetestowanie poprawności zaimplementowanych metod i funkcji. W tej sekcji znajduje się również opis przykładowego scenariusza testów akceptacyjnych. Testy integracyjne na tym etapie nie są jeszcze przewidziane. Moduł zbierania danych jest pierwszym modułem i bez implementacji pozostałych ich wykonanie jest niemożliwe.

2.1.7.1 Testy jednostkowe

Aby uruchomić testy jednostkowe w konsoli należy wpisać polecenie `pytest`.

Poniżej przedstawiamy listę plików z testami jednostkowymi oraz opis poszczególnych funkcji lub metod:

- `test_pracuj_item_loader.py` - plik z testami klasy `PracujItemLoader`.
 - `test_taking_first_from_each_field` - test sprawdza czy żaden z przetworzonych pól nie jest listą (Scrapy domyślnie zwraca wszystkie elementy jako listy, nawet tylko gdy znajduje się w nich jeden element).
 - `test_offer_id_properly_extracted` - test sprawdza czy pole `offer_id` jest odpowiednio wydobywane z adresu URL strony.

- `test_remove_html_tags_from_employer_and_job_title_fields` - test sprawdza czy znaczniki HTML zostały poprawnie usunięte z wartości pól *employer* oraz *job_title*.
- `test_jobsbrowser_pipeline.py` - plik z testami klasy `JobsBrowserPipeline`.
 - `test_jobsbrowser_pipeline_process_item_send_request_to_db_module` - test sprawdza czy w metodzie `process_item` wykonywane jest zapytanie HTTP POST do serwera z działającym modulem bazy danych.
- `test_pracuj_spider.py` - plik z testami klasy `PracujSpider`.
 - `test_parse_item` - test sprawdza czy metoda prawidłowo wyciąga dane z adresu URL oraz treści strony, a następnie zwraca je w atrybutach obiektu typu `PracujItem`.
 - `test_parse_on_page_with_multiple_next_pages` - test sprawdza czy metoda `parse` prawidłowo znajduje link do następnej strony z ofertami. Testowany w tej metodzie jest przypadek gdy istnieją kolejne strony.
 - `test_parse_on_last_page` - test sprawdza czy metoda `parse` prawidłowo zachowuje się na ostatniej stronie z listingami ofert, czyli nie znajduje żadnych nowych linków, tym samym kończąc zbieranie danych.

2.1.7.2 Testy akceptacyjne

Moduł zbierania danych zajmuje się, jak mówi sama nazwa, jedynie ich zbieraniem. Domyślnie nie są one nigdzie przechowywane, ani zapisywane. Podejście takie utrudnia nieco przygotowanie testów akceptacyjnych obejmujących wyłącznie ten komponent, ale nie uniemożliwia, co zaraz wykazemy.

Zasada działania programu (uruchamianego przez `run.sh`) jak opisano już wcześniej sprowadza się do zbierania ofert i wysyłania ich do kolejnego modułu systemu (z wypisaniem stosownego komunikatu, jeśli ten nie odpowiada). Bez tego komponentu, nie zobaczymy nigdzie zebranych ofert, ani też nie dostarczymy scraperowi listy już zebranych, co będzie skutkowało zebraniem wszystkich. Nie są to warunki idealne na testy akceptacyjne - gdzie przecież chcemy upewnić się że komponent faktycznie działa. Wykorzystamy jednak fakt że skrypt uruchamiający przekazuje swoje parametry do procesu scrapera, co pozwala na nadpisanie jego ustawień na czas uruchomienia. Dzięki temu ograniczymy zbiór przetwarzanych ofert (do jednej strony) i zapiszemy je na dysku, aby przekonać się że interesujące nas elementy faktycznie zostały z ofert wyluskane.

Aby wykonać takie polecenie testujące sprawność scrapera, do skryptu musimy przekazać kilka dodatkowych argumentów:

```
./run.sh -s DEPTH_LIMIT=1 -o oferty.json
```

Oznaczają one odpowiednio:

- `-s DEPTH_LIMIT=1` - nadpisanie ustawień scrapera dotyczących maksymalnej “głębokości” na jaką może się zapuścić. W naszym przypadku oznacza to liczbę przetworzonych stron
- `-o oferty.json` - wymusza zapis przetworzonych obiektów do pliku `oferty.json`

Ponadto uruchomieniu towarzyszyć będą wypisywane w terminalu komunikaty informujące o przetworzeniu danej oferty oraz próbie wysłania jej do sąsiedniego komponentu. Mówią one użytkownikowi czym program aktualnie się zajmuje i na jakie trafia problemy. Po zakończeniu w katalogu w którym znajduje się wywołany skrypt znajdziemy plik `oferty.json` który zawiera zebrane oferty. Z powodu przechowywania w obiekcie również surowej wersji przetwarzanej strony (w postaci kodu HTML) nie jest on szczególnie czytelny dla człowieka, jednak bez problemu można wykonywać na nim dowolne operacje, np. z poziomu innego programu.

```

2017-11-23 00:42:11 bsdell pracuj[17637] INFO Offer 5570127 scraped. Sending to storage service...
2017-11-23 00:42:11 bsdell pracuj[17637] WARNING Sending offer 5570127 failed. Service storage is unavailable.
2017-11-23 00:42:11 bsdell pracuj[17637] INFO Offer 5570005 scraped. Sending to storage service...
2017-11-23 00:42:11 bsdell pracuj[17637] WARNING Sending offer 5570005 failed. Service storage is unavailable.
2017-11-23 00:42:13 bsdell pracuj[17637] INFO Offer 5570061 scraped. Sending to storage service...
2017-11-23 00:42:13 bsdell pracuj[17637] WARNING Sending offer 5570061 failed. Service storage is unavailable.
2017-11-23 00:42:13 bsdell scrapy.core.engine[17637] INFO Closing spider (finished)
2017-11-23 00:42:13 bsdell scrapy.extensions.feedexport[17637] INFO Stored json feed (47 items) in: oferty.json

```

Rysunek 2.5: Komunikaty w oknie terminala

```

In [18]: with open("oferty.json") as f:
...:     offers = json.load(f)
...:

In [19]: assert len(offers) == 47

In [20]: for offer in offers[:5]:
...:     print(offer['job_title'])
...:
...:
Programista ASP.NET MVC 6
Specjalista ds. Marketingu
Programme and Project Services Officer
Analityk Hurtowni Danych - Konsultant BI
Oracle ERP Cloud PM

```

Rysunek 2.6: Przykładowy dostęp do wynikowego pliku JSON

2.2 Moduł przechowujący dane

Kod źródłowy znajduje się pod adresem: github.com/jobsbrowser/pipeline.

Kolejnym komponentem systemu jest moduł odpowiedzialny za przechowywanie danych. Przez dane rozumiemy w tym przypadku oferty w postaci “pół-surowej”. Tzn. będące już po wstępnym, prostym procesie przetwarzania w module zbierania danych, ale jeszcze przed najbardziej znaczącym procesem ekstrakcji kluczy. Są to więc dane które są bazą dla przyszłych działań systemu, ale w obecnej formie nie dostarczają wielu informacji.

2.2.1 WYMAGANIA

Moduł z pozostałymi komponentami systemu łączy się dwiema drogami:

- Przez interfejs HTTP wykorzystywany przez moduł zbierania danych (poprzedni komponent)
- Poprzez dodawanie nowych zadań do kolejki, zbieranych i wykonywanych przez moduł ekstrakcji kluczy (następny komponent)

Wymagania funkcjonalne modułu sprowadzają się więc do obsłużenia obydwu kierunków komunikacji. Interfejs HTTP powinien pozwalać na dwie operacje:

- Dodanie oferty do bazy (nadpisując jeśli oferta z takim adresem URL już istnieje)
- Pobranie listy adresów URL ofert zapisanych już w bazie
- Uaktualnienie wybranej oferty z bazy danych

Natomiast na drodze komunikacji z kolejnym modułem:

- Rozpoczęcie nowego zadania Celery z dokumentem oferty przekazanym jako argument, po zapisie oferty do bazy.

Wymagania нефункционалне modułu sprowadzają się natomiast do:

- **niezawodności** - usługa powinna być dostępna możliwie cały czas. Nie jest to jednak kwestia kluczowa, ponieważ stosunkowo krótkie braki w dostępności (rzędu maksymalnie kilku dni) nie ciągną za sobą konsekwencji. Jeżeli moduł zbierający dane nie uzyska odpowiedzi od modułu zajmującego się ich przechowywaniem, informacje o tej ofercie nie

zostaną nigdzie zapisane. Kiedy usługa przechowywania będzie ponownie dostępna, na zapytanie scrapera o listę ofert będących już w bazie zwróci tę sprzed awarii, wszystkie pominięte oferty zostaną więc ostatecznie dodane.

- **wydajności** - liczba nadchodzących ofert może być potencjalnie duża, proces zapisu do bazy powinien być więc jak najmniej skomplikowany i efektywny, aby uniknąć spadków na wydajności z powodu niewydajnych zapytań. Podobnie jak w kwestii niezawodności, nie jest to jednak wymaganie kluczowe. Spadki w wydajności nie będą bowiem objawiać się wolniejszym działaniem aplikacji przeznaczonej dla użytkownika końcowego, a jedynie późniejszym pojawianiem się w niej nowych ofert.

2.2.2 INTERFEJS

Interfejs HTTP zaimplementowany jest przy użyciu micro-frameworka Flask(Flask n.d.). Flask jest wykorzystywany do tworzenia stron internetowych oraz REST API. Zdecydowaliśmy się na niego ze względu na to, że jest to framework dojrzały, idealny do małych lub średnich projektów, a w razie wzrostu skali projektu umożliwia wygodne skalowanie. Flask zawiera również wiele świetnych rozszerzeń, np. rozszerzenie integrujące go z bazą MongoDB z której korzystamy w projekcie.

2.2.3 BAZA DANYCH

Wykorzystanym silnikiem bazy danych jest MongoDB(MongoDB n.d.). Zdecydowaliśmy się na bazę relacyjną z kilku powodów. Po pierwsze tak naprawdę jedynym typem trzymanych w niej danych są same oferty. Oznacza to że w bazie relacyjnej mielibyśmy tylko jedną tabelę - bez żadnych relacji czy potrzeby zachowania spójności. Nie ma potrzeby również stosowania mechanizmu transakcji, czy skomplikowanych zapytań. Tym czego faktycznie oczekujemy od bazy jest wydajność, dostępność oraz ewentualna skalowalność. Wybór był więc prosty. Struktura dokumentów przechowywanych w bazie jest identyczna jak struktura zebranego ogłoszenia. Dla przypomnienia, pola jakie wyróżniamy w dokumencie oferty to:

- Adres URL
- Czas w którym pobrano ofertę
- Kod HTML strony z ofertą
- ID oferty w systemie pracuj.pl
- Data dodania
- Data ważności
- Podmiot dodający

- Tytuł oferty
- Miejsce pracy
- Kategorie oferty
- Kod HTML treści oferty (rozbity na opis, kwalifikacje oraz benefity - wg struktury Pracuj.pl)

2.2.4 INTEGRACJA Z KOLEJNYM MODUŁEM

Początkowe plany zakładały użycie w tym miejscu (po dodaniu oferty do bazy) frameworka *Luigi*. Jest to stworzone przez twórców aplikacji *Spotify* narzędzie do łączenia ze sobą kolejnych funkcji / etapów tworząc łańcuch przetwarzania (wspomniany w kilku miejscach tzw. *Pipeline*). Okazało się jednak, że przewidziane zastosowania narzędzia obsługują etapy nieco innego typu niż te do zaimplementowania w module ekstrakcji kluczy. Luigi przeznaczony jest bowiem do łączenia bardzo wymagających zadań, angażujących wiele zewnętrznych usług czy języków programowania i wykonujących się nawet kilka dni. W efekcie nie udostępnia chociażby tak podstawowej w mniejszych zastosowaniach możliwości, jak uruchamianie zadania z poziomu kodu źródłowego. W grę wchodzi jedynie linia poleceń. Zdecydowaliśmy się więc na porzucenie go, na rzecz frameworka *Celery* którego obsługa zadań jest tym czego potrzebujemy. Stracimy co prawda na odporności na awarie (Luigi zapisuje stan po każdym zadaniu, i wraca do niego po awarii), lecz zdecydowanie zyskamy na łatwości implementacji.

