http://dx.doi.org/10.18778/7969-549-2.11

Krzysztof Tomanek
Grzegorz Bryda
Uniwersytet Jagielloński

Odkrywanie wiedzy w wypowiedziach tekstowych Metoda budowy słownika klasyfikacyjnego

Streszczenie. Wykorzystywanie wiedzy o składni języka, semantyce i logice powiązań pomiędzy elementami wypowiedzi to atrakcyjny obszar w eksploracji danych oraz analizach tekstowych. Jak dotąd metody analizy i klasyfikacji tekstów nie zawsze wykorzystują oferowane we wspomnianych obszarach osiągnięcia. Celem artykułu jest pokazanie metody, która integruje rozwiązania zaczerpnięte z różnych obszarów wiedzy naukowej. Zadania, jakie stawiają przed sobą autorzy, to: (a) wykorzystanie wiedzy z zakresu: językoznawstwa, NLP, logiki, statystyki w celu budowy rzetelnego narzędzia analitycznego w środowisku CAQDAS; (b) zastosowanie przewidzianych w CAQDAS rozwiązań oraz dodanie do nich nowych technik wspierających budowę narzędzi klasyfikacyjnych; (c) ocena zastosowanego rozwiązania. Zdaniem autorów metoda budowy słowników analitycznych, której wynikiem ma być narzędzie dokonujące trafnej klasyfikacji, wymaga syntezy wielu rozwiązań. Z jednej strony konieczna jest znajomość podstaw języka wyszukiwania treści, z drugiej – owocne okazuje się wykorzystanie narzędzi zbudowanych przez językoznawców (Thesaurus, słowniki synonimów, słowników relacji leksykalnych), badaczy jakościowych (lista przymiotnikowa ACL) oraz metodologów (indeks podobieństwa, proces deduplikacji oparty na mechanizmie machine learning, miara trafności klasyfikacji). W ramach proponowanego podejścia autorzy opisują krok po kroku proces budowy słownika kategoryzacyjnego, akcentują pułapki i ważne decyzje, jakie w ramach tego procesu napotyka analityk.

Słowa kluczowe: Text Mining, CAQDAS, słownik klasyfikacyjny, słownik analityczny, analiza tematyczna, przetwarzanie języka naturalnego, NLP, Thesaurus, CRISP-DM, lista przymiotnikowa ACL, Słowosieć, odkrywanie wiedzy w danych tekstowych, KDT.

Wprowadzenie

Wśród badaczy spotkać można przekonanie, jakoby analiza danych jakościowych była procesem niejasnym (Lofland i in. 2009; Silverman 2007), kryła w sobie niejawne decyzje, którymi analityk nie chce podzielić się ze światem naukowym, z różnych zresztą powodów. W polskim piśmiennictwie poświęconym zagadnieniom CAQDAS i komputerowej analizie danych jakościowych (KADJ) spotykamy

stanowisko, które z kolei pomija momenty decyzyjne w analizie, a skupia się jedynie na opisie i prezentacji możliwości, jakie dają oprogramowania do analiz danych jakościowych (Niedbalski 2013; Brosz 2012). W obu przypadkach widoczny jest brak transparentności procesu analiz, jakich dokonuje się na danych jakościowych. Przez proces ten rozumiemy: uporządkowaną sekwencję działań, jakie wykonuje analityk w swojej pracy (są to zarówno: dylematy, jakie analityk spotyka opracowując dane jakościowe, ale też rozwiązań, jakie przyjmuje). Tymczasem w literaturze poświęconej KADJ widoczne jest oczekiwanie, aby proces analiz i metody stosowane w trakcie pracy z danymi jakościowymi były transparentne, ergo niosły ze sobą systematyzację praktyk analitycznych (Kordasiewicz, Haratyk 2013). Zdarzają się już artykuły pokazujące taki sposób prowadzenia analiz (Tomanek 2014a). Kontynuując tę praktykę, poświęcamy ten artykuł prezentacji metody pozwalającej na zbudowanie słownika analitycznego służącego klasyfikacji wypowiedzi. Metodę opiszemy, pokazując całościowy proces analityczny, a także opisując krok po kroku działania wymagające dodatkowych decyzji natury metodologicznej. Budowa słownika opisana będzie przy użyciu oprogramowania QDA Miner i Wordstat. Podkreślamy jednak, że sama metoda może być realizowana niezależnie od rozwiązania IT, po jakie sięgamy. Dlatego też wiedza, jaką przekazujemy, traktowana jest przez nas jako systematyczny wykład z zakresu metodologii analiz danych jakościowych[1].

Podstawy koncepcyjne

Podstawy dla analiz językowych, z których czerpiemy inspiracje, pochodzą przede wszystkim z dorobku: epistemologów, językoznawców oraz badaczy zjawisk społecznych i kulturowych.

Pierwsi zbudowali różnorodne modele rozumienia języka. Willard Van Orman Quine przekonywał, że nasze zrozumienie świata zależy od stopnia opanowania języka, którym ten świat opisujemy (Quine 1999: 29). Hilary Putnam akcentował związek języka ze światem realnym (Putnam 1990), a Charles Sanders Peirce zbudował model obrazujący związek pomiędzy: słowem–znaczeniem–przedmiotem (Peirce 1931: 35). Saul Kripke podkreśla, że związek języka ze światem jest *de facto* związkiem języka z wieloma możliwymi światami. Oznacza to, że jedno słowo (wypowiedź) za sprawą wielu możliwych interpretacji i odmiennych użyć może się odnosić do wielu obiektów zarówno tych istniejących w świecie realnym, jak i tych żyjących w wyobraźni autora wypowiedzi (Kripke 2001). Konsekwencji przywołanych

[1] Artykuł ten jest rozwinięciem sposobu analiz, jaki jeden z autorów przyjął we wcześniejszych pracach poświęconych analizie danych tekstowych z zastosowaniem słowników klasyfikacyjnych (por. Tomanek 2014).

refleksji dla naszej pracy będzie kilka. Są to: wrażliwość na wieloznaczność terminów; wiedza o wielu odniesieniach przedmiotowych, jakie mogą przysługiwać jednemu słowu; użycie ontologii w celu rozpoznania świata i obiektów tego świata w danym języku; zastosowanie modelu znaku dla rekonstrukcji typów wypowiedzi.

Od językoznawców czerpiemy wiedzę o: zasadach budowy wypowiedzi (składnia), treści i znaczeniu wyrazów (semantyka), zależności znaczeń od kontekstu (pragmatyka). Te trzy wybrane obszary językoznawstwa wykorzystamy, budując słownik klasyfikacyjny, w którym zdefiniujemy:

a) **reguły znaczeniowe** identyfikujące wypowiedzi podobne semantycznie (synonimia) i biegunowo różne znaczeniowo (antynomia),

b) **reguły pragmatyczne** pozwalające na określenie kontekstów występowania wypowiedzi, dzięki czemu możliwe staje się między innymi określenie przedmiotu lub przedmiotów wypowiedzi,

c) **reguły składniowe**, których zadaniem będzie określenie usytuowania wyrazów i fraz w analizowanej wypowiedzi.

W osiągnięciach nauk społecznych znajdujemy uzasadnienie dla kierunku analiz językowych, który obieramy. Pierre Bourdieu podkreśla, że sposób użycia języka oraz zawartość wypowiedzi językowych wskazywać mogą na typ kapitału językowego jednostki. Język dla francuskiego socjologa odzwierciedla nabyte dyspozycje do określonych sposobów i schematów myślowych oraz ocen (Bourdieu 2009). Basil Bernstein zwraca uwagę, że habitus możemy określić, odwołując się między innymi do stosowanego przez podmiot kodu językowego (Bernstein 1971). Natomiast Dan Sperber i Deirdre Wilson informują, w jaki sposób zidentyfikować w wypowiedzi pisanej: informacje, fakty, założenia i sądy, idee i emocje (Sperber, Wilson 2004).

Posługując się Rorty'ego ideą pragmatycznego traktowania metodologii jako skrzynki z narzędziami (Rorty 1996), wykorzystamy przywołane powyżej idee dla naszego celu. Jest nim zbudowanie narzędzia klasyfikującego wypowiedzi zgodnie z regułami, które uwzględniają osiągnięcia z różnych dziedzin wiedzy.

Ramy metodologiczne

Prezentowane w artykule podejście metodologiczne czerpie z kilku tradycji. Pierwsza z nich, łącząc standardy analiz jakościowych i ilościowych, sytuuje naszą metodę w paradygmacie Mixed Methods (MM). Zgodnie z użyteczną definicją Greene'a podejście to opiera się na (Greene 2006):

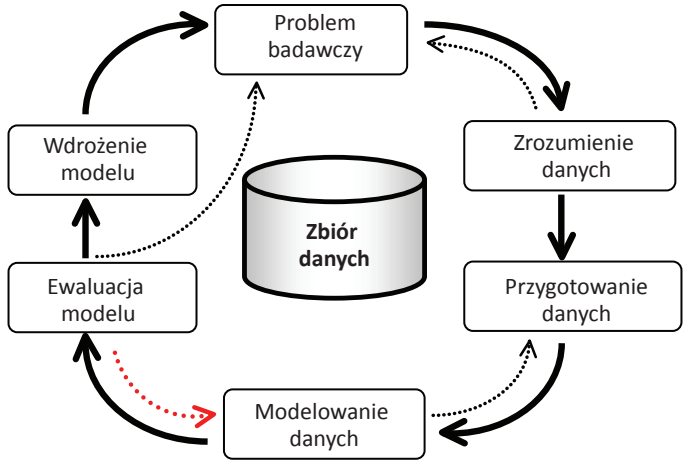
a) **założeniach filozoficznych i epistemologicznych**: pragmatyzm w podejściu do metody (Rorty 1996), realizm pragmatyczny w sposobie traktowania znaczenia (Putnam 1990), względność odniesień wypowiedzi (Quine 1999),

b) **logice postępowania naukowego**: zastosujemy dwa podejścia: bottom-up oraz logikę top-down, których celem jest odkrycie wiedzy w danych tekstowych (Skvoretz 1998),

c) **logice analizy danych**: analizy prowadzimy w środowisku CAQDAS (w większości przypadków stosując oprogramowanie QDA Miner i Wordstat) w sposób zgodny z wytycznymi zawartymi w metodologii CRISP-DM (Chapman, Clinton i in. 2000),

d) **proponowane rozwiązanie adresujemy do** socjologów, analityków posługujących się komputerowo wspomaganą analizą danych jakościowych (KADJ), analityków zainteresowanych metodami z zakresu Text Mining.

Wspomnianą powyżej metodologię projektową CRISP-DM stosujemy w dwojaki sposób. Po pierwsze służy nam ona jako schemat organizujący przebieg prowadzonych analiz. Po drugie metodologię tę modyfikujemy. Rozwijamy ją tak, aby pozwalała na iteracyjność procesu adaptacji i budowy narzędzia analitycznego, jakim jest słownik klasyfikacyjny. Oznacza to, że w trakcie budowy narzędzia analitycznego jego elementy składowe poddajemy ewaluacji i weryfikujemy poprawność ich działania. CRISP-DM stosujemy więc nie tylko w skali makro (jako metodologię dla całego projektu, jakim jest budowa słownika), lecz także w skali mikro (np. na etapie modelowania danych wielokrotnie poddajemy ocenie i jeśli zachodzi taka potrzeba, modyfikujemy takie elementy modelu, jak: definicje rdzeni, trafność klasyfikacji w oparciu o reguły syntaktyczne, poprawność budowy ontologii słownika). W efekcie tych działań z fazy ewaluacji modelu wracamy do etapu modelowania danych – działanie to obrazuje wykropkowana strzałka w lewej dolnej części ilustr. 1. W podejściu CRISP-DM akcentujemy potrzebę iteracyjnego działania podczas budowy słownika. Wizualizację procesów w ramach CRISP-DM pokazuje ilustr. 1.



