Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Studia Podyplomowe
„Metody statystyczne w biznesie.
Warsztaty z oprogramowaniem SAS”

mgr Adam Dudo

Analiza Text Mining powiązania treści opisów na portalu LinkedIn.com z parametrami opisującymi przebieg kariery zawodowej

Praca dyplomowa
napisana pod kierunkiem
dr Karoliny Kuligowskiej
Katedra Informatyki Gospodarczej
i Analiz Ekonomicznych

Warszawa, wrzesień 2016

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona wszystkie wymogi formalne i merytoryczne pracy studiów podyplomowych

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

Praca podejmuje problematykę dotyczącą budowy tzw. *e-wizerunku* (przedstawienia się w środowisku opartym na Internecie) oraz jego skuteczności w szukaniu pracy. Autor pracy przeanalizował wybrane opisy kariery osób, które zamieściły te informacje na międzynarodowym serwisie społecznościowym LinkedIn.com, który specjalizuje się w promowaniu kontaktów zawodowo-biznesowych. Badanie zostało przeprowadzone za pomocą narzędzi Text Mininig firmy SAS na niereprezentatywnej próbie 123 opisów karier. Punktem odniesienia były możliwe do uzyskania dane jakościowe i ilościowe dotyczące kariery zawodowej poszczególnych osób.

Słowa kluczowe

kariera zawodowa, e-wizerunek, text mining, portal społecznościowy, praca, LinkedIn

Dziedzina pracy (kody wg programu Socrates-Erasmus)

Ekonomia (14300)

Spis treści

Wstęp

Przedmiot badań

Badaniu została poddana treść profili zawodowych wybranej grupy osób umieszczonych na portalu LinkedIn, czyli jednym z najpopularniejszych na świecie portalu służącym nawiązywaniu i utrzymywaniu kontaktów zawodowych[1]. Analizie podlegała zawartość tekstowa tych profili i ewentualna korelacja z takimi parametrami opisującymi dany profil, jak: branża, w której pracuje dana osoba, poziom stanowiska (kierownicze, specjalistyczne) oraz długość kariery zawodowej i częstość zmian pracy.

Cel pracy

Celem pracy było ustalenie, czy istnieje związek pomiędzy zawartością tekstową poszczególnych opisów profili zawodowych w wybranej grupie badawczej, a przebiegiem kariery zawodowej poszczególnych osób dla kilku parametrów, które były możliwe do uzyskania na podstawie treści profili. Odkrycie takiego związku mogłoby być przydatną wskazówką dla pracodawców, jak zwiększyć efektywność analizy danych o potencjalnych kandydatach do pracy. Badanie dotyczy portalu LinkedIn (obecnie największego międzynarodowego portalu społecznościowego z biznesowymi profilami kandydatów połączonymi siecią kontaktów zawodowych), ale wnioski mają charakter bardziej uniwersalny – odnoszony do innych tego typu portali.

Charakterystyka pracy

W Rozdziale 1 przedstawiono szereg informacji na temat obecnego wykorzystania portali społecznościowych do kontaktów zawodowych. Zaprezentowane zostały argumenty wskazujące na znaczenie tego typu portali w efektywnym, zakończonym sukcesem, szukaniu pracy. Poruszono aspekt analizy umieszczonych tam danych.

Rozdział 2 opisuje zakres analizy przeprowadzonej w ramach niniejszej pracy, przyjęte założenia i metodę badawczą, specyfikację danych oraz szczegółowe wyniki.

Zakończenie zawiera wnioski uzyskane w wyniku analizy danych. Wnioski dotyczą zarówno samego badania, jak i możliwości Text Mining w zakresie tego typu badań.

[1] Lee Hecht Harrison DBM Polska (dane z 2015 r.)

Rozdział 1

W firmach działających obecnie na rynku w proces rekrutacji pracowników angażowanych jest coraz więcej narzędzi elektronicznych udostępnianych w środowisku internetowym. Bezsporną zaletą tych narzędzi jest łatwa dostępność dla potencjalnych kandydatów do pracy, co przekłada się na znacznie większy wolumen danych o kandydatach, jaki trzeba przeanalizować, w porównaniu do procesów rekrutacji sprzed 10-u czy 20-u lat. Rośnie zatem zapotrzebowanie na szybkość analizowania tych danych pod kątem potrzeb poszczególnych procesów rekrutacyjnych z zachowaniem możliwie jak najwyższej jakości wyników takich analiz.

Portale społecznościowe i e-wizerunek

Portal Praca.pl, będący profesjonalną platformą rekrutacyjną, opublikował w 2011 roku wyniki badań ankietowych dotyczących tzw. e-wizerunku w odniesieniu do procesu szukania pracy[2]. Okazuje się, że tylko około 16% ankietowanych osób stwierdziło, iż świadomie buduje swój wizerunek w Internecie. Podobny udział (15%) badanych jest przekonany, że nie należy się przejmować własnym wizerunkiem. Autor artykułu zwraca uwagę, że jest to o tyle nierozsądne, że coraz częściej szuka się partnerów biznesowych poprzez kontakty networkingowe (sieć znajomych) i coraz więcej rekruterów sprawdza kandydatów w sieci.

Zgodnie Raportem przygotowanym przez Lee Hecht Harrison – firmy, która zajmuje się profesjonalnym wsparciem kariery zawodowej – 97% osób zajmujących się profesjonalnie rekrutacją w Polsce korzysta z mediów społecznościowych, a 79% ankietowanych rekruterów stwierdziło, że media społecznościowe pomogły im w zatrudnieniu nowego pracownika[3]. W tej grupie osób 89% korzysta z portalu LinkedIn, natomiast z kolejnych dwóch narzędzi, czyli Goldenline.pl i Facebook.com – odpowiednio 77% i 32%. Liczba użytkowników LinkedIn w Polsce rośnie i zbliża się do liczby użytkowników jego polskojęzycznego odpowiednika, czyli Goldenline. Obecnie LinkedIn liczy około 1,7 miliona użytkowników, natomiast Goldenline – około 2 milionów[4].

[2] Wizerunek w Internecie a szukanie pracy, Praca.pl, listopad 2011

[3] Media Społecznościowe w Rekrutacji, Raport 2015, Lee Hecht Harrison DBM

[4] Warsztat „E-wizerunek" przeprowadzony przez firmę Lee Hecht Harrison DBM Polska w marcu 2016 roku.

Autorowi niniejszej pracy nie udało się natrafić na opracowania bezpośrednio odnoszące się do podjętego tematu badań. Istnieje natomiast szereg opracowań, które odnoszą się do „skuteczności" danego profilu w mediach społecznościowych w odniesieniu do szukania pracy. Przy czym „skuteczność" jest tu najczęściej rozumiana, jako spowodowanie, że dany profil (dana osoba) będzie łatwiejszy do odnalezienia przez potencjalnego pracodawcę lub rekrutera.

Skuteczność internetowego profilu zawodowego

Joe Chernov na portalu blog.hubspot.com podzielił się wynikami eksperymentu, jaki przeprowadził z użyciem własnego profilu na LinkedIn[5]. Po porównaniu zawartości i liczby „odwiedzin" swojego profilu z profilem innej znanej sobie osoby, dokonał zmian w swoim profilu. Po zmianach ponownie sprawdził, jak dużo osób odwiedziło jego profil na LinkedIn w takim samym, co wcześniej, przedziale czasu (LinkedIn pozwala uzyskać takie statystyki w odniesieniu do swojego profilu). Okazało się, że liczba wyświetleń jego profilu wzrosła o ponad 25%. W wyniku tych zmian autor artykułu sformułował kilka zaleceń odnoszących się do podniesienia „skuteczności" profilu na LinkedIn. Poniżej te z nich, które potencjalnie mogą podlegać dalszej analizie Text Mining:

- Używanie „znormalizowanych" nazw stanowisk – w przeciwieństwie do nazw specyficznych tylko dla jednej firmy.
- Użycie słów kluczowych w polu „Professional Headline", czyli krótkim opisie kompetencji danej osoby – zwiększa skuteczność wyszukiwania.
- Właściwe sformułowanie pola podsumowania kariery („Summary") – opis kompetencji, a nie na przykład lista otrzymanych nagród.
- Rozszerzenie sekcji „Skills" – zgodnie z uzyskanymi kompetencjami.

Z kolei Andy Foote w swoim artykule na stronie LinkedInSights.com sugeruje, że podstawą skutecznego wyszukiwania danego profilu jest w pierwszej kolejności jego kompletność[6]. Zdaniem autora algorytm wyszukiwania w LinkedIn w pierwszej kolejności bierze pod uwagę kompletność profilu (stopień wypełnienia poszczególnych sekcji), wspólne kontakty

[5] „How I Easily Got 25% More Views on My LinkedIn Profile", Joe Chernov

[6] „Why You Should COMPLETE Your LinkedIn PROFILE", Andy Foote

i odległość mierzoną najmniejszą liczbą połączeń w LinkedIn pomiędzy osobą szukaną i osobą szukającą, a na koniec – liczbę współdzielonych grup w ramach LinkedIn.
Co ciekawe, autor artykułu przywołuje dane podawane przez administratorów portalu LinkedIn: mianowicie zgodnie z ich statystykami tylko około 50% użytkowników portalu ma w 100% wypełniony swój profil. Autor nie odnosi się do treści poszczególnych sekcji i pól. W związku z tym przedmiot niniejszej pracy nie został w artykule Foote'a bezpośrednio poruszony.

Najbliższa tematowi niniejszej pracy dyplomowej jest analiza[7] opisana na blogu firmy AYLIEN, która zajmuje się opracowywaniem narzędzi do analizy tekstów w Internecie[8]. Autorzy artykułu postawili sobie za cel wskazanie, która sekcja profilu na LinkedIn najbardziej wpływa na rolę zawodową danej osoby. W ramach eksperymentu powstał model predykcyjny, który miał za zadanie przewidywanie nazwy stanowiska/roli danej osoby na podstawie zawartości jej profilu, a właściwie – poszczególnych jego sekcji.

Pod uwagę brane były cztery sekcje: słowa kluczowe („Headline"), podsumowanie („Summary"), nazwy stanowisk („Experience") i kompetencje („Skills & Endorsements"). Badaniu podlegały wybrane przez autorów profile osób dobrze znanych badaczom. Wyniki predykcji mogły więc być porównywane z faktycznym stanowiskiem/rolą pełnioną przez właściciela profilu.

Autorzy w podsumowaniu zwrócili uwagę, że skuteczność działania modelu predykcyjnego – w przypadku każdej z wybranych sekcji profilu – w dużym stopniu zależała od poziomu wypełnienia treścią danej sekcji. Dlatego ocena oddziaływania danej sekcji na zmienną zależną, jaką było stanowisko/rola danej osoby, zmieniała się w zależności od analizowanego profilu. Wynikiem analizy było stwierdzenie, że na właściwe przypisanie roli zawodowej danej osoby w największym stopniu wpływa zawartość sekcji „Skills & Endorsements".