Użycie frameworka Celery jest częścią kolejnego modułu, tj. modułu ekstrakcji kluczy, więc to tam znajdzie się dotycząca tej kwestii dokumentacja.

2.2.5 URUCHOMIENIE

Do uruchomienia API korzystamy z polecenia `python manage.py runserver`. Aby uruchomić usługę z odpowiednimi ustawieniami trzeba ustawić zmienną środowiskową `APP_CONFIG` na wartość `PRODUCTION`. Możemy to zrobić korzystając z 1 z poniższych poleceń:

- `export APP_CONFIG="production" && python manage.py runserver`
- `APP_CONFIG="PRODUCTION" python manage.py runserver`

2.2.6 STRUKTURA KODU

Poniżej prezentujemy drzewo katalogów oraz plików modułu wraz z krótkim omówieniem.

```

jobsbrowser
    api                                # katalog modułu API
        __init__.py
        resources.py                 # definicja endpointów API
        settings.py                  # ustawienia API
        spec.yml                      # specyfikacja API w standardzie OpenAPI(swagger)
    pipeline                           # moduł pipeline
        __init__.py
manage.py                             # skrypt z pożytecznymi komendami dotyczącymi API
requirements.txt                       # wymagane biblioteki modułu
setup.py
tests
    api                                # katalog z testami modułu API
        __init__.py
        test_api_resources.py        # testy endpointów API
    pipeline                           # katalog z testami modułu Pipeline
tox.ini                               # konfiguracja narzędzia używanego do testowania

```

2.2.7 GŁÓWNE KLASY MODUŁU

- `add_offer` - funkcja implementująca dodawanie otrzymanej oferty do bazy danych.
- `get_offers` - funkcja implementująca pobieranie aktualnych ofert (takich których data w polu *valid_through* jest większa od daty dzisiejszej) znajdujących się w bazie danych.
- `update_offer` - funkcja implementująca uaktualnianie oferty z bazy.
- `init_app` - funkcja tworząca aplikację z podaną konfiguracją (jako parametr *config_name* lub zmienna środowiskowa `APP_CONFIG`).
- **BaseConfig** - klasa z bazową, fundamentalną konfiguracją.
- **DevConfig** - klasa przechowująca konfigurację developerską.
- **ProductionConfig** - klasa przechowująca konfigurację produkcyjną.
- **TestingConfig** - klasa przechowująca konfigurację testową.

2.2.8 TESTY

2.2.8.1 Testy jednostkowe

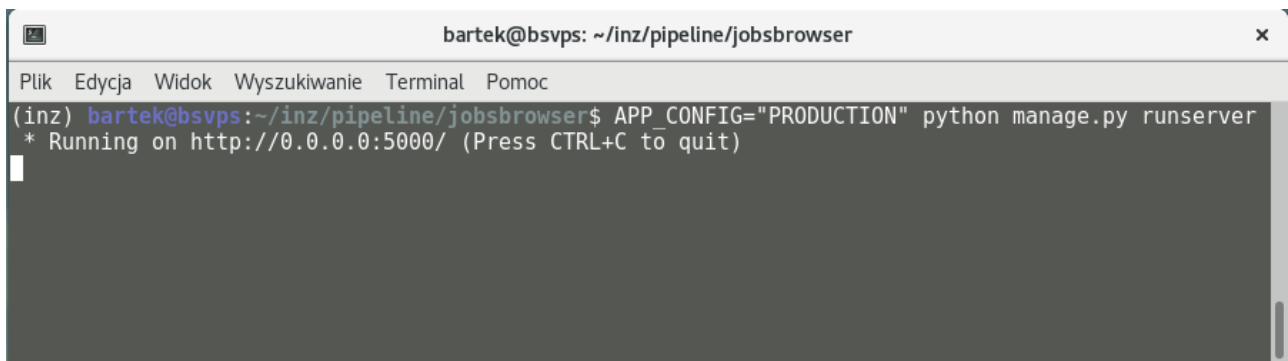
Aby uruchomić testy jednostkowe, w konsoli należy wpisać polecenie `tox`. Poniżej przedstawiamy opis testów znajdujących się w pliku `jobsbrowser/tests/test_api_resources.py`:

- `test_ping_resource_returns_pong` - test sprawdza czy endpoint `/pong` (który jest wykorzystywany do sprawdzania czy usługa API jest aktywna) zwraca odpowiedź ze statusem 200 oraz wartością "pong".
- `test_add_offer_resource_try_add_offer_to_mongo_db` - test sprawdza czy wysłana poprzez zapytanie POST oferta próbuje być zapisana do bazy danych.
- `test_get_offers_resource_query_mongo_db` - test sprawdza czy po wykonaniu zapytania HTTP GET na endpoint `/offers` wykonywana jest próba pobrania danych z bazy danych.
- `test_update_offer_resource_query_mongo_db` - test sprawdza czy po wykonaniu zapytania HTTP PUT na endpoint `/offer` wykonywana jest próba pobrania oferty oraz uaktualnienia jej w bazie.

2.2.8.2 Testy akceptacyjne

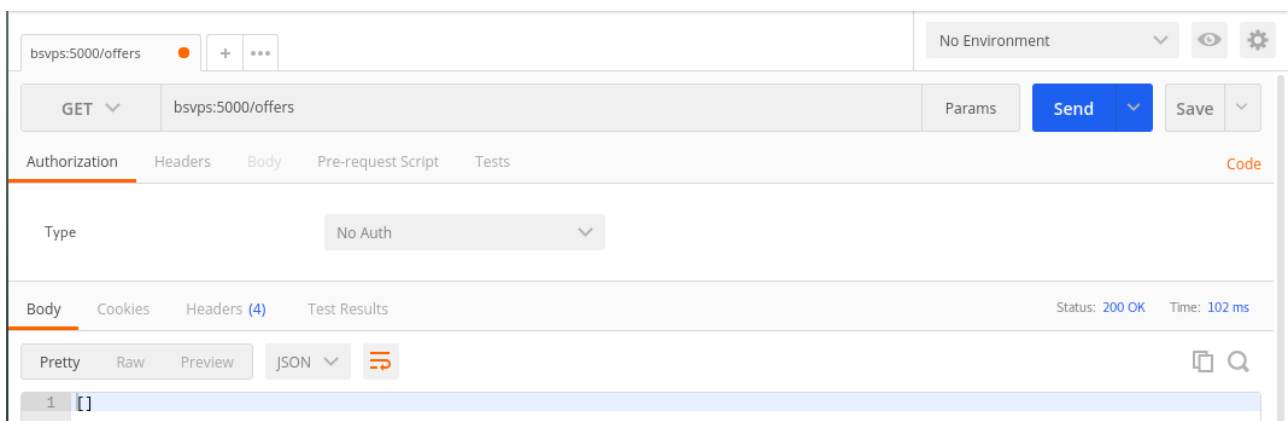
Interfejs modułu zbierania danych jest dość ubogi. Zajmuje się on bowiem jedynie dodawaniem ofert oraz zwracaniem informacji o adresach URL tych już istniejących. Test jaki możemy przeprowadzić aby upewnić się moduł działa poprawnie może więc polegać na:

- Uruchomieniu usługi wg instrukcji z dokumentacji
- Wykonaniu zapytania o listę ofert (powinna być pusta)
- Wykonaniu żądania dodającego nową ofertę
- Wykonaniu zapytania o listę ofert raz jeszcze. Powinniśmy otrzymać adres URL przesłany w poprzednim kroku.

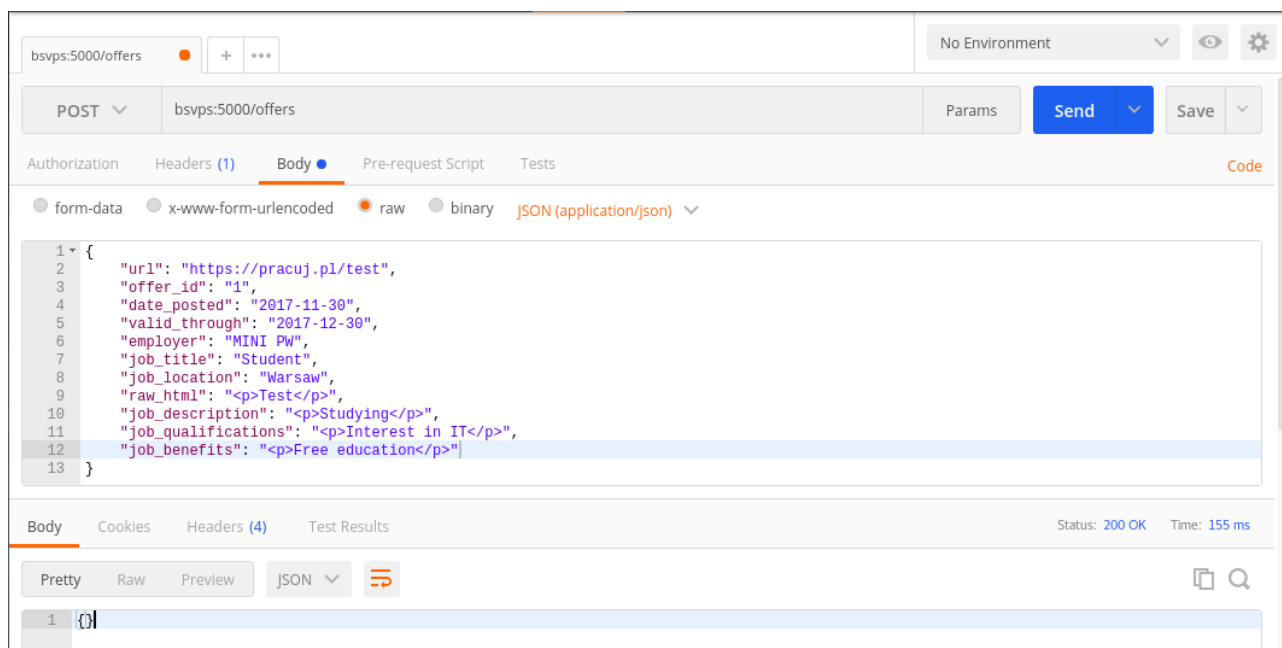


A terminal window titled "bartek@bsvps: ~/inz/pipeline/jobbrowser" with a menu bar (Plik, Edycja, Widok, Wyszukiwanie, Terminal, Pomoc). The command executed is `(inz) bartek@bsvps:~/inz/pipeline/jobbrowser$ APP_CONFIG="PRODUCTION" python manage.py runserver`. The output shows the server running on `http://0.0.0.0:5000/` and prompts to press `CTRL+C` to quit.