Ilustr. 1. Model CRISP-DM z działaniami iteracyjnymi na poszczególnych etapach analiz

Źródło: opracowanie własne

W procesie analitycznym będziemy stosowali, poza wspomnianymi dotychczas, techniki analizy tekstu z obszaru Text Mining[2]. Wśród nich znajdą się takie techniki, jak: analiza tematyczna, POS (*part-of-speach tagging*) i inne (Hotho, Nürnberger, Paaß 2005).

Każdą ze stosowanych technik będziemy opisywali, wskazując jej rolę w procesie budowy słownika.

Budowa narzędzi analitycznych i proces analiz

W dalszej pracy zamierzamy osiągnąć dwa cele. Pierwszym jest cel metodologiczny: budowa słownika klasyfikacyjnego z zastosowaniem metodologii CRISP-DM. Jako przykład wykorzystamy zbiór wypowiedzi zarejestrowanych w systemie oceny zajęć uniwersyteckich USOS. Drugi cel ma wymiar merytoryczny: jest to wykorzystanie słownika analitycznego do rekonstrukcji schematów ocen wykładowców. Pokażemy próbkę takiej analizy, akcentując rolę słownika jako narzędzia klasyfikacyjnego do analizy tego rodzaju.

Zbiór danych, na którym pracujemy, zawiera 65 535 wypowiedzi zarejestrowanych w uniwersyteckim systemie USOS w latach 2008–2013. Opinie zawarte w bazie danych posiadają dwa atrybuty, które wykorzystamy w dalszych analizach. Są to:

a) typ zajęć,
b) rok rejestracji wypowiedzi.

Zgodnie z logiką CRSIP-DM pierwszym krokiem w toku analiz jest rozpoznanie jakości danych. Zaczniemy więc od diagnozy danych zawartych w naszym zbiorze.

Krok 1. Diagnostyka i preprocessing danych

Etap 1. Diagnostyka

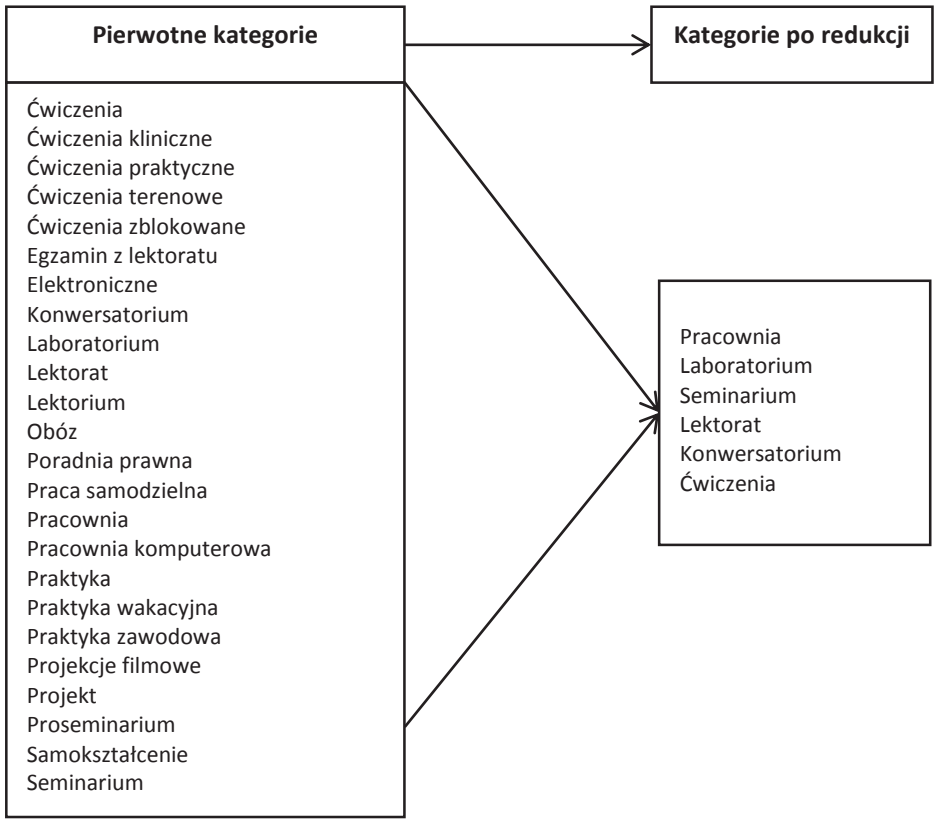
Proces przygotowania zbioru danych do analiz pokażemy, wykonując pracę na tekście, wykorzystamy także przywołane dwie zmienne opisujące wypowiedzi (typ wypowiedzi, rok rejestracji wypowiedzi). Zaczniemy od opisania operacji na zmiennych, a w następnym kroku pokażemy działania w ramach preprocessingu danych tekstowych.

Zarejestrowane w zbiorze wypowiedzi dotyczą 24 typów zajęć. Modyfikujemy dane wyjściowe, redukując przestrzeń własności tej zmiennej. Na potrzeby tej analizy stosujemy zasadę redukcji danych początkowych do najliczniej

[2] Szerzej na ten temat w: Bryda, Tomanek (2014).

występujących danych w ramach jednej kategorii. W ten sposób identyfikujemy 6 najliczniej ocenianych typów zajęć, które stanowią 96,6% zawartości zbioru[3]. Opisany zabieg został zaprezentowany w tab. 1.

Tabela 1. Redukcja kategorii w ramach zmiennej: typ zajęć



Źródło: opracowanie własne.

Rok rejestracji wypowiedzi jest zmienną, której wartości nadane są przez system USOS. Zmienna ta jest zdefiniowana przez wartości określone w ramach systemu USOS. Dla tej zmiennej mamy komplet danych (nie odnotowujemy braków danych).

Treści wypowiedzi to ciągi znaków od krótkich wypowiedzi (jedno- i dwuwyrazowych) do wypowiedzi zawierających 100 i więcej słów. Dalsze operacje wykonujemy na wszystkich tych tekstach.

[3] Wykluczone z analizy kategorie stanowiły przedmiot osobnych analiz. Na potrzeby prezentowanej tu metody pomijamy je.

Etap 2. Preprocessing danych

Dane, które poddajemy analizie, otrzymaliśmy w kilku plikach. Niezbędne jest więc przeprowadzenie ich **integracji**. Każdy ze zbiorów danych zawiera taką samą liczbę zmiennych. Są to: treść komentarza, kod zajęć (nazwany przez nas typem zajęć), kod cyklu dydaktycznego (nazwany przez nas rokiem rejestracji wypowiedzi). Dane w zbiorach są od siebie niezależne. Oznacza to, że żaden ze zbiorów nie posiada zmiennej jednoznacznie identyfikującej treść komentarza. A więc łączenie plików odbywa się zgodnie z zasadą: dodawania obserwacji do istniejącego zbioru z zachowaniem przyporządkowania danych do odpowiednich zmiennych. W wyniku integracji danych otrzymujemy jeden zbiór wypowiedzi. Posługujemy się kilkoma technikami pozwalającymi na ocenę jakości danych. Przeglądanie i czytanie próbek danych pokazuje, że analizowane teksty mają formę nieustrukturyzowanych wypowiedzi, które charakteryzują się:

a) nieprzestrzeganiem reguł ortograficznych,
b) niejednorodnym zapisem nazwisk, nazw własnych, liczebników,
c) niekonsekwencją w stosowaniu znaków interpunkcyjnych,
d) niejednokrotnie brakiem znaków diakrytycznych,
e) brakiem struktury wypowiedzi ze względu na poprawność stosowania znaków przystankowych.

Zidentyfikowane cechy wymagają zastosowania procedur opracowania tekstu. Celem takiego opracowania jest zbudowanie korpusu wypowiedzi zapisanych w sposób znormalizowany. Chcemy więc, aby słowa i znaki w wypowiedziach były zapisane w sposób jednorodny. Do tego celu wykorzystamy techniki czyszczenia danych.

Etap 3. Czyszczenie danych

Chcemy, aby analiza zwracała wszystkie poszukiwane przez nas informacje. W tym celu posłużymy się technikami **normalizacji zapisów wypowiedzi**:

a) ujednolicenie zapisów za pomocą małych i wielkich liter na rzecz jednorodnego zapisu, np. za pomocą małych liter. W ten sposób zapewniamy sobie, że wyszukiwanie tekstów zwróci wyniki bez względu na formę zapisu,
b) ujednolicenie zapisów dat: wszystkie zapisy dat sprowadzamy do jednej formy:
   – dla danych wystandaryzowanych przyjmujemy jako standard zapis DD-MM-YYYY,
   – dla danych nieustrukturyzowanych przyjmujemy jednorodny zapisy dni i miesięcy, sprowadzając je do jednej formy, którą jest pełna nazwa dnia, miesiąca,
c) zapisy liczb w formacie słownym i literowym sprowadzamy do formatu liczbowego.

Chcemy uwzględniać w analizie wypowiedzi istotne z punktu widzenia założonych celów (są to oceny wykładowców). W związku z tym redukujemy korpus tekstów, eliminując z niego: znaki, słowa, frazy nieistotne. Posługujemy się techniką **identyfikacji danych redundantnych**. Do tego celu stosujemy stop listę[4], która:

a) eliminuje nieistotne informacje, takie jak: spójniki, pojedyncze litery,
b) zawęża zbiór analizowanych danych,
c) przyspiesza proces analizy treści.

Analiza **częstości występowania wypowiedzi** podpowiada nam, jakie ciągi znaków nie zostały uwzględnione na stop liście. Rozbudowujemy zatem to narzędzie, dodając do niego wielokrotnie występujące w zbiorze znaki, które nie są istotne dla naszej analizy. Są to: wielokropek, spacja, znak „/", emotikony „☺", „☹", „;)", „:D".

Zakładamy, że zbiór może zawierać wypowiedź o tej samej treści zapisaną wielokrotnie (wypowiedź pochodzącą od jednego autora lub od różnych autorów). Dla tego celu posługujemy się techniką **identyfikacji duplikatów**. Wyszukiwanie wypowiedzi zduplikowanych daje następujący wynik:

a) nie wskazuje na powtórzenia w wypowiedziach od jednego autora (nie powtarza się wśród identyfikatorów wypowiedzi taki sam ciąg znaków – identyfikator wypowiedzi zawiera ciąg oznaczający respondenta + ciąg znaków oznaczający oceniany przedmiot/wykładowcę),

b) daje informację o zduplikowanych wypowiedziach pochodzących od wielu autorów. Takie zduplikowane wypowiedzi poddajemy dodatkowej analizie. Wśród nich znajdują się takie, które dla naszego celu merytorycznego są nieistotne (przykładowe zduplikowane wypowiedzi to: „OK!"; „ogólnie fajnie", „polecam tą Panią", „polecam", „1,5 godziny w tygodniu to zdecydowanie za mało!", „ankieta dotyczy osoby, która nie prowadziła tych zajęć") oraz takie, które dla budowy profilu wykładowcy w opinii studentów mają znaczenie (przykładowe duplikaty: „brak materiałów dydaktycznych", „absolutnie genialny", „bardzo ciekawa forma zajęć"). Te pierwsze usuwamy, drugie pozostawiamy do dalszych analiz.