[7] „What section of a LinkedIn profile best represents a candidate's job function?", AYLIEN

[8] Zobacz http://aylien.com/about

Analiza treści CV

Zagadnieniem bardzo bliskim tematowi niniejszej pracy dyplomowej jest analiza życiorysów kandydatów do pracy w postaci dokumentów Curriculum Vitae (CV). Tutaj również analizowane są dane dotyczące kariery zawodowej kandydata. Z punktu widzenia automatyzacji analizy dokumentów dane te są dostarczane w postaci dokumentu, którego strukturę trzeba dopiero odkryć. W przypadku dokumentów dostarczanych w formie papierowej konieczne jest poprzedzenie automatycznej analizy ich zawartości procesem rozpoznania tekstu Optical Caracter Recognition (ang. OCR). Coraz częściej jednak życiorysy zawodowe w procesach rekrutacji są dostarczane od razu w postaci elektronicznej.

Kolejnym podobieństwem pomiędzy analizą CV, a analizą profili na LinkedIn jest zazwyczaj duża liczba dokumentów wymagających obróbki. Dlatego w obu przypadkach automatyzacja, albo chociaż pół-automatyzacja tego procesu jest czymś, co może przynieść duże korzyści – szczególnie w przypadku osób, które zawodowo zajmują się rekrutacją pracowników.

Temat analizy zawartości dokumentów CV porusza materiał opracowany przez firmę Koltech, opisujący funkcje programu CV Distiller Software tej firmy[9]. Działanie tego oprogramowania obejmuje następujące etapy procesu rekrutacyjnego: pozyskanie dokumentów z różnych źródeł, odczytanie zawartości dokumentów (dostarczanych w różnych formatach), ustalenie języka dokumentu, ustalenie typu dokumentu (CV, list motywacyjny, inne), ekstrakcja informacji z dokumentu CV w zakresie wymaganym przez system HR danej firmy, kwalifikowanie (standaryzacja) tych informacji, wnioskowanie, generowanie wyników analizy. Działanie programu opiera się na przetwarzaniu lingwistycznym dokumentu i analizie semantycznej tekstu. Docelowo narzędzie ma pozwalać na automatyczne łączenie dokumentów CV z ofertami pracy. Ma też obsługiwać więcej języków – obecnie jest to głównie język francuski.

Rekrutacja poprzez media społecznościowe

Przy tej okazji warto zauważyć, że część funkcji realizowanych przez CV Distiller Software obsługuje mniej specjalizowane narzędzie, jakim jest Text Miner firmy SAS, który został wykorzystany w części ćwiczeniowej niniejszej pracy dyplomowej. Jednakże trudno dokonywać dokładnych porównań tych dwóch rozwiązań, ponieważ firma Koltech nie

[9] Automatic Analysis of Curriculum Vitae, a Case Study: the CV Distiller Software, F. Gire, S. Kolodziejczyk

ujawnia szczegółów przyjętego przez nią podejścia analitycznego analizy zawartości dokumentów.

Szczegółowy opis działania oprogramowania SAS Text Miner jest dostępny na stronie internetowej firmy SAS[10]. Dostępne są też liczne pozycje książkowe z tego zakresu. W niniejszej pracy posłużono się podręcznikiem w polskiej wersji językowej[11], opisującym zarówno działanie narzędzia, jak i podstawy teoretyczne i przyjęte metody obliczeniowe.

Ciekawe zjawisko można zaobserwować na podstawie badania przeprowadzonego przez firmę Bullhorn Reach w sierpniu 2014 roku[12]. Badanie dotyczyło co prawda wybranych rynków angielskojęzycznych (głównie Stany Zjednoczone, Anglia, Indie, Kanada), ale biorąc pod uwagę stale postępującą globalizację – w szczególności w odniesieniu do mediów społecznościowych – wydaje się, że wnioski w zakresie istniejących trendów można odnieść również do Polski. Otóż z przedstawionego przez Bullhorn Reach raportu aktywności rekrutacyjnych poprzez media społecznościowe wynika, że coraz większego znaczenia w obszarze poszukiwania pracy i poszukiwaniu pracowników nabiera Facebook. Na rynku amerykańskim jedynie w przypadku Facebooka zaobserwowano wzrost liczby rekruterów korzystających z tej platformy komunikacji. Niemniej jednak udział procentowy rekruterów korzystających z LinkedIn (97%) wciąż znacznie przewyższa taki sam udział w przypadku Facebooka (ok. 19%).

Należy założyć, że wykorzystanie do procesów rekrutacyjnych mediów społecznościowych dostępnych w Internecie, innych niż LinkedIn, będzie rosło. Co z kolei pociąga za sobą konieczność rozwoju narzędzi do analizy danych umieszczanych na tych platformach. O ile w przypadku rozwiązań typu LinkedIn duża część wprowadzanych przez użytkowników danych ma charakter ustrukturyzowany (chociaż przeciętny użytkownik może wyeksportować dane tylko do postaci ciągłego dokumentu tekstowego z rozszerzeniem PDF), to w przypadku

[10] http://support.sas.com/documentation/index.html

[11] Text Mining – Metody, narzędzia, zastosowania, D. Spinczyk, M. Dzieciątko

[12] 2014 Global Social Recriuting Activity Report, Understanding Social Media Use in Recruiting, Bullhorn Reach, August 2014

zwykłych mediów społecznościowych konieczne jest stosowanie narzędzi z obszaru Text Mininig.

Rozdział 2

Badaniu zostały poddane profile zawodowe zamieszczone na portalu LinkedIn. Założeniem wyjściowym do badania była następująca hipoteza: istnieje zależność pomiędzy przebiegiem kariery zawodowej, a opisem tej kariery umieszczonym na portalu LinkedIn; zależność tę można wychwycić metodami Text Mining oraz dostępnymi w pakiecie SAS narzędziami statystycznymi (korelacja, analiza regresji itp.).

Przygotowanie danych

Na podstawie danych dostępnych na LinkedIn zostały wygenerowane pliki w formacie PDF zawierające informacje umieszczone przez właściciela danego profilu na portalu. LinkedIn pozwala wygenerować pojedynczy plik z danymi dla każdej osoby posiadającej profil zawodowy. Odrzucono bardzo ubogie profile (np. tylko imię i nazwisko oraz aktualne stanowisko) oraz profile opisane w języku polskim (założeniem analizy było użycie mechanizmów Text Mining w odniesieniu do języka angielskiego ze względu na ich większą dojrzałość w porównaniu z mechanizmami analizującymi język polski). Po dokonaniu takiej selekcji pozostały 123 dokumenty w formacie PDF – po jednym na każdą osobę/profil zawodowy.

Nie założono z góry konkretnej metody wybierania profili do badań, poza ich powiązaniem z profilem zawodowym na LinkedIn autora badania. Dlatego też nie należy traktować wyników dalej opisanego badania jako reprezentatywnych dla szerszej populacji. Mogą one jednak wskazywać na potencjał zastosowanego podejścia do badań na dużo liczniejszej grupie dokumentów.

Co prawda dane używane do analizy nie podlegają obowiązkowym prawnym ograniczeniom dotyczącym ich poufności, ale przed rozpoczęciem ich przetwarzania zostały poddane podstawowej anonimizacji poprzez ręczne usunięcie z plików imion i nazwisk oraz danych kontaktowych właścicieli profili. Zmiana została wykonana przy użyciu programu MS Word. W wyniku tej zmiany powstały 123 pliki w formacie MS Word zawierające te same informacje, ale pozbawione informacji bezpośrednio identyfikujących osoby, których te dokumenty dotyczą.

Równolegle z przygotowywaniem danych w postaci plików zawierających poszczególne profile zawodowe powstawała lista w formacie MS Excel zawierająca następujące kolumny:

NAME – nazwa pliku zawierającego dany profil,

Branza – nazwa branży zawodowej właściciela danego profilu określona przez autora badania na podstawie szczegółowych opisów doświadczeń zawodowych danej osoby przedstawionych na Rys.1:

Rys.1. Hasła opisujące doświadczenie zawodowe danej osoby

Analityk
Finanse
HR
Inżynier
IT
Marketing
Med
Moto
Ryzyko
Szkolenia
Telekomunikacja

Źródło: opracowanie własne.

Poziom_Stanowiska – podział na Manager, Specjalista,

Pocz_Kariery – rok rozpoczęcia kariery zawodowej,

Liczba_Rol – liczba ról zawodowych/stanowisk opisanych w profilu danej osoby,

Liczba_Firm – liczba firm, w których dana osoba pracowała.

Dalsza obróbka danych i ich analiza była przeprowadzana przy pomocy pakietu SAS Enterprise Miner 14.1 dostępnego w ramach licencji studenckiej.

Wczytanie i wstępna obróbka danych

Pliki zawierające profile zawodowe zostały wczytane do SAS przy użyciu węzła „Wczytanie profili” (standardowy węzeł „Kod SAS-owy”) zawierającego standardowe makro TMFILTER, jak to widać na Rys.2:

Rys. 2. Kod makra %tmfilter

```
libname tabela 'C:\Users\Adam\Documents\Statystyka\PracaDyplomowa\Tabela';

%tmfilter(dataset=tabela.profile, dir=C:\Users\Adam\Documents\Statystyka\PracaDyplomowa\Profile,
        ext=docx, host=localhost, numchars=32000, language=english);
```

Źródło: opracowanie własne.

Wcześniej została utworzona biblioteka TABELA odpowiadająca folderowi dyskowemu o tej samej nazwie, a pliki źródłowe zostały umieszczone w folderze dyskowym Profile.

W wyniku działania tego węzła została utworzona w bibliotece TABELA tablica SAS-owa PROFILE, zawierająca między innymi zmienne wykorzystywane do późniejszej analizy:

NAME – nazwa pliku źródłowego,

TEXT – zawartość tekstowa wczytanego pliku.

Tablica ta została wskazana jako *Źródło danych* do dalszych analiz.

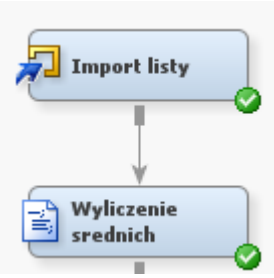
Lista w formacie Excel, o której mowa wcześniej, została wczytana do SAS za pomocą węzła „Import listy” (standardowa nazwa „Import pliku”). Uzyskana w ten sposób tabela SAS została uzupełniona w kolejnym węźle („Wyliczenie średnich”) o wyliczone pola za pomocą kodu SAS 4GL:

staz – różnica między bieżacym rokiem, a rokiem początku kariery,

lata_na_firme – średnia liczba lat przepracowana w jednej firmie,

lata_na_role – średnia liczba lat przepracowana w jednej roli (na jednym stanowisku).

