UNIWERSYTET EKONOMICZNY W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Adam Bień

Zastosowanie ukrytej alokacji Dirichleta w analizie ofert pracy w służbie cywilnej

The application of the latent Dirichlet allocation in the analysis of job offers in the civil service

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Specjalność: Analityka gospodarcza

Poznań 2019

Spis treści

Wstęp

Rynek pracy jest jednym z najważniejszych rynków działających w gospodarce i odpowiadających za jej stan. Badanie go i dokładne zrozumienie poprzez analizy jest kluczowe, by móc podejmować przemyślane decyzje biznesowe. Jednym z ważniejszych aspektów dotyczących rynku pracy jest popyt na pracę, który nie tylko świadczy o kierunku rozwoju gospodarczego, ale także dotyczy każdego obywatela ubiegającego się o pracę. Poprzez analizę popytu na pracę dowiadujemy się informacji jakie kompetencje u pracowników poszukiwane są przez pracodawców i w jakiej wielkości. Te informacje pomogą dostosować się pracownikom np. poprzez podejmowanie właściwych decyzji o kształceniu się.

Istnieje wiele źródeł informacji o popycie na rynku pracy w Polsce. Badania przeprowadzane z różną częstotliwością i przez różne instytucje – te publiczne jak i prywatne – dostarczają wielu cennych informacji, i, w zależności od charakterystyki badania, przychodzą ze swoimi zaletami jak i wadami. Jednymi z największych wad konwencjonalnych badań jest czas ich trwania, koszt i w wybranych przypadkach czas między realizacją badania, a publikacją wyników.

Z perspektywy osoby poszukującej zatrudnienia, od wielu lat wyszukiwarki internetowe ofert pracy górują nad innymi metodami szukania pracy. Jest to bardzo wygodne narzędzie zarówno dla osoby ubiegającej się o prace, jak i dla pracodawcy oferującego stanowisko. Pracownik jest w stanie za pośrednictwem internetu przejrzeć setki ofert pracy z okolic swojego zamieszkania lub poszerzyć zakres poszukiwań na cały kraj i nawet świat. Może on także zawęzić spectrum poszukiwań tylko do takich, które go interesują np. ograniczając poszukiwania tylko do stanowisk, na które dana osoba jest kwalifikowana. Dla pracodawcy zaś publikowanie ofert na stronach internetowych jest bardzo tanim sposobem, aby dotrzeć do potencjalnie ogromnego grona odbiorców niskim kosztem.

W tej pracy analizowane są oferty pracy na stanowiska w służbie cywilnej umieszczane na stronie naborów Kancelarii Prezesa Rady Ministrów poprzez ich masowe pobieranie metodami web-scrapingu i przybliżenie zapotrzebowania na różne dziedziny działalności gospodarczej po-

przez zastosowanie modelu alokacji ukrytej Dichirleta. Model ten pozwala na utworzenie grup utożsamianych z tematyką ofert pracy bazując na analizie tekstu zawartego w ich opisie. Jest to badanie przeprowadzane niskim kosztem i pozwalające na przeanalizowanie dużej próby, przez co wyniki są bardziej prawdopodobne by być wiarygodne. Wyniki te mogą dostarczyć wielu cennych informacji, takich jak zmiana zapotrzebowania na usługi konkretnego typu w czasie, a także wyznaczenie działalności cieszących się największym zapotrzebowaniem.

Praca składa się z trzech rozdziałów. Pierwszy rozdział ma na celu przybliżyć czytelnikowi mechanizm działania rynku pracy, a także opisane są w nim różne źródła informacji, z których można korzystać w celu analizy zagadnień dotyczących rynku. Drugi rozdział poświęcony jest przedstawieniu głównych metod wykorzystywanych w przeprowadzanym badaniu – web-srapingu oraz modelu ukrytej alokacji Dichirleta. Trzeci rozdział poświęcony jest komentarzowi przebiegu badania oraz analizie otrzymanych z niego rezultatów. Praca wzbogacona jest także w użyty w trakcie badania kod języka R i liczne wykresy pomagające w interpretacji wyników.

Rozdział 1

Rynek pracy

1.1 Rynek pracy w ujęciu teoretycznym

Gospodarka jest złożonym systemem reguł i mechanizmów funkcjonujących w określonych warunkach. W jej obrębie operuje jednocześnie wiele powiązanych ze sobą rynków, których prawidłowe funkcjonowanie determinuje rozwój gospodarczy kraju. Do najbardziej istotnych rynków, z ekonomicznego punktu widzenia, możemy zaliczyć (Dębek i in., 2016): rynek kapitałowy, pieniężny, produktów, usług czy też rynek pracy, któremu poświęcone będą rozważania w tym rozdziale.

1.1.1 Definicja rynku i rynku pracy

Aby dobrze zrozumieć istotę rynku pracy i panujące na nim zasady, należy najpierw zdefiniować bardziej ogólne pojęcie, jakim jest sam rynek. Ten utożsamić można z wymianą dóbr między sprzedawcami i nabywcami. Wzajemne oddziaływanie na siebie sprzedawców, chcących osiągnąć maksymalne zyski ze sprzedaży, oraz konsumentów, chcących zaspokoić swoja potrzebę na dany produkt po jak najkorzystniejszej cenie, jest fundamentem działającego na rynku mechanizmu – mechanizmu rynkowego. Ów mechanizm prowadzi do ustalenia takiej ceny, po której obie strony gotowe są do przystąpienia do wymiany towarowej. Określana jest także podaż rozumiana jako wielkość oferowanego przez sprzedawców towaru, oraz popyt utożsamiany z ilością towaru, który nabywcy są gotowi zakupić za daną cenę. W trakcie transakcji budowana jest relacja pomiędzy nabywcą, a sprzedawcą. W gospodarce rynkowej, czyli takiej, w której dominuje własność prywatna, mechanizm rynkowy prowadzi do walki konkurencyjnej między producentami. To zaś zapewnia podnoszenie jakości produktu oferowanego przez

sprzedawców, zaś konsumentom pozwala na porównanie ofert i dokonywanie najkorzystniejszego dla nich zakupu, tym samym zwiększając ich zadowolenie. Ostatecznie można stwierdzić, że mechanizm rynkowy prowadzi do uzyskania pożytku obu stronom transakcji (Milewski & Wydawnictwo Naukowe, 1994; Mizia & Latocha, 2019).

Na rynku pracy funkcję sprzedawcy pełni pracobiorca, czyli osoba gotowa do jej podjęcia, tym samym oferując swoje usługi, będące na tym rynku utożsamiane z towarem. Nabywcą, chcącym pozyskać pracowników, jest pracodawca, zaś ceną wynikającą z relacji między podażą a popytem na pracę - wynagrodzenie. Takie, za które pracownik zgodzi się pracować u pracodawcy. Innymi słowy, rynek pracy można rozumieć jako miejsce, w którym w ustalonych warunkach zachodzą transakcje wymiany usług pracy pomiędzy pracobiorcą, a pracodawcą. Wyznaczane są na nim rozmiary tych transakcji oraz ich wartość wyrażona w płacy. Warty zauważenia jest fakt, że kwalifikacje i umiejętności dostarczane przez pracobiorcę postrzegane są przez pracodawcę jako warunki pozwalające pracownikowi do sprostania wykonywania pracy. Nie są one zaś przedmiotem wymiany. Usługą jest czas poświęcony przez pracownika na wykonanie zleconych zadań. Dodatkowo, pracownik bardzo często poszerza swoje kwalifikacje wykonując pracę, co byłoby sprzeczne rozumiejąc je jako element składowy przedmiotu wymiany. (Kryńska & Kwiatkowski, 2013)

1.1.2 Podaż i popyt na rynku pracy

Podaż pracy, wymiennie nazywana siłą roboczą, tworzy ludność aktywna zawodowo, czyli liczba osób zatrudnionych w wieku produkcyjnym, jak i liczba bezrobotnych, czyli osób w wieku produkcyjnym nie wykonujących żadnej pracy, lecz aktywnie jej poszukujące i gotowe do jej podjęcia. Pozostałą część społeczeństwa określamy mianem biernych zawodowo. Ponieważ jednak nie każdy pracownik zatrudniony jest w identycznym wymiarze czasu, możemy podaż identyfikować z liczbą roboczogodzin, którą ludzie są gotowi przeznaczyć na wykonywanie pracy. Na wielkość siły roboczej wpływa wiele czynników, najważniejszym będąca wysokość płacy za wykonywane usługi, jak i wysokość płacy zawodów pokrewnych, wymagających od pracownika podobnych kwalifikacji. Dodatkowo wymienić można czynniki demograficzne, takie jak stan ludności mieszkającej na terenie obejmowanym przez rynek pracy, jej struktura wiekowa, płci i wykształcenia, model rodziny, przyrost naturalny czy wielkość i kierunek migracji. Do czynników prawnych zaliczyć można granicę nabywania uprawnień emerytalnych, obowiązkowy okres przebywania w szkole i obowiązujący wymiar czasu pracy. Czynniki ekonomiczne

wpływające na podaż to np. wspomniana wcześniej płaca, ale także poziom cen w kraju albo inne źródła przychodu jak świadczenia społeczne. Ostatnim typem czynników wpływających na podaż są czynniki społeczne rozumiane poprzez tradycje i obyczaje, które mogą wpływać na podejście ludności do pracy (Kryńska & Kwiatkowski, 2013).

Wielkość popytu na pracę reprezentowana jest przez liczbę osób, jakie pracodawcy chętni są zatrudnić za ustaloną płacę, albo liczba godzin potrzebnych na wykonanie pracy. Ponieważ zwiększenie zatrudnienia przez pracodawcę jest przykładem prawa malejących przychodów, pracodawca będzie gotowy do zatrudnienia tylko takiej liczby pracowników, których wytwarzane przychody będą większe niż generowane koszty. Co więcej, popyt na pracę jest nierozłącznie związany z popytem na produkty wytwarzane przez pracowników, ale także z innymi czynnikami takimi jak wielkość oczekiwanego przez pracowników wynagrodzenia czy postęp technologiczny zastępujący pracownika w procesie produkcji (Węc & Lelakowska, 2019).

Pracodawcą może być organizacja bądź osoba fizyczna, która chętna jest zatrudnić inne organizacje i/lub osoby fizyczne po odpowiedniej cenie. Ze względu na wielkość przedsiębiorstwa na rynku pracy możemy wyróżnić (Szaban, 2013):

- Mikroprzedsiębiorstwo
  - zatrudniające do 9-ciu pracowników,
  - roczny obrót nie przekracza 2 milionów euro.
- Małe przedsiębiorstwo
  - zatrudniające do 50-ciu pracowników,
  - obrót nie przekracza 10 milionów euro.

- Średnie przedsiębiorstwo
  - zatrudniające do 250 pracowników,
  - obrót nie przekracza 43 milionów euro.
- Duże przedsiębiorstwo
  - zatrudniające więcej niż 250 pracowników
  - nie będące ograniczone od góry granicą generowanych obrotów.

1.1.3 Różne perspektywy rynku pracy

Na rynek pracy składa się wiele aspektów, wymiarów i perspektyw, które należy zidentyfikować w celu przeprowadzenia analizy. Podstawowym wymiarem występującym na rynku pracy jest wymiar przestrzenny określający skalę geograficzną, w której operuje rynek pracy. Rynkiem ogólnokrajowym nazwiemy rynek pracy obejmujący swoja skalą cały kraj, wszystkich pracodawców i pracowników. Analiza krajowego rynku pracy pozwala na porównanie tego rynku z innymi rynkami krajowymi jak i z okresami poprzednimi w danym państwie. Zmniejszając skalę mówić będziemy o regionalnym rynku pracy. Jednym z założeń tej perspektywy jest to, że pracownicy są w stanie pracować bez konieczności zmiany miejsca zamieszkania poprzez codzienne dojazdy do miejsca pracy. Często pracodawcy zatrudniają do pracy ludzi z danego regionu, przez co znacząco redukują koszty rekrutacji. Ponadto w regionalnych rynkach pracy bardzo często występuje pewna specyfika pracy związana zazwyczaj z dobrze rozwiniętą dziedziną gospodarki funkcjonującą w tym regionie. W Polsce opis powyższego regionu najbardziej oddaje podział administracyjny na województwa. Rynek lokalny jest najmniejszym wyodrębnieniem pod względem geograficznym. Dotyczy on powierzchni znacznie mniejszej niż rynek regionalny, chociaż pod względem cech charakterystycznych różni sie on tylko ich siłą nasycenia. W rynkach lokalnych mamy do czynienia z bardzo niskimi kosztami dojazdów i szukania pracy, zaś informacja dotycząca miejsc i pracowników bardzo dobra. Podziałem administracyjnym dobrze oddającym realia rynków lokalnych w Polsce są powiaty i gminy (Kryńska & Kwiatkowski, 2013).

Innym aspektem rynku pracy, który często stosowany jest do jego podziału jest wymiar kwalifikacyjno-zawody, rozumiany jako zbiór zadań wykonywanych przez osoby z odpowiednimi umiejętnościami, wiedzą i uprawnieniami. Rynek pracy można podzielić pod względem

zawodów występujących na nim, których wykonanie nie jest możliwe bez uprzedniego nabycia umiejętności. Zapewnia to izolację rynku danego zawodu utożsamianą z niskim przepływem osób między rynkami. Tworzy to niezależną podaż siły roboczej zdolnej do wykonywania danej pracy, tak samo jak pewien popyt pracodawców zainteresowanych w usługach ludzi o skonkretyzowanych umiejętnościach (Kryńska & Kwiatkowski, 2013).

1.2 Źródła danych o rynku pracy

Wiedza jest istotnym zasobem w gospodarce, umożliwiającym monitorowanie zachodzących procesów i dostosowywanie się do nich poprzez podejmowanie właściwych decyzji. Pozyskiwanie informacji przydatnych i rzetelnych jest zatem niezwykle ważne. Dane, które można wykorzystać w opracowaniach statystycznych można uzyskać z rozmaitych źródeł. Kluczowym aspektem determinującym, z którego źródła powinniśmy czerpać informacje, jest jakość tych danych, a także ich reprezentatywność względem interesującego nas zjawiska. W tym podrozdziale zamierzam opisać różne źródła informacji dostarczające wiedzy o rynku pracy, a w szczególności o popycie na nią, co będzie tematem rozważań w dalszej części pracy.

1.2.1 Informacje dotyczące rynku pracy udostępniane przez urzędy

Źródłem informacji zapewniającym wszechstronne dane dotyczące rynku pracy, w tym popytu na pracę, są urzędy statystyczne, posiadające duży zasób informacji pierwotnych, czyli uzyskiwanych z bezpośrednich pomiarów danego zjawiska, pochodzących m.in. z:

- badań przedsiębiorstw,
- badań ankietowych ludności,
- sprawozdań instytucji finansów publicznych,
- sprawozdań ZUS i KRUS,
- zbiorczych opracowań urzędów pracy.

Innym źródłem informacji są urzędy pracy, które zbierają i wykorzystują w swojej bieżącej działalności informacje dotyczące rynku pracy. Atutami posiłkowania się informacjami pochodzącymi z urzędów jest ich wiarygodność, dostępność i ich darmowy dostęp. Wadą jednak

jest fakt, że wyniki badań publikowane są często z opóźnieniem. W analizie zjawisk mających tendencję do częstych zmian (np. branże szybko rozwijające się) takie opóźnienie może spowodować różność między stanem faktycznym, a tym wynikającym z badań (Góra & Sztanderska, 2006).

1.2.1.1 Badanie Aktywności Ekonomicznej Ludności

Kluczowym badaniem GUS o rynku pracy jest Badania Aktywności Ekonomicznej Ludności (BAEL). Jest to reprezentacyjne badanie gospodarstw domowych i osób w wieku 15 lat i więcej przeprowadzane metodą wywiadów bezpośrednich mające na celu pozyskanie informacji o wielkości zasobów siły roboczej, ich struktury według podstawowych cech demograficznych i społecznych, przestrzennego rozmieszczenia zasobów siły roboczej i statusu na runku pracy. Wyniki tego badania służą do wyliczania wskaźników z zakresu rynku pracy jak i do porównań międzynarodowych. Metoda ankietowa zakłada profesjonalne przygotowanie kwestionariusza, próby losowej oraz obróbki uzyskanych danych. BAEL przeprowadzone jest przy użyciu dwóch kwestionariuszy. Jednego wspólnego dla wszystkich członków wylosowanego gospodarstwa, oraz kwestionariusza indywidualnego dla każdego z członków w wieku 15 lat i więcej. Podstawowym problemem metody ankietowej jest koszt przeprowadzenia badania, zaś atutem wysoka wartość poznawcza (Góra & Sztanderska, 2006; Główny Urząd Statystyczny, 2019).

1.2.1.2 Badanie Popyt na Pracę

Informacji o popycie na pracę dostarcza badanie przeprowadzane przez GUS pod tytułem Popyt na Pracę. Ma ono na celu dostarczenie informacji o popycie zrealizowanym oraz niezrealizowanym, czyli o liczbie pracujących i liczbie wolnych miejsc pracy według zawodów. Informuje także o liczbie nowo utworzonych i zlikwidowanych miejsc pracy według cech charakteryzujących zakłady pracy: rozmieszczenia przestrzennego, sektorów własności, rodzajów działalności oraz wielkości jednostek według liczby pracujących. Badanie realizowane jest metodą reprezentacyjną co kwartał i obejmuje podmioty gospodarki narodowej o liczbie zatrudnionych 1 lub więcej osób (Główny Urząd Statystyczny, 2018).