Podsumowując podkreślmy, że **proces deduplikacji** składał się z trzech etapów: identyfikacji, analizy treści, usuwania zduplikowanych i nieistotnych dla dalszych analiz tekstów. Posłużenie się tą techniką dało nam zbiór mniejszy o 2176 wypowiedzi. Przeszliśmy zatem kolejny etap na drodze do przygotowania danych do analizy. Była to **identyfikacja danych relewantnych** dla naszego problemy badawczego.

Opisane powyżej zabiegi związane z przygotowaniem danych do analizy pozwalają nam przejść do kolejnego etapu pracy. Jest nim budowa słownika klasyfikacyjnego.

[4] Dodajmy, że stop lista jest narzędziem, którego stosowanie powinno być poprzedzone analizą jego zawartości. Edycja stop listy polega na: usuwaniu lub dodawaniu do niej znaków i wyrazów istotnych z punktu widzenia problemu, dla którego analiza jest realizowana. Jeśli na przykład wyszukujemy w wypowiedziach pytań, usuwamy ze stop listy znak zapytania. Jeśli szukamy w tekście emocji, stop lista nie powinna wykluczać z analiz wykrzyknika, ciągów znaków składających się na emotikony itp.

Krok 2. Budowa słownika klasyfikacyjnego

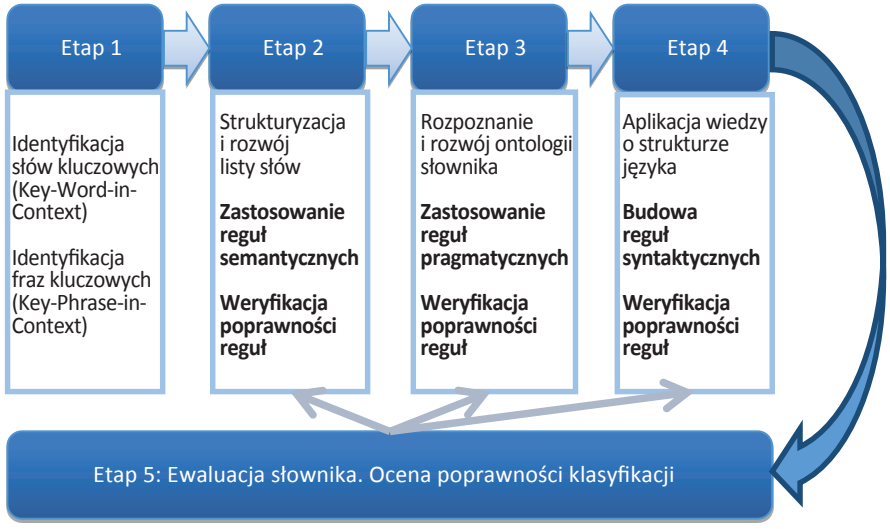
Istnieje kilka podejść stosowanych w budowie słowników analitycznych. Ogólnie można wskazać dwie podstawowe strategie:

a) podejście bottom-up, zgodnie z którym słowniki budowane są na bazie określonego korpusu języka lub korpusu tekstów. W tym przypadku stosuje się najczęściej logikę analizy frekwencyjnej,

b) podejście top-down to sytuacja, w której określony model koncepcyjny jest podstawą do budowy narzędzia klasyfikacyjnego. W tym przypadku stosowana jest logika kwalifikacji do analiz tych wypowiedzi, które spełniają kryterium istotności dla danego modelu koncepcyjnego. Ta strategia posługuje się także regułami leksykalnymi specyficznymi dla danego języka, uwzględniając je w konstruowaniu reguł klasyfikacji.

Nasze podejście mieści się w ramach MM. Oznacza to, że będziemy stosować obie wskazane logiki. Z jednej strony posłużymy się istniejącym modelem koncepcyjnym (jest nim słownik ACL), a z drugiej – wiele kategorii analitycznych wyodrębnimy z analizowanych wypowiedzi, „pozwalając danym mówić za siebie" (bottom-up). Szczegółowo zabiegi te opiszemy w kolejnych akapitach tekstu.

Proces budowy słownika pokazujemy, podkreślając konieczność ewaluacji osiągniętych wyników pracy oraz akcentując powtarzalność analiz, a więc konieczność działania iteracyjnego. Na ilustr. 2 zostały przedstawione kolejne etapy pracy nad słownikiem.



Ilustr. 2. Etapy pracy nad budową słownika klasyfikacyjnego

Źródło: opracowanie własne

Etap 1. Identyfikacja słów kluczowych (Key-Word-in-Context) i fraz kluczowych (Key-Phrase-in-Context)

Za cel merytoryczny stawiamy sobie rekonstrukcję profilu wykładowcy zawartego w ocenach studenckich. Ten cel osiągniemy, stosując metody nazywane syntetycznie *knowledge discovery in databse* (KDD). Za Fayyadem KDD rozumiemy jako: „Odkrywanie wiedzy w bazach danych (KDD) jest nietrywialnym procesem identyfikowania ważnych, nowych, potencjalnie przydatnych, a także zrozumiałych wzorców w danych" (Fayyad, Piatetsky-Shapiro, Smyth 1996).

Budowa profilu to w tym przypadku proces rekonstruowania z tekstu tych jego fragmentów, które są opiniami o wykładowcy. Zakładamy, że wypowiedzi oceniające będą wyrażane za pomocą określonych części mowy. Są to:

a) rzeczowniki abstrakcyjne: określają desygnaty nieuchwytne zmysłami – uczucia, spostrzeżenia, pojęcia, nazwy cech i czynności (np. tęsknota, nuda, zaangażowanie itp.),

b) rzeczowniki zmysłowe: określają zjawiska i przedmioty rozpoznawalne przez człowieka za pomocą zmysłów (np. egzamin, wykład, ćwiczenia itp.),

c) przymiotniki semantyczne: dookreślają rzeczowniki (np. wyczerpująca wypowiedź), nazywają cechy jakościowe lub relacyjne przedmiotów, osób, zjawisk, pojęć (np. uzależniony),

d) przymiotniki gramatyczne: powstające przez dodanie do podstawy słowotwórczej przedrostka (np. prze-, przed-) lub przyrostka (np. -owy, -cki),

e) czasowniki: część mowy określająca czynności, stany rzeczy (np. przekonuje, motywuje),

f) bezokoliczniki: forma czasownika wyrażająca czynności lub stany w sposób abstrakcyjny i w trybie dokonanym (np. zrobić) lub niedokonanym (np. robić),

g) liczebniki: część mowy określająca liczbę, ilość, liczebność, wielokrotność lub kolejność (np. pierwszy, dwukrotnie, podwójny).

Chcąc zidentyfikować wskazane powyżej elementy, rozpoczynamy kolejny etap analiz. Posłużymy się tu metodą identyfikacji istotnych informacji (*information extraction* – IE[5]). Zidentyfikowane fragmenty wypowiedzi będziemy kodowali, aby możliwa była ich szybka identyfikacja na dalszym etapie analiz. Zastosujemy do tego celu technikę *tagowania* części mowy. Metoda zwana *part of speech tagging* (POS) polega na identyfikacji i oznaczaniu (*tagowaniu*) w tekście poszukiwanych części wypowiedzi. Technikę tę możemy zastosować na dwa sposoby:

a) bottom-up: ta strategia nakazuje identyfikację wyrazów niezbędnych dla potrzeby analizy w wypowiedziach, nad którymi pracujemy. Jest to podejście bliskie teorii ugruntowanej. W przypadku naszej analizy unikamy interpretacji

[5] Szerzej o technikach odkrywania wiedzy w: Bryda, Tomanek (2014).

tekstu, wyszukując w nim słowa, które wprost określają cechy wykładowcy (wspomniane powyżej: rzeczowniki, przymiotniki, czasowniki, bezokoliczniki, liczebniki[6]),

b) top-down approach: ta strategia polega na zastosowaniu przygotowanej wcześniej listy słów. Taka lista referencyjna przydatna dla naszego celu powinna zawierać określenia cech osobowości. To podejście posługuje się więc istniejącą teorią, koncepcją, modelem teoretycznym, który wykorzystywany jest jako punkt wyjścia w analizie (w naszym przypadku jest to lista ACL).

Aby rozpocząć eksplorację danych ukierunkowaną na identyfikację treści istotnych dla problemu badawczego (budowa profilu wykładowcy), musimy podjąć jeszcze jedną decyzję. Bez względu na zastosowane podejście (*top-down* czy *bottom-up*) musimy wybrać kryterium kwalifikacji treści do analizy. Jednostkami analizy są w naszym przypadku: słowo, fraza. Do wyboru mamy kilka kryteriów:

a) kryterium minimalnej częstotliwości: to rozwiązanie mówi, ile razy słowo musi wystąpić, aby było zakwalifikowane do analizy,

b) kryterium pokrycia: to kryterium mówi, ile procent wystąpień musi mieć słowo w ramach całego korpusu słów, aby było kwalifikowane do analizy,

c) kryterium TF*IDF (TF – *term frequency*, IDF – *inverse document frequency*): ta propozycja opiera się na metodzie wskazującej wagę słowa w korpusie[7],

d) kryterium CTF-IDF (CTF – *category term frequency*, IDF – *inverse document frequency*): to kryterium odwołuje się do metody pozwalającej na wyodrębnienie często występujących słów w kategoriach i w konsekwencji odrzucanie słów o najmniejszym współczynniku wystąpień w wyróżnionych kategoriach[8],

e) kryterium WCTF-IDF (WCTF – *weighted category term frequency*, IDF – *inverse document frequency*): to podejście odwołuje się do metody ważenia słów w wyodrębnionych kategoriach[9].

---

[6] Dla tych potrzeb posługujemy się listami zawierającymi wskazane części mowy.

[7] Algorytm TFIDF definiowany jest w sposób następujący: $(TF\text{-}IDF)_{i,j} = TF_{i,j} * IDF_{i,j}$, gdzie: $TF_{i,j}$ to (*term frequency*) częstotliwość występowania słowa wyrażona wzorem: $TF_{i,j} = n_{ij}/\sum_k n_{k,j}$, gdzie $n_{i,j}$ jest liczbą wystąpień słowa T_i w dokumencie D_j, a mianownik jest sumą liczby wystąpień wszystkich słów w dokumencie d_j. IDF to (*inverse document frequency*) odwrócona częstotliwość w dokumentach wyrażona wzorem: $IDFi = \log |D|/|d \epsilon D : t \epsilon D|$. Wartość TF-IDF zwiększa się proporcjonalnie do tego, ile razy słowo pojawia się w dokumencie/wypowiedzi; miara ta jest kompensowana przez częstotliwość występowania słowa w korpusie dokumentów/wypowiedzi, co pomaga kontrolować fakt, że niektóre słowa są bardziej popularne niż inne (por. Ramos 2014).

[8] Algorytm CTF-IDF dla danego słowa *w* jest określony za pomocą wzoru: CTF-IDF(*w*) = liczebność(*w*)/$|T_w|$, gdzie liczebność (*w*) jest liczbą wystąpień słowa *w* we wszystkich analizowanych dokumentach/wypowiedziach. CTF-IDF jako słowa istotne oznacza te, które występują często, ale w małej liczbie dokumentów (por. Kobos, Mańdziuk 2008).

[9] Algorytm WCTF-IDF określony jest wzorem: $WC_{t,i} = WCTF_{t,i} * IDF_{t,i}$. $IDF_t = \log(a) * \log(b) * \log(c)$.