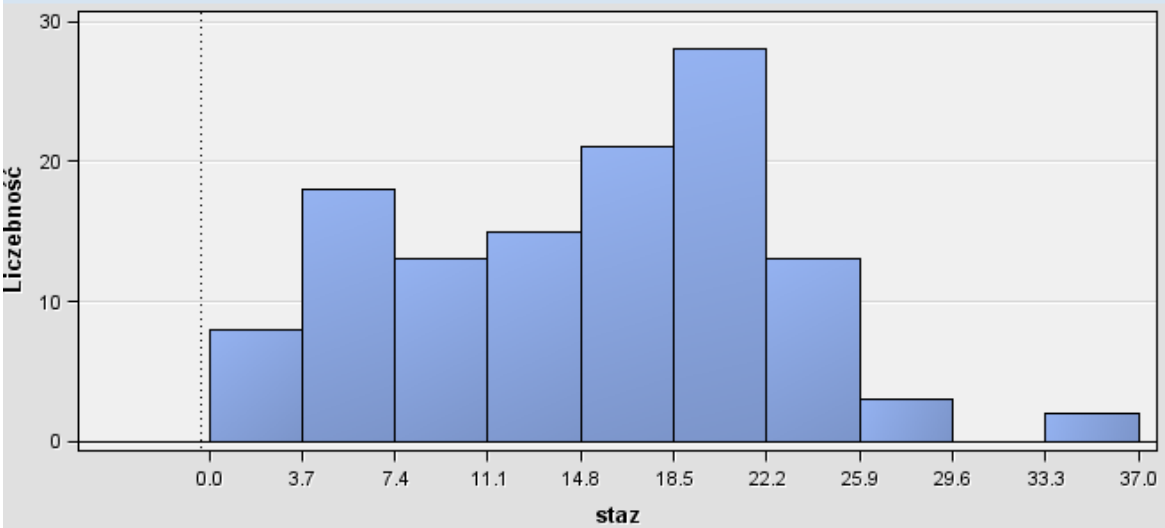
Rys. 3. Węzły „Import pliku” i „Wyliczenie średnich” w SAS Data Miner



Źródło: opracowanie własne.

Wstępna eksploracja uzyskanych w ten sposób danych pokazała, że najwięcej osób ma staż pracy zbliżony do 20 lat, co przedstawia Rys. 4:

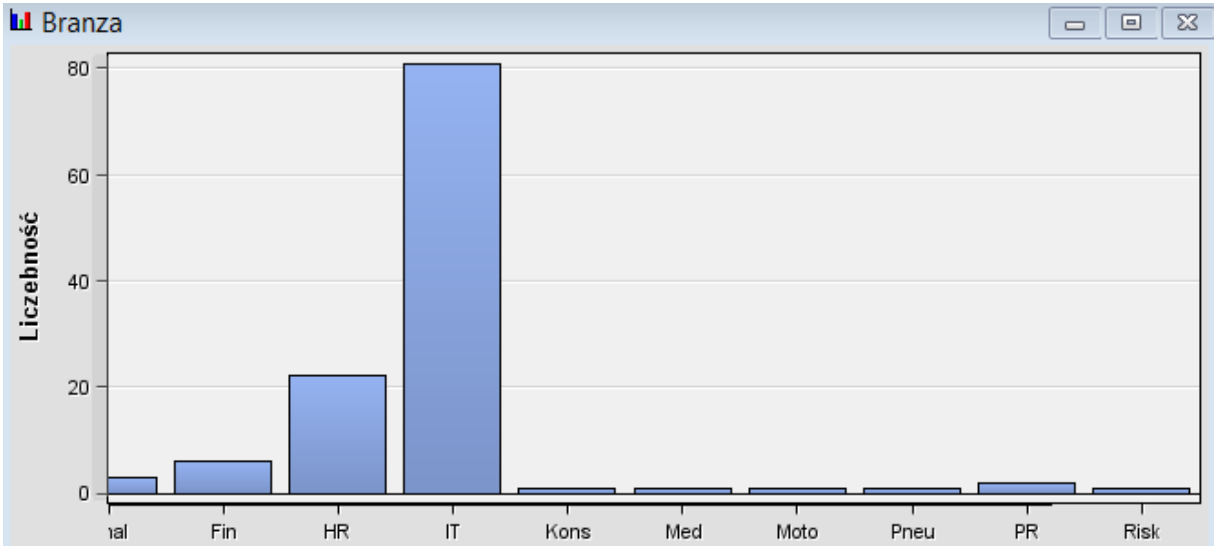
Rys. 4. Rozkład liczebności profili w zależności od długości stażu pracy



Źródło: opracowanie własne.

Najwięcej osób pracuje w branży IT, a w następnej kolejności: HR, finanse. Pozostałe branże reprezentowane są po 1 – 2 osoby jak to ilustruje Rys. 5. Według autora niniejszej pracy analizy w odniesieniu do tych branż nie będą raczej wiarygodne.

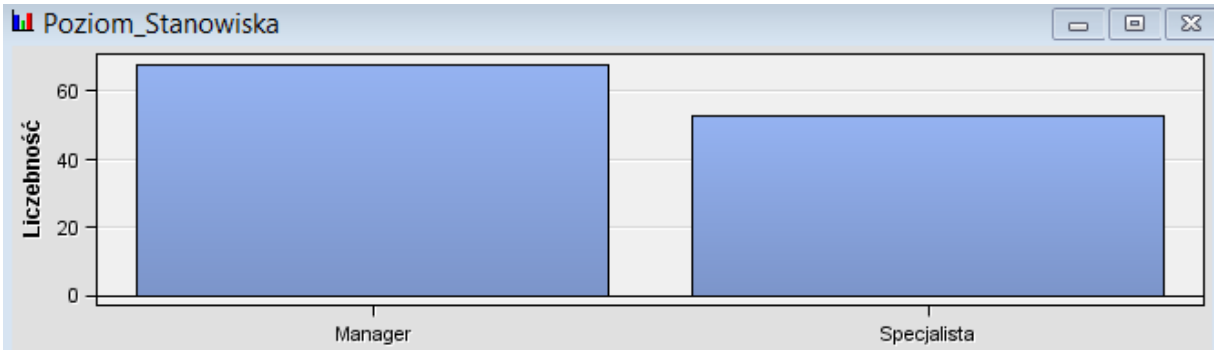
Rys. 5. Rozkład liczebności profili w zależności od branży



Źródło: opracowanie własne.

Nieco więcej jest osób w rolach managerskich, niż w rolach specjalistów, co obrazuje Rys. 6:

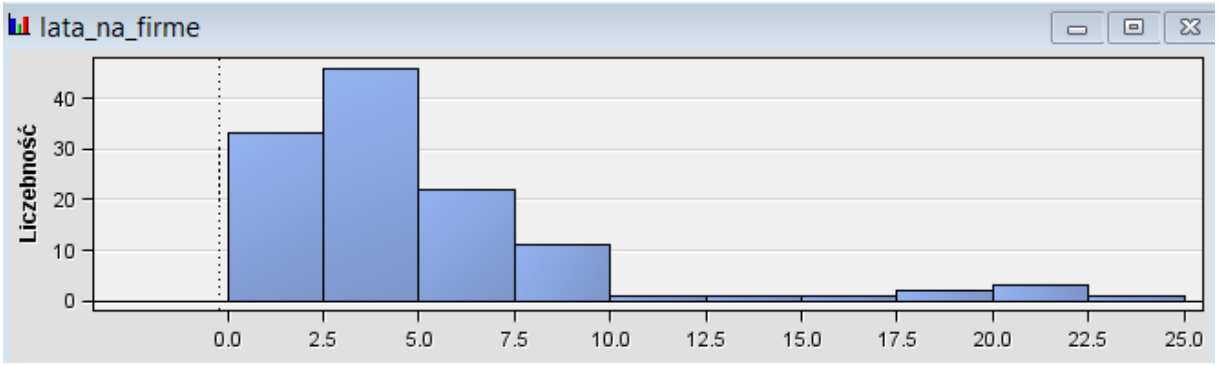
Rys. 6. Rozkład liczebności profili w zależności od poziomu stanowiska



Źródło: opracowanie własne.

Średni staż pracy w jednej firmie to najczęściej 3 – 4 lata, co obrazuje Rys. 7:

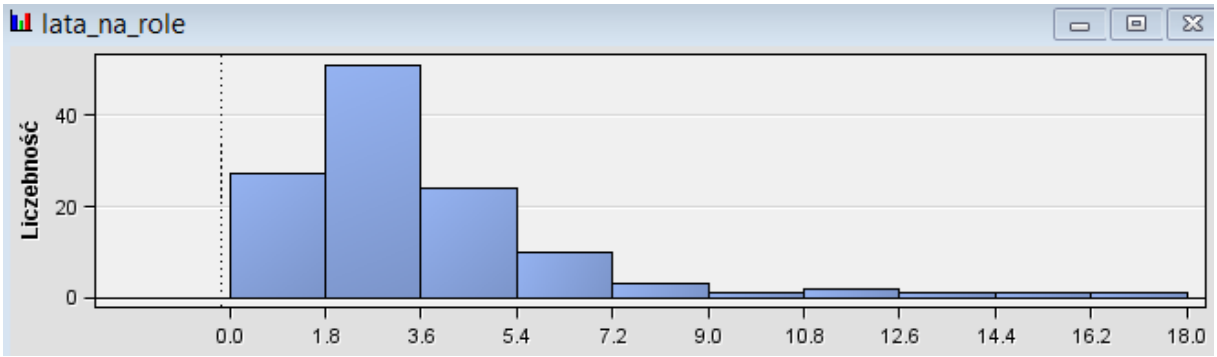
Rys. 7. Rozkład liczebności profili w zależności od średniego stażu pracy w firmie



Źródło: opracowanie własne.

Średni staż pracy w jednej roli/stanowisku to najczęściej 2 – 3 lata, co widać na Rys. 8:

Rys. 8. Rozkład liczebności profili w zależności od średniego stażu pracy w jednej roli

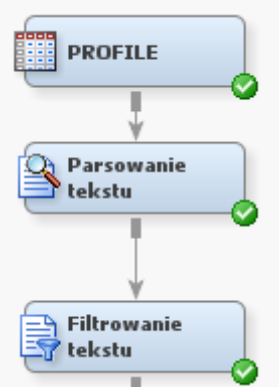


Źródło: opracowanie własne.

Parsowanie i filtrowanie danych tekstowych

Analiza tekstowa została przeprowadzona zgodnie z poniższym schematem węzłów:

Rys. 9. Tabela PROFILE i węzły „Parsowanie tekstu” i „Filtrowanie teksu” w SAS Data Miner

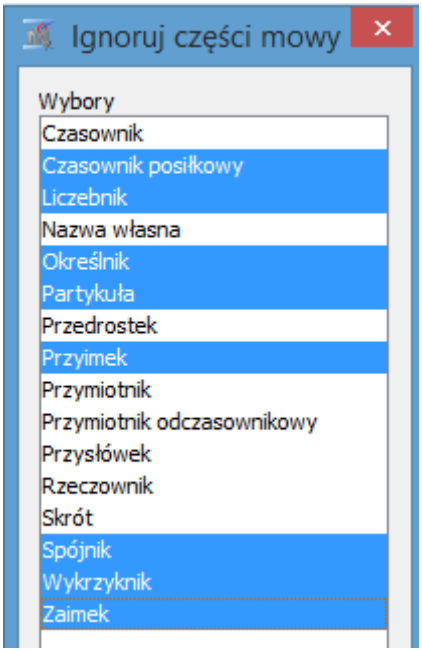


Źródło: opracowanie własne.