1.2.1.3 Oferty pracy publikowane przez Powiatowe Urzędy Pracy

Szukając informacji na temat popytu na pracę warto przeanalizować dane publikowane przez urzędy pracy. Urzędy pracy pełnią funkcję pośredniczącą między pracodawcami i potencjal-

nymi pracownikami, ale także zajmują się badaniem i analizowaniem rynku pracy. Wyróżniamy powiatowe i wojewódzkie urzędy pracy, jednak nie są one sobie podległe – realizują inny zakres zadań.

Powiatowe urzędy pracy przede wszystkim zajmują się przyjmowaniem oraz upowszechnianiem ofert pracy. Pracodawca zgłaszający ofertę pracy do urzędu może oczekiwać odpłatnego poszukiwania kandydatów na terenie Polski jak i Europejskiego Obszaru Gospodarczego.

Aby ogłosić swoją ofertę pracy, pracodawca może to zrobić odwiedzając urząd pracy, ale także wykorzystując e-formularz. Wymagane jest dostarczenie informacji dotyczących m.in:

- pracodawcy (nazwa, adres, numer kontaktowy, numer identyfikacji podatkowej, informacje dotyczące ukarania),
- miejsca i warunków pracy (nazwa stanowiska, liczba wolnych miejsc pracy, zakres obowiązków, rodzaj umowy)
- wymagań, które spełnić muszą potencjalni pracownicy (poziom wykształcenia, umiejętności, uprawnienia, doświadczenie zawodowe, znajomość języków obcych).

Po zweryfikowaniu poprawności i kompletności oferty pracy, urząd publikuje ją w siedzibie odpowiedniego powiatowego urzędu pracy, w Centralnej Bazie Ofert Pracy oraz w bazie ofert pracy European Employment Services (EURES) (Adamowicz, 2019)

Centralna Baza Ofert Pracy jest serwisem, będącym wspólną bazą wszystkich powiatowych urzędów pracy. Są tam publikowane oferty pracy dla osób jej poszukujących, zarejestrowanych w urzędach w całej Polsce. Prowadzi ją Ministerstwo Rodziny, Pracy i Polityki Społecznej. Na stronie znaleźć można również informacje na temat targów pracy, giełd pracy i szkoleń organizowanych przez powiatowe i wojewódzkie urzędy pracy (Kuraś, 2017).

EURES jest siecią współpracy Publicznych Służb Zatrudnienia, która ułatwia swobodny przepływ pracowników pomiędzy państwami wchodzącymi w skład Europejskiego Obszaru Gospodarczego, tj. państw członkowskich Unii Europejskiej, Szwajcarii, Islandii, Lichtensteinu i Norwegii. W ramach usług EURES dostępna jest baza ofert pracy ze wszystkich krajów EOG.

1.2.1.4 Oferty pracy w Służbie Cywilnej

Innym miejscem do szukania zatrudnienia są ogłoszenia o naborach w pracy w Służbie Cywilnej publikowane na stronie Kancelarii Prezesa Rady Ministrów (KPRM). Na oficjalnej stronie czytamy, że ”w bazie znajdują się ogłoszenia o naborach na stanowiska urzędnicze w administracji

rządowej, z wyłączeniem wyższych stanowisk w służbie cywilnej (np. dyrektorów departamentów), które są obsadzane na podstawie powołania". Aby ubiegać się o pracę w Służbie Cywilnej, kandydat powinien spełniać następujące wymogi:

- posiadać polskie obywatelstwo (z wyjątkiem wybranych stanowisk),
- korzystać z pełni praw publicznych,
- spełniać wymagania żądane na danym stanowisku,
- być osobą, która nie została skazana prawomocnym wyrokiem.

1.2.2 Pozostałe źródła informacji o rynku pracy

1.2.2.1 Bilans Kapitału Ludzkiego

Projekt *Bilans Kapitału Ludzkiego* prowadzony przez Polską Agencję Rozwoju Przedsiębiorczości we współpracy z Uniwersytetem Jagiellońskim ma na celu zbadanie i śledzenie zapotrzebowania na kompetencje na rynku pracy. Pierwsza edycja projektu uskuteczniona w latach 2011-2014 obejmowała coroczne badanie ludności, pracodawców, analizę ofert pracy, badanie firm i instytucji szkoleniowych oraz uczniów i studentów. II edycja BKL przeprowadzana będzie w latach 2017-2022. W ramach drugiej osłony projektu egzekwowane będą 3 fale badań przekrojowych oraz 3 fale badań panelowych, jak i badania branżowe z sektorów: finansów, turystyki i IT. Badanie i analiza ofert pracy dostarcza nam wygląd na zgłaszane preferencje kompetencji u pracowników przez pracodawców. Porównanie ich ze stanem rzeczywistym na rynku pracy pozwala na określenie ewentualnych braków kompetencyjnych w konkretnych branżach i sektorach. (Centrum Ewaluacji i Analiz Polityk Publicznych, 2014) (Polska Agencja Rozwoju Przedsiębiorczości, 2018)

Badanie Popytu na Pracę w projekcie BKL oparte było głównie na zebraniu informacji od pracodawców oraz z ofert pracy.

Pracodawca sprecyzowany był jako podmiot gospodarczy zatrudniający na czas prowadzonego badania przynajmniej jednego pracownika. Pominięte zostały osoby samozatrudnione – te ujęte zostały w badaniu ludności. Co więcej, wykluczone z badania zostały podmioty z kilku sekcji Polskiej Klasyfikacji Działalności 2007. Kontakt z respondentami przeprowadzany był na różny sposób: osobisty, telefoniczny i internetowy, z czego kontakt telefoniczny wykorzystany został do zebrania zdecydowanej większości obserwacji.

Oferta pracy objaśniona została jako unikalne ogłoszenie o pracę na terenie 16 województw kraju. ”Unikalna”kojarzona jest z opublikowanym określonego dnia, na pojedyncze stanowisko pracy ogłoszeniem, które nie może się powtarzać w różnych źródłach. Wykluczone zostały oferty staży i praktyk, oraz oferty pracy poza granicami kraju. Tak zdefiniowane oferty pracy zbierane były z Powiatowych Urzędów Pracy oraz z internetowego portalu `careerjet.pl` – ogólnokrajowego portalu pośrednictwa pracy.

1.3 Internetowe źródła informacji

Coraz popularniejszym źródłem informacji staje się Internet. Praktycznie nieograniczone zasoby danych z każdej dziedziny sprawiają, że bardzo łatwo oraz szybko jest zgromadzić interesujące nas dane niskim kosztem. Istotną wadą w bazowaniu na internetowych źródłach danych jest sporna celność i wiarygodność tych danych. Wstawiane przez użytkowników informacje nie zawsze są wystarczająco szczegółowo weryfikowane, a także nie są ustrukturyzowane.

1.3.1 Wyszukiwarki pracy

Publikowane w sieci informacje mogą służyć do analizy rynku pracy – w szczególności popytu na pracę. Internet jest bowiem najczęściej wybieranym narzędziem w poszukiwaniu pracy, oraz jest najczęściej wykorzystywanym źródłem poszukiwania pracowników przez pracodawców. Istnieje bardzo dużo portali internetowych zamieszczających oferty pracy. Do najbardziej popularnych zaliczyć można `olx.pl`, `pracuj.pl`, `goldenline.pl`, `linkedin.com`, ale także wcześniej wspomniany `careerjet.pl`. Różne wyszukiwarki pracy charakteryzują odmienne ogłoszenia np. na portalu Gumtree większość ofert pracy dotyczy gastronomii, recepcji i call center, zaś pod adresem `[nabory.kprm.gov.pl` znajdziemy oferty pracy na stanowiska urzędnicze w administracji rządowej. Warto mieć na uwadze taką specyfikę przeprowadzając analizę posługując się źródłem internetowym.

1.3.1.1 Portal Służby Cywilnej

Konstytucyjnym zwierzchnikiem służby cywilnej jest Prezes Rady Ministrów, który powołuje i nadzoruje pracę szefa służby cywilnej. Szef służby cywilnej jest centralnym organem administracji rządowej właściwym w sprawach służby cywilnej. Obsługę Szefa Służby Cywilnej oraz Rady Służby Publicznej zapewnia Kancelaria Prezesa Rady Ministrów.

Celem pracy pracowników w służbie cywilnej, jak wyjaśnione zostało na oficjalnej stronie, to ”budowa nowoczesnego państwa, wzrost efektywności działania jego organów i przede wszystkim zadowolenie społeczeństwa”. Stanowiska, a tym samym zadania, jakie realizowane są w służbie cywilnej są różnorodne, zaczynając od budowy dróg i kolei, na pomocy Polakom za granicą kończąc.

Nabór do pracy przebiega w warunkach otwartej konkurencji. Powoływany na stanowisko zostaje najlepszy kandydat z puli wszystkich osób, które spełniają wymagania formalne i złożyły dokumenty rekrutacyjne. Oferty pracy publikowane są na oficjalnej stronie KPRM, ale także w Biuletynie Informacji Publicznej urzędu,k który poszukuje pracownika, a także w siedzibie tego urzędu.

Portal oferujący stanowiska w służbie cywilnej – `https://nabory.kprm.gov.pl` – będzie wykorzystywany w dalszej cześci pracy.

1.3.2 Barometr Ofert Pracy

Badaniem, o którym warto wspomnieć mówiąc o analizowaniu popytu na pracę wykorzystując źródła internetowe jest Barometr Ofert Pracy prowadzony przez Biuro Inwestycji i Cykli Ekonomicznych przy współpracy z Instytutem Gospodarki Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie. Barometr przedstawiany jest w postaci indeksu w punktach procentowych, określając zmiany zapotrzebowania na nowych pracowników. Powstaje on na podstawie ofert pracy zbieranych z portali internetowych. Z danych eliminowany jest czynnik sezonowy, co umożliwia porównywalność rezultatów pomiędzy poszczególnymi porami roku oraz wyłącza wypływ zmian ogłoszeń sezonowych, a więc części ofert pracy dotyczących pracy tymczasowej. Poziom tego badania, wraz z innymi miarami, pozwala na określenie w jakiej fazie znajduje się rynek ofert pracy czy wskazać zmiany strukturalne. Przewagą tego indeksu jest to, że posiada on właściwości wyprzedzające w stosunku do zmian zatrudnienia i bezrobocia. (Biuro Inwestycji i Cykli Ekonomicznych, 2019)

1.4 Podsumowanie

W powyższym rozdziale przedstawiono podstawowe pojęcia pozwalające zrozumieć istotę oraz źródła informacji o rynku pracy. W dalszych rozdziałach przedstawione zostaną analizy ofert pracy publikowanych na stronach KPRM, w szczególności uwzględniając informacje o zakresie

obowiązków i kompetencji. W kolejnym rozdziale przedstawione zostanie podejście wykorzystane do analizy danych z KPRM, w szczególności w zakresie analizy danych tekstowych.

Rozdział 2

Model ukrytej alokacji Dirichleta w analizie tekstu

2.1 Ekstrakcja danych z Internetu – web scrapping

2.1.1 Istota web scrapingu

Pojęcie *web scraping* odnosi się do zautomatyzowanego procesu ekstrakcji danych zamieszczanych na witrynach internetowych w celu ich przechowywania i/lub dalszego przetwarzania.

Oprogramowanie służące do scrapowania nawiązuje połączenie z siecią WWW poprzez protokół HTTP bądź przez przeglądarkę i symuluje eksplorację strony taką, jaką wykonuje użytkownik przeglądając daną witrynę. Następnie program, często określany mianem *web crawlera*, kopiuje i zapisuje określone dane zawarte na stronach internetowych np. na dysku lub w bazie danych.

Web crawler to program bądź zautomatyzowany skrypt, który przegląda witryny sieci Web w usystematyzowany sposób. Może on wykonywać inne funkcje, w zależności na jaki użytek został on stworzony. Przykładowo web crawlery wykorzystywane są przez wyszukiwarki internetowe w celu indeksowania stron. W procesie zbierania danych, web crawler szuka konkretnych informacji zawartych na stronie internetowej, które następnie zbiera, agreguje i zapisuje. Kluczową częścią procesu pozyskiwania danych z witryn internetowych jest transformacja zawartych na nich danych. Dane, które pozyskujemy z innych źródeł, są zazwyczaj ustrukturyzowane, co oznacza, że są przystosowane do przetwarzania przez narzędzia komputerowe. Przeciwieństwem są dane zawarte na stronach internetowych. Te są najczęściej nieustrukturyzowane.

Bloki tekstu wyświetlające się użytkownikowi są z łatwością odczytywane i rozumiane przez człowieka, lecz sprawiają kłopot próbującymi przyswoić je komputerowi. Web crawler uzyskuje dostęp do danej strony internetowej, odnajduje pożądane fragmenty informacji, kopiuje je, aby dalej je przetransformować. Po zebraniu wszystkich interesujących fragmentów strony, zapisywany jest ustrukturyzowany zbiór danych gotowy do dalszej obróbki.

Proces web scrapingu podzielić można na dwie części.

- uzyskanie dostępu do witryny
- ekstrakcja danych

Pierwszy z nich to pobranie zawartości strony, tak jak robi to przeglądarka internetowa przy jej przeglądaniu. Gdy treść została wczytana, może nastąpić ekstrakcja wybranych fragmentów strony. Zawartość strony może być przeszukana, przeformatowana, przekopiowana, można na niej zastosować analizę składniową, zapisana itd.

2.1.1.1 Różne techniki web-scrapingu

Istnieje wiele metod i technik przeprowadzenia procesu web-scrapingu. Niektóre wymagają wysiłku użytkownika, by funkcjonować, inne zapewniają w pełni zautomatyzowane pobranie całej zawartości strony.

- „Ręczne" kopiowanie – Pomimo braku wykorzystania inteligentnego oprogramowania, które wykonuje żmudną pracę, wciąż zaliczamy ten sposób do metod web scrapingu. Tworzone są bowiem zbiory danych, których źródłem jest internet. Nie jest to sposób pożądany. Bywa jednak, że strony posiadają zabezpieczenia chroniące je przed przeglądającymi je botami. W takich wypadkach rozwiązanie to sprawdza się, jeśli do przekopiowania nie jest zlecony duży zbiór danych.
- Przeszukiwanie zawartości przy użyciu wzorców tekstowych – Bazujące na dopasowywaniu wyrażeń regularnych funkcji języków programowania, odnajdowane są interesujące użytkowania treści, które następnie można pobrać i zapisać.
- Analizowanie składniowe kodu źródłowego HTML – Dane podobnego typu składowane są na stronach internetowych w tak samo oznaczonych szablonach. Programy, których zadaniem jest wykrywanie istniejących szablonów, wyciągnięcie z nich zawartości i przetransformowanie jej nazywamy wrapperami.

- Odnajdywanie danych przy użyciu selektora Xpath/CSS – Selektory pozwalają na wyodrębnienie ścieżki dostępu do danych określonej kategorii na stronie internetowej, a następnie na przekazanie tej ścieżki dostępu funkcji pobierającej zawartość jakiegoś języka programowania.

2.1.2 Narzędzia do web scrapingu

Istnieje wiele narzędzi pozwalających na pobieranie i zapisywanie danych ze źródeł internetowych w ustrukturyzowany sposób. Możemy je podzielić na trzy główne kategorie:

- biblioteki wykorzystywane przez języki programowania,
- szkielety/frameworki,
- aplikacje komputerowe.

2.1.2.1 Biblioteki

Jednym z najbardziej powszechnie stosowanych środków do web scrapowania jest tworzenie własnego programu pobierającego dane przy użyciu wybranego języka programowania. Do tych tworzone są biblioteki przez firmy trzecie umożliwiające web scrapowanie np. poprzez dodanie funkcjonalności uzyskania dostępu do strony poprzez zaimplementowanie protokołu HTTP strony użytkownika. Same dane przetwarzane są przez funkcje dostarczane przez język programowania bądź inne pakiety.

Przykładami takich bibliotek są:

- Biblioteka cURL wykorzysytwana przez wiele języków programowania – Wspiera ona najistotniejsze funkcjonalności protokołu HTTP takie jak certyfikaty SSL, HTTP POST, HTTP PUT, zmienne proxy, zarządzanie plikami cookies, autoryzację HTTP
- Moduł WWW::Mechanize Web języka Perl – Pozwala na m.in interakcję z hiperłączami i formularzami, wspiera HTTPS, zarządzanie plikami cookies, autoryzację HTTP, zarządzanie historią, a także posiłkowanie się ścieżkami XPath.
- pakiet Apache HttpClient języka Java – Wspiera najważniejsze funkcje protokołu http, czyli wszystkie żądania, zarządzanie plikami cookies, autoryzację SSL i HTTP, co więcej, sama Java wspiera funkcjonalność ścieżek XPath.

- BeautifulSoup języka Python zapewnia analizę składniową HTML pozwalając na przeprowadzanie operacji na zawartości strony.

2.1.2.2 Frameworki

Korzystanie z robotów do scrapowania danych napisanych w językach programowania posiada wady. Często trzeba jednocześnie posiłkować się wieloma bibliotekami. Dodatkowo tak stworzone programy są wrażliwe na zmiany w kodzie źródłowym HTML strony internetowej, z czym wiąże się potrzeba ciągłego utrzymywania i pielęgnowania kodu. Zmiany wprowadzone w robocie wymagają ponownej kompilacji wszystkich elementów składowych aplikacji. Frameworki do scrapowania dostarczają bardziej zintegrowane rozwiązanie np. framework języka Python o nazwie Scrapy, w którym roboty scrapujące są zdefiniowane jako klasy dziedziczące funkcjonalność od klasy nadrzędnej BaseSpider. W niej opisane są zbiór początkowych adresów url oraz funkcja parse pozwalająca na ich odczytanie.