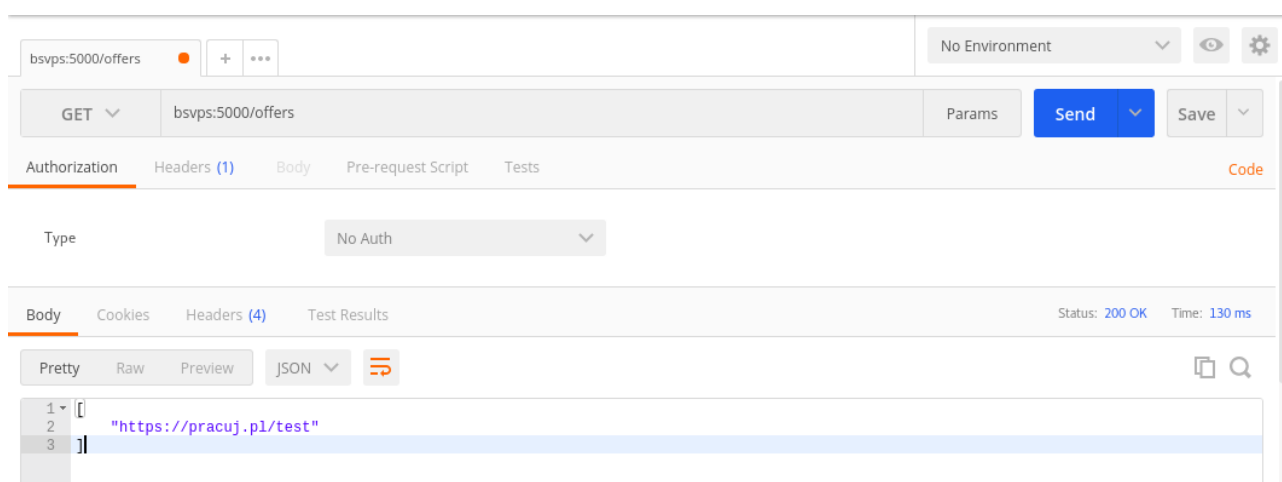
Rysunek 2.7: Uruchomienie serwera



Rysunek 2.8: Pierwsze zapytanie o listę



Rysunek 2.9: Dodanie nowej oferty



Rysunek 2.10: Ponowne zapytanie o listę

2.3 Moduł ekstrakcji kluczy

Kod źródłowy znajduje się pod adresem: github.com/jobsbrowser/pipeline.

Modułem systemu do którego trafiają przetwarzane oferty w następnej kolejności jest moduł ekstrakcji kluczy - czyli wymaganych na stanowisko którego dotyczy ogłoszenie umiejętności i technologii. To tutaj odbywa się, kluczowy z punktu widzenia biznesowego zastosowania projektu, proces wyciągania z “surowych” ofert wartościowych informacji. Wejściem modułu są zapisane w bazie oferty, natomiast wyjściem te same obiekty, ale uzupełnione o listę kluczy wykrytych w ich treści.

2.3.1 PRZEZNACZENIE MODUŁU

Dzięki opisanym wcześniej modułom w bazie danych systemu znajdują się zbierane z serwisu pracuj.pl oferty. Dla przypomnienia, informacje jakie uzyskiwane są bezpośrednio w trakcie ich pozyskiwania to:

- tytuł
- ID
- treść (w postaci kodu HTML)

oraz kilka mniej znaczących z punktu widzenia tego komponentu, a szerzej opisanych w części dotyczącej modułu zbierania danych. Informacje te same w sobie nie są szczególnie istotne dla użytkownika końcowego naszej aplikacji. Są to bowiem te same dane które zobaczyć możemy korzystając bezpośrednio z serwisu pracuj.pl. Nie niosą więc za sobą póki co żadnych dodatkowych wartości.

Sednem naszej aplikacji jest przekształcenie dużego zbioru ofert w statystyki dotyczące ich wspólnych elementów - czyli umiejętności i technologii które są w nich wymieniane. Moduł ten pozwala właśnie na to - uzyskanie z pojedynczej oferty listy takich kluczy, które są zawarte w jej treści.

2.3.2 ARCHITEKTURA I ALGORYTM

Z technicznego punktu widzenia moduł jest ciągiem funkcji, wykonywanych w ustalonej kolejności, gdzie wyjście każdej z nich przekazywane jest jako argument do następnej. Stąd powtarzający się w kodzie termin *pipeline*. Argumentem pierwszej funkcji jest zapisana do bazy danych chwilę po pobraniu z serwisu zewnętrznego oferta, natomiast wynikiem ostatniej uproszczony obiekt oferty zawierający listę znalezionych kluczy. Na koniec oferta jest wyszukiwana w bazie,

dodawana jest do niej wspomniana lista (dzięki zastosowaniu nierelacyjnej bazy danych nie wiąże się to z potrzebą jej przebudowy).

Do obsługi łańcucha, czyli utrzymania poprawnej kolejności wykonywania funkcji, oraz asynchronicznego przetwarzania wielu ofert jednocześnie korzystamy z frameworka **Celery** (Celery n.d.), którego opis znajduje się w dalszej części tego rozdziału.

2.3.2.1 Rozpoznawanie umiejętności i technologii

Uznaliśmy, że aby w miarę poprawnie rozpoznawać klucze, powinniśmy w pierwszym kroku zaopatrzyć się w miarę obszerny i sprawdzony ich zbiór.

Podejście takie, jak później się okazało wydaje się być całkiem słuszne, ponieważ stosowane jest także w projektach o znacznie szerszym zasięgu i złożoności niż nasz. Dla przykładu, powstająca w momencie pisania tej pracy platforma **Google Cloud Job Discovery**, zajmująca się automatycznym dopasowywaniem ofert pracy do CV potencjalnych pracowników, do działania wykorzystuje zbudowaną przez zespół Google ontologię zawierającą, jak podają, ok. 50 tys. umiejętności z różnych pól zawodowych.

Z racji tego, że nasz projekt skupia się na ofertach pracy z branży IT, postanowiliśmy skorzystać z faktu że istotne, oczekiwane przez pracodawców umiejętności, pokrywają się z nazwami technologii informatycznych, czy języków programowania. Jako źródło takich danych, postanowiliśmy wykorzystać niesamowicie popularny wśród ludzi zainteresowanych IT portal **Stack Overflow**. API tego serwisu pozwala na pobranie używanych przez jego użytkowników kluczy, posortowanych według popularności. Wszystkich jest blisko 40 tys.

Na potrzeby naszego projektu, korzystamy ze zbioru dziesięciu tysięcy najpopularniejszych kluczy. Znajdują się wśród nich wszelakie technologie, frameworki, wzorce projektowe i języki programowania. Mając taką listę, możemy każdą przychodzącą ofertę przeszukać pod względem występowania niektórych z nich. Proces ten przeprowadzamy w następując sposób:

1. Po zapisaniu całej oferty do bazy, upraszczamy obiekt przekazując dalej tylko istotne z punktu widzenia komponentu dane, tj.
 - ID
 - tytuł
 - treść
2. Z treści oferty usuwamy tagi HTML
3. Treść oferty rozbijamy na *tokens*, czyli wyrazy oraz znaki interpunkcyjne

4. Sprawdzamy czy oferta jest w języku polskim, bo tylko takie akceptujemy
5. Usuwamy zbędne z punktu widzenia analizy językowej tokeny, takie jak przyimki, spójniki czy zaimki.
6. Dla listy uzyskanych tokenów sprawdzamy które z nich znajdują się na liście znanych nam kluczy, i te zapisujemy.

W ten sposób każdej ofercie z osobna przypisujemy listę technologii które znaleźliśmy w jej treści. Śledząc opis algorytmu, zauważyć można że wynikowo będą one posortowane jedynie ze względu na kolejność występowania w treści oferty. Właściwością tą zajmie się kolejny moduł systemu, który znalezionym tutaj kluczom nada odpowiednie priorytety, biorąc pod uwagę nie tylko kolejność ich występowania w tekście, ale także sąsiedztwo wyrazów w jakim się znajdują, czy poprzedzające je frazy.

2.3.2.2 Uzasadnienie wyboru powyżej opisanego algorytmu

Dużo czasu poświęciliśmy na testowanie wielu algorytmów do wyznaczania słów kluczowych z tekstu. Wiele z nich dawało niezadowalające wyniki na zbiorach ogłoszeń napisanych w języku polskim. Poniżej opisujemy krótko algorytmy które przetestowaliśmy wraz z przykładowymi wynikami jakie dawały. Wszystkie przedstawione poniżej wyniki zostały uzyskane z poniższej oferty:

Wystarczy, że:

- posiadasz doświadczenie w podobnej roli
- znasz na poziomie przynajmniej dobrym język Python (idealnie Python 3) wraz z frameworkiem Django
- znasz również JavaScript (nie tylko jQuery)
- posiadasz dobrą wiedzę na temat relacyjnych baz danych (mile widziana znajomość PostgreSQL)
- sprawnie poruszasz się w HTML5 i CSS3
- cechuje Cię ponadprzeciętna samodzielność oraz umiejętność samoorganizacji
- potrafisz pisać wysokiej jakości kod
- posiadasz umiejętność dekompozycji zadań oraz dostarczania działających rozwiązań
- jesteś skrupulatny (lub przynajmniej pracujesz nad tym, aby takim się stać)

to z dużym prawdopodobieństwem jesteś osobą, której szukamy.

Jeżeli dodatkowo:

- potrafisz pisać testy jednostkowe
- znasz dowolny, inny niż wyżej wymienione, język programowania
- możesz wykazać się znajomością protokołów sieciowych

to z tym większą niecierpliwością oczekujemy na możliwość spotkania z Tobą.