Dane wczytane wcześniej do tabeli PROFILE zostały poddane parsowaniu zgodnie z poniższymi założeniami:

1. Parsowaniu podlagała tylko zmienna TEXT.
2. Językiem parsowania był język angielski (chociaż w niektórych profilach wystąpiły też sporadycznie słowa w języku polskim (najczęściej nazwy własne)).
3. Zostały wskazane części mowy, które były ignorowane w procesie parsowania (Rys. 10).

Rys 10. Wskazanie części mowy, które będą ignorowane w procesie parsowania



Źródło: opracowanie własne.

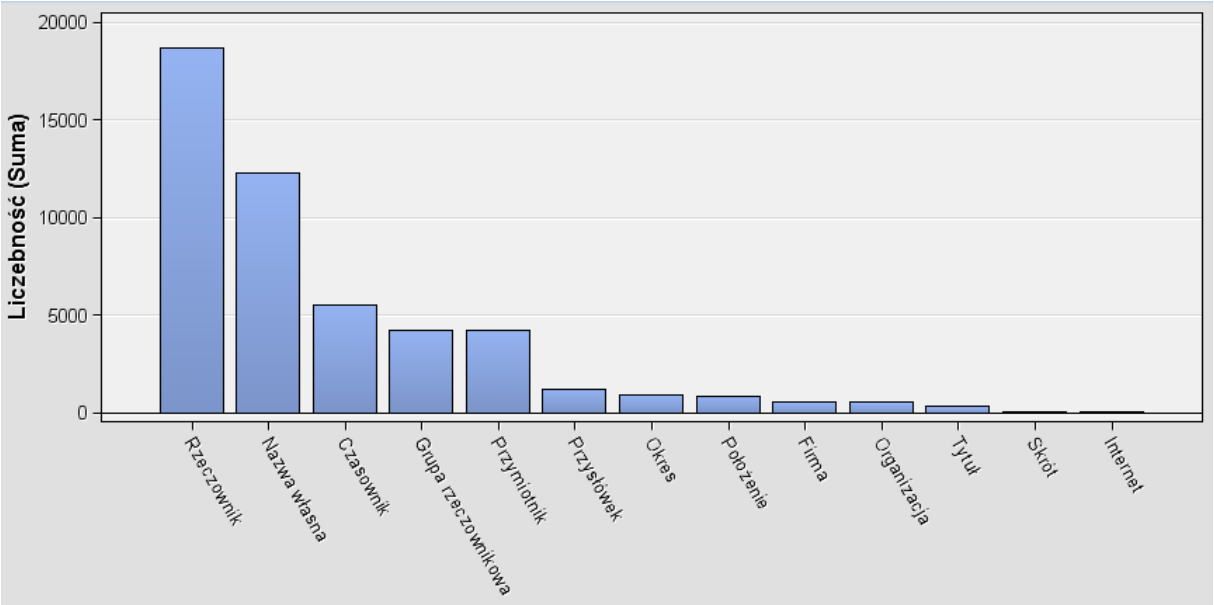
W związku z tym analizie podlegały: czasowniki, nazwy własne, przedrostki, przymiotniki, przysłówki, rzeczowniki, skróty.

4. W celu wyeliminowania ewentualnych pozostałych danych osobowych jako ignorowane typy obiektów specjalnych zaznaczono między innymi: *Inne nazwy własne, Osoba, PESEL, Telefon*. Również obiekty liczbowe (*Data, Czas, Miara, Procent, Waluta*) zostały wyłączone, ponieważ właściwe dane liczbowe (powiązane ze stażem pracy) zostały ustalone w badaniu w ramach osobnego procesu.

5. Inne ustawienia węzła (np. *Lista zatrzymań* i *Synonimy*) pozostawiono bez zmian w stosunku do wartości domyślnych.

W analizowanych dokumentach najwięcej jest rzeczowników, następnie – nazw własnych, czasowników, grup rzeczownikowych i przymiotników. Inne części mowy są już znacznie gorzej reprezentowane, co pokazuje Rys. 11.

Rys 11. Rozkład liczebności części mowy w badanych dokumentach



Źródło: opracowanie własne.

Najbardziej liczne wyrazy lub grupy wyrazów (10 pierwszych) pokazano w Tabeli 1:

Tabela 1. Liczebność wyrazów lub grup wyrazów w badanych dokumentach łącznie

Termin	Rola	Liczebność ▼	Atrybut
sap	Rzeczownik	859	Alfanumeryczne
+ be	Czasownik	672	Alfanumeryczne
management	Nazwa własna	552	Alfanumeryczne
+ year	Rzeczownik	534	Alfanumeryczne
business	Nazwa własna	413	Alfanumeryczne
+ month	Rzeczownik	400	Alfanumeryczne
+ work	Czasownik	380	Alfanumeryczne
+ business	Rzeczownik	362	Alfanumeryczne
o	Rzeczownik	329	Alfanumeryczne
+ project	Rzeczownik	279	Alfanumeryczne

Źródło: opracowanie własne.

Terminy występujące w największej liczbie dokumentów (10 pierwszych) pokazano w Tabeli 2.

Tabela 2. Liczba dokumentów, w jakich występują poszczególne wyrazy lub grupy wyrazów

Termin	Rola	L. dokumentów ▼	Atrybut
page1 Termin	... Nazwa własna	123	Mieszane
+ experience	... Rzeczownik	107	Alfanumeryczne
+ skill	... Rzeczownik	103	Alfanumeryczne
+ year	... Rzeczownik	102	Alfanumeryczne
management	... Nazwa własna	99	Alfanumeryczne
business	... Nazwa własna	93	Alfanumeryczne
expertise	... Nazwa własna	90	Alfanumeryczne
present	... Przymiotnik	90	Alfanumeryczne
+ month	... Rzeczownik	89	Alfanumeryczne
education	... Nazwa własna	83	Alfanumeryczne

Źródło: opracowanie własne.

Z czego część (np. „page1", „be", „o") nie niesie ze sobą istotnej treści z punktu widzenia opisu kariery, dlatego w kolejnym kroku powinny zostać wyeliminowane z analizy.

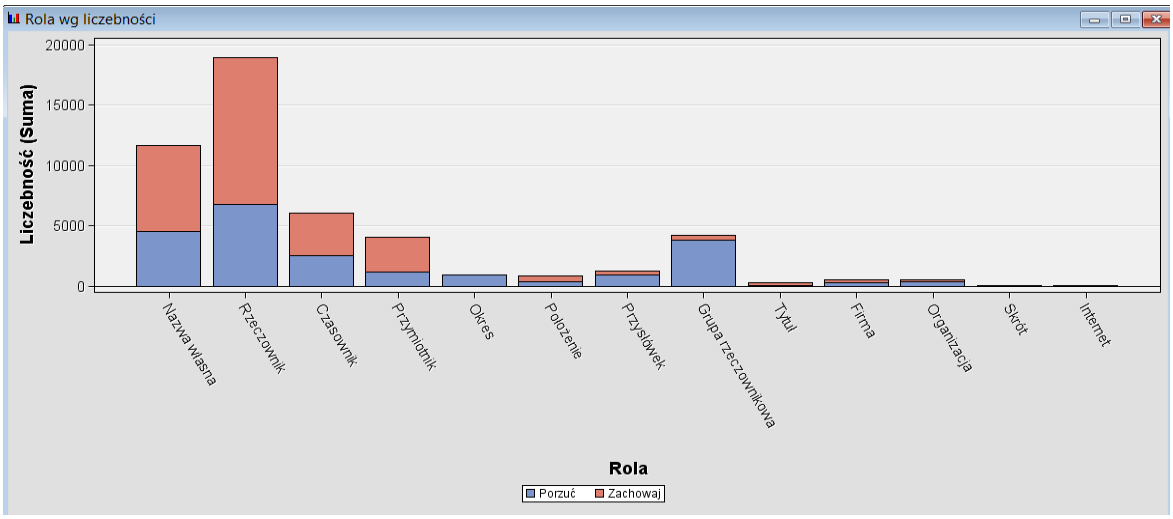
Kolejny węzeł diagramu („Filtrowanie tekstu") został uruchomiony w pierwszym momencie przy całkowicie standardowych ustawieniach. Okazało się jednak, że konieczne są ingerencje ręczne w działanie filtra, ponieważ nie zostały wyeliminowane wszystkie słowa, które w intencji autora miały zostać usunięte w tym kroku przebiegu procesu.

Dlatego została użyta opcja „Przeglądarka filtrów" dostępna z poziomu Właściwości węzła. Dokonano następujących ingerencji w standardowe działanie filtra:

- Usunięcie wyrazów lub ciągów wyrazowych odnoszących się do czasu (z powodów opisanych wyżej),
- Usunięcie pozostałych imion i nazwisk lub ich fragmentów,
- Połączenie tych samych wyrazów w grupy obejmujące synonimy (np. w niektórych przypadkach system potraktował ten sam wyraz jako rzeczownik pospolity, a innym razem jako nazwę własną – prawdopodobnie z powodu pisowni dużą literą),
- Usunięcie fragmentów polskich wyrazów, które były błędnie interpretowane ze względu na polskie znaki diaktryczne.

Po ponownym uruchomieniu węzła „Filtrowanie tekstów" uzyskano następujące rezultaty (Rys. 12):

Rys 12. Rozkład liczebności części mowy w podziale na wyrazy zachowane i odrzucone



Źródło: opracowanie własne.

Wśród zachowanych części mowy nadal największą grupę stanowią *Rzeczowniki*, następnie *Nazwy własne*, *Czasowniki* i *Przymiotniki*. Pozostałe części mowy stanowią teraz niewielką część zachowanego tekstu. W szczególności bardzo mocno zmniejszyła się *Grupa rzeczownikowa*, a wyrazy opisujące *Okres* zostały całkowicie wyeliminowane, co było zgodne z zamierzeniami autora.

Tabela analizowanych terminów potwierdza teraz, że nastąpiło odfiltrowanie wyrazów lub ich grup zgodnie z przyjętymi wcześniej założeniami. Poniżej pokazany jest tylko fragment tabeli:

Tabela 3. Filtrowanie wyrazów lub ich grup zgodnie z przyjętymi założeniami

Termin	Rola	Atrybut	Status	Waga	Liczebność importowanych ▼
sap ...	Rzeczownik	Alfanumery...	Zachowaj	0.236	859
+ be ...	Czasownik	Alfanumery...	Porzuć	0.000	672
+ year ...	Rzeczownik	Alfanumery...	Porzuć	0.000	534
business ...	Nazwa wła...	Alfanumery...	Zachowaj	0.120	413
+ month ...	Rzeczownik	Alfanumery...	Porzuć	0.000	400
+ work ...	Czasownik	Alfanumery...	Porzuć	0.000	380
+ business ...	Rzeczownik	Alfanumery...	Zachowaj	0.231	362
o ...	Rzeczownik	Alfanumery...	Porzuć	0.000	329
s ...	Rzeczownik	Alfanumery...	Porzuć	0.000	240
page1 ...	Nazwa wła...	Mieszane	Porzuć	0.000	227

Źródło: opracowanie własne.