2.1.2.3 Aplikacje komputerowe

Aplikacje komputerowe stanowią alternatywę dla użytkowników niebędących swobodnymi w programowaniu. Użytkowanie wspomagane jest przez graficzny interfejs, pomagający w kreacji i utrzymaniu programu. Program pozwala użytkownikowi na poruszanie się w przeglądarce i zaznaczanie fragmentów odwiedzanych stron, które mają zostać pobrane. Użytkownik korzystając z takiej aplikacji nie musi się martwić o sprecyzowanie ścieżek XPath czy wyrażeń regularnych ani innych specyfikacji technicznych. Wybierany jest następnie sposób zapisu zabranego zbioru danych. Mogą to być np. pliki w rozszerzeniu CSV, XML lub xls/xlsx. Wadami takiego rozwiązania są koszt aplikacji i ograniczony dostęp do API. Przykłady takich narzędzi: Visual Web Ripper, Newbie, Mozenda, Screen-Scraper, FMiner.

2.1.3 Zastosowania web scrapingu w biznesie

Internetowe źródła informacji posiadają wielki potencjał, by być niezwykle zyskownymi. Obecnie web scraping używa się powszechnie do („Scraping Agent”, 2019; Jeffery, 2017; Patel, 2018):

- Zbierania danych kontaktowych – Stosowane w sprzedaży i marketingu przez wiele firm. Zbierane są adresy skrzynek mailowych leadów/ potencjalnych klientów dla danej spółki,

następnie rozsyłane są na nie hurtowo wiadomości email z propozycjami współpracy, reklamami produktów itd.

- Zbierania opinii i recenzji dotyczących produktów (swoich bądź wyprodukowanych przez konkurentów) – Dostarcza to informacje dotyczące zalet i wad produktu, ale także można wykorzystać te informacje porównując swoją pozycję na rynku względem konkurentów.

- Zbierania danych z portali społecznościowych – Media, osoby publiczne ale także duże firmy mogą wykorzystać wiedzę na temat tego, co aktualnie staje się popularne na dużą skalę np. w akcjach marketingowych, w budowaniu wizerunku itd.

- Budowania modeli uczenia maszynowego – Te, mające za zadanie przewidywanie rezultatów i odnajdywanie trendów, mają zastosowanie w wielu sferach biznesu. Dobrze zbudowany model opiera się na dużej ilości danych będących wzorcem do podejmowania przez model decyzji.

- Sprawdzania cen na podobne produkty – Przedsiębiorcy zajmujący się sprzedażą detaliczną mogą wspomóc się w ustalaniu cen na swoje produkty porównując je i dostosowując do cen na substytuty oferowane przez konkurencję.

- Znajdywania pracowników – Rekruterzy wykorzystują web scraping by odnaleźć profile osób pasujących na dane stanowisko.

- Zbierania i łącznego publikowania ofert – Może to dotyczyć np. ogłoszeń o pracę lub ofert sprzedaży nieruchomości. Powstają portale internetowe, które zbierają z różnych źródeł oferty i umieszczają je zagregowane w jednym miejscu.

- Zbieranie danych w celach badawczych – Internet oferuje potencjalnie bardzo dużymi zbiorami danych na cele badawcze niskim kosztem (w porównaniu do tradycyjnych metod zbierania obserwacji).

2.1.4 Kwestie legalne

Kwestie legalności masowego pobierania zawartości stron internetowych oraz wykorzystywania ich do własnych celów są tematem sporym, zwłaszcza, że w nie istnieją zapisy w prawie dotyczące precyzyjnie web-scrapingu. Wiele sporów prawnych w przeszłości opartych było na

prawach dotyczących naruszenia praw autorskich, naruszenia umowy korzystania z serwisu, czy naruszenia praw do własności ruchomej.

1. Warunki korzystania – Właściciele stron internetowych, chcąc zapobiec masowemu pobieraniu zawartości publikowanej na tych stronach, często uwzględniają taki zakaz w warunkach korzystania z serwisu. Niestosowanie się do zasad narzuconych przez ten dokument może prowadzić do naruszenia umowy korzystania z serwisu i może być to podstawą do działania prawnego. Problem stanowi fakt, że warunki prawne muszą zostać jednoznacznie zaakceptowane przez użytkownika, aby zobowiązywały go do ich przestrzegania (np. poprzez zaznaczenie wyświetlanego pola wyboru po wejściu na witrynę). W związku z tym, jeśli strona publikuje dane niewymagające wcześniejszej zgody korzystania z serwisu, nie ma podstaw, by oprogramowanie nie mogło ich pobierać i zapisywać.

2. Zawartość chroniona prawem autorskim – Pobieranie i publikowanie treści należącej do właściciela portalu internetowego i wyraźnie chronionej prawem autorskim może prowadzić do naruszenia praw autorskich. Jednak zawartość danego serwisu nie zawsze jest własnością osoby zarządzającej portalem. Przykładem mogą być materiały generowane przez innych użytkowników strony internetowej takie jak komentarze, wpisy na forum czy recenzje. Co więcej, prawa autorskie mogą być nadane na konkretny sposób reprezentacji pewnego pomysłu, nie zaś na sam pomysł. Tak więc korzystanie z danych w celu ich definitywnej transformacji stanowiącej inną treść niż ich pierwowzór nie podlega pod naruszenie praw autorskich. Ostatecznie, użytkownicy mogą także korzystać z treści chronionych prawem autorskim, bez zgody podmiotu praw autorskich, przestrzegając zasad dozwolonego użytku.

3. Prawo do własności ruchomej – Prawo do własności ruchomej może zostać wykorzystane przez właściciela serwera, jeśli będzie on w stanie udowodnić w sposób jednoznaczny i mierzalny, że użytkownik scrapujący stronę internetową wyrządził szkody materialne serwerowi. Jeśli takie dowody zostaną dostarczone, właściciel uszkodzonego serwera może ubiegać się o odszkodowanie finansowe za zniszczenia.

4. Etykieta w web-scrapingu – Użytkownik chcący korzystać z technik web-scrapingu jest zobowiązany do przestrzegania prawa, ale także powinien stosować się do dobrych praktyk etycznych, zakładających postawę niosącą korzyść, a nie szkodę innym. Web-scraping, nawet jeśli wykorzystywany legalnie, może nieintencjonalnie wyrządzać krzywdę innym

internautom - właścicielom stron internetowych, innym użytkownikom danej strony lub firmom związanym z działalnością internetową. Przykładami takich działań mogą być (Krotov & Silva, 2018):

- Produkt stworzony na bazie pobranych informacji i przetransformowany (np.: do postaci raportu bądź opracowania) może stać się konkurencją dla danych oferowanych przez oryginalne źródło tych informacji i prowadzić do strat finansowych.
- Użytkownik pobierający i korzystający z zawartości strony omija w ten sposób elementy strony takie jak reklamy czy ankiety, które zostały tam umieszczone, by nieść pożytek właścicielowi portalu.
- Opracowania bazujące na zawartości stron internetowych mogą ujawnić poufne informacje właściciela strony. Przykładowo zliczenie liczby reklam pojawiających się na witrynie biznesu opartego na przychodzie z reklam, może przybliżyć zarobki z nich płynące.

2.2 Text Mining

Za sprawą bardzo szybkiego rozwoju technologicznego ostatnich lat, żyjemy obecnie w czasach, w których generowane są ogromne ilości danych. Większość z nich przechowywana jest w Internecie i ma charakter tekstowy np.: artykuły, książki, komentarze czy mejle. W związku z chęcią wykorzystywania nowo powstałych źródeł informacji powstało zapotrzebowanie na techniki umożliwiające cyfrowe przetwarzanie tych danych. Te techniki, określane jako techniki text miningu, znajdują zastosowania w zdecydowanej większości dziedzin działalności człowieka.

Text mining to pojęcie określające wiele technologii i zastosowań, wszystkie stworzone do analizy i przetwarzania nieustrukturyzowanych danych tekstowych. Celem tych zabiegów jest przemiana tekstu do postaci, w której maszyny mogą swobodnie na nich operować. Taki zabieg wymaga zastosowania różnych technik do przetwarzania tekstu, zaczynając od radzenia sobie z pojedynczymi słowami, do zarządzania całymi dokumentami. W książce *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* autorzy rozróżniają siedem zastosowań Text-miningu ze względu na charakter każdego z rozwiązań.

1. Wyszukiwanie i odzyskiwanie informacji – Do tego zastosowania zalicza się przechowywanie, wyszukiwanie i odzyskiwanie dokumentów ze względu na ich zawartość. Świetnym

tego przykładem są wyszukiwarki, takie, jakie stosowane są np. w przeglądarce.

2. Klastrowanie dokumentów – Grupowanie i kategoryzowanie dowolnej długości ciągów słów ze względu na podobieństwa między obserwacjami.

3. Klasyfikacja dokumentów – grupowanie i kategoryzowanie dokumentów. Klasyfikacja różni się od klastrowania tym, że w klasyfikacji wiemy, że istnieją etykiety, do których dokumenty powinny być przypisane i jesteśmy w stanie zweryfikować błędne zaklasyfikowanie. Klastrowanie sprowadza się do wytworzenia skupień ze względu na podobieństwa między obserwacjami, ale nie ma możliwości porównania wyników algorytmu z poprawnym przypisaniem.

4. Web-mining – Wykorzystywanie technik text-miningu na danych w Internecie.

5. Ekstrakcja informacji – proces strukturyzowania nieustrukturyzowanych danych tekstowych poprzez identyfikację i wydobycie kluczowych informacji i zależności w tekście.

6. Natrual language processing – czyli przetwarzanie języka naturalnego, to zbiór technik mających na celu zrozumienie języka pisanego (bądź mówionego) przez maszynę

7. Wydobywanie pojęć – grupowanie słów i fraz w semantycznie spójne grupy.

W praktyce, podczas pracy z tekstem, często w ramach jednej analizy stosuje się podejścia łączące wiele z wymienionych wyżej zastosowań. Text-mining zazwyczaj wykorzystywany jest w dziedzinach takich jak: statystyka, sztuczna inteligencja, uczenie maszynowe, lingwistyka komputerowa, informatologia, bazy danych (Miner i in., 2012).

2.3 Latent Dirichlet Allocation

Jednym z najważniejszych aspektów przetwarzania danych tekstowych jest znalezienie słów kluczowych dokumentu. Z takiego zastosowania korzystamy na co dzień wpisując frazy w wyszukiwarce internetowej i oczekując listy proponowanych materiałów związanych z wprowadzonym zapytaniem, albo przeglądając proponowane media i produkty wybrane i wyświetlane dla nas na podstawie naszych wcześniejszych wyborów.

Model ukrytej alokacji Dirichleta (ang. Latent Dirichlet Allocation; LDA) jest algorytmem stworzonym przez Davida Blei'a i powstał przez rozszerzenie probabilistycznego modelu probabilistic latent semantic indexin (pLSI). Zalicza się on do rodziny modeli, których zadaniem jest

określenie podobieństw między dokumentami, a w szczególności określenie, czy treści danych dokumentów dotyczą podobnych bądź odmiennych tematów.

Model pLSI – probabilistic latent semantic indexing – jest bardzo podobny w założeniach co model LDA. Przewagą tego drugiego jest zamienienie paramteru θ z rozkładu wielomianowego na rozkład Dichirleta, co sprawia, że model ten lepiej generalizuje się dla nowych, nie widzianych wcześniej w zbiorze treningowym, obserwacji. Model LDA zakłada, że w korpusie składającym się z dokumentów istnieje wiele tematów. Tematy oznaczają tutaj pewne ukryte relacje (rozkłady) pomiędzy zmiennymi modelu jakimi są słowa występujące w korpusie dokumentów. Rozkłady te łączą słowa i ich występowanie w dokumentach by ostatecznie nadać każdemu z dokumentów kompozycję tematów jakie są poruszane w ich zawartości.

Zastosowanie modelu LDA początkowo zakładało jedynie modelowanie danych tekstowych, lecz obecnie algorytm ten wykorzystywany jest także dla obrazów oraz materiałów wideo. Dalej w pracy opisywać będę jedynie zastosowanie dla danych tekstowych.

2.3.1 Podstawy teoretyczne modelu LDA

Modelowanie LDA wymaga dostarczenia zbioru dokumentów. Dokument stanowi dowolnej długości ciąg słów. Może nim być w takim razie artykuł, książka, przepis kucharski, a także komentarz na stronie internetowej. Pełną kolekcję dokumentów nazywamy ciałem (ang. *corpus*). Każdy dokument składa się ze słów, zaś wszystkie unikatowe słowa występujące w ciele stanowią słownik (ang. *vocabulary*). Następnie ze słów w słowniku i dokumentów z ciała tworzona jest macierz wyznaczająca liczbę wystąpień danego słowa w danym dokumencie. Macierz ta nazywana jest w angielskiej nomenklaturze jako Document-Term Matrix i często skracana jest do DTM. DTM to przykład podejścia ”worka ze słowami”(ang. *Bag of Words*), w którym kolejność, w jakiej ułożone są słowa w dokumencie, jest wymienna i w związku z tym nieistotna. Poniższa tabela 2.1 przedstawia przykładową macierz DTM dla ciała o wielkości trzech dokumentów i słownika składającego się z 5 wyrazów.

Tabela 2.1. Przykładowy układ dokumentów oraz słów kluczowych na potrzeby LDA

	Wyraz 1	Wyraz 2	Wyraz 3	Wyraz 4	Wyraz 5=V
Dokument 1	2	1	0	0	1
Dokument 2	1	0	2	3	1
Dokument 3=D	0	2	4	0	0

Źródło: Opracowanie własne na podstawie *Latent Dirichlet allocation in R*, 2012; Martin Ponweiser.

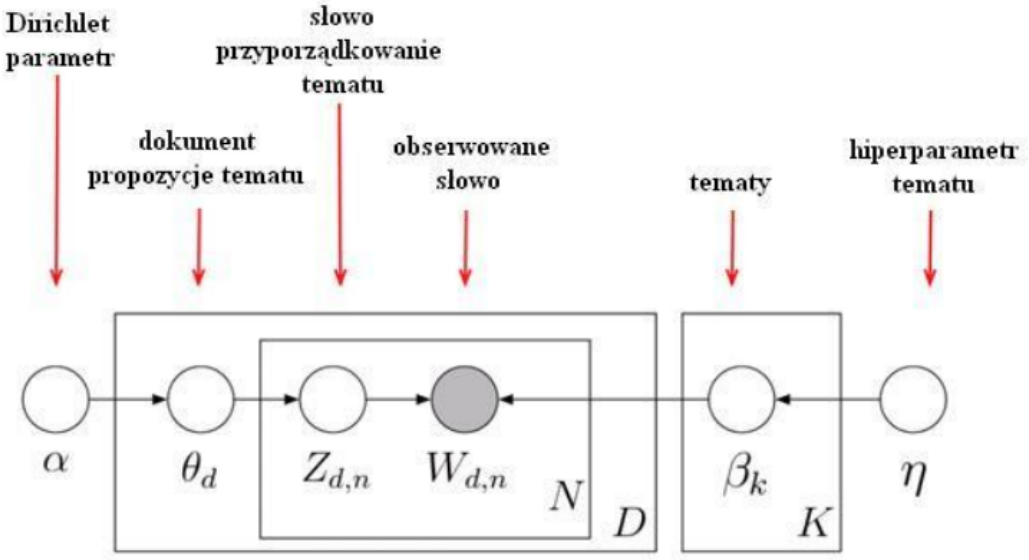
Model LDA postuluje, że cechy charakterystyczne tematów i dokumentów są czerpane z rozkładu prawdopodobieństwa Dirichleta. Rozkład Dirichleta jest wielowymiarową generalizacją rozkładu beta i określany jest funkcją gęstości prawdopodobieństwa (2.1):

$$p(x|a_1,...,a_K) = \frac{\Gamma(\sum_{i=1}^{K} a_i)}{\prod_{i=1}^{K} \Gamma(a_i)} \prod_{i=1}^{K} x_i^{a_i-1}, \tag{2.1}$$

gdzie: α oznacza dodatni wektor K (liczby tematów), Γ oznacza funkcję gęstości rozkładu Gamma daną wzorem (2.2)

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt. \tag{2.2}$$

Model LDA jako przykład hierarchicznego modelu bayesowskiego może zostać przedstawiony przy pomocy notacji tablicowej (ang. *plate notation*). Poniżej na grafice 2.1 widać schemat działania modelu przedstawiony graficznie, gdzie prostokąty reprezentują zbiory elementów.



Rysunek 2.1. Graficzna reprezentacja działania parametrów estymowanych w modelu LDA

Źródło: opracowanie na podstawie *Wykorzystanie metody opartej na ukrytej alokacji Dirichleta do automatycznej identyfikacji słów kluczowych w dokumentach*, 2014; Anna Gładysz

.

Graficzna prezentacja estymowanych parametrów przedstawiona na rysunku 2.1 może być zapisana w postaci równania wspólnego rozkładu prawdopodobieństwa prezentującego się na-

stępująco (2.3):

$$p(w,z,\theta,\phi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi). \tag{2.3}$$

Pierwszym czynnikiem iloczynu prawej strony równania (2.3) jest rozkład tematów dla każdego z dokumentów. Czerpie on z rozkładu Dirichleta przy założonym parametrze α i jest dany wzorem (2.4). Tabela 2.2 przedstawia przykład przypisania prawdopodobieństw czterech tematów do trzech dokumentów.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\theta_1^{\alpha_k-1}...\theta_K^{\alpha_K-1} = \frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \tag{2.4}$$

Tabela 2.2. Rozkład prawdopodobieństwa wystąpienia tematów w dokumentach

	Temat 1	Temat 2	Temat 3	Temat 4
Dokument 1 $\theta_{d=1}$	0,5	0,1	0,3	0,1
Dokument 2 $\theta_{d=2}$	0,0	0,9	0,1	0,0
Dokument 3 $\theta_{d=3}$	0,02	0,48	0,25	0,25

Źródło: Opracowanie własne na podstawie *Latent Dirichlet allocation in R*, 2012; Martin Ponweiser.