Na potrzeby naszej analizy posłużymy się dwoma kryteriami. TF*IDF powie nam, jakie słowa są specyficzne dla wszystkich wypowiedzi, a kryterium minimalnej częstotliwości pozwoli wykluczyć z analizy słowa mało liczne.

Ad a) Zestawienie 1 pokazuje częstotliwości występowania słów wraz z wartościami TF*IDF. Na liście zaznaczamy słowa zakwalifikowane do dalszych analiz.

Included | Leftover words | Unknown words

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
BARDZO	7749	2,4%	2,0%	1,4%	5742	39,5%	3127,0
ZAJĘCIA	7170	2,3%	1,8%	1,3%	5778	39,7%	2873,9
ZAJĘĆ	4241	1,3%	1,1%	0,8%	3407	23,4%	2672,8
STUDENTÓW	3491	1,1%	0,9%	0,6%	2955	20,3%	2415,9
BYŁY	2863	0,9%	0,7%	0,5%	2360	16,2%	2260,9
PANI	2813	0,9%	0,7%	0,5%	2138	14,7%	2342,1
BYŁO	2780	0,9%	0,7%	0,5%	2196	15,1%	2282,3
SPOSÓB	2271	0,7%	0,6%	0,4%	1994	13,7%	1959,6
PROWADZĄCY	2207	0,7%	0,6%	0,4%	1898	13,1%	1951,6
ZAJĘCIACH	2153	0,7%	0,5%	0,4%	1884	13,0%	1910,8
DR	1951	0,6%	0,5%	0,4%	1467	10,1%	1943,5
ZE	1737	0,5%	0,4%	0,3%	1480	10,2%	1723,7
PROWADZONE	1663	0,5%	0,4%	0,3%	1600	11,0%	1593,9
ĆWICZENIA	1543	0,5%	0,4%	0,3%	1342	9,2%	1596,8
TYLKO	1496	0,5%	0,4%	0,3%	1342	9,2%	1548,1
TEGO	1448	0,5%	0,4%	0,3%	1258	8,7%	1539,1
CIEKAWE	1384	0,4%	0,4%	0,3%	1352	9,3%	1427,8
WIEDZY	1367	0,4%	0,3%	0,3%	1227	8,4%	1467,8
ZAWSZE	1361	0,4%	0,3%	0,3%	1200	8,3%	1474,5
PROWADZĄCA	1360	0,4%	0,3%	0,3%	1197	8,2%	1474,9
PRACY	1321	0,4%	0,3%	0,2%	1159	8,0%	1451,1
KTÓRE	1277	0,4%	0,3%	0,2%	1130	7,8%	1416,9
DOKTOR	1236	0,4%	0,3%	0,2%	968	6,7%	1454,4

Zestawienie 1. Częstotliwość występowania słów w korpusie analizowanych tekstów

Źródło: opracowanie własne z zastosowaniem programu Wordstat

Powyższe zestawienie obejmuje 23 słowa występujące najczęściej w korpusie[10]. Niektóre z nich nie są istotne dla naszych analiz i zostają wyeliminowane. Słowa „prowadzący", „zajęcia" identyfikujemy i podkreślamy już teraz jako potencjalnie istotne na dalszym etapie budowy słownika (wrócimy do tego wątku na Etapie 3). Słowo „zawsze" zwraca naszą uwagę, ponieważ może wskazać na pewne tendencje lub stałe cechy czy elementy dotyczące przebiegu zajęć, sposobu zachowania wykładowcy. Aby przekonać się, czy ta intuicja analityczna jest trafna, sięgniemy na dalszym etapie analizy po dwie techniki pozwalające na poznanie kontekstu, w którym występuje słowo „zawsze" (*Key-Word-in-Context*

[10] Jest to zestawienie przykładowe dla top 23 słów. Analizie poddaliśmy 200 najczęściej występujących słów. Dla uproszczenia odwołujemy się tu tylko do wybranej grupy słów.

KWIC pozwoli nam zidentyfikować konteksty występowania słowa, a *Key-Phrase-in-Context* KPIC pokaże nam: frazy, w jakich pojawia się analizowane słowo oraz konteksty, w jakich frazy te występują).

Idąc tropem strategii *bottom-up*, identyfikujemy listę słów, które kwalifikujemy do dalszych analiz. Lista składa się z:

a) cech określających osobę, przedmiot, zjawisko (przymiotniki semantyczne, przymiotniki gramatyczne),

b) słów identyfikujących przedmioty abstrakcyjne i zmysłowe (rzeczowniki abstrakcyjne, rzeczowniki zmysłowe),

c) wyrazów opisujących czynności (czasowniki),

d) wyrazów wymagających dodatkowych analiz KWIC, KPIC (liczebniki, bezokoliczniki),

e) słów pozwalających na strukturyzację słownika klasyfikacyjnego (ćwiczenia, zajęcia, prowadzący, wykładowca).

Ad b) W strategii *top-down* wykorzystujemy istniejące w nauce rozwiązania. Są to istniejące w języku polskim listy: czasowników, bezokoliczników, rzeczowników. Narzędzie, które interesuje nas najbardziej, to inna lista. Mianowicie lista przymiotnikowa ACL (*Adjective Check List*) określająca wybrane cechy osobowości (Martowska 2012). Listę tę traktujemy jako punkt wyjścia do celów identyfikacji słów opisujących cechy osobowości w wypowiedziach. ACL stosowana jest w badaniach naukowych w analizie archetypów kulturowych, identyfikacji stereotypów. W naszym przypadku jest użytecznym słownikiem wychwytującym w analizowanych opiniach interesujące nas słowa.

ACL jest listą zawierającą cechy opisane przez słowa w formie podstawowej. Tak skonstruowana lista nie może nam posłużyć jako słownik wyszukujący. Trzeba poddać ją koniecznie kilku zabiegom. Po pierwsze musimy sprawić, aby słowa były odporne na zmianę:

a) czasu,

b) trybu,

c) liczby,

d) osoby i rodzaju.

W tym celu zastosujemy procedurę stemmingu, czyli wydobywania z wybranego wyrazu tzw. rdzenia, a więc tej jego części, która jest odporna na odmiany przez przyimki, rodzaje itp. Oto kilka przykładów zapisu rdzeni słów w formacie, w jakim zapis ten jest akceptowalny w programach QDA Miner i Wordstat (* zapis z gwiazdką oznacza, że po tym symbolu może wystąpić jakikolwiek ciąg znaków) (zestawienie 2):

a) asymetryczne zastosowanie rdzeniowania:
   - autokrat* dla np.: autokratyczna, autokrata,
   - kultural* dla np.: kulturalna, kulturalnie;

b) symetryczne zastosowanie rdzeniowania:
   – *ambit* dla np.: przeambitna, przeambitny, ambitnie,
   – *agresyw* dla np.: autoagresywny, autoagresywnie i *agresj* dla np.: autoagresja, agresji;
c) budowa kilku rdzeni dla słów podobnych znaczeniowo:
   – dla słów o podobnym znaczeniu, ale które nie posiadają wspólnego rdzenia, budujemy wiele rdzeni. Przykładem takiej sytuacji jest: rados* (radosna, radosnej – np. dla frazy takiej jak: w radosnej atmosferze), radoś* (radośnie, radość);
d) słowa w wersji bez rdzeniowania:
   – miła, miły: tworzenie rdzenia dla tych dwóch słów jest ryzykowne, ponieważ kwalifikowałoby wiele wyrazów, których możemy nie chcieć stosować w dalszych analizach (przykład rdzenia dla miła, miły to mił* – rdzeń ten klasyfikuje zatem takie słowa, jak: miły, miła, miłość, miłościwy, a także marka piwa „Miłosław").

Included | Leftover words | Unknown words

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
CIEKAW*	3011	21,8%	0,7%	0,5%	2713	18,6%	2201,5
PRZYJEMN*	897	6,5%	0,2%	0,2%	869	5,9%	1099,3
PRAKTYCZN*	869	6,3%	0,2%	0,2%	798	5,5%	1097,2
PEWN*	662	4,8%	0,2%	0,1%	603	4,1%	916,4
ZDECYDOWAN*	639	4,6%	0,2%	0,1%	592	4,1%	889,7
SYMPATYCZN*	638	4,6%	0,2%	0,1%	628	4,3%	871,9
PROST*	488	3,5%	0,1%	0,1%	461	3,2%	732,4
DOBRY	446	3,2%	0,1%	0,1%	436	3,0%	680,2
AKTYWN*	445	3,2%	0,1%	0,1%	389	2,7%	700,7
SAMODZIELN*	414	3,0%	0,1%	0,1%	390	2,7%	651,4
SZYBK*	298	2,2%	0,1%	0,1%	279	1,9%	512,3
PRZYJAZN*	288	2,1%	0,1%	0,1%	284	1,9%	492,8
OTWART*	281	2,0%	0,1%	0,1%	270	1,8%	487,0
POTRZEBN*	252	1,8%	0,1%	0,0%	242	1,7%	448,8
DOSKONA*	232	1,7%	0,1%	0,0%	221	1,5%	422,3
CIERPLIW*	216	1,6%	0,1%	0,0%	214	1,5%	396,2
DOBRA	158	1,1%	0,0%	0,0%	156	1,1%	311,5

Zestawienie 2. Częstotliwość występowania przymiotników z listy ACL w korpusie analizowanych tekstów

Źródło: opracowanie własne z zastosowaniem programu Wordstat

Lista przymiotnikowa ACL w wersji oryginalnej zawiera 300 słów. Do naszej analizy kwalifikujemy niektóre z nich. Podobnie jak wcześniej posługujemy się dwoma kryteriami. TF*IDF powie nam, które cechy z listy ACL są specyficzne dla wszystkich opinii, a kryterium minimalnej częstotliwości pozwoli wykluczyć z analizy słowa mało liczne.

Etap 2. Strukturyzacja i rozwój listy słów. Zastosowanie reguł semantycznych

Bez względu na wybraną strategię analiz (*bottom-up*, *top-down*) kolejny etap w procesie budowy słownika klasyfikacyjnego wiąże się z pytaniami o: (a) możliwości rozwoju słownika i (b) strukturę słownika, czyli sposób kategoryzacji słów.

Istnieje więcej niż jeden sposób rozwoju słowników klasyfikacyjnych. Oto kilka przykładowych metod poszerzania zawartości słowników:

a) w ramach podejścia *top-down*: dodanie nowych narzędzi analitycznych (np. lista czasowników odprzymiotnikowych, lista wulgaryzmów),

b) w ramach podejścia *bottom-up*: wykorzystanie Thesaurusa, a więc słownika wyrazów bliskoznacznych lub słów o podobnej budowie. Słownik ten podpowiada słowa, które występują w tekście, a które posiadają podobne znaczenie lub ten sam rdzeń co słowo kluczowe, które kwalifikujemy do analiz,

c) w ramach podejścia *mixed approach*: rozwój poszczególnych kategorii słownikowych. Rozwój ten może być realizowany na różne sposoby, np. poprzez dodanie wyrazów związanych z kluczowym dla kategorii słowem (Thesaurus), poprzez rozwój pola semantycznego dla słowa/słów w kategorii (nowe techniki).

W naszym przypadku sięgamy po rozwiązania opisane w punktach b) oraz c).

Ad b) Wiemy już, na czym polega dodanie listy referencyjnej (w naszym przypadku jest to lista ACL). Przejdźmy teraz do dwóch pozostałych metod. Aby zidentyfikować słowa podobne do tych, które zakwalifikowaliśmy do analizy, skorzystamy z Thesaurusa zaimplementowanego w programie Wordstat. Thesaurus posługuje się techniką wyszukiwania podobnych słów. W budowie słownika klasyfikacyjnego narzędzie, jakim jest Thesaurus, spełni kilka użytecznych ról.