Warto przy tej okazji zwrócić uwagę, że nie w każdym przypadku ma sens ręczne korygowanie automatycznych przypisań takiego samego wyrazu do różnych ról. Na przykład wyraz „business" powyżej występuje w profilach zawodowych zarówno jako zwykły rzeczownik odnoszący się do pracy/działalności, ale też jako część nazwy własnej (np. firmy lub działu). Dlatego potraktowanie tych pozycji jako synonimów byłoby błędem.

Klastrowanie

Dane tekstowe przygotowane w poprzednich krokach (opisanych powyżej) zostały poddane klastrowaniu:

- według algorytmu *maksymalizacji oczekiwań*,
- według algorytmu klastrowania *hierarchicznego*.

I w jednym i w drugim przypadku określono maksymalną liczbę skupień na 40 (wartość domyślna). Analizie została poddana tylko zmienna TEXT, ponieważ druga zmienna tekstowa (URL) nie niosła ze sobą treści merytorycznych. Pozostałe parametry określające działanie węzła pozostały bez zmiany.

W wyniku klastrowania algorytmem *maksymalizacji oczekiwań* uzyskano 5 skupień opisanych następującymi wyrazami (Tabela 4):

Tabela 4. Liczebności poszczególnych skupień

Id. skupienia	Terminy opisowe	Liczebność	Procent
1	enterprise pre-sales solution strategy sales software +architecture present +manager intelligence ...	30	24%
2	financial polish +report key +service +implementation +process +team integration +contract ...	23	18%
3	search recruitment +acquisition recruiting executive +contract consulting technical +consultant financial ...	21	17%
4	public +plan marketing +report key polish english +business +warsaw +implementation ...	24	19%
5	migration netweaver sap abap erp data +consultant integration business solution ...	28	22%

Źródło: opracowanie własne.

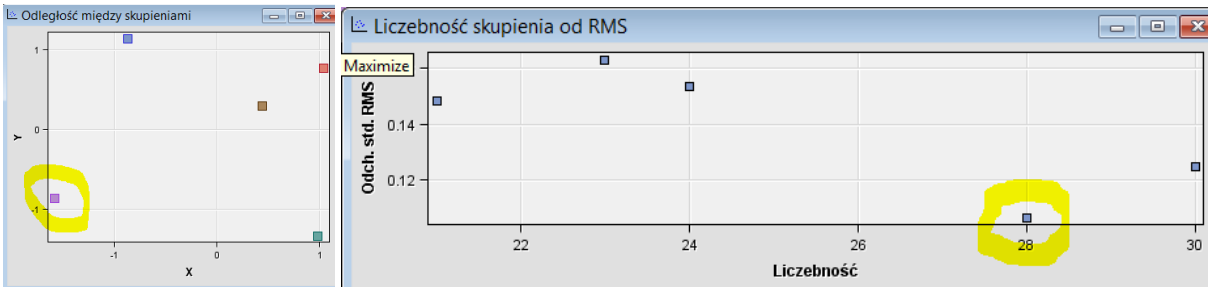
Skupienia są prawie równoliczne, co obrazuje wykres *Liczebności skupień* (Rys. 13).

Rys 13. Liczebności skupień

Źródło: opracowanie własne.

Wykresy *Odległość między skupieniami* i *Liczebność skupienia od RMS* wskazują łącznie, że najlepiej wyodrębnione i najbardziej spójne jest skupienie 5 (wyróżnione na wykresach poniżej).

Rys 14 i 15. Odległość między skupieniami i Liczebność skupienia od RMS



Źródło: opracowanie własne.

I rzeczywiście to skupienie najłatwiej zinterpretować – jako „Konsultant/deweloper SAP ERP". W przypadku pozostałych skupień interpretacja jest również możliwa, chociaż nie tak jednoznaczna:

Skupienie 1: Konsultant/architekt rozwiązań w obszarze oprogramowania dla przedsiębiorstw.

Skupienie 2: Osoba mająca do czynienia z finansami.

Skupienie 3: Rekruter.

Skupienie 4: Pracownik działu marketingu lub PR.

Klastrowanie algorytmem *hierarchicznym* dało w wyniku również 5 klastrów, ale nieco inaczej zdefiniowanych, co obrazuje Tabela 5.

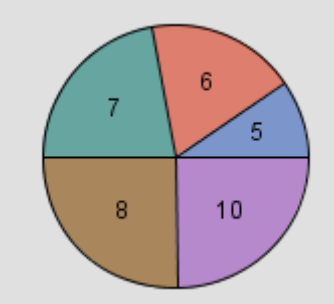
Tabela 5. Liczebności poszczególnych skupień

Id. skupienia	Terminy opisowe	Liczebność	Procent
5	support +warsaw systems financial university marketing analysis planning projects +master intelligence sap international business education	12	10%
6	strategy strategic service architecture technical +development +project management development systems integration project poland marketing key	23	18%
7	managed present key proficiency working +experience expertise best professional sales managing solutions team services software	28	22%
8	recruitment recruiting specialist financial international planning +company skills languages +team +master expertise education +business consulting	32	25%
10	sap netweaver migration erp abap data implementation business process solution integration requirements expertise intelligence analysis	31	25%

Źródło: opracowanie własne.

Nie ma dużych różnic w liczebności skupień (Rys. 16), ale jednak są one wyraźniejsze, niż w przypadku poprzedniej metody:

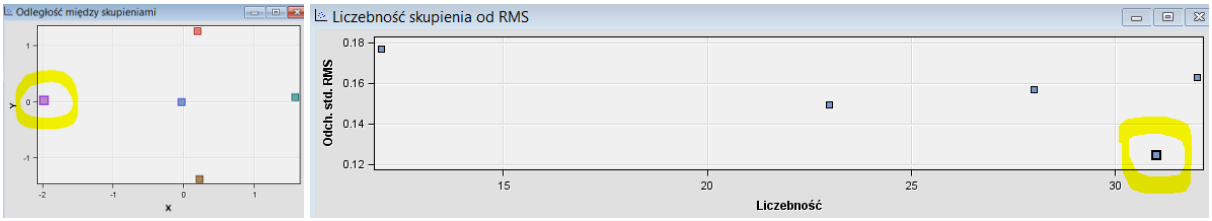
Rys 16. Liczebności skupień



Źródło: opracowanie własne.

Podobnie jak wcześniej, tutaj również wydaje się, że dosyć łatwo można zinterpretować jeden z klastrów (nr 10). Interpretacja może być zresztą podobna do wcześniejszej: „Konsultant/deweloper SAP ERP". To skupienie jest również najbardziej wyodrębnione i spójne, co potwierdzają poniższe wykresy (Rys. 17 i 19):

Rys 17 i 18. Odległość między skupieniami i Liczebność skupienia od RMS

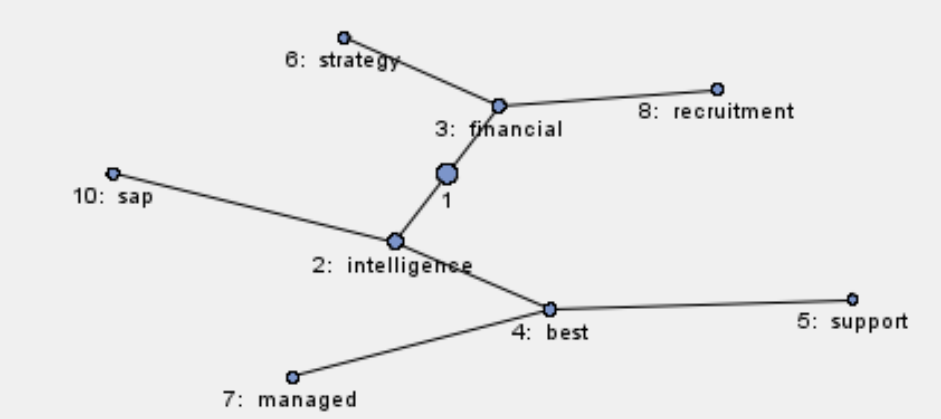


Źródło: opracowanie własne.

Klaster nr 8 można kojarzyć z rekrutacją (jak klaster numer 3 uzyskany wcześniejszym algorytmem).

Natomiast klastry 5, 6 i 7 są nieco trudniej interpretowalne. Dlatego w tym przypadku warto skorzystać z wykresu *Hierarchii skupień* (Rys. 19).

Rys 19. Hierarchia skupień



Źródło: opracowanie własne.

I tym razem interpretacja skupień nie jest jednoznaczna, ale wydaje się, że można:

- potwierdzić jednoznaczność interpretacji klastra nr 10 (jak wyżej),
- przyjąć, że węzeł nr 8 odnosi się do rekruterów,
- zinterpretować węzeł nr 6 jako dotyczący strategii lub zarządzania na poziomie strategicznym,
- potraktować węzły nr 5 i 7 jako odnoszące się bardziej do merytoryki, a w związku z tym:
- uznać węzeł nr 7 jako odnoszący się do zarządzania operacyjnego,
- uznać węzeł nr 5 jako węzeł zawierający opisy kompetencji specjalistycznych.

Niemniej jednak trzeba zauważyć, że taka interpretacja możliwa jest dzięki ogólnej interpretacji autora niniejszej pracy co do zawartości badanych profili zawodowych jego znajomych.