Krótsza w zapisie forma równania (2.4) korzysta z operatora $(.)$ w indeksie zmiennej. Oznacza on sumowanie po wszystkich wartościach zmiennej.

Drugim czynnikiem wspólnego rozkładu prawdopodobieństwa (2.3) jest rozkład przyporządkowujący tematy do słów w ciele dokumentów. Każdemu słowu w_i w dokumencie przyporządkowana jest liczba $(1,...,K)$ co przedstawia tabela poniżej 2.3.

Tabela 2.3. Przyporządkowanie słów do określonych tematów w badanych dokumentach

	Wyraz w_1	Wyraz w_2	Wyraz w_3	Wyraz w_4	Wyraz w_5	Wyraz w_6
Dokument 1 $z_{d=1}$	Temat k=2	Temat k=1	Temat k=1	Temat k=4	Temat k=3	Temat k=3
Dokument 2 $z_{d=2}$	Temat k=2	Temat k=3	Temat k=2	Temat k=2	Temat k=2	Temat k=3
Dokument 3 $z_{d=3}$	Temat k=4	Temat k=2	Temat k=2	Temat k=4	Temat k=3	Temat k=3

Źródło: Opracowanie własne na podstawie *Latent Dirichlet allocation in R*, 2012; Martin Ponweiser.

Ostatecznie funkcja prawdopodobieństwa przypisująca tematy do słów dla wszystkich dokumentów i tematów ma postać (oznaczana z) (2.5):

$$p(z|\theta) = \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{n_{d,k,.}}, \tag{2.5}$$

gdzie $d = 1,...,D$ oznacza dokument, $k = 1,...,k$ określony temat.

Trzecim czynnikiem iloczynu (2.3) jest rozkład wyrazów ze słownika do tematów, który został zaprezentowany w poniższej tabeli 2.4

Tabela 2.4. Rozkład prawdopodobieństwa słów w okeślonych tematch

	Wyraz 1	Wyraz 2	Wyraz 3	Wyraz 4	Wyraz 5
Temat 1 $\phi_{k=1}$	0,1	0,1	0	0,7	0,1
Temat 2 $\phi_{k=2}$	0,2	0,1	0,2	0,2	0,3
Temat 3 $\phi_{k=3}$	0,01	0,2	0,39	0,3	0,1
Temat 4 $\phi_{k=4}$	0,0	0,0	0,5	0,3	0,2

Źródło: Opracowanie własne na podstawie *Latent Dirichlet allocation in R*, 2012, Martina Ponweisera.

Rozkład ten określa prawdopodobieństwo, że wyraz ze słownika v jest dobierany, gdy wybrany jest temat k. Funkcją prawdopodobieństwa ϕ dla wszystkich tematów i wszystkich słów ze słownika prezentuje się następująco (2.6):

$$p(\phi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1} \tag{2.6}$$

Ostatnim czynnikiem iloczynu jest prawdopodobieństwo ciała w pod warunkiem z oraz ϕ przyjmuje postać (2.7):

$$p(w|z,\phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{.,k,v}} \tag{2.7}$$

gdzie $n_{.,k,v}$ oznacza liczbę razy, ile temat k został przypisany do słowa v w całym korpusie dokumentu.

Finalnie równanie (2.3) można zapisać w postaci (2.8):

$$\begin{aligned} p(w,z,\theta,\phi|\alpha,\beta) &= p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi) = \\ &= \left(\prod_{d=1}^{D} \frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k-1}\right)\left(\prod_{d=1}^{D}\prod_{k=1}^{K} \theta_{d,k}^{n_{d,k,.}}\right) \times \\ &\left(\prod_{k=1}^{K} \frac{\Gamma(\beta_{k,.})}{\prod_{k=1}^{V}\Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{n_{.,k,v}}\right) = \\ &= \left(\prod_{d=1}^{D} \frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k+n_{d,k,.}-1}\right) \times \left(\prod_{k=1}^{K} \frac{\Gamma(\beta_{k.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}+n_{.,k,v}} - 1\right) \end{aligned} \tag{2.8}$$

Następnie wyznaczana jest funkcja prawdopodobieństwa modelu dla ustalonego ciała w

i hiperparametrów α i β w celu estymacji parametrów funkcji największej wiarygodności (ang. *maximum-likelihood function*) i wyznaczenia rozkładu zmiennych ukrytych. (2.9)

$$p(w|\alpha,\beta) = \int_{\phi}\int_{\theta}\sum_{z}\left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+n_{d,k,.}-1}\right)\times \left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k.})}{\prod_{k=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+n_{.,k,v}-1}\right)d\theta d\phi \tag{2.9}$$

Sumowanie wszystkich możliwych kombinacji wyznaczanych tematów $z(n_{d,k,.},n_{.,k,v})$ jest bardzo skomplikowane obliczeniowo przez co nie można wykorzystać w tym celu standardowego algorytmu maksymalizacji estymacji. Aby uzyskać dobre przybliżenie brzegowych rozkładów prawdopodobieństwa korzysta się z technik uczenia maszynowego.

2.3.2 Próbnik Gibbsa

Wyliczenie rozkładu brzegowego wymaga ewaluacji rozkładu prawdopodobieństwa na dużej przestrzeni dyskretnej stanu (ang. *large discrete state space*).

Po raz pierwszy metoda próbnika Gibbsa dla zmiennych ukrytych modelu LDA została zaproponowana w *Probabilistic topic models* M. Steyversa T. Griffithsa. Algorytm próbnika Gibbsa jest przykładem metody Monte Carlo łańcuchami Markowa, które umożliwiają dobór próby dla skomplikowanych rozkładów prawdopodobieństwa przy użyciu losowych liczb.

Metoda Gibbsa symuluje wielowymiarowe rozkłady poprzez próbkowanie podzbiorów zmiennych o mniejszej liczbie wymiarów, gdzie każdy podzbiór jest uwarunkowany każdym pozostałym. Próbkowanie postępuje sekwencyjnie dopóki wartości nie stanowią odpowiedniego przybliżenia docelowego rozkładu.

2.3.2.1 Metoda Gibbsa dla modeli LDA

Do wykorzystania tej metody w modelu LDA potrzebne jest określenie prawdopodobieństwa przypisania tematu $z_{a,b}$ do b-tego słowa a-tego dokumentu $w_{a,b}$, pod warunkiem zdarzenia przypisania każdego innego tematu do wszystkich innych słów $z_{-(a,b)}$ co opisuje następujące równanie (2.10):

$$p(z_{z,b}|z_{-(a,b),w,\alpha,\beta}) \propto p(w,z|\alpha,\beta) = \int_{\theta}\int_{\phi} p(w,z,\theta,\phi|\alpha,\beta)d\theta d\phi \tag{2.10}$$

Równanie to zostaje rozszerzone i przetransformowane w pracy B. Carpentera do równania nieznormalizowanego prawdopodobieństwa warunkowego (2.11):

$$p(z_{a,b}|z_{-(a,b)},w,\alpha,\beta) \propto \frac{\left(n_{z_{z,b},a,.}^{-(a,b)}+\alpha_{z_{a,b}}\right)\times\left(n_{z_{z,b},.,w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}\right)}{n_{z_{a,b},.,.}^{-(a,b)}+\sum_{j=1}^{J}\beta_j} \tag{2.11}$$

Pierwszy element mnożenia w liczniku ułamka, $n_{z_{a,b},a,.}^{-(a,b)}+\alpha_{z_{a,b}}$, oznacza liczbę innych słów w dokumencie a, które zostały przypisane do tematu $z_{a,b}$ oraz wcześniejszego tematu. Drugi elementem licznika, $n_{z_{a,b},.,w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}$, oznacza liczbę razy, jaką słowo $w_{a,b}$ oraz słowo poprzednie zostało przypisane do tematu $z_{a,b}$. Mianownik tego ułamka normalizuje drugą część do prawdopodobieństwa.

Znormalizowane prawdopodobieństwo warunkowe przybiera postać (2.12):

$$p(z_{a,b}|z_{-(a,b)},w\ ,\alpha,\beta) = \frac{\left(\frac{\left(n_{z_{a,b},a,.}^{-(a,b)}+\alpha_{z_{a,b}}\right)\times\left(n_{z_{a,b},.,w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}\right)}{n_{z_{a,b},.,.}^{-(a,b)}+\sum_{j=1}^{J}\beta_j}\right)}{\left(\sum_{k=1}^{K}\frac{\left(n_{z_{a,b},a,.}^{-(a,b)}+\alpha_{z_{a,b}}\right)\times\left(n_{z_{a,b},.,w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}\right)}{n_{z_{a,b},.,.}^{-(a,b)}+\sum_{j=1}^{J}\beta_j}\right)} \tag{2.12}$$

Działanie algorytmu można przedstawić za pomocą następujących kroków:

1. przydziel losowo tematy ze zbioru $\{1,..,K\}$ do słów.
2. dla każdej próbki Gibbsa:
   (a) Dla każdego wyrazu, macierze wystąpień $n^{(a,b)}$ są pomniejszane o jeden dla zajść nawiązujących dla danego przypisywanego tematu.
   (b) Pobierana jest próbka dla nowego tematu według wzoru rozkładu prawdopodobieństwa powyżej (**??**).
   (c) Macierze wystąpień są uaktualniane poprzez zwiększanie o jeden dla nowego przydziału tematu.
3. Gdy łańcuch Markowa osiągnie stan stacjonarny, na podstawie rozkładu aposteriori można przypisać prawdopodobieństwo, że dany temat jest obserwowany w określonym dokumencie.

Zmienne θ (2.13) oraz ϕ (2.14) można oszacować z pojedynczej próbki Gibbsa z

$$\hat{\theta}_{d,k} = \frac{\alpha_k + n_{d,k,}}{\alpha_. + n_{d,.,.}} \tag{2.13}$$

$$\hat{\phi}_{k,v} = \frac{\beta_{k,v} + n_{.,k,v}}{\beta_{k,.} + n_{.,k,.}} \tag{2.14}$$

2.3.3 Ocena modelu LDA

Ewaluacja modelu, czyli zmierzenie jego wydajności, jest wymagana, aby upewnić się, że model będzie działał prawidłowo i dostarczał przydatne wyniki dla obserwacji spoza zbioru próbnego. Możliwość pomiaru dopasowania modelu przydaje się także w doborze parametrów. W modelu LDA, taki problem występuje w trakcie wskazywania liczby tematów, gdy ta nie jest określona a priori. Istnieje wiele metryk, które pozwalają porównać ze sobą modele o różnych parametrach i pomagają dobrać najlepszy. Do tych metryk zaliczyć można:

- miara trudności predykcji (ang. *perplexity*) – w dalszej części będzie używane pojęcie *trudność* w odniesieniu do tej miary,
- rozkład prawdopodobieństwa empirycznego,
- rozkład prawdopodobieństwa brzegowaego.

Rozkład prawdopodobieństwa brzegowe można przybliżyć za pomocą kilku metod, z których wybrałem do szerszego przeanalizowania metodę średniej harmonicznej (ang. *harmonic mean method*) opisanej dokładnie w pracy T. Griffithsa i M. Steyversa *Finding scientific topics*.

2.3.3.1 Metoda średniej harmonicznej

Metoda ta została po raz pierwszy zastosowana dla modelu LDA w 2004 roku przez T Griffithsa i M. Steyversa do rozwiązania problemu najlepszego doboru parametru K odpowiadającego za liczbę tematów w modelu. Zyskała ona na popularności ze względu na swoją prostotę oraz wydajność obliczeniową. W wyżej wymienionej pracy można przeczytać o metodzie:

> Zbiór danych stanowią słowa w ciele w, zaś model opisany jest za pomocą liczby tematów K. Pragniemy wyznaczyć prawdopodobieństwo warunkowe $p(w|K)$. To zaś wymaga sumowania wszystkich możliwych przydziałów słów do tematów z (tj.

$p(w|K) = \int p(w|z,K)p(z)dz$). Jednakże można to przybliżyć wyznaczając średnią harmoniczną dla wartości $p(w|z,K)$, gdy z jest próbką dobraną z prawdopodobieństwa a posteriori $p(z|w,K)$. Algorytm doboru próbki Gibbsa dostarcza takich próbek, a wartość $p(w|z,K)$ można wyliczyć korzystając ze wzoru (2.15):

$$P(w|z) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_V \Gamma\left(n_k^{(w)} + \beta\right)}{\Gamma\left(n_k^{(\cdot)} + V\beta\right)} \tag{2.15}$$

gdzie n_k^w oznacza liczbę razy słowo w zostało przypisane do tematu k w wektorze przypisań z, a $\Gamma(.)$ jest funkcją Gamma.

2.3.3.2 Ewaluacja metodą ekspercką

Ostatecznie aparatura modelująca tematy powinna przede wszystkim zostać zaakceptowana przez człowieka bazując na jego osądzie i kierując się logiką. Nie zawsze dobór optymalnych parametrów sugerowanych przez algorytmy będzie skutkować najlepiej dopasowanym modelem pod względem interpretacji wyników. Przykładowo, algorytm może nalegać, by zwiększać liczbę tematów K. Zbyt wysoki parametr może utrudnić jednak zidentyfikowanie i interpretację utworzonych grup, bądź sprawić, że granice pomiędzy grupami zaczną się zacierać. Metody optymalizacji prawdopodobieństwa powinny służyć jedynie jako pomoc do dokonania wyboru. By zaakceptować model powinno się przede wszystkim sprawdzić (Ponweiser, 2012; Gładysz, 2014):

- spójność semantyczną tematów – czyli czy słowa przydzielone tematom ugrupowane są w sposób, w jaki pogrupował by je człowiek.
- czy kompozycja tematów przydzielona dokumentom odpowiada tematom, z jakimi skojarzyłby ten dokument człowiek.

2.4 Podsumowanie

W powyższym rozdziale opisano metody ekstrakcji danych z wykorzystaniem web scrapping oraz model ukrytej alokacji Dichirleta – które odgrywać będą kluczową rolę w przeprowadzanym badaniu empirycznym. Dzięki zastosowaniu web scrappingu pobrano oferty pracy publikowanych na stronie Kancelarii Prezesa Rady Ministrów, zaś model LDA pomoże mi przybliżyć

na ich podstawie tematy zawarte ogłoszeniach o pracę w służbie cywilnej.

Rozdział 3

Analiza zawartości tematycznej ofert pracy w służbie cywilnej

3.1 Ekstrakcja danych z Internetu

3.1.1 Pobranie danych z portalu Służby Cywilnej

Celem badania jest analiza tematów ofert pracy na portalu 'nabory.kprm.gov.pl'. Dlatego pierwszym krokiem tej analizy jest pozyskanie informacji dotyczących wszystkich obecnych, jak i archiwalnych, ofert pracy, które można znaleźć na wyżej wymienionej stronie w sposób umożliwiający przeprowadzanie na nich operacji w środowisku R (R Core Team, 2013).

Jest do tego potrzebny dodatkowy pakiet o nazwie `rvest` (Wickham, 2016) pozwalający na zbudowanie crawlera do web scrapingu. Wykorzystany będzie także pakiet `tidyverse` (Wickham, 2017) będący w rzeczywistości zbiorem najpopularniejszych pakietów przeznaczonych do manipulacji i wizualizacji danych. W językach programowania przyjęto, że symbolem # oznacza się komentarze opisujące poszczególne linie kodu. Poniższy kod 3.1 służy do instalacji, a następnie do uruchomienia dodatkowych pakietów w środowisku R.

```
#komenda do instalacji ópakietw
    install.packages(c("rvest","tidyverse"))
#komendy łąwywoujce zainstalowane świczeniej pakiety
    library(rvest)
    library(tidyverse)
```

Program 3.1. Instalacja i uruchamianie pakietów dodatkowych

Dodatkowo niezwykle pomocnym narzędziem okaże się rozszerzenie dla przeglądarki Go-

ogle Chrome o nazwie *Selector Gadget*. Po uaktywnieniu go, najeżdżając kursorem myszy na dowolny element strony internetowej wyświetla on ścieżkę CSS i XPath danego segmentu strony. Po zbudowaniu crawlera pozwoli to na wskazanie drogi dostępu wybranych elementów i ich pobieranie.

Po przygotowaniu wymaganych narzędzi stworzona zostaje lista wszystkich adresów url listujących oferty pracy. Na stronie związanej z jednym adresem url wyświetlane jest (domyślnie) po dziesięć ofert pracy, po czym użytkownik zmuszony jest do przejścia na następną stronę w celu wyświetlenia kolejnych dziesięciu. Na stronie KPRM desygnowane są osobno zbiory dla ofert aktualnych i archiwum. Proces pobierania ofert aktualnych jak i historycznych jest do siebie bardzo podobny, dlatego przedstawione w pracy są tylko kroki do pobrania ofert aktualnych 3.2.

```
#ęTworz ązmienn ąąprzechowujc adres pierwszej strony ńwyszukiwa ofert pracy
    url <- "https://nabory.kprm.gov.pl/?Ad%5BisAdvancedMode%5D=&Ad%5Bsort%5
       D=1&Ad%5BpagesCnt%5D=10&Ad%5Bid_province%5D=&Ad%5Bid_city%5D=&Ad%5
       Bid_institution%5D=&Ad%5Bphrase%5D=&Ad%5Beducation%5D=&Ad%5Bid_
       institution_position%5D=&Ad%5Bis_disability%5D=0&Ad%5Bis_first_
       foreigner%5D=0&Ad%5Bis_replacement%5D=0&Ad%5Bdate_publication%5D=&Ad
       %5Bdate_expiration%5D=&Ad%5Bprocess_state%5D=1&search-button=&page=1
       &per-page=10"
#ęTworz ęlist do przechowywania óadresw url
    sub_url <- c()
#ęTworz ązmienn ąąprzechowujc ęliczb wszystkich ofert pracy na portalu.
    n_ofert <- read_html(url) %>% html_node(".h2 b") %>%
    html_text() %>% as.numeric()
#ęTworz ązmienn ąąprzechowujc ęliczb stron.
    if(n_ofert%%10 == 0)
    {
      n_stron <- n_ofert%/%10
    } else
    {
      n_stron <- n_ofert%/%10+1
    }
#łWypeniam ęlist adresami url
    for(i in 1:n_stron){
      sub_url[i] <- paste0("https://nabory.kprm.gov.pl/?Ad%5BisAdvancedMode
         %5D=&Ad%5Bsort%5D=1&Ad%5BpagesCnt%5D=10&Ad%5Bid_province%5D=&Ad%5
         Bid_city%5D=&Ad%5Bid_institution%5D=&Ad%5Bphrase%5D=&Ad%5
         Beducation%5D=&Ad%5Bid_institution_position%5D=&Ad%5Bis_disability
         %5D=0&Ad%5Bis_first_foreigner%5D=0&Ad%5Bis_replacement%5D=0&Ad%5
         Bdate_publication%5D=&Ad%5Bdate_expiration%5D=&Ad%5Bprocess_state
         %5D=1&search-button=&page=", i, "&per-page=10")
    }
```

Program 3.2. Tworzenie listy zawierającej adresy url wyszukiwarki ofert pracy

Liczba stron wyznaczona jest w prostym wyrażeniu warunkowym, w którym sprawdzone zostaje czy liczba ofert pracy jest liczbą podzielną przez 10. Jeśli tak, do zmiennej przypisywana jest wartość tej operacji. Jeśli nie, przypisywana jest wartość ilorazu bez reszty + 1.