B1. Po pierwsze weryfikuje trafność zdefiniowanych przez badacza rdzeni. A więc pokazując słowa podobne, pozwala badaczowi upewnić się, czy rdzeń, jaki samodzielnie zdefiniował w danym korpusie słów, na pewno kwalifikuje te słowa, na których badaczowi zależy (zestawienie 3).

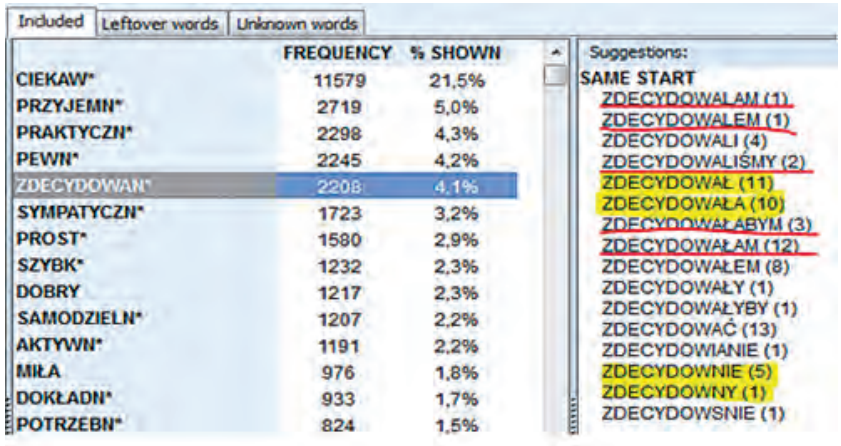
B2. Po drugie podpowiada słowa związane z tymi, które zdefiniował w analizie badacz. Tym samym Thesaurus pozwala na szybki rozwój kategorii analitycznych zdefiniowanych w analizie (zestawienie 4).

Poniżej podajemy przykłady zastosowania Thesaurusa i komentarz do zaprezentowanego zestawienia.

Ad B1. Thesaurus – narzędzie weryfikacji rdzeni. W zestawieniu 3 pokazujemy dwie sytuacje. Pierwsza to ta, w której Thesaurus pozwala na weryfikację niepoprawnie skonstruowanego rdzenia. Rdzeń zdecydow* kwalifikuje, jak widzimy, takie słowa, jak:

a) zdecydował, zdecydowała, zdecydowny (błędny zapis podpowiadający przypuszczalnie poprawne słowo: zdecydowany – to słowo jest potencjalnie użyteczne w naszej analizie i wymaga jedynie sprawdzenia kontekstu, w jakim występuje). Te słowa oznaczamy w zestawieniu kolorem żółtym,

b) zdecydowałam, zdecydowałabym, zdecydowalem, zdecydowalam – te słowa nie pasują jednak do celu, jaki postawiliśmy w projekcie. Słowa te identyfikują konteksty, w których podmiot wypowiada się o swoich działaniach, decyzjach, a nie o cechach wykładowców. Te słowa podkreślamy na czerwono.



Zestawienie 3. Wykorzystanie Thesaurusa wbudowanego w oprogramowaniu Wordstat jako narzędzia oceny poprawności definicji rdzeni

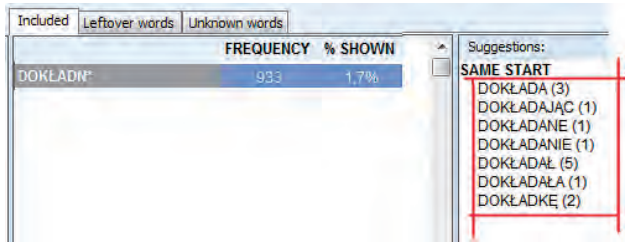
Źródło: opracowanie własne z zastosowaniem programu Wordstat

Jak widzimy, rdzeń zdecydow* wymaga poprawy. Ponieważ te słowa, na których nam zależy, nie mają wspólnego rdzenia, który jednoznacznie by je identyfikował, musimy problem rozwiązać w inny sposób. Najprostszym rozwiązaniem jest rozbudowanie słownika o konkretne wyrazy, których szukamy w naszej analizie. W naszym przypadku taka kategoria słów dla cechy „zdecydowanie" mogłaby wyglądać następująco:

a) poprawne formy wyrazowe: zdecydowany, zdecydowana, zdecydowanie,

b) niepoprawne formy wyrazowe zauważone przez analizę z zastosowaniem Thesaurusa: zdecydowny, zdecydowna, zdecydownie.

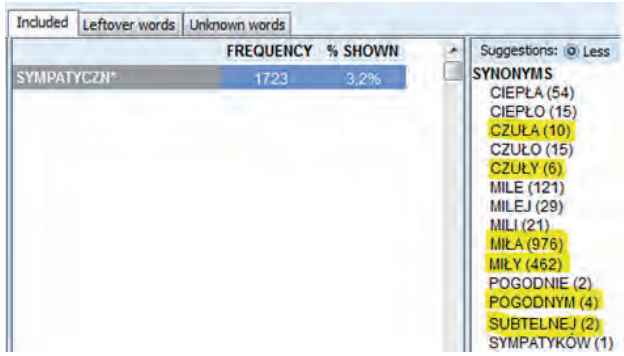
Ad B2. Thesaurus – narzędzie rozbudowy kategorii słownikowych. Technika rozbudowywania słów specyficznych dla danej kategorii za pomocą Thesaurusa opiera się na wyszukiwaniu słów podobnych (znaczeniowo lub o wspólnym rdzeniu). Spójrzmy na przykład podpowiedzi trafnych (zestawienie 4.1) i nietrafnych (zestawienie 4.2) formułowanych przez Thesaurusa.



Zestawienie 4.1. Rozwój słownika z wykorzystaniem Thesaurusa wbudowanego w programie Wordstat – podpowiedzi nietrafne

Źródło: opracowanie własne z zastosowaniem programu Wordstat

Jak widzimy, Thesaurus podpowiada niepoprawne dla naszej analizy słowa. Wynika to z zastosowania innego rdzenia, a mianowicie dokład*.



Zestawienie 4.2. Rozwój słownika z wykorzystaniem Thesaurusa wbudowanego w programie Wordstat – podpowiedzi trafne

Źródło: opracowanie własne z zastosowaniem programu Wordstat

Tym razem Thesaurus podpowiada słowa trafne dla naszej analizy. Kluczem dla podpowiedzi, jakie tym razem formułuje narzędzie, jest podobieństwo znaczeniowe.

Ad c) Trzecia z metod rozwoju słownika korzysta z osiągnięć obu wcześniej zarysowanych podejść. Polega na rozszerzeniu zawartości słownika w oparciu o: wykorzystanie reguł semantycznych (korzystamy z istniejącego narzędzia, np. słownika wyrazów bliskoznacznych) oraz weryfikację zastosowanych reguł za pomocą korpusu analizowanych słów (podejście *bottom-up* pozwala nam stwierdzić, które z rozszerzeń słownika daje dodatkowe rezultaty wyszukiwania, a które nie zwraca nowych informacji w procesie wyszukiwania).

Nakreślmy schemat metody, w którym kategoria słownikowa zawiera jeden wyraz. W schemacie tym podamy również przykład rozbudowy jednej kategorii słownikowej. Na potrzeby analizy wykorzystujemy dwa istniejące narzędzia wspomagające opisaną tu analizę. Pierwsze to słownik WordNet w polskiej wersji językowej. Oryginalna wersja WordNeta została zbudowana i jest rozwijana od ponad 25 lat na Uniwersytecie Princeton (Miller 1995). Polska wersja WordNeta nazywa się Słowosieć i jest rozwijana na Uniwersytecie Wrocławskim[11].

Idea leżąca u podstaw tego słownika daje się opisać w następujący sposób. Jest to leksykalna baza wiedzy, na którą składają się:

a) słownik wyrazów bliskoznacznych,
b) opis relacji semantycznych między wyrazami,
c) źródło definicji znaczeń,
d) hierarchia pojęć.

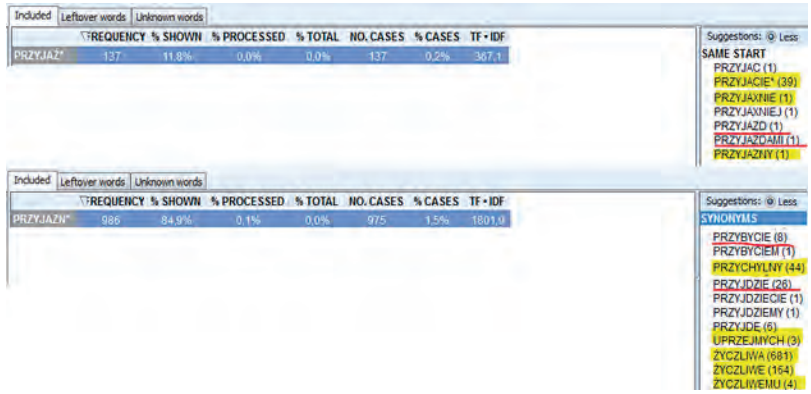
Słowosieć posłuży nam do ustalenia relacji leksykalnych pomiędzy słowem, które zakwalifikowaliśmy do słownika analitycznego i innymi słowami. Relacje, na jakich będziemy się skupiali, to:

a) synonimia: relacja pomiędzy słowami oparta na równoważności znaczeniowej,
b) antonimia: relacja oparta na przeciwieństwie znaczeniowym,
c) homonimia: relacja, w której różne znaczenia wyrażane są za pomocą identycznych form językowych,
d) hiperonimia: relacja oparta na nadrzędności wyrazów wobec innych wyrazów,
e) meronimia: asymetryczna relacja słowa do innego słowa (np. szprycha jest meronimem koła rowerowego).

Drugim narzędziem, po jakie sięgamy, jest słownik *Gdy Ci słowa zabraknie* (Broniarek 2010). Jest to słownik synonimów rekomendowany przez Jerzego Bralczyka. Posłuży on nam jako narzędzie weryfikacji trafności łączenia słów w jedną kategorię analityczną w oparciu o relację synonimii. Powodem, dla którego słownik Broniarka wykorzystujemy w roli narzędzia weryfikacyjnego, jest fakt, iż słownik ten jest zakorzeniony w języku polskim (nie jest tłumaczeniem innego słownika, językiem rdzennym autora jest język polski).

Za punkt wyjścia przyjmujemy kategorię, która ma w sobie jedno słowo znalezione w korpusie analizowanych opinii: przyjazny. Słowa kwalifikowane do dalszej analizy zakreślamy na żółto. Niepoprawne podpowiedzi podkreślamy na czerwono.

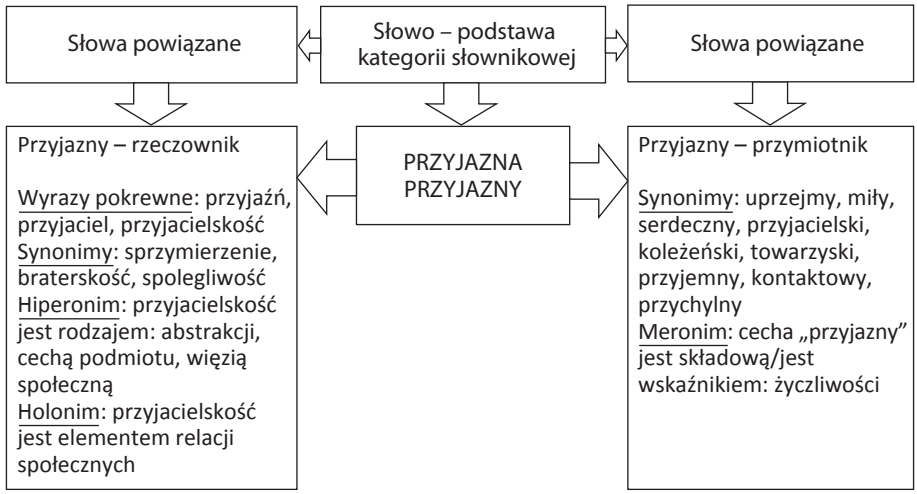
[11] Opis polskiej wersji słownika znaleźć można na stronie: http://nlp.pwr.wroc.pl/projekty/slowosiec2.