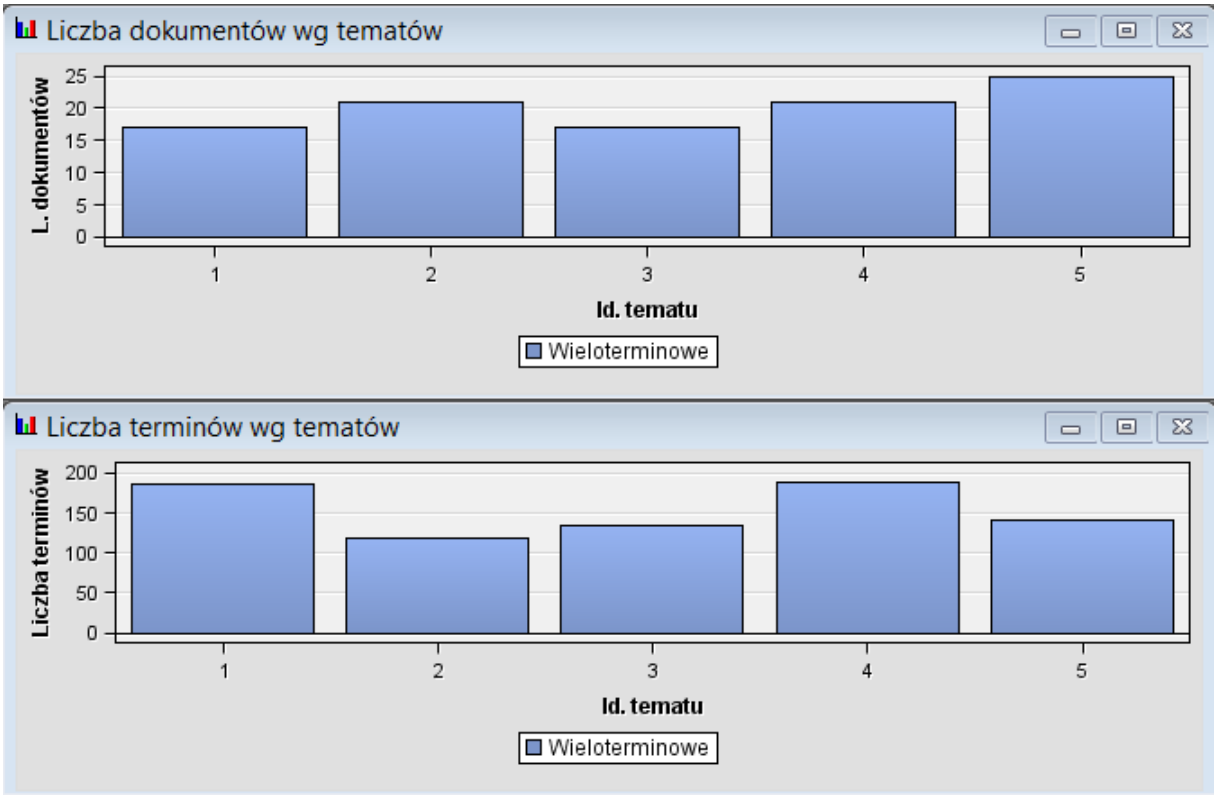
Przypisanie tematów tekstów

Równolegle do klastrowania zastosowany został węzeł Temat tekstu, który również pozwala na grupowanie badanych dokumentów poprzez przypisanie im tematów.

Żeby móc porównać wyniki działania tego węzła z klastrowaniem opisanym wyżej przyjęto początkowo tę samą liczbę tematów, jak liczba klastrów uzyskana w sposób automatyczny z klastrowania.

Do poszczególnych tematów system zakwalifikował zbliżoną liczbę dokumentów i zbliżoną liczbę terminów, co obrazują poniższe wykresy (Rys. 20 i 21):

Rys. 20 i 21. Liczba dokumentów wg tematów i Liczba terminów wg tematów



Źródło: opracowanie własne.

Natomiast definicje tematów tylko częściowo pokryły się z grupami wyznaczonymi poprzez klastrowanie (Tabela 6).

Tabela 6. Liczba terminów i liczba dokumentów dla poszczególnych tematów

Id. tematu	Temat	Liczba terminów	L. dokumentów
1	+great,+person,+well,excellent,+recommend	186	17
2	sap,abap,netweaver,data,+architect	118	21
3	recruitment,+candidate,recruitment,recruiting,hr	133	17
4	+accounting,+prepare,+tax,financial,financial	187	21
5	sales,+sale,marketing,marketing,channel	141	25

Źródło: opracowanie własne.

Temat nr 2 można utożsamiać z wcześniej wyznaczonymi klastrami nazwanymi w tej pracy „Konsultant/deweloper SAP ERP".

Temat nr 3 to prawdopodobnie ta sama lub zbliżona grupa dokumentów/profili zawodowych, do wcześniejszej odnoszącej się do rekrutacji.

Ale pozostałych tematów nie można już łatwo przypisać do wcześniej uzyskanych klastrów. Występują wyrazy wspólne, ale łączna interpretacja może już być inna:

- Temat nr 1 nie odnosi się do konkretnych umiejętności lub branż, ale do zestawu wyrażeń opisujących przymioty danej osoby lub jej doświadczeń (wspaniały, dobry, rekomendowany).
- Temat nr 4 skupia się wyraźnie na kwestiach dotyczących księgowości i finansów, więc najbliżej mu być może do definicja klastra nr 2 uzyskanego algorytmem *maksymalizacji oczekiwań*.
- Temat nr 5 opisuje zagadnienia związane ze sprzedażą i marketingiem, więc najbliżej mu być może do definicja klastra nr 4 uzyskanego algorytmem *maksymalizacji oczekiwań*.

Wydaje się, że w porównaniu do klastrowania przypisanie tematów dla badanych dokumentów pozwoliło na utworzenie bardziej jednoznacznych definicji grup. Wyjątkiem jest być może temat nr 1. Próby dla innej *Liczby tematów wieloterminowych* powodowały jedynie podział tematów na bardziej szczegółowe grupy, ale z zachowaniem logiki widocznej w przypadku pięciu tematów. W szczególności zawsze odrębnym tematem pozostawał temat nr 1. Poniżej przykład tematów dla wybranej liczby tematów równej 10:

Tabela 7. Liczba terminów i liczba dokumentów dla poszczególnych tematów

Id. tematu	Temat	Liczba terminów	L. dokumentów
1	+great,excellent,pleasure,+good,able	173	18
2	service,delivery,+architecture,enterprise,+project	118	19
3	recruitment,+candidate,recruitment,recruiting,search	115	17
4	+accounting,+tax,financial,financial,reporting	117	13
5	sales,channel,partner,+sale,selling	147	15
6	sap,sd,abap,netweaver,bw	94	22
7	si,polska,zarz,hr,em	129	10
8	pa,dziernik,bieg,sierpie,dwuj	144	15
9	marketing,+communication,marketing,pr,+market	164	14
10	+system,security,+network,energy,security	183	16

Źródło: opracowanie własne.

Przy zwiększaniu liczby tematów ujawniają się też przypadki niedoskonałości odfiltrowania niektórych wyrazów. W powyższym przykładzie tematy nr 7 i 8 pokazują, że zachowały się jeszcze w analizowanych danych wyrazy lub fragmenty wyrazów w języku polskim, które miały być odfiltrowane we wcześniejszych krokach. Daje to podstawę do ewentualnego dalszego czyszczenia danych. Konieczny wtedy byłby powrót do węzła „Filtrowanie tekstu".

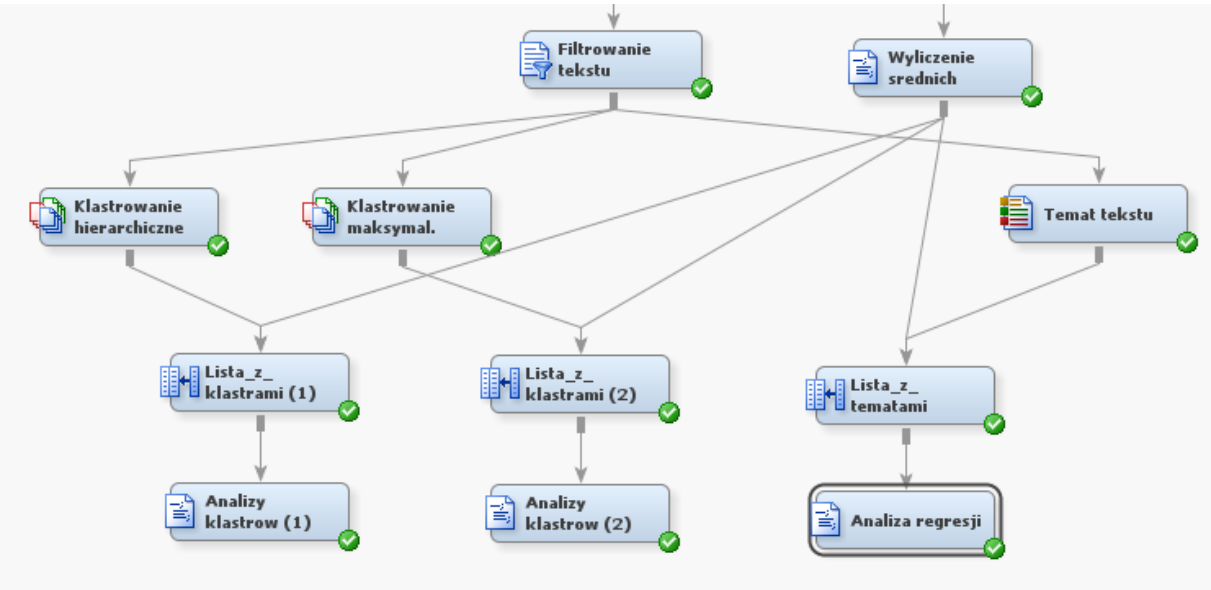
Łączenie zbiorów

W następnym kroku zostały przygotowane trzy zbiory danych będących wynikiem połączenia:

- wcześniej przygotowanego zbioru z listą badanych profili uzupełnionego o wyliczone średnie: staż pracy w jednej firmie, staż pracy na jednym stanowisku,
- każdego z trzech zbiorów będących rezultatem klastrowania lub wyznaczania tematów.

Następnie każdy z trzech uzyskanych w ten sposób zbiorów został poddany analizom korelacji lub regresji. Schemat przepływu danych obrazuje poniższy schemat węzłów Text Minera (Rys. 22):

Rys 22. Schemat węzłów Text Miner



Źródło: opracowanie własne.

Przy czym węzły „Lista_z_…" to standardowe węzły *Scalanie* (*Merge*), natomiast węzły „Analizy klastrów…" lub „Analiza regresji" to węzły zawierające kod SAS 4GL, który zostanie opisany poniżej.

Analizy współzależności danych

W ramach węzła „Analizy klastrów (1)" (odpowiada metodzie *hierarchicznej* klastrowania opisanej wyżej) została wygenerowana tablica częstości (FREQ) dla klastrów przypisanych przez program do poszczególnych dokumentów w powiązaniu ze stanowiskiem oraz z branżą.

Rys 23. Kod programu SAS 4GL w węźle „Analiza klastrów (1)"

```
proc freq data = &EM_IMPORT_DATA;
   tables Poziom_Stanowiska * TextCluster2_cluster_
          Branza * TextCluster2_cluster_ / chisq;
run;
```

Źródło: opracowanie własne.

Wyniki zaprezentowano na Rys. 23 i 24.

Rys. 23. Tablica częstości FREQ (stanowisko – klaster)

Tabela Poziom_Stanowiska od TextCluster2_cluster_

Poziom_Stanowiska(Poziom_Stanowiska)
TextCluster2_cluster_(Cluster ID)

Liczebność Procent Proc. wier. Proc. kol.	5	6	7	8	10	Razem
Manager	7 5.79 10.29 58.33	18 14.88 26.47 78.26	21 17.36 30.88 75.00	13 10.74 19.12 44.83	9 7.44 13.24 31.03	68 56.20
Specjalista	5 4.13 9.43 41.67	5 4.13 9.43 21.74	7 5.79 13.21 25.00	16 13.22 30.19 55.17	20 16.53 37.74 68.97	53 43.80
Razem	12 9.92	23 19.01	28 23.14	29 23.97	29 23.97	121 100.00

Liczebność braków danych = 5

Źródło: opracowanie własne.