Lista zapełniana jest w pętli. Pętla będzie iterowała tyle razy, ile wynosi wartość zmiennej

n_stron, zaś w każdej iteracji dla $i-tego$ elementu listy przypisywany jest adres url, w którym zmieniana jest wartość wskazująca numer strony na odpowiedni.

Pod każdym zapisanym adresem url listowane są oferty pracy. Dla każdej z ofert wyświetlane są podstawowe informacje takie jak numer identyfikacyjny oferty, nazwa itd. Zainteresowany ofertą użytkownik może przejść na stronę danej oferty i wyświetlić jej szczegóły po skorzystaniu z hiperłącza ustawionego przy skrócie oferty. Celem jest pobranie informacji o wszystkich ofertach pracy, więc wymagane jest stworzenie listy adresów wszystkich ofert pracy 3.3:

```
#ęTworz ęlist do przechowywania óadresw url ofert pracy
    job_url = list()
#łWypeniam ęlist adresami url ofert pracy
    for (i in 1:length(sub_url)) {
      kprm_web <- read_html(sub_url[i])
    job_url[[i]] <- kprm_web %>% html_nodes("a.single") %>%
              html_attr("href")
    }
    job_url <- unlist(job_url) %>% paste0("https://nabory.kprm.gov.pl",.)
```

Program 3.3. Tworzenie listy zawierającej adresy url ofert pracy

Lista przechowująca adresy url wszystkich ofert pracy wypełniana jest w pętli. Dla każdego adresu url strony listującej oferty pracy zapisywane są hiperłącza kryjące się pod ofertami po wskazaniu odpowiedniej ścieżki CSS.

Teraz, gdy adresy przekierowujące na stronę każdej z ofert pracy są dostępne w środowisku, rozpoczynane zostaje pobieranie informacji potrzebnych do badania 3.4.

```
#ęTworz macierz do przechowywania informacji o~wszystkich ofertach pracy
    job_offer <- data.frame()
#ęTworz macierz ąpomocnicz
    df <- data.frame(matrix(ncol = 7))
#łWypeniam macierz informacjami ądotyczcymi ofert pracy
    for (i in job_url)
    {
      page <- read_html(i)

        df[,1] <- page %>% html_nodes(".so-h h4, .so-h h3 , h1") %>%
                           html_text() %>% paste(., collapse = " ")
        df[,2] <- page %>% html_nodes(".bor p , .bor strong") %>%
                           html_text() %>% paste(., collapse = " ")
        df[,3] <- page %>% html_nodes(".ar div") %>%
                           html_text() %>% str_remove_all(.,"•-[=-]")
                              %>%
                           str_squish()
        df[,4] <- page %>% html_nodes("section:nth-child(2) li") %>%
                           html_text() %>% str_remove_all(.,"•-[=-]")
                              %>%
                           paste(., collapse = " ") %>% str_squish()
        df[,5] <- page %>% html_nodes("p:nth-child(5) span") %>%
                               html_text() %>% paste(., collapse = " "
                                  )
        df[,6] <- page %>% html_nodes("section:nth-child(3) li") %>%
                          html_text() %>% str_remove_all(.,"•-[=-]")
```

```
                                %>%
                               paste(., collapse = " ") %>% str_squish()
        df[,7] <- page %>% html_nodes(".id") %>% html_text() %>% str_squish
            ()

        job_offer <- rbind.data.frame(job_offer, df)
    }
    job_offer <- cbind(job_offer, job_url)
#ęNadaj kolumnom macierzy etykiety
    colnames(job_offer) <- c("łtytu", "termin łskadania ódokumentw", "
        warunki pracy", "opis stanowiska", "data zamieszczenia oferty", "
        wymagania", "ID", "url")
```

Program 3.4. Iteracyjne pobieranie wybranych danych z ofert pracy

Macierz wypełniona zostaje informacjami w sposób iteracyjny w pętli, dla każdego adresu url oferty pracy, zapisując w kolejnych kolumnach macierzy odpowiednie segmenty oferty wyselekcjonowane za pomocą odpowiedniej ścieżki CSS. Ostatecznie doklejona zostaje dodatkowa kolumna przechowującą adres url danej oferty i nadane zostają etykiety wszystkim kolumnom. Pobrane zostały następujące informacje: tytuł oferty, termin składania dokumentów, warunki pracy, opis stanowiska, data zamieszczenia oferty, wymagania i numer identyfikacyjny. Nie wszystkie dane będą wykorzystywane w badaniu, lecz na moment ich zbierania mogły one okazać się przydatne w przyszłości.

Ostatecznie zapisywana zostaje stworzoną macierz przechowująca informacje dotyczące wszystkich aktualnych ofert pracy na dysk twardy w pliku z rozszerzeniem *.csv* 3.5.

```
    write.csv(x = job_offer, file = "oferty-pracy-kprm.csv")
```

Program 3.5. Zapisywanie pliku na dysk twardy

Proces ten powtarzany jest dla zbioru ofert archiwalnych podając inny startowy adres url. Ostatecznie łączone są dwa powstałe zbiory w jeden 3.6.

```
#ęWczytuj dane z~dysku twardego
    dane_ar <- read.csv("oferty-pracy-kprm_archiwum.csv")
    dane_cu <- read.csv("oferty-pracy-kprm.csv")
#Łączcz dwa zbiory w~jeden
    dane <- bind_rows("aktualne" = dane_cu, "archiwum" = dane_ar, .id = "
        łzrodo")
```

Program 3.6. Łączenie zbiorów danych ofert przebiegajłacych w toku i ofert archiwalnych

Podczas łączenia zbiorów tworzona zostaje dodatkowa kolumna *źródło*, w której dla obserwacji przypisywana jest wartość *archiwum* jeśli oferta jest archiwalna i *aktualne* jeśli rekrutacja na daną ofertę jest dalej w toku.

3.1.2 Czyszczenie danych

Po pobraniu potrzebnych do badania danych powinny przejść one jeszcze proces czyszczenia. Czyszczenie danych ma na celu poprawienie jakości danych w zbiorze. Pomoże to nimi operować, poprawi ich czytelność dla człowieka oraz poprawi wiarygodności wyników. Proces ten zakłada m.in.:

- Parsowanie – czyli rozbicie pojedynczego złożonego pola na kilka pól z pojedynczą informacją w oparciu o kontekst (np.: rozbicie pola z imieniem i nazwiskiem na dwa pola, jednym dla imienia, drugim dla nazwiska)
- Usunięcie błędnych rekordów – czyli rekordów, które nie będą pomocne w przeprowadzanej analizie. Może być to spowodowane np.: niekompletnością rekordu (brakującymi informacjami), bądź gdy rekord jest duplikatem (identycznym, bądź podobnym) do innego w tym samym zbiorze.
- Standaryzację – Czyli zamianę wystąpień bliskoznacznych na jedną zdefiniowaną formę zapisu (np.: zamiana skrótów *wlkp.*, *wielkopol.*, *w.-pol.* na *Wielkopolska*).
- Sformatowanie zmiennych – czyli nadanie zmiennym w zbiorze danych odpowiednich typów (np.: typu *Date* zmiennej oznaczającej datę, aby maszyna rozpoznawała tą wartość jako datę, a nie tekst)

Czyszczenie danych rozpoczęte zostaje poprzez nadanie odpowiednich typów danych dla zmiennych w zbiorze. W kolumnie *termin składania dokumentów* zamienione zostają także odmienione nazwy miesięcy w języku polskim na odpowiednie kody, aby stworzyć daty w zapisie *DD.MM.YYYY* 3.7.

```
#Zmieniam typ źółrda z~character na factor
    dane <- dane %>% mutate_at(., vars("zrodlo"), funs(as.factor)) %>%
#Zmieniam śćkolejno kolumn, usuwam ękolumn X
    select(ID, data.zamieszczenia.oferty, termin.skladania.ódokumentw,
        tytul, opis.stanowiska, wymagania, warunki.pracy, zrodlo, url, -X)
           %>%
#Zmieniam typ ID z~character na integer
    mutate_at(., vars("ID"), funs(as.integer(str_remove(., "nr "")))) %>%
#Zmieniam typ odpowiednich kolumn z~character na Date
    mutate(termin.skladania.ódokumentw = str_replace_all(termin.skladania.
       ódokumentw,
                                          c(
                                             "stycznia" = "01",
                                             "lutego" = "02",
                                             "marca" = "03",
                                             "kwietnia" = "04",
                                             "maja" = "05",
```

```
                                        "czerwca" = "06",
                                        "lipca" = "07",
                                        "sierpnia" = "08",
                                        "śwrzenia" = "09",
                                        "źpadziernika" = "10",
                                        "listopada" = "11",
                                        "grudnia" = "12")) %>%
    str_replace_all(" ", ".")) %>%
    mutate_at(., vars("data.zamieszczenia.oferty", "termin.skladania.
       ódokumentw"),
            funs(dmy(str_sub(., -10, -1))))
```

Program 3.7. Nadanie odpowiednim elementom zbioru danych właściwych typ'ow

Dokonane zmiany zostają zapisane jako nowy plik na dysku twardym z rozszerzeniem *.rds*, aby nie musieć ich wprowadzać na nowo w przyszłości 3.8.

```
    saveRDS(dane, "dane.rds")
```

Program 3.8. Zapisanie danych na dysku twardym

3.1.2.1 Usunięcie liter i słów nieistotnych

Kolejnym krokiem jest usunięcie *stop words*. Mianem *stop words* określa się słowa/litery w danym języku, które nie noszą za sobą istotnych treści z punktu widzenia przekazu. Są to słowa, które ze względu na swoją powszechność lub niewielkie znaczenie nie będą użyteczne w analizie dokumentu. Przykładami takich słów/liter są np. spójniki, zaimki, przyimki czy liczebniki. *Stop words* można usunąć z dokumentu posługując się specjalnie utworzonym słownikiem bądź podejściem statystycznym, w którym usuwane są słowa, których częstość występowania mieści się w przyjętym przedziale. Usunięcie ich zredukuje wielkość zbioru danych, przyśpieszy czas obliczeniowy i poprawi jakość wyników.

W projekcie *stop words* usuwane są metodą słownikową, pobierając przygotowaną listę słów do usunięcia ze strony `https://github.com/MarcinKosinski/trigeR5/tree/master/dicts`. Dodatkowo musi zostać wcześniej odpowiednio przygotowany zbiór danych. W tym celu zainstalowane i wczytane są kolejne pakiety dodatkowe do środowiska R 3.9.

```
    install.packages(c("tidytext", "stringi"))
    library(tidytext)
    library(stringi)
```

Program 3.9. Instalacja i uruchamianie pakiet/'ow

Pakiet *tidytext* (Silge & Robinson, 2016), za pomocą funkcji *unnest_tokens()*, pozwoli na rozbicie tekstu na pojedyncze słowa, zaś *stringi* (Gagolewski, 2018) pozwoli na usunięcie polskich

znaków. Są one usuwane, ponieważ nie ma pewności, że wszystkie wyrazy napisane dla oferty pracy zostały napisane poprawnie. Może wystąpić sytuacja, w której ktoś błędnie napisał *az* zamiast *aż*. W takim wypadku słowo *az* nie zostało by usunięte. Z analogicznego powodu zamieniona zostaje także wielkość wszystkich liter na małe przy użyciu wcześniej załadowanego pakietu *stringr* (Wickham, 2017) (wchodzącego w skład pakietu *tidyverse*). Załadowana zostaje ostatnia wersja danych 3.10.

```
kprm_raw <- readRDS("dane.rds")
```

Program 3.10. Wczytywanie zbioru danych z dysku twardego

W tym momencie ograniczony zostaje zakres badania jedynie dla opisów ofert pracy, pomijając warunki czy wymagania. Dlatego wybierane zostają odpowiednie kolumny: ID, datę zamieszczenia oferty i opis stanowiska 3.11.

```
kprm_raw <- kprm_raw %>%
        select(ID, data.zamieszczenia.oferty, opis.stanowiska)
```

Program 3.11. Ograniczenie zbioru danych poprzez wybór określonych kolumn

Ładowany zostaje do środowiska R wcześniej pobrany słownik *stop words* i plik ten zostaje przypisany do zmiennej 3.12.

```
stopwords <- read_lines("morfologik/stopwords.txt")
```

Program 3.12. Wczytanie słownika *stop words* z dysku twardego

Przy pomocy funkcji *unnest_tokens* rozdzielone zostają opisy ofert pracy na pojedyncze słowa 3.13. Wynikiem takiego działania jest tabela (pierwsze 10 wierszy) 3.1:

```
kprm_tidy <- kprm_raw %>%
    unnest_tokens(output = word,
                  input = opis.stanowiska,
                  to_lower = TRUE)
```

Program 3.13. Rozdzielenie opisów ofert pracy na pojedyncze słowa

Przy wykorzystaniu pakietu stringi usuwane są polskie akcenty ze zbioru danych jak i ze słownika *stop words* 3.14.

```
kprm_ascii <- kprm_tidy %>%
    mutate(word = stri_trans_general(word,
                                     "Latin-ASCII"))
stopwords_ascii <- stri_trans_general(stopwords,
                                      "Latin-ASCII")
```

Program 3.14. Zmiana kodowania znaków z UTF-8 na ASCII

Tabela 3.1. Przykładowy długi układ danych po zastosowaniu funkcji `unnest_tokens`

ID	Data zamieszczenia oferty	Wyraz
45863	2019-04-10	prowadzenie
45863	2019-04-10	rachunkowosci
45863	2019-04-10	jednostki
45863	2019-04-10	w
45863	2019-04-10	oparciu
45863	2019-04-10	o
45863	2019-04-10	zasady
45863	2019-04-10	polityke
45863	2019-04-10	rachunkowosci
45863	2019-04-10	zgodnie

Źródło: Opracowanie własne.

Wykorzystana ponownie jest funkcja *unnest_tokens()* w celu stworzenia podobnej macierzy, tym razem dla słownika *stop words*. Gdy oba zbiory są podobnej postaci można wykorzystać funkcję *anti_join()* pakietu *dplyr* (wchodzącego w skład wcześniej załadowanego *tidyverse*), aby znaleźć w zbiorze danych słowa odpowiadające tym, z listy *stop words*, i usunąć je 3.15.

```
stopwords_ascii <- stopwords_ascii %>% str_split(", ", simplify = F) %>%
    unlist() %>% tibble()
dane <- kprm_ascii %>%
    anti_join(stopwords_ascii, by = c("word" = "."))
```

Program 3.15. Usunięcie ze zbioru danych słów ze słownika *stop words*

3.1.2.2 Stemming

W celu zmniejszenia wielkości zbioru danych, a także dla uzyskania lepszych wyników, przeprowadzany będzie *stemming* na zbiorze danych. *Stemming* to proces, w którym słowa są doprowadzane do swojego tematu poprzez ucinanie końcówki fleksyjnej (np.: przedrostków i przyrostków). Maszyna przeprowadzająca analizę dokumentu będzie brała osobno pod uwagę każdą odmianę słowa przez przypadki i czasy, gdy w rzeczywistości można dla uproszczenia przyjąć, że wszystkie odmiany słowa niosą za sobą to samo znaczenie pod względem logicznym. Przykładowo zbiór słów (biegnący, biegnąca, biegać) przekształcany jest w zbiór (biegać, biegać, biegać). Efektywnie zmniejszy to rozmiary zbioru danych co przełoży się na czas obliczeniowy potrzebny do wykonania analizy. Dodatkowo wyniki będą prostsze w analizowaniu.