Chcemy powierzyć Ci następujące obowiązki:

- projektowanie oraz implementowanie aplikacji internetowych, w tym wspierających płatności, uwierzytelnianie SMS, jak również aplikacji wyszukujących informacje na wielu stronach
 - implementowanie nowych funkcji i poprawek w istniejących aplikacjach internetowych (utrzymanie i rozwój narzędzi stworzonych na potrzeby wewnętrzne)
 - ścisłą współpracę z doświadczonymi architektami oprogramowania, wraz z wpływem na kształt oraz sposób działania tworzonych narzędzi
-

Dodatkowo oferujemy:

- kontakt z najnowszymi technologiami oraz niespotykaną różnorodność tematyczną projektów
 - możliwość realizowania własnych, innowacyjnych projektów
 - elastyczne godziny pracy
 - współpracę z ekspertami z wieloletnim doświadczeniem, którzy chętnie dzielą się swoją wiedzą
 - szkolenia wewnętrzne, udział w spotkaniach, seminariach oraz konferencjach
 - bardzo dobre warunki pracy, nowoczesny sprzęt, świeże owoce i pyszną kawę
 - dostęp do dodatkowych świadczeń (Multisport, Medicovert)
 - swobodną atmosferę, brak dres code'u, nieformalne relacje
 - mnóstwo ciekawych wyzwań i szans rozwoju
 - nowoczesne, doskonale skomunikowane biuro w centrum Warszawy
-

Rysunek 2.11: Oferta wykorzystywana do uzyskania przykładowego wyniku

- RAKE(Rapid Automatic Keyword Extraction)(Rake-NLTK n.d.) Bardzo popularny na githubie stosunkowo prosty algorytm, niestety nie przyniósł zadowalających efektów. Poniżej prezentujemy listę 10 najważniejszych według algorytmu RAKE fraz/kluczy:
 - umiejętność samoorganizacji potrafisz pisać wysokiej jakości kod posiadasz umiejętność dekompozycji zadań
 - potrafisz pisać testy jednostkowe znasz dowolny
 - dostarczania działających rozwiązań jesteś skrupulatny
 - poziomie przynajmniej dobrym język python
 - język programowania możesz wykazać
 - dużym prawdopodobieństwem jesteś osobą
 - temat relacyjnych baz danych
 - mile widziana znajomość postgresql
 - znajomością protokołów sieciowych to
 - posiadasz dobrą wiedzę
- TF-IDF(Term Frequency - Inverse Document Frequency)(TF-IDF n.d.) Jeden z najbardziej popularnych i najszerzej wykorzystywanych algorytmów w świecie przetwarzania języka naturalnego, również nie dawał zadowalających wyników.
 - pisać
 - implementowanie
 - python
 - posiadasz
 - wyszukiwujących
 - znasz
 - uwierzytelnianie
 - dres
 - niecierpliwością
 - niespotykaną

Z racji bardzo szczególnego podzbioru języka polskiego używanego przy pisaniu ogłoszeń powyższe algorytmy dawały bardzo dziwne wyniki, dlatego też zdecydowaliśmy się na korzystanie z algorytmu opartego na tagach pobranych z serwisu stackoverflow.

2.3.2.3 Znajdowanie podobnych kluczy za pomocą modelu Word2Vec

W celu znalezienia kluczy podobnych do wybranego przez użytkownika zestawu, postanowiliśmy skorzystać z modelu Word2Vec (Yoav Goldberg n.d.). Word2vec to grupa modeli reprezentujących słowa jako tzw. zanurzenia słów (Bengio Yoshua n.d.), czyli słowa reprezentowane

jako wektory w wielowymiarowej przestrzeni. W naszym przypadku zdecydowaliśmy się wytrenować model word2vec nie na całych korpusach (wszystkich ogłoszeniach), ale na liście kluczy przypisanych do każdego ogłoszenia przez wcześniejszy algorytm. Dzięki takiemu podejściu dla każdego klucza otrzymujemy klucze jedynie z naszego ustalonego zbioru kluczy (pobranego z serwisu stackoverflow). Do wytrenowania modelu skorzystaliśmy z biblioteki gensim (Gensim n.d.). Poniżej prezentujemy wyniki otrzymane dla klucza **JAVA**:

- java-ee
- maven
- junit
- hibernate
- groovy
- soa
- tomcat
- jsp
- jenkins
- eclipse

Jak widać wyniki są zadowalające, wszystkie klucze znajdujące się na powyższej liście mają wiele wspólnego z językiem programowania JAVA. Jednak z powodu stosunkowo małego zbioru treningowego rezultaty nie są tak dobre dla kluczy mających mało wystąpień. Poniżej prezentujemy taki przykład na kluczu **Vue.JS**:

- node.js
- elasticsearch
- jasmine
- testy jednostkowe
- bitbucket
- api
- nosql
- webpack
- mongodb
- user-interface

2.3.3 CELERY

Celery(Celery n.d.) jest asynchroniczną kolejką zadań. Zadania są dystrybuowane do kolejki z różnych źródeł, np. z aplikacji webowej. Celery jest przeznaczone głównie do operacji zleczanych

oraz wykonywanych w czasie rzeczywistym. Jako kolejkę lub bazę na zlecone zadania możliwe jest wykorzystanie wielu backend'ów np. redis czy rabbitmq. W aplikacji korzystamy z Celery do wykonywania przetwarzania pobranych z zewnętrznych serwisów (pracuj.pl) ofert.

2.3.4 STRUKTURA KODU

Poniżej prezentujemy drzewo katalogów oraz plików modułu wraz z krótkim omówieniem.

```
jobsbrowser
  api                                # moduł przechowywania danych
  pipeline                          # moduł łańcucha przetwarzania danych
    _app.py
    chains.py                      # konstrukcja łańcucha przetwarzania danych
    __init__.py
    settings.py                   # ustawienia Celery oraz MongoDB
    tasks                          # moduł poszczególnych etapów pipeline
      bases.py                    # bazowe klasy etapów
      exceptions.py              # wyjątki modułu etapów pipeline
      __init__.py
      postprocess.py             # etapy wykonywane na koniec pipeline
      preprocess.py              # etapy przygotowujące ofertę do ekstrakcji cech
      process.py                 # etapy ekstraktujące cechy, klucze z oferty
      utils.py                   # przydatne, mniejsze funkcje
    __init__.py
  requirements.txt
  setup.py
  tests                             # katalog z testami modułu
    api                          # katalog z testami modułu przechowywania danych
    pipeline                     # katalog z testami modułu pipeline
      __init__.py
      tasks                      # katalog z testami modułów etapów pipeline
        __init__.py
        test_process.py          # testy etapów ekstrakcji pipeline
        test_postprocess.py      # testy etapów końcowych pipeline
        test_preprocess.py       # testy etapów początkowych pipeline
  tox.ini                         # konfiguracja narzędzia używanego do testowania
```

2.3.5 GŁÓWNE KLASY MODUŁU

- `MongoDBTask` - klasa bazowa dla etapów łańcucha przetwarzania danych korzystających z danych znajdujących się w bazie MongoDB.
- `TagsFindingTask` - klasa bazowa dla etapu pipeline'a, który znajduje technologie(tagi) w ofercie.
- `LanguageNotSupported` - klasa błędu, rzucanego w przypadku próby przetwarzania oferty napisanej w języku innym niż polski.
- `prepare` - etap przygotowujący dane do dalszego przetwarzania.
- `strip_html_tags` - etap usuwający tagi HTML z ofert.
- `tokenize` - etap rozbijający opis oferty na tokeny.
- `detect_language` - etap przeprowadzający detekcję języka oraz kończący przetwarzanie w przypadku detekcji języka innego niż polski.
- `remove_stopwords` - etap usuwający tokeny składające się ze słów nieistotnych.
- `find_tags` - etap znajdujący tagi z technologiami spośród tokenów oferty.
- `save_to_mongodb` - etap zapisujący ofertę do bazy MongoDB.
- `pracuj_pipeline` - pipeline wykorzystywany do ekstrakcji z oferty niezbędnych cech, takich jak tagi z technologiami wymienionymi w opisie oferty.

2.3.6 TESTY

W kodzie źródłowym modułu znajdują się testy jednostkowe pozwalające na przetestowanie poprawności zaimplementowanych metod oraz funkcji. W tej sekcji znajduje się również opis przykładowego scenariusza testów akceptacyjnych.

2.3.6.1 Testy jednostkowe

Aby uruchomić testy jednostkowe w konsoli należy wpisać polecenie `tox`.

Poniżej przedstawiamy listę plików z testami jednostkowymi oraz opis poszczególnych funkcji lub metod:

- `pipeline.tasks.test_preprocess.py` - plik z testami etapów przygotowujących oferty do ekstrakcji kluczy.
 - `TestDetectLanguage` - klasa zawierająca metody testujące etap detekcji języka w którym napisane jest ogłoszenie.
 - `test_prepare_extract_proper_fields_from_offer` - test sprawdzający czy etap przygotowujący ofertę, tworzy oraz wybiera odpowiednie pola z całej oferty.

- `test_remove_stopwords_return_tokens_without_stopwords` - test weryfikujący poprawne usunięcie tokenów będących słowami nieistotnymi.
 - `test_strip_html_tags_return_tokens_without_html_tags` - test sprawdzający czy etap poprawnie usuwa tagi HTML.
 - `test_tokenize_return_list_of_lowercase_tokens` - test sprawdzający czy etap zwraca odpowiednią listę tokenów składających się z małych liter.
- `pipeline.tasks.test_process.py` - plik z testami etapów wykonujących ekstrakcję kluczy z technologiami z oferty.
 - `pipeline.tasks.test_postprocess.py` - plik z testami etapów wykonujących zadania po wykonaniu ekstrakcji kluczy.

2.3.6.2 Testy akceptacyjne

Funkcje modułu podczas normalnej pracy systemu działają w tle, uzupełniając przychodzące z modułu zbierania danych oferty o listę wykrytych kluczy. Poprawne jego działanie, można więc przetestować wywołując łańcuch przetwarzania ręcznie, dla wybranej z bazy oferty. Proces wygląda następująco:

- W linii poleceń pobieramy jedną z ofert z bazy
- Wywołujemy na niej łańcuch przetwarzania
- Podglądamy listę wykrytych kluczy

```

In [25]: def test_pipeline(offer_id):
...:     # Get offer from database
...:     db = MongoClient(ProductionConfig.MONGO_URI).get_database()
...:     offer = db['offers'].find_one({'offer_id': str(offer_id)}, {'_id': False})
...:
...:     # Run pipeline
...:     result = pracuj_pipeline.apply((offer,)).get()
...:
...:     # Print list of tags
...:     pprint(result['tags'])
...:

```

Rysunek 2.12: Przykładowy skrypt testujący

```

In [31]: print_content('5708181')

```

Zadania:

analiza wymagań i projektowanie elementów rozwiązania,
tworzenie oprogramowania, wraz z zapewnieniem odpowiedniej jakości poprzez tworzenie testów
jednostkowych/integracyjnych,
wsparcie w procesie utrzymania.

Od kandydatów oczekujemy:

dobrej znajomości języka programowania Python,
doświadczenia w aplikacjach webowych,
znajomości relacyjnych baz danych i języka SQL
znajomości usług sieciowych i protokołów: HTTP, SOAP,
znajomości języka angielskiego,
umiejętność pracy w zespole.

Dodatkowym atutem będzie znajomość:

wzorców projektowych, systemu Linux, systemu kontroli wersji (SVN),
języków: PHP, JavaScript, HTML, CSS,
usług sieciowych: REST,
bibliotek/oprogramowania: Django, DOJO.

Rysunek 2.13: Tekst przykładowej oferty


```
In [24]: test_pipeline('5708181')
['python',
 'sql',
 'http',
 'soap',
 'linux',
 'svn',
 'php',
 'javascript',
 'html',
 'css',
 'rest',
 'django',
 'dojo']
```

Rysunek 2.14: Rezultat

2.4 Moduł statystyk i wyszukiwarki

Kod źródłowy znajduje się pod adresem: github.com/jobsbrowser/backend.