Widać, że rozkład procentowy udziału poszczególnych klastrów dla poszczególnych poziomów stanowisk nie odpowiada rozkładowi łącznemu. Widać to szczególnie w przypadku klastra nr 10 (specjalista SAP), gdzie wśród specjalistów jest takich profili ok. 38%, a w całej analizowanej próbie ok. 24%. Ale również w przypadku pozostałych klastrów widać te niezgodności. Wyjątkiem jest klaster nr 5, gdzie proporcje udziałów są bardzo bliskie udziałom dla całej próby.

Statystyka Chi-Kwadrat potwierdza, że kolumny i wiersze tabeli nie są niezależne.

Rys. 24. Statystyki Chi-Kwadrat

Statystyki dla tabeli Poziom_Stanowiska od TextCluster2_cluster_

Statystyka	DF	Wartość	Prawd.
Chi-kwadrat	4	17.5745	0.0015

Źródło: opracowanie własne.

Natomiast kolejne wyniki (Rys. 25 i 26) potwierdzają zaobserwowany już wcześniej nierównomierny rozkład udziału poszczególnych branż w analizowanych danych. Niektóre branże są bardzo słabo reprezentowane (są to z reguły pojedyncze profile). Ale dla tych bardziej licznych można zaobserwować różnice.

Rys. 25 – część 1. Tablica częstości FREQ (branża – klaster)

Tabela Branza od TextCluster2_cluster_

Branza(Branza)	TextCluster2_cluster_(Cluster ID)					
Liczebność Procent Proc. wier. Proc. kol.	5	6	7	8	10	Razem
Analityk	1 0.83 33.33 8.33	0 0.00 0.00 0.00	2 1.65 66.67 7.14	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 2.48
Finanse	1 0.83 16.67 8.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 3.31 66.67 13.79	1 0.83 16.67 3.45	6 4.96
HR	3 2.48 13.64 25.00	0 0.00 0.00 0.00	5 4.13 22.73 17.86	12 9.92 54.55 41.38	2 1.65 9.09 6.90	22 18.18
IT	7 5.79 8.54 58.33	18 14.88 21.95 78.26	20 16.53 24.39 71.43	12 9.92 14.63 41.38	25 20.66 30.49 86.21	82 67.77

Rys. 25 – część 2. Tablica częstości FREQ (branża – klaster)

Razem	12 9.92	23 19.01	28 23.14	29 23.97	29 23.97	121 100.00

Liczebność braków danych = 5

Źródło: opracowanie własne.

I w tym przypadku statystyka Chi-Kwadrat (Rys. 26) potwierdza zależność pomiędzy zmiennymi, ale pojawia się też komunikat odnoszący się do małych ilości obserwacji

w niektórych grupach.

Rys. 26. Statystyka Chi-Kwadrat

Statystyki dla tabeli Branza od TextCluster2_cluster_

Statystyka	DF	Wartość	Prawd.
Chi-kwadrat	40	64.7222	0.0080
Chi-kw. ilorazu wiarygodn.	40	63.3667	0.0107
Chi-kwadrat Mantela-Haenszela	1	0.1276	0.7209
Współczynnik FI		0.7314	
Współczynnik kontyngencji		0.5903	
V Cramera		0.3657	

OSTRZEŻENIE: 85% komórek ma oczekiwane liczby wyst. mniejsze niż 5. Chi-kwadrat może nie być właściwym testem.

Źródło: opracowanie własne.

Klastrowanie metodą *maksymalizacji oczekiwań* daje większe możliwości dalszej analizy danych. W wyniku klastrowania uzyskuje się nie tylko przypisanie obserwacji (dokumentu) do jednego klastra, ale również liczbową reprezentację prawdopodobieństwa przypisania danej obserwacji do każdego z utworzonych klastrów. Pozwala to na bezpośrednie analizy ilościowe danych (korelacja, regresja).

W ramach niniejszej pracy wykonano następujące obliczenia za pomocą kodu SAS 4GL (Rys. 27, 28, 29):

Rys 27. Kod programu SAS 4GL w węźle „Analiza klastrów (2)" – część 1

```
proc freq data = &EM_IMPORT_DATA;
   tables Poziom_Stanowiska * TextCluster_cluster_
          Branza * TextCluster_cluster_ / chisq;
run;
```

Źródło: opracowanie własne.

Uzyskano analogiczne do wcześniejszych wyniki działania procedury FREQ. Statystyka Chi-Kwadrat potwierdziła istnienie zależności pomiędzy zmiennymi.

Analiza korelacji pomiędzy zmiennymi reprezentującymi prawdopodobieństwo przypisania do danego klastra (zmienne „TextCluster_prob…") pokazała, że wszystkie te zmienne są wzajemnie ze sobą skorelowane.

Rys 28. Kod programu SAS 4GL w węźle „Analiza klastrów (2)” – część 2

```
proc corr data = &EM_IMPORT_DATA;
   var TextCluster_probl TextCluster_prob2 TextCluster_prob3 TextCluster_prob4 TextCluster_prob5;
run;
```

Źródło: opracowanie własne.

W związku z tym należy spodziewać się, że analiza regresji, której zmiennymi objaśniającymi będą powyższe zmienne pokaże, że wystarczy jedna zmienna do zbudowania modelu.

I rzeczywiście uruchomienie modelu regresji zmiennej „lata_na_firme” objaśnianej przez powyższe zmienne pokazało, że model jest istotny statystycznie, ale tylko jedna ze zmiennych objaśniających w tym modelu jest istotna statystycznie. Co więcej program wykrył, że jedna ze zmiennych objaśniających jest kombinacją liniową pozostałych i ją wyeliminował z modelu.

Kod SAS 4GL (początkowy i po wybraniu zmiennej istotnej statystycznie) pokazuje Rys. 29:

Rys 29. Kod programu SAS 4GL w węźle „Analiza klastrów (2)” – część 3

```
proc reg data = &EM_IMPORT_DATA;
   model lata_na_firme = TextCluster_probl TextCluster_prob2 TextCluster_prob3 TextCluster_prob4 TextCluster_prob5;
run;

proc reg data = &EM_IMPORT_DATA;
   model lata_na_firme =  TextCluster_prob3;
run;
```

Źródło: opracowanie własne.

Uzyskane wyniki dla wybranej zmiennej zobrazowane są na Rys. 30.

Rys. 30. Wyniki regresji

Oceny parametrów

Zmienna	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	1	5.58347	0.43273	12.90	<.0001
TextCluster_prob3	1	-3.24698	1.04630	-3.10	0.0024

Źródło: opracowanie własne.

Ocena parametru przy zmiennej objaśniającej wskazuje na ujemną zależność pomiędzy czasem pracy w danej firmie, a prawdopodobieństwem przynależności do klastra nr 3 (profil osoby pracującej w obszarze związanym z rekrutacją pracowników).

Wykorzystując analogiczny sposób postępowania można uzyskać statystyczne zależności między prawdopodobieństwem przypisania danego profilu zawodowego do danego każdego klastra, a średnim czasem pracy w jednej firmie.

Do analogicznych wniosków prowadzi analiza zależności zmiennej „lata_na_role", opisującej średni czas pracy danej osoby w danej roli (na danym stanowisku).

Użyty kod SAS 4GL (początkowy i po wybraniu zmiennej istotnej statystycznie) został pokazany na Rys. 31:

Rys 31. Kod programu SAS 4GL w węźle „Analiza klastrów (2)" – część 4

```
proc reg data = &EM_IMPORT_DATA;
   model lata_na_role = TextCluster_prob1 TextCluster_prob2 TextCluster_prob3 TextCluster_prob4 TextCluster_prob5;
run;
```

```
proc reg data = &EM_IMPORT_DATA;
   model lata_na_role =  TextCluster_prob3;
run;
```

Źródło: opracowanie własne.

Wyniki obrazują Rysunki 32 i 33:

Rys 32. Wyniki analizy regresji – część 1

```
Procedura REG
Model: MODEL1
Zmienna zależna: lata_na_role

Wczytano obserwacji                       126
Użyto obserwacji                          121
Liczba obserwacji z brakami danych          5
```

Źródło: opracowanie własne.

Rys 33. Wyniki analizy regresji – część 2

Analiza wariancji

Źródło	DF	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	1	59.29840	59.29840	7.32	0.0078
Błąd	119	963.66090	8.09799		
Razem skorygowane	120	1022.95929			

Pierw. z MSE	2.84570	R-kwadrat	0.0580
Średnia zależna	3.54137	Skor. R-kw.	0.0501
Wsp. zmienności	80.35585		

Oceny parametrów

Zmienna	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	1	3.86283	0.28467	13.57	<.0001
TextCluster_prob3	1	-1.86259	0.68831	-2.71	0.0078

Źródło: opracowanie własne.

Ocena parametru przy zmiennej objaśniającej wskazuje na ujemną zależność pomiędzy czasem pracy w jednej roli (na jednym stanowisku), a prawdopodobieństwem przynależności do klastra nr 3 (profil osoby pracującej w obszarze związanym z rekrutacją pracowników).

Podobną analizę można przeprowadzić dla tabeli uzyskanej w wyniku połączenia tabel z *tematami tekstów* przypisanych do poszczególnych dokumentów i innymi zmiennymi opisującymi te same dokumenty.

W strukturze tabeli wynikowej zawierają się zmienne zero-jedynkowe od „TextTopic_1” do „TextTopic_5” wskazujące na przypisanie danej obserwacji do danego tematu (od 1 do 5) oraz zmienne od „TextTopic_raw_1” do „TextTopic_raw_5” określające siłę przypisania danej obserwacji (profilu zawodowego) do danego tematu.

Przy pomocy kodu SAS 4GL zostały wykonane następujące analizy tych danych:

1. Model regresji liniowej zmiennej „lata_na_firme” objaśnianej zmiennymi od „TextTopic_raw_1” do „TextTopic_raw_5”. Model okazał się istotny statystycznie, ale tylko zmienne objaśniające „TextTopic_raw_4” i „TextTopic_raw_5” były w tym modelu istotne statystycznie, a zmienna „TextTopic_raw_2” była na granicy istotności.

2. Model regresji liniowej zmiennej „lata_na_firme” objaśnianej zmiennymi „TextTopic_raw_2”, „TextTopic_raw_4” i „TextTopic_raw_5”. Wyniki działania tego modelu potwierdziły nieistotność statystyczną zmiennej „TextTopic_raw_2.