Stemming przeprowadzony będzie metodą słownikową pobierając w tym celu słownik ze strony: `https://github.com/MarcinKosinski/trigeR5/tree/master/dicts`. Następnie, podobnie jak w przypadku usuwania *stop words*, słownik musi zostać doprowa-

dzony do odpowiedniej postaci. Wczytany zostaje słownika do *stemmingu* w środowisku R 3.16,

```
    stem_dict <- read_csv2("morfologik/polimorfologik-2.1.txt", col_names =
        c("stem", "word"))
#Zmieniam śćwielko liter na łmae i~usuwam duplikaty
    stem_dict <- stem_dict %>%
        mutate(stem = str_to_lower(stem),
        word = str_to_lower(word)) %>%
        distinct()
```

Program 3.16. Wczytanie słownika morfologicznego do stemmingu z dysku twardego

dalej usuwane są z niego polskie akcenty 3.17,

```
    stem_dict_ascii <- stem_dict %>%
    mutate(stem = stri_trans_general(stem, "Latin-ASCII"),
           word = stri_trans_general(word, "Latin-ASCII"))
```

Program 3.17. Zmiana kodowania znaków z UTF-8 na ASCII

I ostatecznie dodana jest kolumna *stem* do zbioru danych, w której znajdują się rdzenie słowa z kolumny *word* - oryginalnych słów odmienionych w kontekście oferty pracy. Dodatkowo filtrowane są wszystkie te słowa ze zbioru danych, które nie występują w słowniku do *stemmingu*. Są to literówki, słowa obce, liczby czy nazwy własne, które nie będą użyteczne w badaniu 3.18.

```
    dane <- dane %>%
    left_join(stem_unique, by = ("word" = "word")) %>%
    filter(!is.na(stem))
```

Program 3.18. Połączenie zbioru danych ze słownikiem morfologicznym do stemmingu

Efekt pracy widać na załączonej poniżej tabeli (przedstawiającej pierwsze 10 wierszy) w tabeli 3.2.

Tabela 3.2. Przykładowe dane otrzymane po procesie czyszczenia danych

ID	Data zamieszczenia oferty	Wyraz	Stem
45863	2019-04-10	prowadzenie	prowadzenie
45863	2019-04-10	rachunkowosci	rachunkowosc
45863	2019-04-10	jednostki	jednostka
45863	2019-04-10	oparciu	oparcie
45863	2019-04-10	zasady	zasada
45863	2019-04-10	polityke	polityka
45863	2019-04-10	rachunkowosci	rachunkowosc
45863	2019-04-10	zgodnie	zgodnie

Źródło: Opracowanie własne.

3.2 Eksploracyjna analiza danych

Po uzyskaniu danych i doprowadzeniu ich od pożądanej postaci, można zacząć wstępną eksplorację danych. Robi się to głównie poprzez tworzenie wykresów i podsumowań, które mają pomóc zrozumieć zbiór danych, aby dalej podejmować świadome decyzje przy budowaniu modelu i przy interpretowaniu rezultatów.

Wczytane zostają dana oraz instalowane i uruchomiane są potrzebne do pracy pakiety 3.19.

```
install.packages(c("RColorBrewer", "wordcloud", "lubridate"))
library(tidyverse)
library(RColorBrewer)
library(wordcloud)
library(lubridate)
```

Program 3.19. Instalacja i uruchamianie pakietów

Pakiety najwięcej wykorzystywane w tym etapie to *dplyr* do przetwarzania danych oraz *ggplot2* do tworzenia wykresów i grafów. Oba te pakiety wchodzą w skład wcześniej już zainstalowanego pakietu *tidyverse*, dzięki czemu nie trzeba ich instalować, a jedynie uruchomić. Nowe pakiety, które pojawią się w tej części projektu to:

- `RColorBrewer` (Neuwirth, 2014) – oferujący ciekawe palety kolorów przy tworzeniu wykresów.
- `wordcloud` (Fellows, 2018)– umożliwiający tworzenie grafik typu *chmura słów*.
- `lubridate` (Grolemund & Wickham, 2011)– do przetwarzania danych zakodowanych jako daty w czasie.

Wczytuję ostatnią wersję zbioru danych 3.20.

```
dane <- read_rds(path = "dane/dane_lda")
```

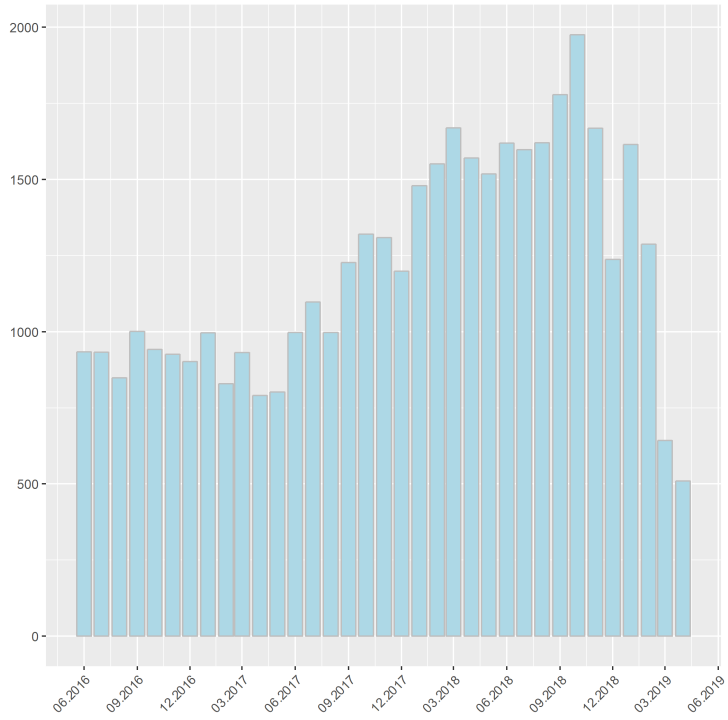
Program 3.20. Wczytanie zbioru danych z dysku twardego

Eksplorację rozpoczyna wizualizacja danych. Na początku tworzony jest wykres pokazujący jak zapotrzebowanie na pracę, wyrażone poprzez liczbę publikowanych ofert pracy, kształtowało się w czasie 3.1. W tym celu rozbita zostaje kolumna zawierającą informacje o dacie publikacji oferty na nowe zmienne: rok, miesiąc i dzień tygodnia 3.21.

```
dane_czas <- dane %>%
mutate(rok = year(data.zamieszczenia.oferty),
       mies = month(data.zamieszczenia.oferty),
       tyg_dzien = factor(wday(data.zamieszczenia.oferty),
                          levels = c(2:7, 1),
```

```
                labels = c("pn", "wt",                                    6
                "sr", "czw", "pt", "sob", "nied")))                       7
```

Program 3.21. Rozbicie zmiennej z datą publikacji oferty pracy na trzy zmienne: rok, miesiąc, dzień tygodnia



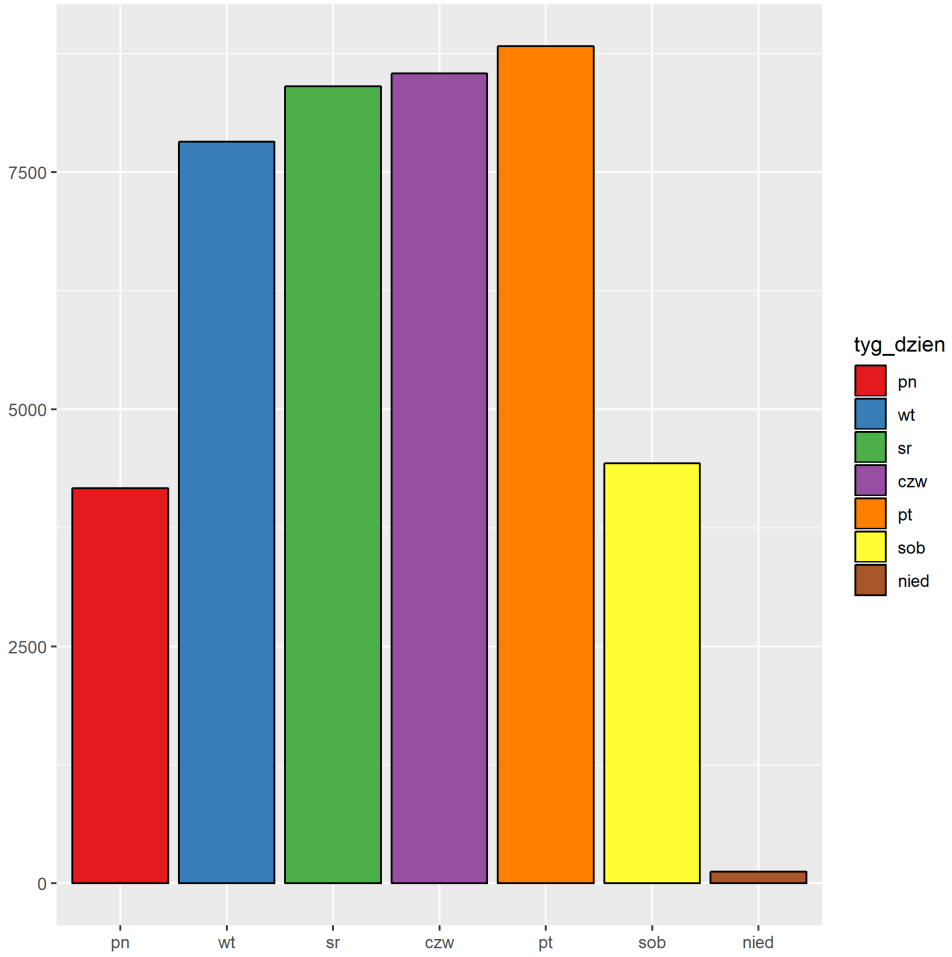
Rysunek 3.1. Liczba ofert pracy publikowanych w serwisie *nabory.kprm.gov.pl* według miesięcy

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

Wysokość słupka na wykresie 3.1 wskazuje ile ofert pracy zostało opublikowanych w danym miesiącu. Warto zauważyć, że dane były pobierane w połowie kwietnia. W związku z tym, nie należy mylnie interpretować niewielkiej liczby publikowanych ofert pracy na portalu KPRM w tym miesiącu. Zaobserwować można dość gwałtowny przyrost zapotrzebowania na pracowników w okresie 06.2017 - 10.2018. W październiku 2018 roku występuje maksymalna liczba publikacji ofert pracy - blisko dwa tysiące. Po październiku ubiegłego roku popyt na pracowników w służbie cywilnej zaczął maleć, aby w marcu 2019 roku osiągnąć minimum globalne.

Następnie warto sprawdzić czy istnieją pewne właściwości dotyczące dnia tygodnia, w którym została opublikowana oferta pracy. Tworzony zostaje w związku z tym wykres przedstawiający liczbę ofert pracy opublikowanych w kolejnych dniach tygodnia 3.2.

Wykres 3.2 przedstawia, że najwięcej ofert pracy było opublikowanych w piątek, oraz bardzo niewielka ich liczba w niedzielę. Co ciekawe, w sobotę publikowanych było więcej ofert pracy



Rysunek 3.2. Liczba ofert pracy publikowanych w serwisie *nabory.kprm.gov.pl* według dnia tygodnia

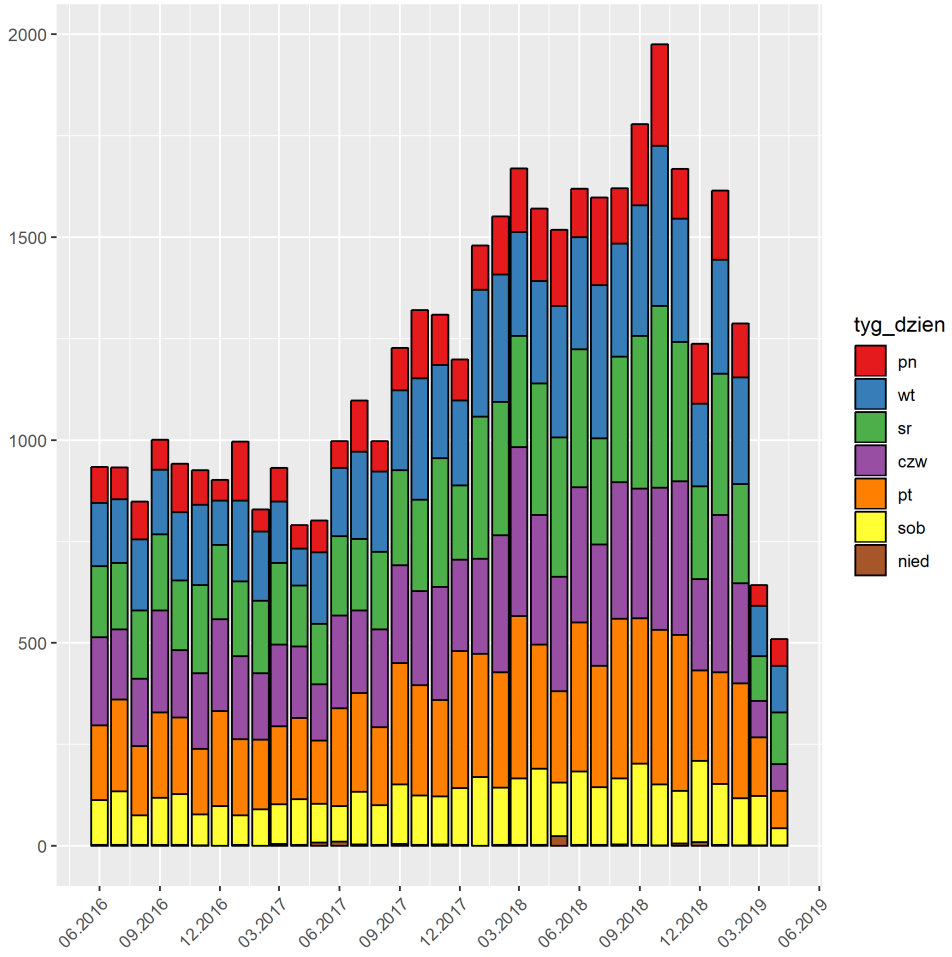
Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

niż w poniedziałek.

Informacje zawarte w dwóch poprzednich wykresach zamieszczone zostają na pojedynczym wykresie słupkowym, w którym przedstawiona jest liczba publikowanych w danym miesiącu ofert pracy z podziałem na dni 3.3.

Warto też zobrazować najczęściej występujące na przestrzeni wszystkich dokumentów (tj. wszystkich opisów ofert pracy) wyrazy 3.4.

Na wykresie 3.4 przedstawionych zostało trzydzieści najczęstszych wyrazów występujących w opisach pracy. Trzema zdecydowanie górującymi nad innymi są: prowadzenie, zakres i cela. Słowo *cela* prawdopodobnie nie odnosi się do cel więziennych, lecz zostało błędnie skorygowane przez algorytm sprowadzający do tematu słowa i oryginalnie oznaczał *cel*, jak w frazie ”w jakimś celu”. Następne trzy słowa to: sprawa, projekt i przygotowywać. Można wnioskować,



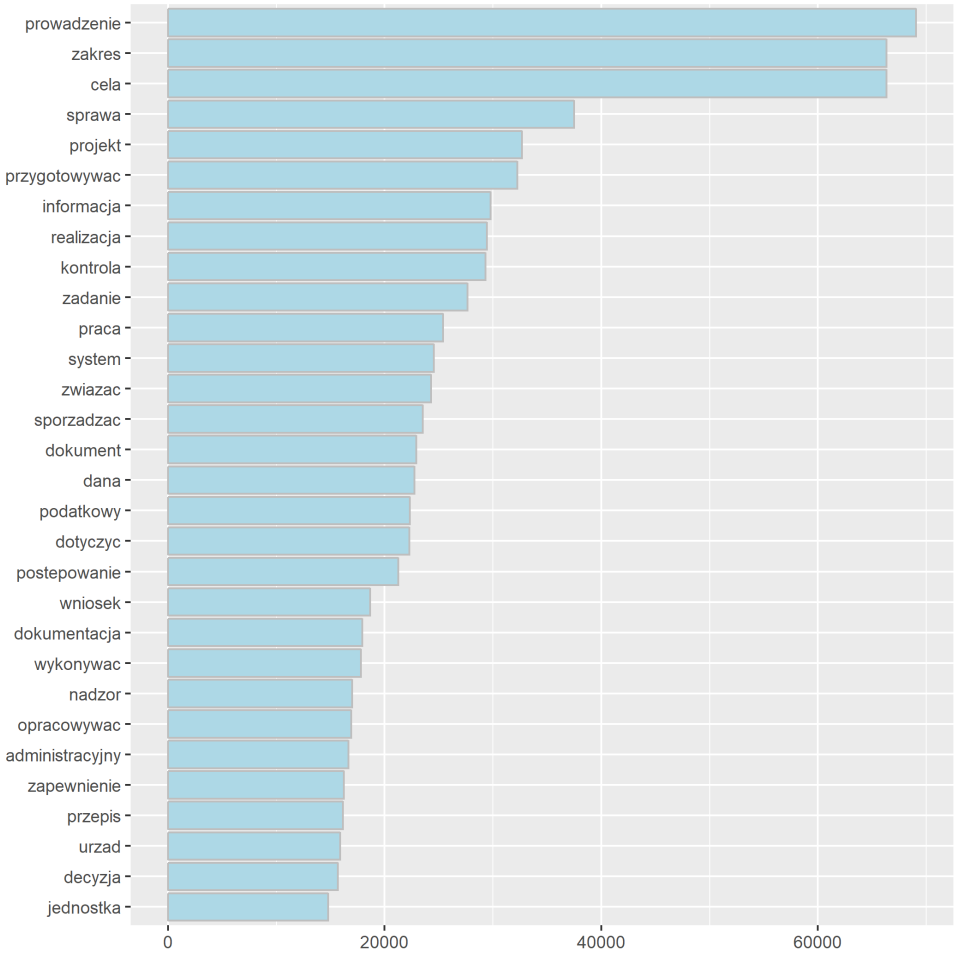
Rysunek 3.3. Liczba ofer pracy publikowanych w serwisie *nabory.kprm.gov.pl* według miesięcy z uwzględnieniem dnia tygodnia

Źródło: Opracowanie własne na postawie danych zebranych z portalu *nabory.kprm.gov.pl*

że najczęściej występujące słowa, które będą miały duży wpływ na działanie modelu alokacji ukrytej Dirichleta, występują prawdopodobnie w zdecydowanej większości ofert, bez względu na kategorię oferowanego stanowiska. Jest to problem, z którym będzie trzeba uporać się podczas budowania modelu.