Zestawienie 5. Podpowiedź Thesaurusa dla rdzenia przyjazn* i przyjaź*

Źródło: opracowanie własne z zastosowaniem programu Wordstat

Thesaurus podpowiada nam dodatkowe słowa występujące w korpusie analizowanych wypowiedzi, które mogą być związane z przyjazn* i przyjaź*. Rozwijamy kategorię o słowo zapisane niepoprawnie: przyjaxnie. Poniżej podajemy rozwój tej jednej kategorii analitycznej, wykorzystując relacje leksykalne, które podpowiada nam Słowosieć. Ilustr. 3 pokazuje słowa, które kwalifikujemy do kategorii: przyjazny/przyjazna.



Ilustr. 3. Słowa znajdujące się w relacjach leksykalnych ze słowem przyjazny/przyjazna

Źródło: opracowanie własne z wykorzystaniem zasobów Słowosieci/WordNetu, słownika Broniarka (2010)

Zastosowanie reguł semantycznych do rozbudowy słownika klasyfikacyjnego rozszerza nam jedną kategorię o słowa:

a) proponowane przez istniejące słowniki,
b) widniejące w tekście i spełniające kryterium zależności semantycznej.

Ten drugi zabieg sprawia, że zaczynamy pracę nad strukturyzacją listy ACL. Nie traktujemy każdego słowa jako osobnej kategorii, ale łączymy je w oparciu o reguły semantyczne, redukując przestrzeń właściwości/cech wyznaczanych przez model ACL.

Zastosowanie opisanej tu logiki rozbudowy kategorii słownikowych pozwala nam dla redukcję listy ACL tak, że zamiast pojedynczych słów otrzymujemy jedną kategorię (tab. 2).

Tabela 2. Słowa zakwalifikowane do kategorii analitycznej w oparciu o wykorzystane w analizie narzędzia (Thesaurus w oprogramowaniu WordStat, WordNet i Słowosieć, słownik Broniarka)

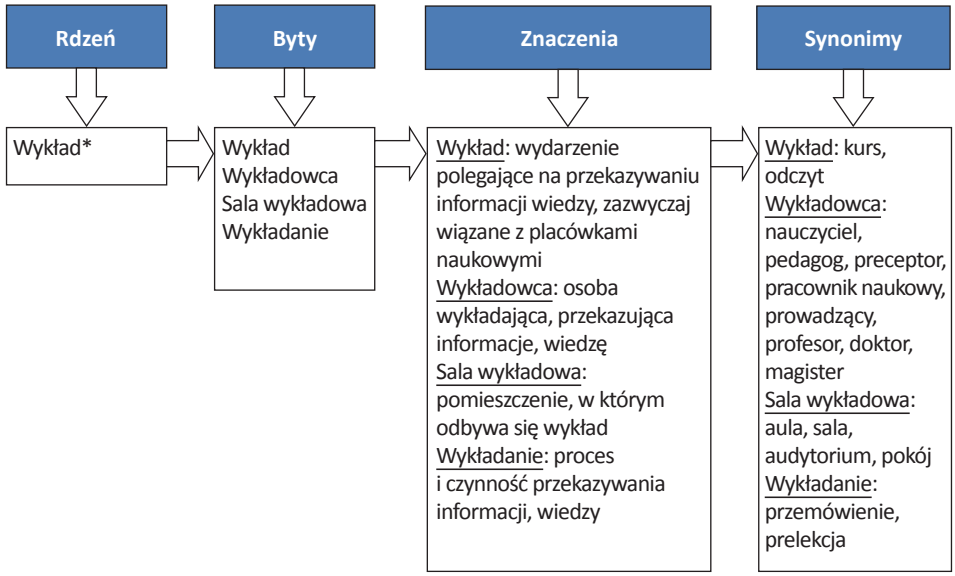
Słowo nadrzędne	Słowa włączone do kategorii
PRZYJAZNA/PRZYJAZNY	Pełna postać słowa: przyjacielsk(a)i, przyjaciel, przyjaciółka, serdeczn(a)y, serdeczność, mił(a)y, życzliwość, życzliw(a)y, przychylność, przychyln(a)y, uprzejmość, uprzejm(a)y Rdzenie: przyjaciel*, przyjaź*, przyjazn*, serdeczn*, życzliw*, przychyln*, uprzejm* Słowa zakwalifikowane w pełnej wersji bez rdzeni: miły, miła

Źródło: opracowanie własne w oparciu o: Wordstat, WordNet, Słowosieć, słownik Broniarka (2010).

Etap 3. Rozpoznanie i rozwój ontologii słownika. Zastosowanie reguł pragmatycznych

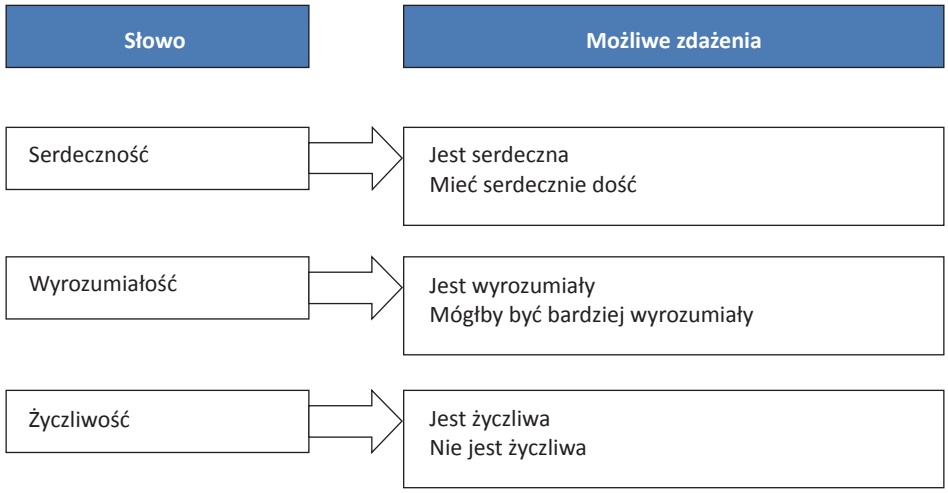
Proces analizy znaczeń słów w tekście pisanym pokazuje nam, że jedno słowo może posiadać więcej niż jedno odniesienie (więcej niż jeden desygnat). O wieloznaczności słów, a także o wielości odniesień przedmiotowych słów pisał Charles S. Peirce, wyróżniając trzy kluczowe relacje: znak–obiekt, idea–obiekt, idea–znak.

Wiedzę tę wykorzystamy na kolejnym etapie budowy słownika. Jest nim rozpoznanie ontologii słownika. Pojęcie ontologia w tym przypadku ma szczególne znaczenie. Oznacza ono albo: „to co jest", albo: „może być" (Hayek 1945). Modalności te wykorzystujemy do trafniejszego zrozumienia relacji znaczeń zawartych w rdzeniach i słowach, do ich odniesień przedmiotowych. Podamy dwa przykłady, w których pokażemy, jak klasyfikacja słów za pomocą rdzeni pozwala na identyfikację różnych obiektów odniesienia (ilustr. 4 i 5).



Ilustr. 4. Wykorzystanie ontologii w znaczeniu „to co jest" do strukturyzacji słownika

Źródło: opracowanie własne



Ilustr. 5. Wykorzystanie ontologii w znaczeniu „może być" do strukturyzacji słownika

Źródło: opracowanie własne

Ilustr. 4 pokazuje nam, że słowa o wielu desygnatach musimy poddać dodatkowej analizie kontekstowej (*keyword in context analysis* – KWIC). Natomiast ilustr. 5 – że kwalifikacja słów do zdefiniowanych kategorii wymaga

zbudowania zbioru warunków logicznych, które będą definiowały kontekst, w jakim dane słowo występuje. W następnym etapie utworzymy reguły słownikowe, które zapewnią trafną klasyfikację wypowiedzi i ułatwią identyfikację poszukiwanych opinii.

Etap 4. Rozwój syntaktycznych reguł słownikowych

Mamy za sobą budowę listy słów, rdzeni. Wiemy też, że słowa nabierają innego znaczenia w zależności od kontekstu, w którym występują. Chcąc zwiększyć poprawność klasyfikacji słownika, wyposażamy go w reguły leksykalne. Wykonają one eksploracyjną analizę kontekstu, w jakim występują poszukiwane przez nas słowa i frazy. Wykorzystamy reguły, które dopuszcza Wordstat i zdefiniujemy warunki początkowe dla tych reguł. Dostępne reguły to operatory logiczne oraz warunki określające położenie słów i fraz. Są to:

a) operatory logiczne: I, LUB, NIE,
b) warunki dotyczące bliskości wystąpienia słów i fraz: BLISKO, NIE BLISKO,
c) warunki określające kolejność słów i fraz: PRZED, NIE PRZED, PO, NIE PO.

Warunki, jakie możemy dodać do analizy, określają obszar, którego dotyczą konstruowane przez nas reguły słownikowe oraz poziom bliskości słów. Są to:

a) obszar obowiązywania reguły: zdanie, paragraf, dokument, obserwacja (może składać się z wielu dokumentów),
b) dystans pomiędzy słowami określony jest za pomocą: liczby słów.

Poniżej prezentujemy różne typy reguł, które zastosowaliśmy w słowniku analitycznym. Do reguły dodajemy warunek początkowy definiujący zakres jej obowiązywania. W ramach reguł łączymy warunki i operatory, tworząc reguły wielopoziomowe. Reguły mogą dotyczyć słów, rdzeni, fraz, kategorii analitycznych. Oto kilka przykładów reguł (operatory zastosowane w regułach piszemy wersalikami):

Reguła 1: „super I wykład*”, w ramach tego samego zdania, na przestrzeni 2 słów.

Reguła 2: „wymagają* PRZED wykładowca”, w ramach tego samego zdania, na przestrzeni 1 słowa (identyfikuje np. sformułowanie wymagający wykładowca) albo „wymagają* PO wykładowca”, w ramach tego samego zdania, na przestrzeni 2 słów (identyfikuje np. sformułowanie wykładowca jest wymagający).

Reguła 3: „miły LUB przychylny I wykładowca I NIE PO negacji”, w ramach tego samego zdania na przestrzeni 3 słów.

Reguła 4: „dość PO serdecznie I mam”, w ramach tego samego zdania na przestrzeni 3 słów.

Reguła 5: „profesjonalny LUB dokładny I wykładowca PRZED negacja”, w ramach tego samego zdania na przestrzeni 3 słów – reguła kwalifikuję frazę do kategorii określających negatywne cechy wykładowcy.

Reguła 6: „wykładowca NIE BLISKO dowolna cecha w zdefiniowanych kategoriach analitycznych" w tym samym zdaniu – wyklucza zdanie z dalszej analizy.