Na tym etapie w naszym systemie znajduje się już na bieżąco powiększana lista ofert wraz z przypasowanymi im kluczami. Jednym z brakujących elementów jest wydajny i łatwy w użyciu interfejs pozwalający na wykonywanie zapytań na tym zbiorze, tj. wyszukiwanie po kluczu oraz generowanie statystyk. Będzie on później wykorzystywany przez aplikację WWW prezentującą wyniki takich zapytań użytkownikowi. Wystawieniem takiego interfejsu zajmie się moduł, któremu poświęcony jest ten rozdział.

2.4.1 WYMAGANIA

Przeznaczeniem modułu jest odpowiadanie na żądania wyszukiwania oraz generowania statystyk. Komunikacja taka powinna odbywać się z wykorzystaniem protokołu HTTP, tak aby łatwo można było zintegrować moduł z aplikacją WWW bądź w przypadku takiej potrzeby innym dowolnym konsumentem.

Z powodu złożoności zapytań (jak np. generowanie statystyk sumarycznych dla wielu dni i wielu kluczy jednocześnie) moduł potrzebuje własnej bazy danych na której wykonywane będą zapytania. Wymusza to dodatkowe wymaganie w postaci możliwości dodawania nowych ofert, z którego korzysta moduł ekstrakcji kluczy (wysyłając na bieżąco w pełni przetworzone oferty).

Wymagania funkcjonalne modułu sprowadzają się więc do obsługi następujących żądań HTTP:

- dodanie nowej oferty do bazy
- wyszukiwanie ofert przy pomocy danego zbioru kluczy
- generowanie statystyk, tj. dla każdego dnia z zakresu od 02.12.2018 do daty wykonania żądania obliczenie ilości aktywnych wówczas w serwisie pracuj.pl ofert zawierających podane klucze
- generowanie statystyk całosciowych - tj. ilości wszystkich ofert w bazie (również we wspomnianym wyżej zakresie) oraz daty dodania ostatniego ogłoszenia

Natomiast wymagania niefunkcjonalne postawione modułowi to:

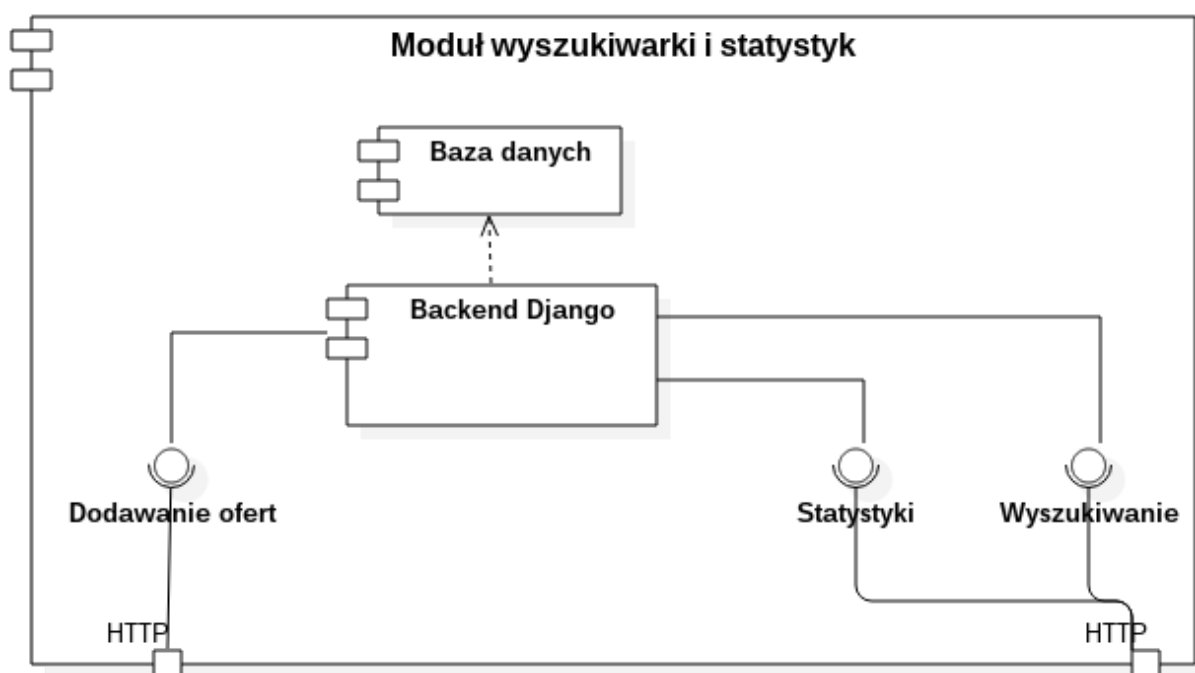
- **niezawodność** - usługa powinna być dostępna cały czas, ponieważ dostarcza ona danych niezbędnych do pokazania wszystkich statystyk w aplikacji z której korzysta użytkownik. Bez niej aplikacja WWW staje się bezużyteczna.

- **wydajność** - liczba przychodzących do modułu danych może być momentami bardzo duża, dlatego zapisywanie do bazy nie może być zrobione niewydajnie. Największy wpływ na wydajność mają zapytania generujące odpowiednie statystyki. Niektóre z nich są skomplikowane, dlatego niezbędna jest ich optymalizacja. Wydłużony czas generowania statystyk obniży znacznie komfort używania aplikacji WWW

2.4.2 INTERFEJS

Moduł z pozostałymi komponentami systemu łączy się przez interfejs HTTP. Dane do statystyk są dostarczane przez moduł ekstrakcji kluczy. Natomiast konsumentem danych generowanych przez moduł statystyk jest aplikacja WWW z której korzysta końcowy użytkownik.

Interfejs HTTP zaimplementowany jest przy użyciu frameworka Django(Django n.d.). Django jest to jeden z największych oraz najprężniej rozwijanych narzędzi Open Source. Zdecydowaliśmy się na niego ze względu na to, że w tym module kluczowa jest wydajność generowania odpowiednich statystyk, ale też duża elastyczność w pisaniu kodu. Django jako projekt niezwykle dojrzały, rozwijany przez bardzo doświadczonych developerów z całego świata jest świetnie zoptymalizowany do tego typu zadań.



Rysunek 2.15: Schemat modułu.

2.4.3 BAZA DANYCH

Wykorzystanym silnikiem bazy danych jest SQLite3(SQLite n.d.). Zdecydowaliśmy się na bazę relacyjną ze względu na to, że wykonujemy skomplikowane zapytania, które silnikom nierelacyjnych baz danych zajmują dużo więcej czasu oraz zasobów serwera, o czym przekonaliśmy się, testując te same zapytania na bazie danych MongoDB(MongoDB n.d.). Struktura dokumentów przechowywanych w bazie jest relacyjnym odzwierciedleniem struktury zebranego ogłoszenia oraz dokumentu przechowywanego przez moduł zbierający dane. Wyróżniliśmy trzy tabele:

- Oferty
 - Adres URL
 - Czas w którym pobrano ofertę
 - Kod HTML strony z ofertą
 - ID oferty w systemie pracuj.pl
 - Data dodania
 - Data ważności
 - Podmiot dodający
 - Tytuł oferty
 - Miejsce pracy
 - Kategorie oferty
 - Kod HTML treści oferty (rozbity na opis, kwalifikacje oraz benefity - wg struktury Pracuj.pl)
- Tagi
 - nazwa tagu
- Dodatkowa tabela reprezentująca relację ofert z tagami - jest to bowiem relacja many-to-many - czyli wiele do wielu.

2.4.3.1 Indeksy

W celu optymalizacji często wykonywanych zapytań skorzystaliśmy z indeksów. W bazie kluczy indeksowaną kolumną jest nazwa tagu, zaś w bazie ofert data dodania oferty oraz data wygaśnięcia oferty.

2.4.4 URUCHOMIENIE

Uruchomiony serwer działa nieprzerwanie nasłuchując nadchodzących połączeń na podanym hoście oraz porcie. Do uruchomienia korzystamy z polecenia `python manage.py runserver`

[host:port] wykonywanego w nadrzędnym katalogu modułu.

2.4.5 STRUKTURA KODU

Poniżej prezentujemy drzewo katalogów oraz plików modułu wraz z krótkim omówieniem. Struktura ta odpowiada szablonowemu projektowi działającemu z użyciem Django. Zdecydowaną większość funkcjonalności (filtrowanie, paginację, serializowanie modeli, wyszukiwanie po atrybucie) rozwiązaliśmy korzystając z gotowych, oferowanych przez framework narzędzi. Główna część logiki stworzonej przez nas znajduje się w pliku `offers/views.py` i jest to zoptymalizowane generowanie statystyk.

```
backend
├── config                                # moduł ustawień (Django)
│   ├── __init__.py
│   ├── settings.py                    # plik z ustawieniami
│   ├── urls.py                        # konfiguracja adresów URL serwera
│   └── wsgi.py                        # aplikacja WSGI (np. do integracji z Apache)
├── offers                               # moduł aplikacji
│   ├── migrations                     # migracje (kod generujący schemat bazy danych na podstawie modelu)
│   ├── __init__.py
│   ├── apps.py
│   ├── filters.py                    # filtrowanie po tagach
│   ├── models.py                     # modele oferty oraz klucza
│   ├── pagination.py                 # paginacja (używana przy wyszukiwaniu ofert)
│   ├── serializers.py                # serializacja modeli
│   ├── tests.py                      # testy jednostkowe
│   └── views.py                      # główna część logiki - implementacja interfejsu
├── manage.py                          # skrypt do zarządzania serwerem i jego uruchamiania
└── requirements.txt                  # lista koniecznych do zainstalowania zależności
```

2.4.6 GŁÓWNE KLASY MODUŁU

Najbardziej istotnymi klasami są:

- **Offer** - Model oferty. Tutaj zadeklarowane są pola stanowiące strukturę tabeli w bazie danych.
- **Tag** - Model pojedynczego klucza. Zawiera jedynie pole na nazwę.
- **OffersListView** - klasa widoku obsługującego wyszukiwanie ofert.

- **OffersStatsView** - klasa widoku obsługującego generowanie statystyk
- **SystemInfoView** - klasa widoku obsługującego statystyki całłościowe

każda z tych klas jest rozszerzeniem funkcjonalności oferowanych przez framework, dziedzicząc po odpowiedniej klasie bazowej. Logika implementowana przez nas znajduje się tylko w klasach **OffersStatsView** oraz **SystemInfoView** w postaci własnych metod. W pozostałych dostosowanie klasy oferowanej przez framework do naszych potrzeb sprowadza się do ustawienia odpowiednich atrybutów.

2.4.7 ALGORYTM GENEROWANIA STATYSTYK

O ile wyszukiwanie ofert na podstawie listy podanych kluczy realizowane jest w wydajny i optymalny sposób przez framework którego używamy, o tyle generowanie statystyk w określony przez wymagania modułu sposób jest nieco bardziej skomplikowane i wymagało implementacji własnego algorytmu.