Warto zauważyć, że na wykresie 3.4 nie występują różne odmiany tego samego słowa, a także brak na nim słów z listy *stop words*. Świadczy to o prawidłowym rezultacie zastosowanych wcześniej metod usuwania słów.

Ostatnim wykresem jaki będzie tworzony podczas eksploracji danych, to grafika typu *chmura słów* (ang. *word cloud*) 3.5. Częstość występowania słowa ze słownika interpretowana jest poprzez wielkość czcionki użytej w tworzeniu grafiki. Słowa występujące stosunkowo rzadko pisane są niewielką czcionką, zaś występujące często i najczęściej - dużą. Wykres można



Rysunek 3.4. Najczęściej występujące słowa w opisach stanowisk ofert pracy publikowanych na portalu *nabory.kprm.gov.pl*

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

wzbogacić o użycie palety kolorów do bardziej intuicyjnej interpretacji, a także, ze względów estetycznych. Wykorzystać można palety skokowe, w których po przekroczeniu pewnego progu następuje zmiana koloru, bądź palety gradientowe, w których częstość występowania słowa oznaczana jest intensywnością barwy. W grafice użyta jest paleta gradientowa przechodząca z koloru zielonego w niebieski.

Eksploracja danych dostarczyła paru cennych informacji, które wykorzystane będą w dalszej pracy nad projektem. Potwierdzone zostało usunięcie słów z listy *stop words* oraz pomyślne sprowadzenie pozostałych słów do ich tematu. Najczęściej występujące słowa niekoniecznie będą istotne w grupowaniu ofert pracy ze względu na ich specyfikę, ponieważ są to słowa pospolite bez względu na charakter stanowiska. Nic także nie wskazuje, aby w zbiorze danych przejawiały się właściwości, które należałoby uwzględnić - takie jak sezonowość. Z takimi wnio-

Rysunek 3.5. Grafika przedstawiająca najczęściej występujące słowa w opisach stanowisk ofert pracy publikowanych na portalu *nabory.kprm.gov.pl*

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

skami rozpoczęty zostaje dalszy fragment badania, będący budową modelu.

3.3 Detekcje tematów z wykorzystaniem algorytmu LDA

Do budowy modelu wykorzystane zostają kolejne dwa, nowe pakiety o nazwach *topicmodels* (Grün & Hornik, 2011) oraz *ldatuning* (Nikita, 2019). Pierwszy z nich umożliwi zbudowanie samego modelu, drugi zaś, dzięki wbudowanym metrykom, znacząco ułatwia dobór parametru K, odpowiadającego za liczbę tematów, które model ma rozpoznać. Dodatkowo pakiet *broom* (Robinson & Hayes, 2018) pomoże uporządkować dane funkcją *tidy()*. Inne pakiety, jakie się pojawią, to *tidyverse* i *tidytext*. Instaluję i ładuję wymienione pakiety 3.22

```
install.packages(c("broom", "topicmodels", "ldatuning"))
library(tidyverse)
library(tidytext)
library(broom)
library(topicmodels)
library(ldatuning)
```

Program 3.22. Instalacja i uruchamianie pakietów

Jednym z wniosków z poprzedniego podrozdziału było to, że najczęściej występujące słowa

w słowniku mają zastosowanie do większości pozycji oferowanych przez KPRM. W związku z tym zapada decyzja na usunięcie po jednym procencie najczęściej, ale także najrzadziej, występujących słów. Usuwając obserwacje odstające znacząco skrócę czas potrzebny na utworzenie modelu, a także podniosę jakoś rezultatów 3.23.

```
#wyznaczam 1% ęśnajczciej ęąwystpujcych ówyrazw
    high_n <- dane %>%
  count(stem, sort = TRUE) %>%
  top_n(round(0.01*nrow(.),0)) %>%
  mutate(stem = reorder(stem, n))
#wyznaczam 1% najrzadziej ęąwystpujcych ówyrazw
    low_n <- dane %>%
  count(stem, sort = TRUE) %>%
  top_n(-round(0.01*nrow(.),0)) %>%
  mutate(stem = reorder(stem, n))
#usuwam je ze zbioru
    dane2 <- dane %>%
  anti_join(high_n, by = "stem") %>%
  anti_join(low_n, by = "stem")
```

Program 3.23. Usuwanie 1% najczęściej i najrzadziej występujących słów

W momencie, w którym zbiór danych nie będzie już w żaden sposób modyfikowany, musi on być przekształcony do postaci Document-Term Matrix (w skrócie DTM) – macierzy rzadkiej, w której kolumny stanowią wyrazy ze słownika, zaś wiersze - dokument z którego one pochodzą 3.24. W polach macierzy zawarta jest liczba wystąpień danego słowa w danym dokumencie. Zdecydowana większość macierzy jest wypełniona zerami, oznaczającymi brak wystąpień tych słów w dokumentach.

```
    dane_dtm2 <- dane2 %>%
  select(ID, stem) %>%
  count(ID, stem, sort = TRUE) %>%
  ungroup() %>%
  cast_dtm(ID, stem, n)
```

Program 3.24. Sprowadzenie zbioru danych do postaci DTM

Wynikiem tego zabiegu jest macierz o rozmiarach 42 321 dokumentów na 8 745 wyrazy. Zera w tej macierzy zapełniają prawie jej całość - około 99.6%. Tylko 1 530 007 na 368 567 138 pól zawiera inne wartości.

Dla tak przygotowanych danych szukana jest wartość parametru K, dla którego model będzie najlepiej grupował oferty pracy 3.25.

```
    system.time(
  k_temat2 <- FindTopicsNumber(
    dtm = dane_dtm2,
    topics = c(seq(2, 9, 1),
               seq(10, 20, 2),
               seq(25, 50, 5)),
    metrics = c("Griffiths2004",
                "CaoJuan2009",
```

```
                "Arun2010",
                "Deveaud2014"),
    method = "Gibbs",
    control = list(seed = 12345),
    mc.cores =  4L,
    verbose = TRUE)
)
```

Program 3.25. Wyliczanie miar pomagających w doborze parametru K

Funkcja ta zwraca macierz, w której zawarte są wartości czterech metryk, nazwanych od artykułów z których pochodzą, *Griffiths2004* (opisana w drugim rozdziale jako metoda średniej harmonicznej), *CaoJuan2009*, *Arun2010* i *Deveaud2014* dla różnych wartości parametru K. Pakiet *ldatuning* dostarcza także funkcję 3.26, która tworzy wykres 3.6 dla wyżej opisanej macierzy w celu łatwiejszej interpretacji wyników.

```
    FindTopicsNumber_plot(k_temat2)
```

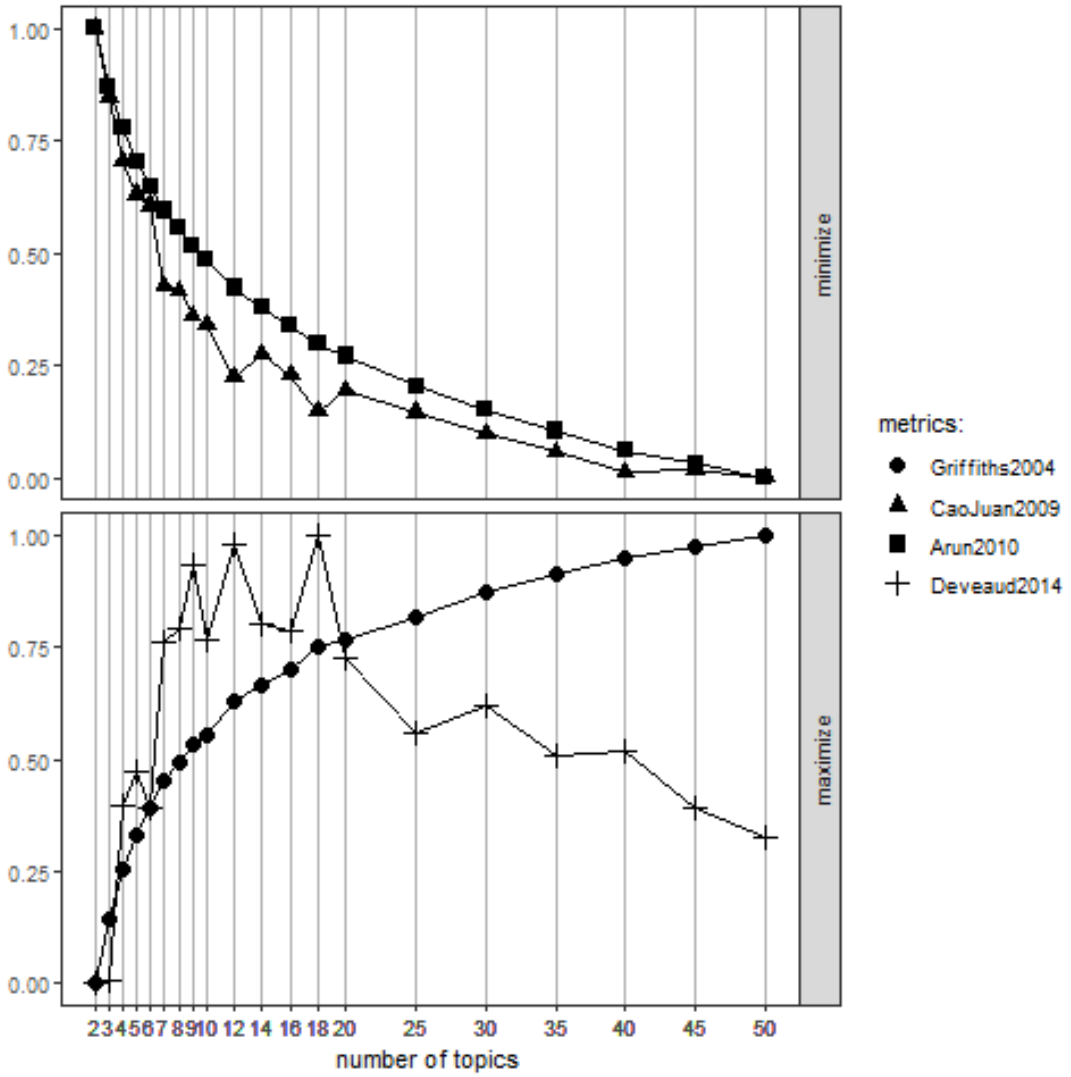
Program 3.26. Kod tworzący wykres

Wykres 3.6 jest podzielony na dwie płaszczyzny. Miary oznaczone *CaoJuan2009* i *Arun2010* powinno się minimalizować, zaś *Griffiths2004* i *Deveaud2014* maksymalizować. Miary *Arun2010* i *Griffiths2004* obie tworzą dość gładkie krzywe, zaś w miarach *CaoJuan2009* i *Deveaud2014* występują skoki wartości związane ze zmianą parametru. Warto też mieć na uwadze, że nie powinno się kierować tylko i wyłącznie miarami matematycznymi. Gdyby tak było, należało by powiększyć zakres poszukiwań z 50 tematów do paruset. Tak duża liczba tematów w modelu okazałaby się jednak niepraktyczna w interpretacji. Zauważyć można, że dla zbioru wartości $k \in (5, 7, 9, 12, 18)$ występują największe zmiany wartości miar. Liczby w tym zbiorze są także na tyle małe, że modele zbudowane z takim parametrem K będą względnie proste w interpretacji. Zbudowane zostaną modele dla każdej z wartości zbioru i poprzez analizę wyników wybrany zostanie najlepszy z nich 3.27.

```
    lda_list <- list()
    i~<- 1
    for (k in c(5, 7, 9, 12, 18))
    {
      lda_list[[i]] <- LDA(dane_dtm2,
                           k = k,
                           control = list(seed = 1234))
      i~<- i+1
    }
```

Program 3.27. Tworzenie modelu LDA dla pięciu wybranych parametrów K

Model o konkretnym parametrze K wybrany będzie metoda ekspercką. W tym celu stworzone zotają dwa nowe obiekty 3.28:



Rysunek 3.6. Reprezentacja graficzna miar optymalizujących dobór parametru K modelu LDA

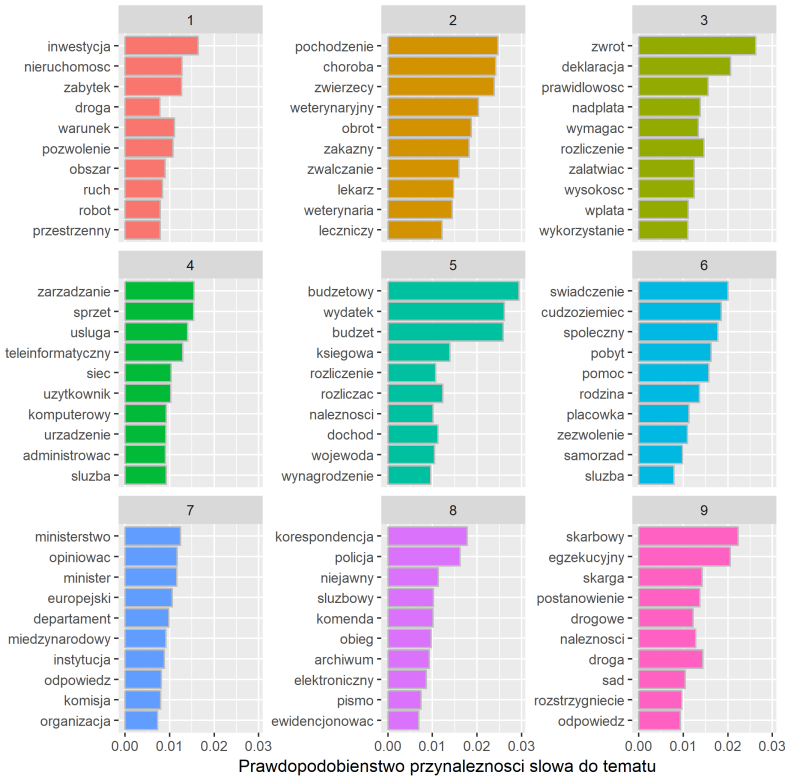
Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

- listę beta – zawierającą prawdopodobieństwa przypisania danego słowa do danego tematu dla każdego z modeli,
- listę gamma – zawierającą kompozycje prawdopodobieństw przynależności danego dokumentu do tematów w modelu dla każdego z modeli.

```
beta_list <- list()
gamma_list <- list()
for(i in 1:length(lda_list))
{
beta_list[[i]] <- tidy(lda_list[[i]], matrix = "beta")
gamma_list[[i]] <- tidy(lda_list[[i]], matrix = "gamma")
}
```

Program 3.28. Wyznaczenie wartości beta i gamma dla zbudowanych modeli

Specjalną uwagę warto poświęcić obiektowi przechowującemu prawdopodobieństwo przynależności danego wyrazu do tematu w modelu. Nakreślić można wykres dla każdego z pięciu modeli ilustrujący dziesięć wyrazów z największym współczynnikiem β dla każdego z tematów. Porównując te słowa dla każdego z modeli, szukany jest taki model z takimi tematami, żeby łatwo i intuicyjnie można było określić charakter wszystkich grup dokumentów. W ramach tematu słowa w nim występujące powinny być klarownie powiązane, a także być odmienne dla zagadnień opisywanych przez inne tematy. Na tej podstawie wybrany zostaje model z parametrem K, oznaczającym liczbę tematów, równym dziewięć. Wykres 3.7 poniżej przedstawia po 10 najbardziej istotnych wyrazów w tematach wyznaczonych metodą alokacji ukrytej Dichirleta.



Rysunek 3.7. Najczęściej występujące słowa w opisach stanowisk ofert pracy publikowanych na portalu *nabory.kprm.gov.pl* według tematów

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

Dokumentom w modelu LDA przypisuje się kompozycję tematów. Oznacza to, że zawartość $i-tego$ dokumentu składa się przykładowo w 50% z tematu pierwszego, w 30% z drugiego i z 20% z trzeciego. W związku z tym wyznaczone zostają dla każdego z tematów po 3 oferty pracy, które są przypisane w największym stopniu do jednego z nich 3.29. Są to obserwacje praktycznie jednoznacznie przypisane, a ich prawdopodobieństwa oscylują wokół jedynki.

```
  stanowiska <- gamma_list[[3]] %>%
left_join(opisy, by = c("document" = "ID")) %>%
arrange(-gamma) %>%
distinct(tytul, .keep_all = TRUE) %>%
group_by(topic) %>%
slice(1:3) %>%
select(topic, tytul)

  for(i in 1:9)
  {
  stanowiska %>%
      filter(topic == i) %>%
      select(tytul) %>%
      print()
  }
```

Program 3.29. Wyznaczenie najbardziej jednoznacznie przypisanych ofert pracy do tematów zbudowanego modelu

Dołączone zostają do tabeli z wartościami γ opisy i tytuły stanowisk najbardziej powiązanych z danymi tematami i weryfikowane zostaje, czy stanowiska te w rzeczywistości łączą się z kluczowymi wyrazami wyznaczonymi dla tematów przez współczynnik β. Następnie, podejmując decyzję o zatrzymaniu modelu z dziewięcioma tematami, nadane zostają tematom w sposó intuicyjny etykiety związane ze specyfiką stanowisk 3.30.

```
  lda_popyt <- gamma_list[[3]] %>%
mutate(topic_new = case_when(
  topic == 1 ~ "infrastruktura",
  topic == 2 ~ "weterynaria, ńbezpieczestwo żśywnoci i~higiena",
  topic == 3 ~ "ewidencja i~rejestracja",
  topic == 4 ~ "informatyka, programowanie i~wsparcie techniczne",
  topic == 5 ~ "finanse, ęśćksigowo i~żbudetowanie",
  topic == 6 ~ "pomoc łspoeczna i~cudzoziemcy",
  topic == 7 ~ "polityka ęmidzynarodowa",
  topic == 8 ~ "ochrona i~ńbezpieczestwo ńpastwa",
  topic == 9 ~ "legislacja"
))
```

Program 3.30. Przypisanie tematom etykiet

Jak wyczytać można z zamieszczonego powyżej kodu, tematy te to:

1. infrastruktura
2. weterynaria, bezpieczeństwo żywności i higiena
3. ewidencja i rejestracja
4. informatyka, programowanie i wsparcie techniczne
5. finanse, księgowość i budżetowanie
6. pomoc społeczna i cudzoziemcy

7. polityka międzynarodowa
8. ochrona i bezpieczeństwo państwa
9. legislacja

Kończy to sekcję samej budowy modelu. Ostatni podrozdział skupiał się będzie na interpretacji i wizualizacji otrzymanych rezultatów.