3. Model regresji liniowej zmiennej „lata_na_firme” objaśnianej zmiennymi „TextTopic_raw_4” i „TextTopic_raw_5”. Model okazał się istotny statystycznie, a oszacowane parametry przy zmiennych objaśniających wskazują na ujemną zależność pomiędzy tymi zmiennymi, a objaśnianą zmienną „lata_na_firme”. Ta zależność jest silniejsza w przypadku profilu zawodowego związanego z rekrutacją, a nieco słabsza (ale też istotna) dla profilu dotyczącego sprzedaży i marketingu (Rys. 34).

Rys 34. Wyniki analizy regresji

Analiza wariancji

Źródło	DF	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	2	442.63582	221.31791	13.29	<.0001
Błąd	118	1964.31614	16.64675		
Razem skorygowane	120	2406.95196			

Oceny parametrów

Zmienna	Etykieta	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	Intercept	1	7.16839	0.59418	12.06	<.0001
TextTopic_raw3	recruitment,+candidate, recruitment,recruiting,hr	1	-14.89025	3.18478	-4.68	<.0001
TextTopic_raw5	sales,+sale,marketing, marketing,channel	1	-8.98558	3.42645	-2.62	0.0099

Źródło: opracowanie własne.

4. Analogiczne kroki do powyższych przeprowadzono dla zmiennej „lata_na_firme” objaśnianej zmiennymi zero-jedynkowymi od „TextTopic_1” do „TextTopic_5”. Uzyskano istotny statystycznie model, w którym zmiennymi objaśniającymi istotnymi statystycznie są „TextTopic_3” i „TextTopic_5”. Oszacowane parametry przy zmiennych objaśniających wskazują na ujemną zależność pomiędzy tymi zmiennymi, a objaśnianą zmienną „lata_na_firme”. I tu również zależność jest silniejsza w przypadku profilu zawodowego związanego z rekrutacją, a nieco słabsza (ale też istotna) dla profilu dotyczącego sprzedaży i marketingu (Rys. 35).

   Rys 35. Wyniki analizy regresji

Źródło	DF	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	2	328.52078	164.26039	9.33	0.0002
Błąd	118	2078.43119	17.61382		
Razem skorygowane	120	2406.95196			

Pierw. z MSE	4.19688	R-kwadrat	0.1365
Średnia zależna	5.02309	Skor. R-kw.	0.1219
Wsp. zmienności	83.55185		

Oceny parametrów

Zmienna	Etykieta	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	Intercept	1	6.14822	0.46659	13.18	<.0001
TextTopic_3	_1_0_recruitment,+candidate, recruitment,recruiting,hr	1	-4.04844	1.11009	-3.65	0.0004
TextTopic_5	_1_0_sales,+sale,marketing, marketing,channel	1	-2.69271	0.95279	-2.83	0.0055

Źródło: opracowanie własne.

5. Analogiczne kroki do powyższych przeprowadzono dla zmiennej „lata_na_role” objaśnianej zmiennymi od „TextTopic_raw1” do „TextTopic_raw5”. Uzyskano istotny statystycznie model, w którym zmiennymi objaśniającymi istotnymi statystycznie są wszystkie zmienne poza „TextTopic_raw1”.

   Oszacowane parametry przy zmiennych objaśniających wskazują na ujemną zależność pomiędzy tymi zmiennymi, a objaśnianą zmienną „lata_na_role”. Zależność jest dużo silniejsza w przypadku profilu zawodowego związanego z rekrutacją, a nieco słabsza (ale też istotna) w przypadku pozostałych profili (Rys. 36).

Rys 36. Wyniki analizy regresji

Źródło	DF	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	4	274.63232	68.65808	10.64	<.0001
Błąd	116	748.32698	6.45109		
Razem skorygowane	120	1022.95929			

Pierw. z MSE	2.53990	R-kwadrat	0.2685
Średnia zależna	3.54137	Skor. R-kw.	0.2432
Wsp. zmienności	71.72087		

Zmienna	Etykieta	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	Intercept	1	6.82179	0.70755	9.64	<.0001
TextTopic_raw2	sap,abap,netweaver,data,+ architect	1	-5.80762	2.57113	-2.26	0.0258
TextTopic_raw3	recruitment,+candidate, recruitment,recruiting,hr	1	-11.48241	2.07580	-5.53	<.0001
TextTopic_raw4	+accounting,+prepare,+tax, financial,financial	1	-6.58931	2.31861	-2.84	0.0053
TextTopic_raw5	sales,+sale,marketing, marketing,channel	1	-7.74360	2.21124	-3.50	0.0007

Źródło: opracowanie własne.

6. Bardzo podobne wyniki uzyskano dla modelu regresji zmiennej „lata_na_role" objaśnianej zmiennymi zero-jedynkowymi od „TextTopic_1" do „TextTopic_5". Uzyskano istotny statystycznie model, w którym zmiennymi objaśniającymi istotnymi statystycznie są wszystkie zmienne poza „TextTopic_1". Podobnie jak wcześniej najsilniejsza negatywna zależność wartości zmiennej „lata_na_role" ma miejsce dla zmiennej objaśniającej powiązanej z profilem rekrutacyjnym, a nieco mniejsza (ale też ujemna i istotna statystycznie) w przypadku pozostałych zmiennych objaśniających (Rys. 37).

Rys 36. Wyniki analizy regresji

Źródło	DF	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	4	212.90467	53.22617	7.62	<.0001
Błąd	116	810.05462	6.98323		
Razem skorygowane	120	1022.95929			

Pierw. z MSE	2.64258	R-kwadrat	0.2081
Średnia zależna	3.54137	Skor. R-kw.	0.1808
Wsp. zmienności	74.62030		

Oceny parametrów

Zmienna	Etykieta	DF	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	Intercept	1	5.01261	0.37283	13.44	<.0001
TextTopic_2	_1_0_sap,abap,netweaver,data,+ architect	1	-1.60329	0.67460	-2.38	0.0191
TextTopic_3	_1_0_recruitment,+candidate, recruitment,recruiting,hr	1	-3.05861	0.71068	-4.30	<.0001
TextTopic_4	_1_0_+accounting,+prepare,+ tax,financial,financial	1	-1.99086	0.65814	-3.02	0.0031
TextTopic_5	_1_0_sales,+sale,marketing, marketing,channel	1	-2.02189	0.62179	-3.25	0.0015

Źródło: opracowanie własne.

Zakończenie

Z całą pewnością na przebieg kariery zawodowej danej osoby wpływa szereg czynników, które nie podlegały badaniu. Zapewne treść opisów na portalu typu LinkedIn nie decyduje o tym, jak łatwo dana osoba znajduje pracę. Z drugiej jednak strony czynniki wpływające na atrakcyjność danego kandydata na rynku pracy (wykształcenie, doświadczenie zawodowe, znajomość języków itp.) powinny znaleźć swoje odzwierciedlenie w profilu zawodowym tego kandydata (pomijając oczywiście sytuacje, kiedy dana osoba niedostatecznie rzetelnie opisze swój profil). Dlatego wydawało się być ciekawym sprawdzenie metodami Text Mining, czy te zależności są możliwe do wychwycenia i czy narzędziami analiz statystycznych można wyciągać wnioski dotyczące „wartości" danego profilu z punktu widzenia szukania pracy.

W niniejszej pracy podjęto próbę oceny możliwości powiązania parametrów mierzalnych opisujących przebieg kariery zawodowej (częstości zmiany pracodawcy opisanej zmienną „lata_na_firme" i częstości zmiany stanowisk/ról zawodowych opisanej zmienną „lata_na_role") z wynikami analizy tekstowej zawartości profili zawodowych udostępnionych na portalu LinkedIn.com. Wyniki testów statystycznych potwierdziły możliwość wyciągania wniosków kierunkowych na podstawie analizy Text Mining dokumentów.

Dla badanej próby profili zawodowych uzyskano wnioski mówiące o większej częstości zmian firm i stanowisk osób z branży rekrutacyjnej. Następną w kolejności grupą z punktu widzenia częstości zmian są osoby zajmujące się zawodowo sprzedażą lub marketingiem. Badanie potwierdziło możliwość wychwycenia takich zależności, jak i możliwość rozpoznania poszczególnych grup zawodowych poprzez analizę profili zawodowych metodami Text Mining.

Wydaje się, że zastosowany w niniejszej pracy sposób wnioskowania można rozszerzyć na inne źródła nieustrukturyzowanych informacji dotyczących przebiegu kariery zawodowej np. dokumentów CV lub danych pochodzących z innych portali zawodowych. Należy równocześnie zwrócić uwagę na niedoskonałości takiego podejścia, których źródłem jest sam sposób tworzenia profili zawodowych. Profile te tworzą poszczególne osoby, których one dotyczą, dlatego należy brać pod uwagę próby manipulowania danymi przez ich autorów. Analiza podatności zastosowanego wyżej podejścia na potencjalne manipulacje mogłaby stanowić ciekawy temat kolejnych badań.

Z drugiej z kolei strony warto zauważyć, że wyniki badań mogłyby być bardziej wiarygodne, gdyby możliwe było powiązanie dostępnych danych tekstowych (profili zawodowych) z danymi ustrukturyzowanymi, które są w dyspozycji osób zarządzających portalami typu LinkedIn. W tym przypadku nie byłoby konieczne żmudne wyliczanie parametrów liczbowych opisujących przebieg kariery zawodowej (długość kariery zawodowej, liczba stanowisk, liczba firm) na podstawie treści profili, ponieważ te dane z pewnością są na wprost dostępne w bazach danych administratorów profili. Tam też dostępne są z pewnością inne dane, które można korelować z treścią profili zawodowych.

Nie zapominając o powyższych uwagach wydaje się, że można na podstawie przykładowych analiz zastosowanych w ramach niniejszej pracy dojść do przekonania, że przyjęte podejście można uznać za wartościowe, możliwe do zastosowania i warte dalszych badań.

Bibliografia

Media Społecznościowe w Rekrutacji, Raport 2015, Lee Hecht Harrison DBM

2014 Global Social Recriuting Activity Report, Understanding Social Media Use In Recruiting, Bullhorn Reach, August 2014

How I Easily Got 25% More Views on My LinkedIn Profile, Joe Chernov, January 14, 2015, http://blog.hubspot.com/marketing/increase-linkedin-profile-views

Wizerunek w Internecie a szukanie pracy, Praca.pl, listopad 2011, http://www.praca.pl/centrum-prasowe/komunikaty-prasowe/wizerunek-w-internecie-a-szukanie-pracy_cp-666.html

Why You Should COMPLETE Your LinkedIn PROFILE, Andy Foote, March 27, 2013, http://www.linkedinsights.com/why-you-should-complete-your-linkedin-profile/

What section of a LinkedIn profile best represents a candidate's job function?, AYLIEN, April 29, 2015, http://blog.aylien.com/post/117696023083/what-section-of-a-linkedin-profile-best-represents

Automatic Analysis of Curriculum Vitae, a Case Study: the CV Distiller Software, F. Gire, S. Kolodziejczyk, www.koltech-group.com

Text Mining – Metody, narzędzia, zastosowania, Wykorzystanie SAS Text Analytics, D. Spinczyk, M. Dzieciątko, PWN 2016