Przypominając wymagania, interfejs statystyk ma działać w następujący sposób:

- Przyjmij listę kluczy
- Dla każdego dnia z przedziału od 02.12.2018 (początek zbierania ofert) do dzisiaj policz ile tego dnia było w serwisie pracuj.pl aktywnych ofert zawierających te klucze oraz jaki procent wszystkich aktywnych ofert z branży IT stanowiły.
- Zwróć wynik

Ofertę uważamy za aktywną danego dnia, jeśli zachodzą dwa warunki: **dzień dodania** \leq **wybrany dzień** oraz **dzień wygaśnięcia** \geq **wybrany dzień**. Mając te dwie daty w bazie danych możemy więc konstruować zapytania wybierające odpowiednie oferty.

Naiwna implementacja algorytmu, sprowadzałaby się do wykonania n zapytań zliczających do bazy danych, gdzie n to ilość dni które upłynęły od daty początkowej. Nie jest to rozwiązanie optymalne, ponieważ zapytania do bazy danych są kosztowne a z każdym kolejnym dniem wygenerowanie statystyk wymagałoby ich więcej.

Postanowiliśmy ograniczyć się do jednego zapytania, a algorytm wygląda następująco:

1. Wybierz oferty aktywne przez choć jeden dzień na zadanym przedziale. Wykorzystane warunki to: **dzień dodania** \leq **koniec przedziału**, **dzień wygaśnięcia** \geq **początek przedziału** oraz warunek zawierania się wszystkich podanych kluczy wśród kluczy oferty.

2. Utwórz n “kubelków”, gdzie każdy kubek odpowiada jednej dacie z przedziału i na początku ma wartość 0
3. Przejdź po liście ofert i dla każdej z nich:
 - Zwiększ wartość kubka odpowiadającego dacie dodania oferty o 1. Może się zdarzyć że data dodania oferty poprzedza początek zakresu, i nie ma takiego kubka. Wtedy zwiększ wartość pierwszego kubka.
 - Jeśli data wygaśnięcia oferty poprzedza koniec zakresu, to zmniejsz wartość w kubku z dniem następnym po dniu wygaśnięcia.
4. Wykonaj operację sumy skumulowanej na kubkach.

Dzięki temu w każdym kubku odpowiadającym każdemu dniu z zakresu znajdzie się ilość pasujących do zapytania, aktywnych tamtego dnia ofert. Do uzyskania wyniku procentowego operację powtarzamy, ale w kroku 1 pomijamy warunek dotyczący kluczy. Wtedy dzielimy jeden wynik przez drugi.

2.4.8 TESTY

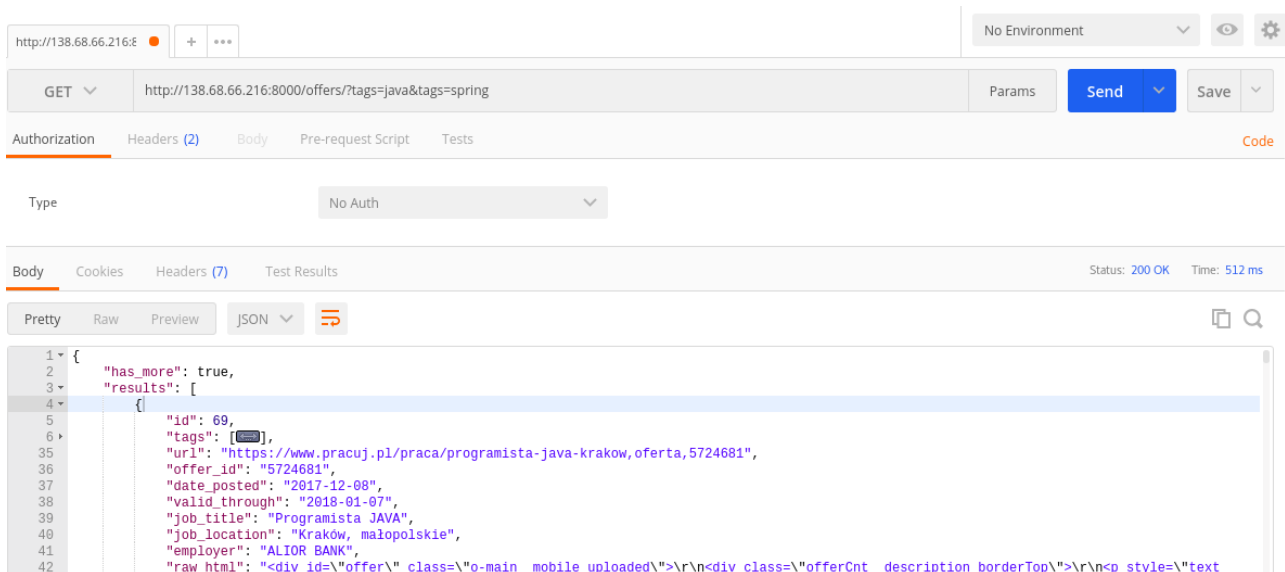
2.4.8.1 Testy jednostkowe

Aby uruchomić testy jednostkowe w konsoli należy wpisać polecenie `python manage.py test`.

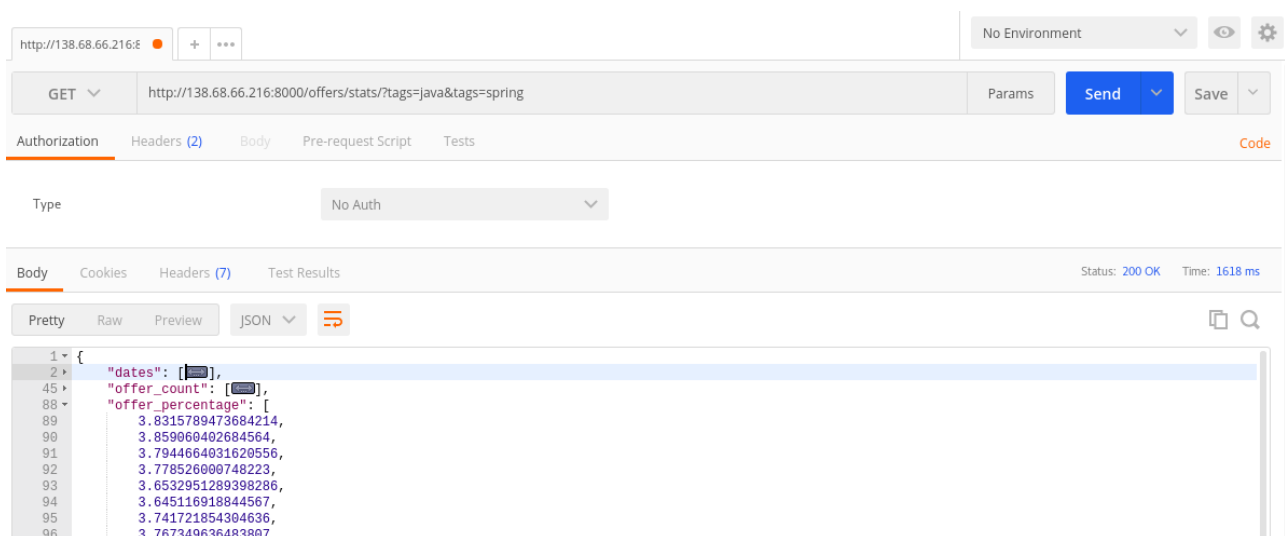
Testy jednostkowe znajdują się w pliku `offers/offers.py` i obejmują klasy oraz metody implementujące logikę widoków znajdujących się w pliku `offers/views.py`.

2.4.8.2 Testy akceptacyjne

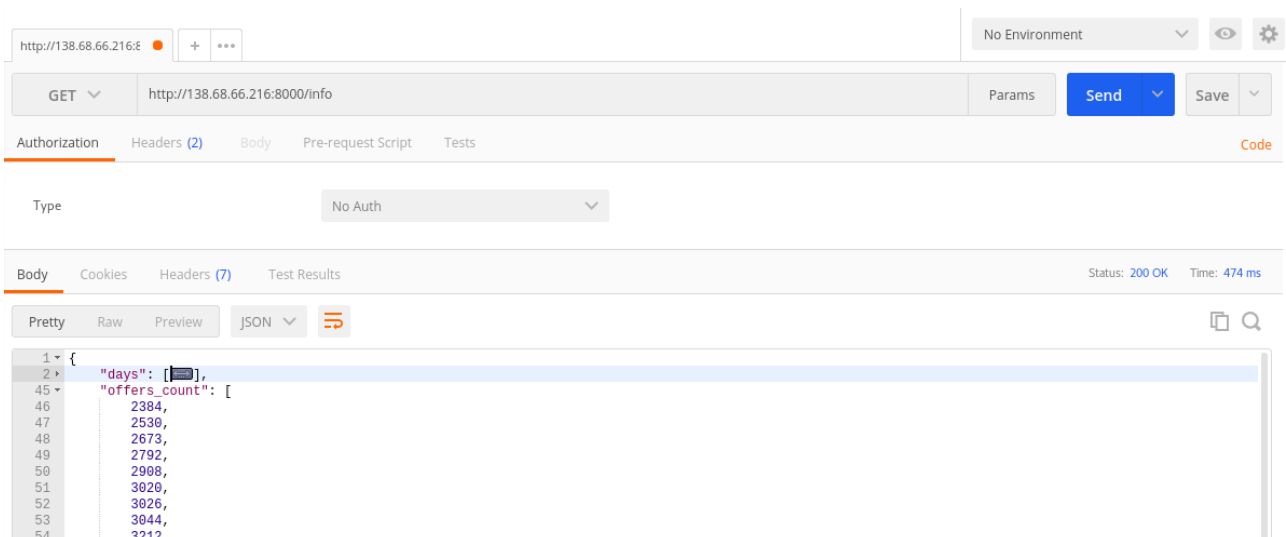
Moduł oferuje funkcjonalność serwera backendowego, brak jest w nim jakiegokolwiek warstwy wizualnej. Możliwość sprawdzenia poprawności działania sprowadza się więc do wykonania zapytań HTTP z odpowiednimi parametrami oraz sprawdzenia odpowiedzi.



Rysunek 2.16: Wyszukiwanie ofert.



Rysunek 2.17: Generowanie statystyk.



Rysunek 2.18: Statystyki zbiorcze.

2.5 Aplikacja webowa

Kod źródłowy znajduje się pod adresem: github.com/jobsbrowser/frontend.

Aplikacja będąca ostatnim komponentem systemu oraz jedynym który wchodzi w bezpośrednią interakcję z użytkownikiem.

2.5.1 WYMAGANIA

Głównym wymaganiem funkcjonalnym aplikacji jest prezentowanie danych oraz wyciągniętych z nich statystyk i informacji użytkownikowi w jak najbardziej przystępny oraz przejrzysty sposób. Rozwiązaniem jest organizacja strony na trzy zakładki pomiędzy którymi użytkownik może się swobodnie przełączać. Zakładki odpowiadają odpowiednio możliwościom:

- wyszukiwania ofert na podstawie wpisanych w polu wyszukiwania kluczy.
- generowania statystyk na podstawie wpisanych w polu wyszukiwania kluczy
- przeglądania strony informacyjnej wraz ze statystykami zbiorczymi

Wymagania niefunkcjonalne sprowadzają się natomiast do:

- **niezawodności** - usługa powinna być dostępna cały czas, ponieważ jest ona usługą końcową z której użytkownicy powinni móc korzystać w każdej chwili.
- **wydajności** - liczba użytkowników korzystających ze strony może być bardzo duża, dlatego też aplikacja powinna być napisana wydajnie, aby jej niedostępność lub zbyt wolne działanie nie przynosiły negatywnych odczuć użytkownikowi.