3.4 Analiza tematów uzyskanych z metody LDA

Po pomyślnym zbudowaniu modelu zostaje już jedynie analiza wyników. Do rozważenia pozostają dwie macierze. Jedna przechowująca prawdopodobieństwa przypisania danego słowa do konkretnego tematu modelu, druga - przechowująca kompozycje tematów przypisanych dokumentom w korpusie. W wyznaczaniu popytu bardziej przydatna okaże się ta druga. Niestety, z założenia modelu LDA, dokumenty z korpusu nie są jednoznacznie klasyfikowane do żadnego z tematów. Co prawda, niektóre dokumenty praktycznie idealnie wpasowują się w wygenerowane przez model tematy, jednak nie jest to podstawa, by opierać wyznaczanie popytu biorąc tylko takie oferty pracy pod uwagę. Rozwiązanie, które zostanie zastosowane do wyznaczenia popytu, zakłada zsumowanie, dla każdego z tematów, prawdopodobieństw przynależności każdego z dokumentów. Suma prawdopodobieństw przynależności do tematów pojedynczego dokumentu równa jest jedności. Tym samym suma wszystkich elementów macierzy równa jest liczbie obserwacji w zbiorze, czyli wszystkim oferowanym, na przestrzeni czasu, ofertom publikowanym na stronie KPRM. Biorąc to pod uwagę, można założyć, że rozkłady brzegowe zbiorowości wyznaczane dla tematów przybliżą liczbę ofert zaliczanych do każdej z kategorii działalności. Dzieląc wyznaczone rozkłady brzegowe przez liczbę wszystkich obserwacji uzyskuje udział każdego z tematów w całej zbiorowości. Wyniki tych operacji prezentują się następująco 3.3:

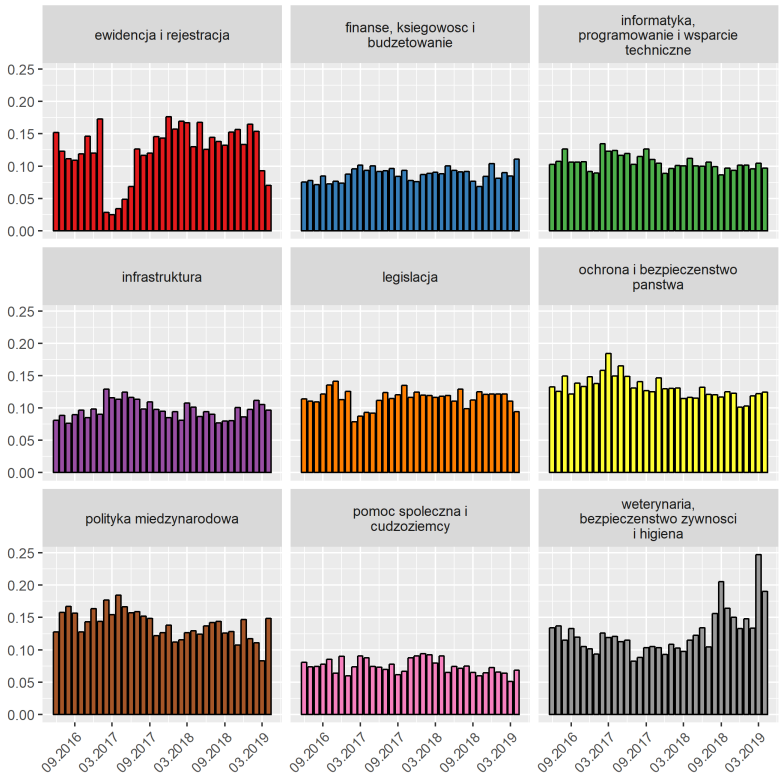
Jak można zauważyć, najwięcej ofert pracy publikowanych na portalu KPRM zostało zakwalifikowanych do działalności związanej z polityką międzynarodową. Najmniej zaś przypisanych zostało do pomocy społecznej i pracy związanej ze sprawami dotyczącymi cudzoziemców. Interesujące okazać się może jak udziały w poszczególnych branżach prezentowały się z perspektywy czasu. Tworzone w związku z tym zostają wykresy prezentujące te zależności. Pierwszy z nich jest wykresem słupkowym, w którym wysokość słupków odpowiada zapotrzebowaniu

Tabela 3.3. Rozkład tematów w badanych ofertach z KPRM

Temat	Liczba ofert	Udział
ewidencja i rejestracja	5576	0.132
finanse, księgowość i budżetowanie	3667	0.0867
informatyka, programowanie i wsparcie techniczne	4401	0.104
infrastruktura	4038	0.0954
legislacja	4931	0.117
ochrona i bezpieczeństwo państwa	5457	0.129
polityka międzynarodowa	5762	0.136
pomoc społeczna i cudzoziemcy	3149	0.0744
weterynaria, bezpieczeństwo żywności i higiena	5340	0.126

Źródło: Opracowanie własne.

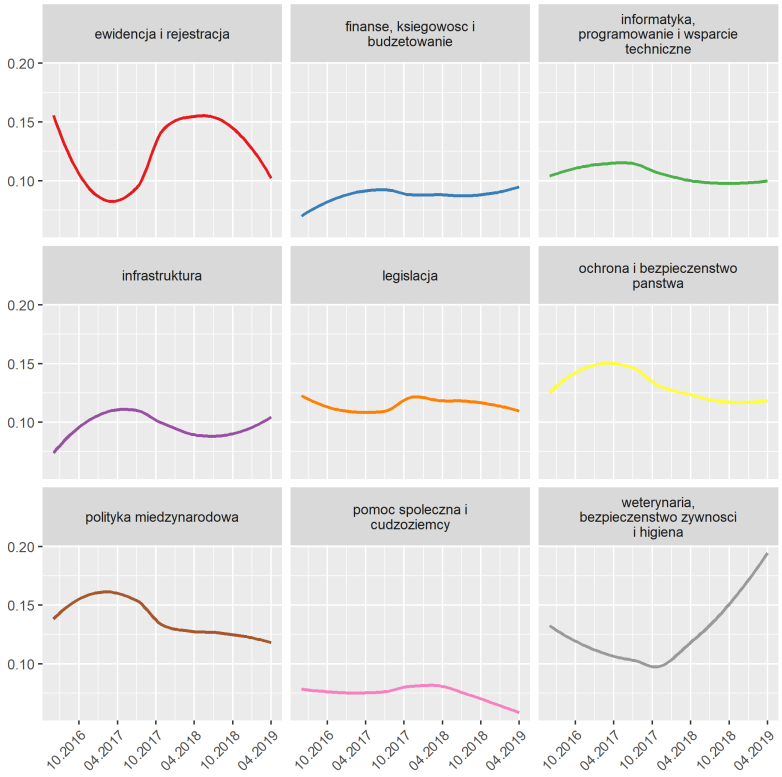
na kategorie działalności w danym miesiącu 3.8. Drugi wykres jest wygładzonym wykresem liniowym 3.9.



Rysunek 3.8. Udział procentowy ofert pracy publikowanych na portalu *nabory.kprm.gov.pl* w czasie według tematów

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

Na tych wykresach odczytać można, że zapotrzebowanie na większość z działalności jest dość stabilne z niewielkimi wzrostami bądź spadkami. Odstającymi branżami wydają się być *ewidencja i rejestracja* i *weterynaria, bezpieczeństwo żywności i higiena*.



Rysunek 3.9. Udział procentowy ofert pracy publikowanych na portalu *nabory.kprm.gov.pl* w czasie według tematów

Źródło: Opracowanie własne na podstawie danych zebranych z portalu *nabory.kprm.gov.pl*

W przypadku *ewidencji i rejestracji* na początku obserwujemy dość drastyczny spadek w lutym 2017 roku. Potem, w kwietniu, zapotrzebowanie na tego typu prace znowu wzrasta i utrzymuje dość stabilny poziom, aby w marcu 2019 roku znowu gwałtownie spaść.

Odmienną sytuację widać w kategorii oznaczonej jako *weterynaria, bezpieczeństwo żywności i higiena*, w której to w marcu 2019 roku nastąpił wzrost popytu by osiągnąć maksymalną wartość dla tej kategorii działalności.

Zmiany te, mniejsze jak i większe, mogą być oczywiście spowodowane niedoskonałością modelu, który dla ofert pracy w wybranych punktach w czasie dokonał interpretacji tekstu dokumentu niezgodnej z realiami, przyporządkowując jej złą kompozycję tematów. Może to być także prawda, będąca uzasadniona nieznanymi powodami, według których zapotrzebowanie rosło bądź malało.

W rozdziale tym przedstawiony został przebieg przeprowadzonego badania, w które miało przybliżyć zapotrzebowanie na różne kategorie działalności stosując model ukrytej alokacji Dichirleta. Badanie to miało także na celu pokazanie potencjału jaki tkwi w niekonwencjonalnych,

internetowych źródłach danych. Te oferują ogromne zasoby informacji mogących służyć badaniom i wyciąganiu interesujących wniosków. Podchodzić należy do nich jednak rozważnie, aby uniknąć problemów prawnych.

Podsumowanie

Celem pracy była analiza ofert pracy na stanowiska w służbie cywilnej publikowanych na stronie naborów Kancelarii Prezesa Rady Ministrów poprzez stworzenie zbioru danych metodami web-scrapingu i zastosowanie na tak stworzonym zbiorze danych modelu alokacji ukrytej Dichirleta. Rezultatem badania jest przybliżone zapotrzebowanie na pracowników wykonujących pracę o różnym charakterze w służbie cywilnej. Web-scraping jest to masowe pobieranie danych wyświetlanych użytkownikowi podczas przeglądania stron internetowych, co pozwala na uzyskanie niskim kosztem dużych wolumenów danych, na których przeprowadzać można analizy. Wykorzystany model alokacji ukrytej Dichirleta to przykład modelu modelującego tematy, co znaczy, że na podstawie dostarczonych danych tekstowych, model ten wyznacza k liczbę tematów, jakie poruszane są przez dane, gdzie k jest parametrem modelu.

Duży nacisk w badaniu, jak i tej pracy, stawiany był na stworzenie i zastosowanie robota służącego do masowego pobierania danych, czyli inaczej poprzez zastosowanie metod web-scrapingu do uzyskania danych – w przybliżeniu aż 42 tysięcy obserwacji zostało skomponowanych z informacji pobranych z ofert pracy publikowanych na stronie z naborami. Metody web-scrapingu są przykładem wykorzystania ogromnych rozmiarów danych zawartych w Internecie, które oferują niezwykłe możliwości przeprowadzającemu badanie niskim kosztem. Zbiór danych wykorzystywany w badaniu został w całości utworzony przez autora pracy przez napisanie kodu pobierającego dane specjalnie w tym celu. Dalszą częścią badania była transformacja danych w celu łatwiejszego korzystania z nich, by ostatecznie móc zastosować model ukrytej alokacji Dichirleta. Z powodów praktycznych, zbiór danych pozbawiony został polskich akcentów, ujednolicona została wielkość liter na małe, usunięte zostały słowa z listy *stop words*, czyli takie, które same w sobie nie noszą żadnej wartości logicznej w tekście i ostatecznie zastosowany został *stemming*, czyli wszystkie słowa sformatowane zostały do tematu słów – usunięta została odmiana gramatyczna każdego słowa, pozostawiając słowo w jego bazowej postaci. Zabieg ten był krytyczny w zbudowaniu modelu, którego zwracane rezultaty były wysokiej jakości

i można je było z łatwością interpretować. Model alokacji ukrytej Dichirleta, będący jednym z algorytmów modelowania tematów, został wykorzystany by wyznaczyć podobieństwa między opisami ofert pracy i ugrupować je. Parametr modelu odpowiadający za liczbę grup, które wygenerować ma model został wybrany na podstawie zbudowania modelów o różnej wartości parametru, wyznaczenia i porównania miar oceny modelu i finalnie weryfikacji najlepszych modeli w sposób ekspercki. Na tej podstawie wyznaczone zostały dziedziny działalności zgłaszane przez pracodawców i przybliżone zostało zapotrzebowanie na pracowników w służbie cywilnej tych dziedzin poprzez odpowiednie operacje matematyczne. Wyznaczonych zostało 9 kategorii usług oferowanych przez pracowników i są to:

- infrastruktura,
- weterynaria, bezpieczeństwo żywności i higiena,
- ewidencja i rejestracja,
- informatyka, programowanie i wsparcie techniczne,
- finanse, księgowość i budżetowanie,
- pomoc społeczna i cudzoziemcym
- polityka międzynarodowa,
- ochrona i bezpieczeństwo państwa,
- legislacja.

W pracy zawarte zostały informacje teoretyczne o rynku pracy, źródłach danych na temat rynku pracy oraz najważniejszych metodach wykorzystywanych w trakcie badania – web-scrapingu oraz modelu alokacji ukrytej Dichirleta. Dodatkowo znaleźć w niej można kod w środowisku R napisany przez autora pracy, który umożliwił realizację badania i wiele tabel i wykresów przedstawiających i pomagających zrozumieć wyniki.

Przedstawione badanie i metody sugerują odmienne podejście do badania rynku pracy, które wyróżnia się bardzo niewielkimi kosztami i szybkim czasem przeprowadzenia badania. Metody opisane w pracy znajdują zastosowanie w wielu dziedzinach życia i nie są one dedykowane jedynie do analiz rynkowych. Web-scraping oferuje ogromny potencjał w zbieraniu danych do celów poznawczych, zaś model ukrytej alokacji Dichirleta pozwala na grupowanie wszelakich dokumentów.

Bibliografia

Adamowicz, S. (2019). Urząd pracy. Dostęp z https://pl.wikipedia.org/wiki/Urz%C4%85d_pracy

Biuro Inwestycji i Cykli Ekonomicznych. (2019). Barometr Ofert Pracy (BOP). Dostęp z http://biec.org/produkty/

Centrum Ewaluacji i Analiz Polityk Publicznych. (2014). Badanie Kapitału Ludzkiego. Dostęp z https://www.parp.gov.pl/component/publications/publications/?series=14#filter-publications

Dębek, S., Kowalik, P., Pujer, K., Dahlke, M., Balicka, A., Potapenko, K., … Czerwińska, K. (2016). *Rynek pracy w Polsce-szanse i zagrożenia*. Exante.

Fellows, I. (2018). *wordcloud: Word Clouds*. R package version 2.6.

Główny Urząd Statystyczny. (2018). Popyt na Pracę w 2017r.

Główny Urząd Statystyczny. (2019). Badanie aktywności ekonomicznej ludności (BAEL).

Gładysz, A. (2014). Wykorzystanie metody opartej na ukrytej alokacji Dirichleta do automatycznej identyfikacji słów kluczowych w dokumentach. *Logistyka*, (3), 2011–2019.

Góra, M. & Sztanderska, U. (2006). *Wprowadzenie do analizy lokalnego rynku pracy*. Przewodnik, MPiPS, Warszawa.

Gagolewski, M. (2018). *R package stringi: Character string processing facilities*.

Grolemund, G. & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25.

Grün, B. & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30. doi:10.18637/jss.v040.i13

Jeffery, S. (2017). Web Scraping Uses: The Epic List of Web Scraping Uses.

Krotov, V. & Silva, L. (2018). Legality and Ethics of Web Scraping.

Kryńska, E. & Kwiatkowski, E. (2013). *Podstawy wiedzy o rynku pracy*. Wydawnictwo Uniwersytetu Łódzkiego.

Kuraś, J. (2017). Centralna Baza Ofert Pracy: pracodawcy łatwiej znajdą pracowników. Rzeczpospolita.

Milewski, R. & Wydawnictwo Naukowe, P. (1994). *Elementarne zagadnienia ekonomii*. Wydaw. Naukowe PWN.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R. & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science.

Mizia, A. & Latocha, A. (2019). Rynek pracy. Dostęp z https://mfiles.pl/pl/index.php/Rynek_pracy

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Nikita, M. (2019). *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. R package version 1.0.0.

Patel, H. (2018). How Web Scraping is Transforming the World with its Applications. Towards Data Science.

Polska Agencja Rozwoju Przedsiębiorczości. (2018). Badanie Kapitału Ludzkiego – wyniki. Dostęp z https://ec.europa.eu/epale/pl/content/bilans-kapitalu-ludzkiego-wyniki-badania

Ponweiser, M. (2012). Latent Dirichlet allocation in R.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Robinson, D. & Hayes, A. (2018). *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.1.

Scraping Agent. (2019). Dostęp z https://www.agenty.com/blog/web-scraping-top-15-ways-to-use-it-for-business

Silge, J. & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). doi:10.21105/joss.00037

Szaban, J. (2013). *Rynek pracy w Polsce i w Unii Europejskiej*. Diffin.

Węc, P. & Lelakowska, J. (2019). Popyt na pracę. Dostęp z https://mfiles.pl/pl/index.php/Popyt_na_prac%C4%99

Wickham, H. (2016). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Spis tabel

Spis rysunków

Spis Programów