Reguła 7: „cecha pozytywna BLISKO cecha negatywna" w ramach tego samego zdania – kwalifikuje do kategorii analitycznej: ocena ambiwalentna.

Kwalifikacja słów i fraz oparta na zbudowanych powyżej regułach wymaga ewaluacji. Jest to poprawianie trafności klasyfikacyjnej modelu poprzez uczenie z nauczycielem. Wynikiem tej weryfikacji może być modyfikacja definicji reguł, budowa nowych reguł. Podajemy przykład analizy ewaluacyjnej, której wynik nakazał modyfikację reguły.

Reguła 7: „miły LUB przychylny I wykładowca I NIE PO negacji", w ramach tego samego zdania na przestrzeni 3 słów. Reguła zakwalifikowała następującą wypowiedź: „nie wiem, ale generalnie miły człowiek, wyrozumiały nawet wobec tych, których nigdy na wykładzie nie widział J".

Reguła 7 po modyfikacji wygląda tak: „miły LUB przychylny I wykładowca I NIE PO negacji LUB NIE PO nie wiem LUB NIE PO nie jestem pewien".

Ewaluację reguł prowadziliśmy dwustopniowo (dwuetapowo). Zbiór reguł podzieliliśmy na trzy równe części. Każdy z autorów samodzielnie oceniał skuteczność jednego zbioru reguł. Następnie przekazywaliśmy ocenione reguły sobie wzajemnie bez komentarza dotyczącego ich efektywności czy konieczności poprawy zapisu reguł. Ten drugi etap oceny kończyliśmy konfrontacją swoich spostrzeżeń oraz wprowadzeniem zmian, co do których istniała między nami zgoda. Etap ten zaowocował precyzyjniejszym sformułowaniem kryteriów oceny reguł słownikowych. Analiza trzeciego zbioru reguł to wspólna praca dwóch autorów posługujących się wypracowanymi kryteriami oceny.

Zbudowany słownik klasyfikacyjny, którego budowa uwzględniała już ocenę poprawności budowy jego elementów, wymaga teraz całościowej oceny. Przejdźmy do tego etapu.

Etap 5. Ewaluacja słownika

Oceny poprawności klasyfikacji można dokonać na wiele sposobów[12]. Na potrzeby tego artykułu posłużyliśmy się dwoma z nich. Pierwszy to wspominana już manualna metoda weryfikacji. W ten sposób sprawdzaliśmy poprawność klasyfikacji wykonywanej za pomocą:

a) rdzeni, słów kluczowych, fraz,
b) reguł syntaktycznych i logicznych.

[12] W tym miejscu nie omawiamy bogatej tematyki pomiaru poprawności klasyfikacji. Dodamy tylko, że problematyka ta jest obszernie opisana na przykład przez Powers (2007/2011).

Sprawdziliśmy trafność reguł klasyfikacji na zbiorze uczącym się. Posłużymy się teraz tymi regułami dla pozostałej części zbioru i ocenimy trafność klasyfikacji. Dla tego celu zastosujemy jedną z najpopularniejszych (Witten, Frank, Hall 2005) miar do oceny poprawności klasyfikacji. Jest nią OSR (*overall success rate*), którą wyraża się wzorem:

$$OSR = \frac{1}{n}\sum_{i=1}^{k} n_{i,i}$$

Wyznaczyliśmy poziom OSR dla poszczególnych kategorii analitycznych w ramach słownika klasyfikacyjnego. Przez kategorię analityczną rozumiemy nazwę kategorii i zbiór/koszyk słów, które wchodzą w jej skład. W szczególnym przypadku kategoria może liczyć jedno słowo. Podajemy przykładowe wyniki dla pięciu najliczniejszych kategorii (tab. 3). Pokazane są syntetyczne klasy.

Tabela 3. Wynik OSR dla pierwszych pięciu klas w słowniku klasyfikacyjnym

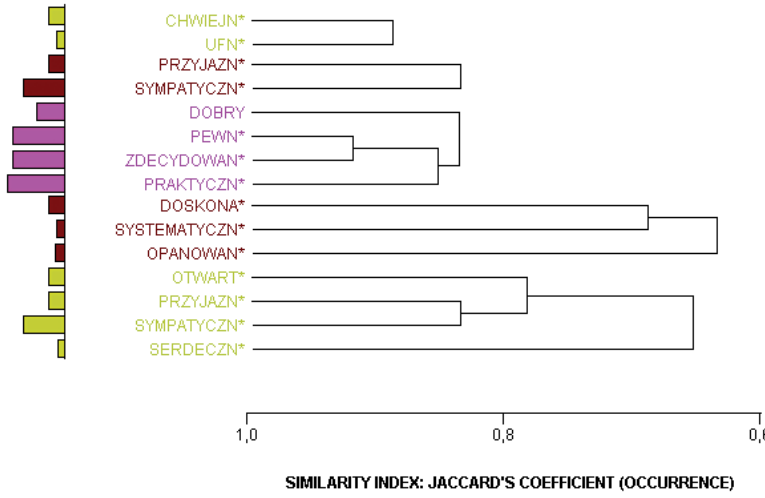
Kategoria słownikowa	Overall Success Rate
CIEKAWY	1.0
DOBRY	1.0
PRAKTYCZNY	0.97
PRZYJACIELSKI	0.95
PROFESJONALNY	0.94

Źródło: opracowanie własne.

Prezentowane powyżej wyniki osiągnięte zostały na etapie drugiej ewaluacji reguł klasyfikacyjnych. Interpretacja wyników OSR jest intuicyjna: im wyższa wartość współczynnika, tym większa poprawność klasyfikacji. Każda z wartości może być poprawiona poprzez doszczegółowienie i rozwój reguł słownikowych. Na tym etapie pozostawiamy miary na zaprezentowanym poziomie[13]. W tym momencie możemy już wykonać analizę podobieństwa wypowiedzi. Naszą jednostką analityczną będzie cecha reprezentowana przez słowo lub frazę.

[13] Dość często przyjmuje się, że wystarczający poziom poprawności klasyfikacji to 80% (por. Tomanek 2014a).

Pierwsza z analiz to analiza współwystępowania zakodowanych (otagowanych) wypowiedzi. W analizie posłużymy się indeksem podobieństwa Jaccarda[14] (dla potrzeb prezentacji dendrogram został przycięty do 5 klastrów). Wyniki analizy wizualizujemy za pomocą dendrogramu (ilustr. 6).



Ilustr. 6. Analiza współwystępowania kategorii analitycznych z zastosowaniem indeksu podobieństwa Jaccarda

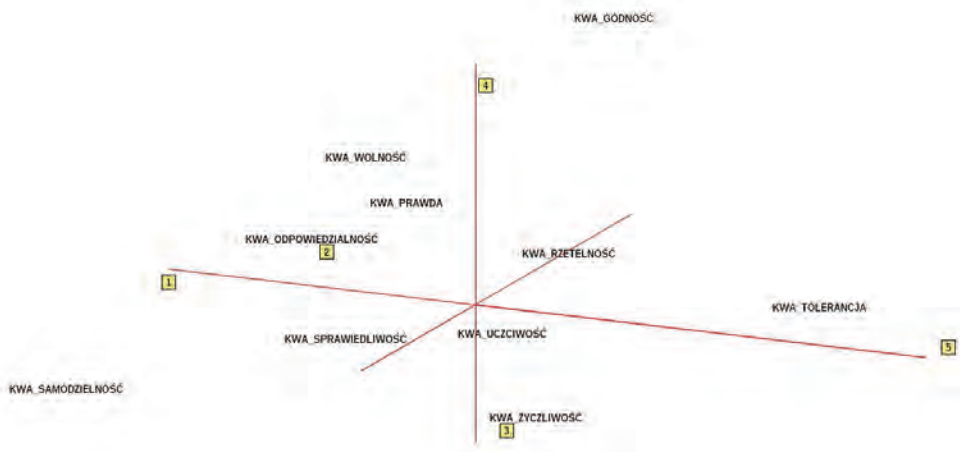
Źródło: opracowanie własne z zastosowaniem oprogramowania Wordstat

W wyodrębnionych wiązkach rdzeni cech/przymiotników widzimy takie, które wydają się tworzyć spójne logicznie i koncepcyjnie cechy (wiązka$_3$: pewny, zdecydowany, wiązka$_4$: doskonały, systematyczny, opanowany) oraz takie, które wymagają zapoznania się z treścią opinii w celu lepszego rozpoznania logiki leżącej u jej podstaw (wiązka$_1$: chwiejny, ufny).

Na tym etapie kończymy pracę związaną z budową słownika oraz reguł słownikowych. Zadanie, które wykonaliśmy, doprowadziło nas do sytuacji, w której nasz słownik cechuje się wysokim poziomem poprawności klasyfikacji wypowiedzi tekstowych. Budowa słownika wedle podanych powyżej zasad daje w rezultacie narzędzie klasyfikacyjne, które można z powodzeniem stosować w analizach zarówno danych tekstowych, jak i danych dźwiękowych oraz obrazów. W tych dwóch ostatnich obszarach zarówno dźwięki, jak i obrazy prezentowane są między innymi za pomocą ciągów zero-jedynkowych. Jednakże budowa słownika

[14] Podobnie jak wcześniej – wybieramy popularną miarę podobieństwa. Poza indeksem Jaccarda mamy do wyboru np.: współczynnik Sorensena, współczynnik Ochiai, współczynnik korelacji Pearsona, miara Kullbacka-Leiblera itd. Przystępne omówienie niektórych z miar można znaleźć w: Yung-Shen, Jung-Yi, Shie-Jue (2013).

nie jest finalnym etapem procesu analizy danych tekstowych. Kolejnym jest jego wykorzystanie w konstruowaniu modelu analitycznego. Kierunek dalszej analizy danych tekstowych wynika bezpośrednio z problemu badawczego. Na kolejnych etapach analizy badacz może zastosować szereg rozwiązań analitycznych (w tym statystycznych) oferowanych w programach CAQDAS – od prostych zestawień tabelarycznych, aż do zaawansowanych technik wielowymiarowych. Przykładowo w dalszej części prowadzonej przez nas analizy danych (wypowiedzi studenckich zarejestrowanych w systemie oceny zajęć uniwersyteckich USOS), wykorzystaliśmy słownik klasyfikacyjny do sprawdzenia zgodności postaw wykładowców akademickich z wzorcem postaw zawartych w kodeksie wartości akademickich na wybranych wydziałach. W tym celu posłużyliśmy się analizą korespondencji. W konsekwencji zastosowanie słownika w analizie korespondencji pozwoliło na sprawdzenie, jak bardzo postawy wykładowców odbiegają od przyjętego na uczelni wzorca, oraz pokazało, jak analizowane wydziały sytuują się w przestrzeni wartości zawartych w kodeksie. Poniżej prezentujemy przykładowy wykres z analizy korespondencji z zastosowaniem słownika klasyfikacyjnego do nieustrukturyzowanych danych tekstowych, wypowiedzi studenckich[15].



Ilustr. 7. Analiza korespondencji z zastosowaniem słownika klasyfikacyjnego

Źródło: opracowanie własne

Na zakończenie chcemy podsumować spostrzeżenia związane z pracą nad procesem budowy słownika klasyfikacyjnego.

[15] Przedstawione na wykresie punkty od 1 do 5 reprezentują pozycje analizowanych wydziałów w przestrzeni geometrycznej. Kodeks wartości akademickich został utworzony w 2003 r. na Uniwersytecie Jagiellońskim, www.uj.edu.pl/c/document_library/get_file?uuid=d63b4be0--5eee-4d94-bd32-3b1ccef396f6&groupId=10172.