2.5.2 VUEJS

Do stworzenia aplikacji zdecydowaliśmy się użyć stosunkowo nowego frameworka VueJS (VueJS n.d.). Jest to framework napisany w języku JavaScript oparty na architekturze MVVVM (Model-View-View-Model). VueJS umożliwia tworzenie lekkich oraz szybkich aplikacji webowych. Budowanie aplikacji polega na tworzeniu coraz to nowych komponentów i składaniu z nich całej aplikacji. VueJS jest świetnym wyborem przy tworzeniu aplikacji, które opierają się na konsumowaniu danych z różnych API oraz prezentowaniu ich użytkownikowi, a więc wydaje się być rozwiązaniem dopasowanym do naszych potrzeb. Jest wydawany na licencji MIT.

2.5.2.1 Dodatkowe biblioteki

Do budowy aplikacji skorzystaliśmy z kilku rozszerzeń usprawniających pracę z Vue. Przedstawiamy je poniżej:

- **vuetify**(Vuetify n.d.) - framework dostarczający wiele pożytecznych komponentów stworzonych w stylu material design.
- **vuex**(Vuex n.d.) - system(wzorzec) zarządzania stanem aplikacji napisanej w VueJS. Zapewnia miejsce na przechowywanie danych, które będą dostępne w każdym komponencie oraz kontroluje ich nadpisanie, zmianę. Jest także przydatny przy debugowaniu aplikacji.
- **chart.js**(Chart.js n.d.) - biblioteka używana przez nas do tworzenia wykresów, prezentowania statystyk.
- **axios**(axios n.d.) - klient HTTP, za pomocą którego w wygodny sposób komunikujemy się z modułem statystyk.

2.5.3 URUCHOMIENIE

Aplikacja jest stroną internetową tak więc do korzystania z niej niezbędna jest odpowiednia przeglądarka internetowa. Poniżej prezentujemy przeglądarki kompatybilne z aplikacją:

- Google Chrome w wersji 49 lub wyższej
- Mozilla Firefox w wersji 52 lub wyższej
- Safari w wersji 10.1 lub wyższej

2.5.3.1 Wersja developerska

W celu uruchomienia wersji developerskiej należy przejść do głównego katalogu projektu oraz wykonać następujące polecenia:

- 1) `npm install`
- 2) `npm run dev`

Wersja developerska aplikacji będzie dostępna pod adresem `http://localhost:8080`.

2.5.4 STRUKTURA KODU

Poniżej prezentujemy drzewo katalogów oraz plików modułu wraz z krótkim omówieniem.

```

index.html          # główny plik HTML
package.json        # plik z zależnościami projektu
package-lock.json
public              # katalog na zasoby niezmieniające się np. zdjęcia, ikony
README.md
src                  # katalog z kodem źródłowym aplikacji
  App.vue            # główny komponent aplikacji
  components          # katalog z komponentami UI
    LineChart.js     # komponent obsługujący rysowanie wykresów liniowych
    Menu.vue          # komponent reprezentujący menu aplikacji
    Offer.vue         # komponent reprezentujący widok oferty
    TagInput.vue      # komponent obsługujący pobieranie kluczy od użytkownika
  main.js             # plik wejściowy dla aplikacji
  pages               # katalog z komponentami reprezentującymi podstrony aplikacji
    Info.vue          # komponent prezentujący informacje o projekcie
    Search.vue         # komponent umożliwiający interaktywne wyszukiwanie ofert
    Stats.vue         # komponent pokazujący statystyki
  router
    index.js          # konfiguracje routowania aplikacji
  store
    index.js          # konfiguracja oraz inicjalizacja vuex-store
webpack.config.js    # ustawienia babel

```

2.5.5 GŁÓWNE KOMPONENTY APLIKACJI

Poniżej prezentujemy najważniejsze komponenty aplikacji wraz z krótkim opisem:

- komponenty UI
 - Menu - komponent generujący widok górnego menu.
 - TagInput - komponent obsługujący pobieranie kluczy (tagów) od użytkownika w interaktywny sposób.
 - Offer - komponent generujący widok listy ofert z odpowiednimi tagami pobranymi od użytkownika poprzez komponent TagInput.
 - LineChart - komponent renderujący wykresy liniowe dla kluczy pobranych dzięki komponentowi TagInput.
- komponenty stron
 - Search - komponent korzystający z komponentów: Menu, TagInput oraz Offer. Pobiera informacje o ofertach z pobranymi tagami i prezentuje je w formie listy.

- **Info** - komponent korzystający z komponentów: **Menu**. Pobiera informacje na temat projektu **JobsBrowser** oraz je prezentuje.
- **Stats** - komponent korzystający z komponentów: **Menu**, **TagInput** oraz **LineChart**. Pobiera statystyki dotyczące wybranych kluczy oraz prezentuje je użytkownikowi w formie wykresów.

2.5.6 INSTRUKCJA UŻYTKOWANIA ORAZ TESTY AKCEPTACYJNE

Z racji tego, że funkcjonalność aplikacji jest bardzo zwięzła instrukcja jej użytkowania może mieć postać dokumentacji testów akceptacyjnych i tak też chcielibyśmy ją przedstawić.

2.5.6.1 Zakładka statystyk

Jest to pierwsza zakładka którą jako użytkownik widzimy po wprowadzeniu w przeglądarce adresu aplikacji. W górnej części strony znajduje się pasek menu pozwalający na zmianę zakładki, a w środkowej pole do wpisywania kluczy.

Po wprowadzeniu klucza i wciśnięciu klawisza Enter zakoloruje się on, a pod spodem pojawiają się dwa wykresy. Wpisany klucz można usunąć lub dodać do niego kolejny pisząc w polu i ponownie wciskając Enter.

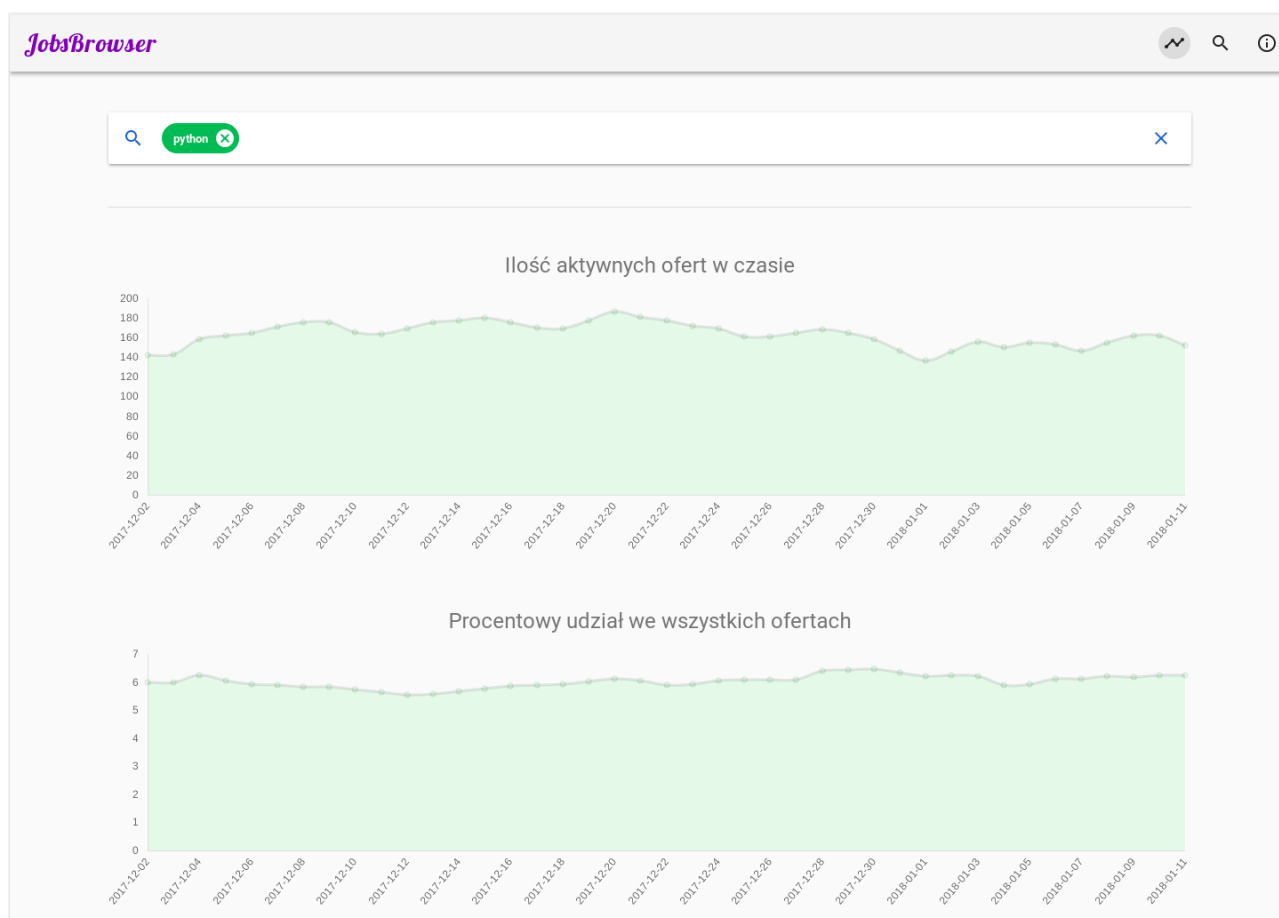
Na pierwszym wykresie przedstawiona będzie ilość aktywnych ofert zawierających dany klucz konkretnego dnia, a na drugim jaką procentowo część wszystkich aktywnych wówczas ofert z branży IT one stanowiły. Wykresy są w pewnym stopniu interaktywne. Tzn. nie pozwalają na zmianę skali czy zakresu, ale pozwalają na dokładne zbadanie wartości w określonym dniu najeżdżając na wykres kursorem myszy.