Podsumowanie – wnioski

Słowniki analityczne pełnią kilka funkcji w odkrywaniu wiedzy. Z punktu widzenia rzetelności zastosowanej metody ważne jest budowanie takiego słownika klasyfikacyjnego, który:

a) klasyfikuje wypowiedzi tekstowe z „wysokim poziomem poprawności",

b) posiada „uniwersalne" dla danego języka reguły syntaktyczne,

c) daje się stosować do różnych materiałów tekstowych w ramach tego samego języka.

W praktyce skonstruowany przez badacza słownik klasyfikacyjny jest zazwyczaj jednym z narzędzi analitycznych i jednym z etapów projektu analitycznego czy badawczego. Może służyć do eksploracji tekstu, klasyfikacji wypowiedzi czy jako model predykcyjny. Aby słownik mógł pełnić te funkcje, konieczne jest wykorzystanie wiedzy z kilku obszarów. Niezbędne są informacje o formalnym i naturalnym sposobie funkcjonowania danego języka. Potrzebna jest:

a) znajomość reguł syntaktycznych,

b) znajomość reguł logiki pierwszorzędowej,

c) świadomość znaczenia wiedzy o modalnościach językowych,

d) umiejętność rozpoznania ontologii języka, którą zakładają analizowane wypowiedzi.

Wydaje się, że proces budowy poprawnie funkcjonującego słownika klasyfikacyjnego powinien uwzględniać kilka zasad tj.:

a) metodologię projektową porządkującą proces budowy i pracy (np. CRISP-DM),

b) definiowanie reguł klasyfikacyjnych opartych na relacjach:
   - semantycznych (pozwalających łączyć w grupy słowa podobne znaczeniowo),
   - pragmatycznych (pozwalających rozróżniać odmienne konteksty użycia słów),
   - syntaktycznych (zwiększających prawdopodobieństwo poprawnej klasyfikacji),

c) procedury ewaluacji stosowane zarówno na etapie budowy elementów składowych słownika, jak i po zakończeniu prac nad słownikiem,

d) iteracyjność procesu weryfikacji i testowania elementów składowych słownika (dotyczy to np. reguł klasyfikacyjnych, definiowania rdzeni),

e) możliwość stosowania istniejących narzędzi klasyfikacyjnych jako elementów inspirujących analizę lub przyspieszających pracę klasyfikacyjną (w przypadku budowy słownika rekonstruującego schematy oceny jest to np. lista przymiotnikowa ACL),

f) konieczność adaptacji reguł wypowiedzi swoistych dla języków specyficznych (np. dla dialektów, subkultur),

g) procedury testowania słownika jak i jego elementów składowych zawsze wtedy, gdy słownik stosowany jest do nowego korpusu wypowiedzi.

Jak pokazaliśmy, słowniki analityczne wspierają pracę w środowisku CAQDAS. W szczególności słowniki klasyfikacyjne dla języków fleksyjnych dają możliwość przenoszenia wypracowanych rozwiązań na inne zadania analityczne w obrębie tego samego języka. Zbudowany przez nas słownik wykorzystujemy wielokrotnie w kolejnych prowadzonych analizach. W szczególności przetestowane przez nas reguły syntaktyczne cechuje przenośność (Micek, Beźnic 2004), a więc możliwość ich stosowania w różnych kontekstach sytuacyjnych. Podobnie zdefiniowane powiązania semantyczne pomiędzy słowami kluczowymi dają się stosować w innych analizach tematycznych. Dla przykładu transformacja (*stemming*) takiego rozwiązania jako lista ACL pozwala nam na stosowanie tego narzędzia do odkrywania schematów ocen w analizie forów dyskusyjnych oraz analizie FGI (Tomanek 2014b).

Bibliografia

Bernstein Basil (1971), *Class, Codes and Control*, Routledge, London.

Bourdieu Pierre (2009), *Doksa i życie codzienne. O habitusie, oświeconej fałszywej świadomości i rapie rozmawiają krytyk ideologii i realista*, rozmowa z Terrym Eagletonem, „Recykling Idei", nr 12.

Broniarek Wojciech (2010), *Gdy Ci słowa zabraknie. Słownik synonimów*, Haroldson, Brwinów.

Brosz Maciej (2012), *Komputerowe wspomaganie badań jakościowych. Zastosowanie pakietu NVivo w analizie materiałów nieustrukturyzowanych*, „Przegląd Socjologii Jakościowej", t. VII, nr 1, s. 98–125.

Bryda Grzegorz, Tomanek Krzysztof (2014), *Od CAQDAS do Text Miningu. Nowe techniki w analizie danych jakościowych*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Chapman Pete, Clinton Julian, Kerber Randy, Khabaza Thomas, Reinartz Thomas, Shearer Colin, Wirth Rüdiger (2000), *CRISP-DM 1.0. A Step-by-step Data Mining Guide*, SPSS, New York.

Fayyad Usama M., Piatetsky-Shapiro Gregory, Smyth Padhraic (1996), *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, "Knowledge Discovery and Data Mining", no. 2–4, s. 82–88; www.facweb.iitkgp.ernet.in/~shamik/autumn2004/dwdm/papers/Knowledge%20Discovery%20and%20Data%20Mining%20Towards%20a%20Unifying%20Framework%20%281996%29.pdf [dostęp: 1.05.2014].

Fellbaum Christiane (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge; http://wordnet.princeton.edu [dostęp: 1.05.2014].

Gadomski Adam Maria (2013), *Meta-Ontological Assumptions: Information, Preferences and Knowledge (IPK): Universal Cognitive Architecture*; http://erg4146.casaccia.enea.it/wwwerg26701/gad-dict.htm [dostęp: 1.05.2014].

Greene Jennifer C. (2006), *Toward a Methodology of Mixed Methods Social Inquiry*, "Research in the Schools", vol. 13, no. 1, s. 93–98.

Hayek Friedrich A. (1945), *The Use of Knowledge in Society*, "The American Economic Review", vol. 35, no. 4, s. 519–530.

Hotho Andreas, Nürnberger Andreas, Paaß Gerhard (2005), *A Brief Survey of Text Mining*, "German Journal for Computer Linguistics and Speech Technology", vol. 20 (1), s. 19–62.

Kobos Mateusz, Mańdziuk Jacek (2008), *Metody sztucznej inteligencji w przewidywaniu wartości indeksu giełdowego z wykorzystaniem artykułów prasowych*, [w:] Cezary Orłowski, Zdzisław Kowalczuk, Edward Szczerbicki (red.), *Zarządzanie wiedzą i technologiami informatycznymi*, Pomorskie Wydawnictwo Naukowo-Techniczne PWNT, Gdańsk.

Kodeks wartości akademickich; www.uj.edu.pl/c/document_library/get_file?uuid=d63b4be0--5eee-4d94-bd32-3b1ccef396f6&groupId=10172 [dostęp: 1.01.2015].

Kordasiewicz Anna, Haratyk Karol (2013), *Między wizerunkiem a praktyką – diagnoza stanu wykorzystania programów komputerowych wspomagających analizę danych jakościowych w Polsce*, „Przegląd Socjologiczny", t. LXII/1, s. 167–187.

Kripke Saul (2001), *Nazywanie a konieczność*, przeł. Bohdan Chwedeńczuk, Fundacja Aletheia, Warszawa.

Lofland John, Snow A. David, Anderson Leon, Lofland Lyn H. (2009), *Analiza układów społecznych. Przewodnik metodologiczny po badaniach jakościowych*, Scholar, Warszawa.

Martowska Katarzyna (2012), *Lista Przymiotnikowa. Harrison G. Gough, Alfred B. Heilbrun Jr. Polska Normalizacja*, Pracownia testów psychologicznych Polskiego Towarzystwa Psychologicznego, Warszawa.

Micek Dorota, Beźnic Szymon (2004), *Jakościowe badania marketingowe – fokusy i wywiady pogłębione – funkcje, zastosowania*; www.cem.pl [dostęp: 1.05.2014].

Miller George A. (1995), *WordNet: A Lexical Database for English*, "Communications of the ACM", vol. 38, no. 11, s. 39–41.

Niedbalski Jakub (2013), *Odkrywanie CAQDAS. Wybrane bezpłatne programy komputerowe wspomagające analizę danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Peirce Charles Sanders (1931–1935), *Collected Papers*, vol. 1–6, Harvard University Press, Cambridge.

Powers David M. (2007/2011), *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*, "Journal of Machine Learning Technologies", vol. 2 (1), s. 37–63.

Provali Research, oprogramowanie; http://provalisresearch.com [dostęp: 1.10.2014].

Putnam Hilary (1990), *Realism with a Human Face*, Harvard University Press, Cambridge.

Quine Van Orman Willard (1999), *Słowo i przedmiot*, przeł. Cezary Cieśliński, Fundacja Aletheia, Warszawa.

Ramos Juan (2014), *Using TF-IDF to Determine Word Relevance in Document Queries*, Rutgers University, Piscataway, New York.

Rorty Richard (1996), *Przygodność, ironia i solidarność*, przeł. Wacław Jan Popowski, Spacja, Warszawa.

Silverman David (2007), *Interpretacja danych jakościowych*, PWN, Warszawa.

Skvoretz John (1998), *Theoretical Models: Sociology's Missing Links*, [w:] Alan Sica (ed.), *What is Social Theory? The Philosophical Debates*, Blackwell, Oxford.

Słowosieć; http://nlp.pwr.wroc.pl/projekty/slowosiec2 [dostęp: 1.10.2014].

Sperber Dan, Wilson Deirdre (2004), *Relevance Theory*, [w:] Gregory Ward, Laurence Horn (eds), *Handbook of Pragmatics*, Blackwell, Oxford.

Tomanek Krzysztof (2014a), *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*, „Przegląd Socjologii Jakościowej", t. 10, nr 2, s. 118–136; www.przegladsocjologiijakosciowej.org [dostęp: 2.01.2015].

Tomanek Krzysztof (2014b), *Jak nauczyć metodę samodzielności? O „samouczących się" metodach analizy treści*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Witten Ian H., Frank Eibe, Hall Mark A. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo.

Yung-Shen Lin, Jung-Yi Jiang, Shie-Jue Lee (2014), *A Similarity Measure for Text Classification and Clustering*, [w:] *IEEE Transactions on Knowledge and Data Engineering* IEEE Transactions on Knowledge and Data Engineering (Impact Factor: 1.89), 07.2014; 26 (7), s. 1575–1590.

Knowledge Discovery in Textual Statements Construction Method of Classification Dictionary

Summary. Using knowledge of syntax, semantics and logic links between elements of expression is an attractive area in Data Mining and text analysis. Methods of text analysis and text classification do not always use resolutions like these. The purpose of this article is to show a method that integrates the solutions taken from different areas of scientific knowledge. The goals authors deal with are: (a) the use of knowledge in the following field: linguistics, NLP, logic, statistics in order to build a reliable analytical tool in CAQDAS environment; (b) the use of application available in CAQDAS solutions and developing them with new techniques for classification tools; (c) assessment of the adopted solution. The method for dictionary building requires a synthesis of many solutions. To build accurate classification dictionary one needs: the basics of the language content search, Thesaurus, synonym dictionary, lexical relations definitions. Authors describe a step-by-step process of building a classification dictionary, accentuate the pitfalls and important decisions, which appears to be important during the analysis process.

Keywords: Text Mining, CAQDAS, Classification Dictionary, thematic analysis, natural language processing, NLP, Thesaurus, CRISP-DM, adjective list ACL, Wordnet, knowledge discovery in textual data, KDT.