2.5.6.2 Zakładka wyszukiwania

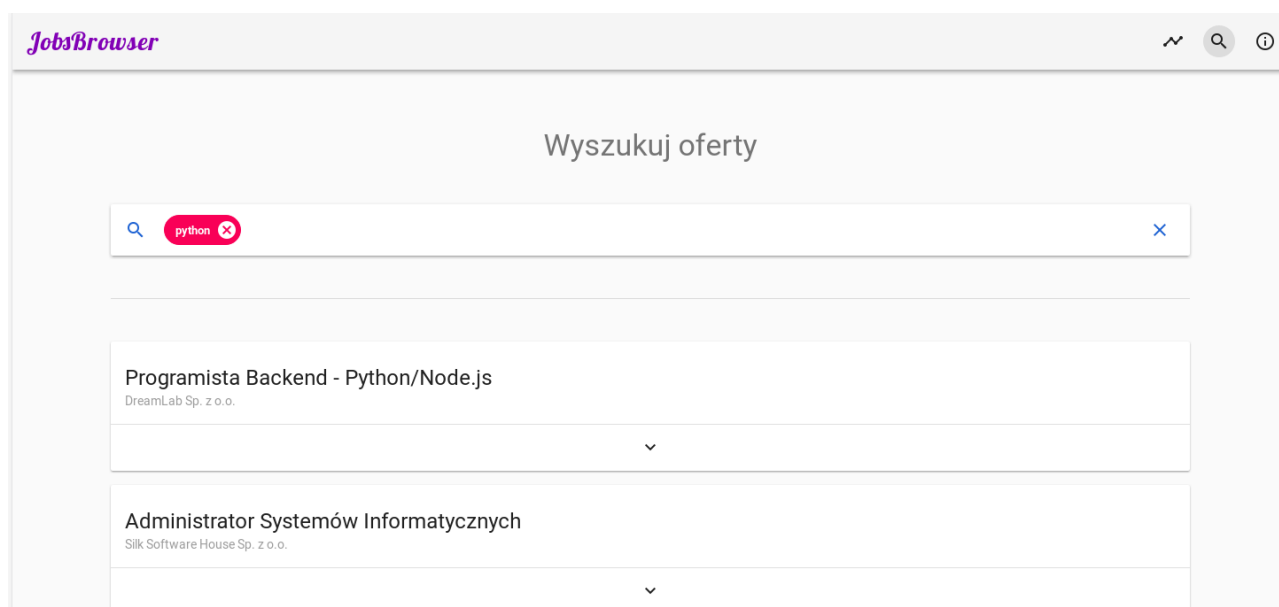
Drugą dostępną z menu zakładką jest ta z widokiem wyszukiwania. Tutaj ponownie zobaczymy identyczne pole do wpisywania kluczy, jednak tym razem pod spodem pojawi nam się lista znalezionych ofert które ten klucz zawierają. Każdą z nich możemy rozwinąć żeby zobaczyć więcej informacji, w tym link do oryginalnego ogłoszenia czy pozostałe wykryte w ofercie klucze. Lista jest paginowana, co zwiększa komfort użytkownika i poprawia szybkość ładowania wyników.

2.5.6.3 Zakładka informacji

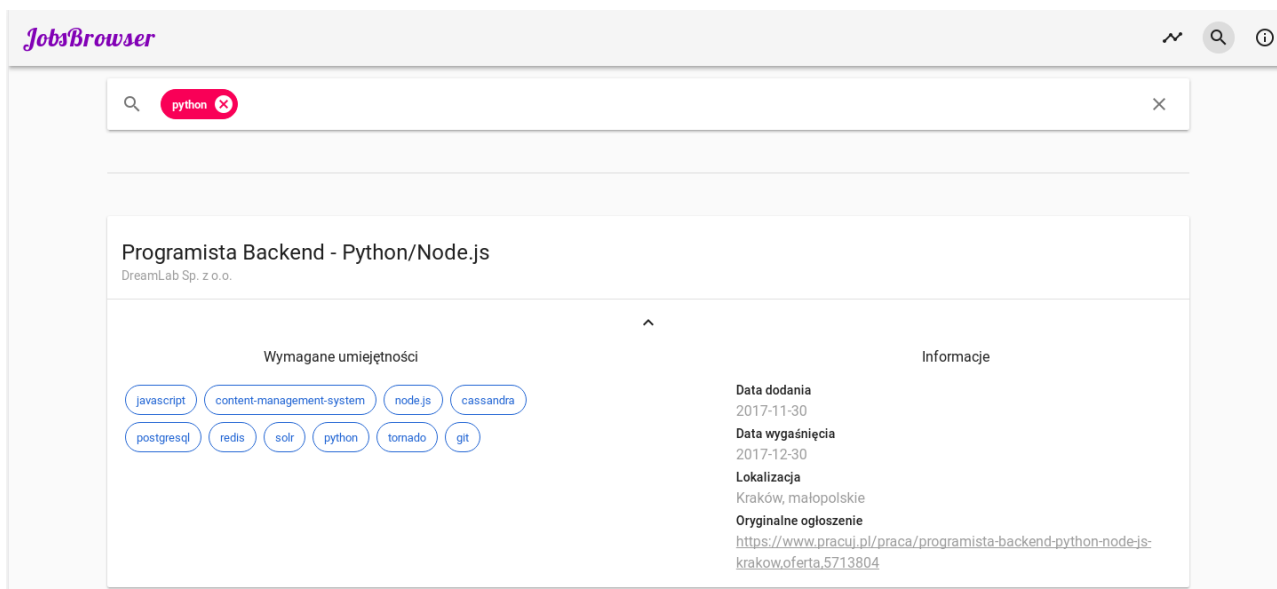
Ostatnią zakładką jest ta poświęcona informacjom o serwisie oraz statystykom zbiorczym. Zobaczymy tam krótki opis projektu oraz wykres ilości wszystkich przetworzonych przez niego ofert względem czasu.



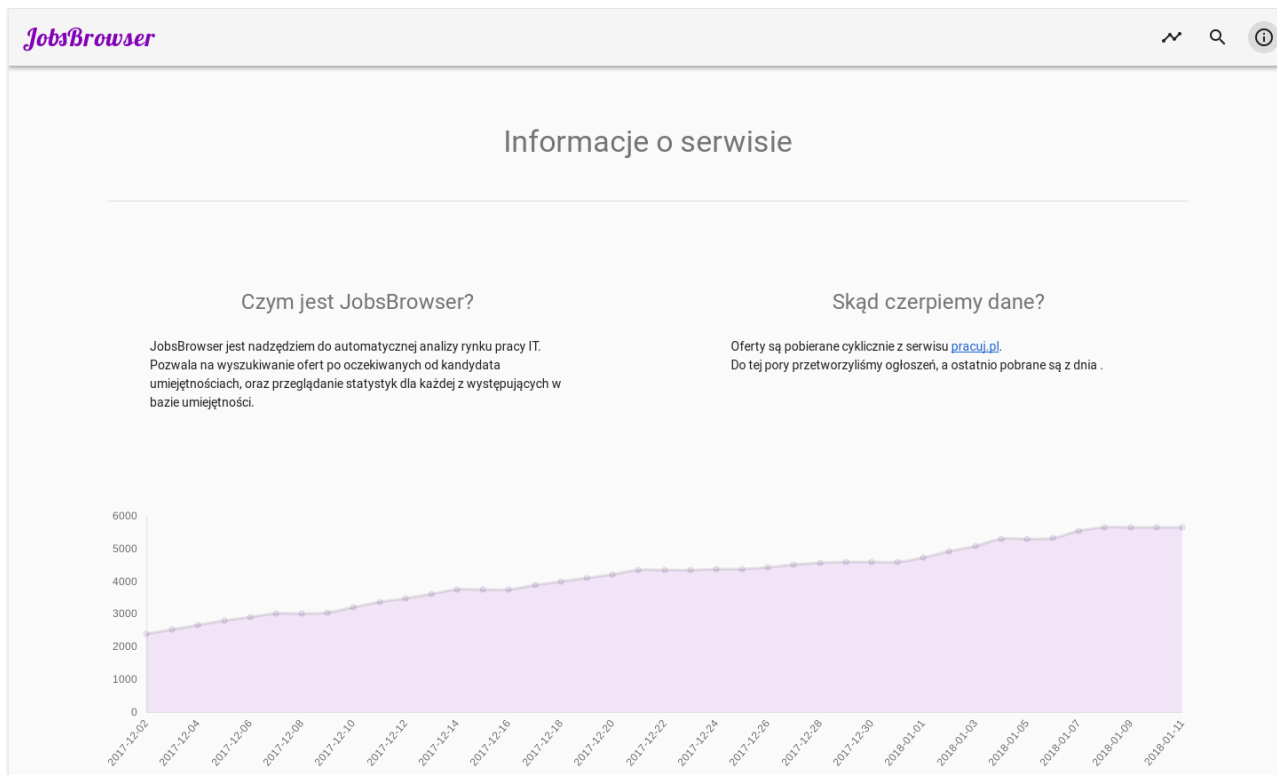
Rysunek 2.19: Zakładka statystyk.



Rysunek 2.20: Zakładka wyszukiwania.



Rysunek 2.21: Rozwinięte ogłoszenie.



Rysunek 2.22: Zakładka informacji.

Historia zmian dokumentu

Tabela 2.1: Historia Zmian

Data	Autor	Opis zmian	Wersja
23.11.2017	Bartłomiej Sielicki	Rozdział 1	0.1
	Łukasz Skarżyński	Rozdział 2 - moduł zbierania danych	

Bibliografia

axios. Klient http napisany w node.js. Dostępne pod adresem: <https://github.com/axios/axios> [2018-01-13].

Bengio Yoshua, V.P., Ducharme Réjean. Advances in neural information processing systems 13 (nips'00). Dostępne pod adresem: <http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/64> [2018-01-27].

Celery. Framework pozwalający na asynchroniczne wykonywanie zadań i łańcuchów przetwarzania (pipeline). Dostępne pod adresem: <http://www.celeryproject.org/> [2017-12-16].

Chart.js. Biblioteka ułatwiająca rysowanie wykresów. Dostępne pod adresem: <http://www.chartjs.org/> [2018-01-13].

Django. Framework wykorzystywany do tworzenia aplikacji webowych oraz rest api. Dostępne pod adresem: <http://www.djangoproject.com/> [2018-01-13].

Flask. Framework wykorzystywany do tworzenia stron internetowych oraz api. Dostępne pod adresem: <http://flask.pocoo.org/> [2017-12-16].

Gensim. Biblioteka przeznaczona do przetwarzania języka naturalnego oraz pozyskiwania informacji w tekście. Dostępne pod adresem: <https://github.com/RaRe-Technologies/gensim> [2018-01-27].

MongoDB. Nierelacyjna baza danych, przechowująca rekordy w formacie bson. Dostępne pod adresem: <https://www.mongodb.com/> [2017-12-16].

Rake-NLTK. Implementacja w języku python (z użyciem nltk) algorytmu rake (en. rapid automatic keyword extraction). wykorzystywany do automatycznej ekstrakcji słów kluczowych z ogłoszeń. Dostępne pod adresem: <https://github.com/csurfer/rake-nltk> [2017-12-16].

Scrapy. Zestaw narzędzi umożliwiający szybką, wydajną ekstrakcję informacji ze stron internetowych. Dostępne pod adresem: <https://scrapy.org> [2017-12-16].

SQLite. System zarządzania relacyjną bazą danych. Dostępne pod adresem: <https://www.sqlite.org/> [2018-01-13].

TF-IDF. Algorytm wyznaczania wagi słów na podstawie ich wystąpień. Dostępne pod adresem: <https://en.wikipedia.org/wiki/Tf-idf> [2018-01-21].

VueJS. Framework wykorzystywany do tworzenia progresywnych aplikacji internetowych. Dostępne pod adresem: <https://vuejs.org> [2017-12-16].

Vuetify. Zestaw komponentów ui do vuejs zbudowany przy użyciu wyglądu material design. Dostępne pod

adresem: <https://vuetifyjs.com/> [2018-01-13].

Vuex. System zarządzania stanem komponentów vuejs. Dostępne pod adresem: <https://vuex.vuejs.org> [2018-01-13].

Yoav Goldberg, O.L. Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. Dostępne pod adresem: <https://arxiv.org/abs/1402.3722> [2018-01-27].