



Politechnika Wrocławska

Wydział Informatyki i Zarządzania

kierunek studiów: Inżynieria Zarządzania

specjalność: Zastosowanie IT w biznesie

Praca dyplomowa – inżynierska

PROJEKT APLIKACJI WSPOMAGAJĄCEJ ANALIZĘ OPINII PRACOWNIKÓW FIRMY AMAZON

Mateusz Guściora

słowa kluczowe:

krótkie streszczenie:

opiekun pracy
dyplom owej	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis
Ostateczna ocena za pracę dyplomową			
Przewodnicz ący Komisji egzaminu dyplomowego Tytuł/stopień naukowy/imię i nazwisko ocena podpis

Do celów archiwalnych pracę dyplomową zakwalifikowano do:*

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

* niepotrzebne skreślić

pieczętka wydziałowa

Wrocław
[2020]

DEKLARACJA REALIZACJI PRACY DYPLOMOWEJ

Temat pracy dyplomowej w języku polskim:

Projekt aplikacji wspomagającej analizę opinii pracowników firmy Amazon.

Temat pracy dyplomowej w języku angielskim:

Project of an application supporting analysis of Amazon employees' reviews.

Język realizacji pracy: Język polski

Opiekun pracy dyplomowej: dr inż. Marek Lubicz

Cel pracy, a także zakres pracy lub sformułowanie problemów praktycznych/zawodowych:

Problemem jest duża liczba opinii i ocen negatywnych pracowników Amazon (25% ocen negatywnych na portalu Indeed) oraz ich wielojęzyczność. Przekłada się to na trudności w analizie danych, część danych jest w formie słownej, a część jako ocena w skali 1-5. Utrudnienia to identyfikacje negatywnych przyczyn opinii i ocen, w konsekwencji, potencjalne podejmowanie decyzji, przez menedżerów zarządzających kapitałem ludzkim oraz menedżerów wyższych szczebli. Spada też średnia wartość ocen pracy w Amazon co może oznaczać konieczność podjęcia działań mogących zmienić ten trend.

Celem pracy jest projekt aplikacji, która wspomogę menedżerów zarządzających zasobami ludzkimi oraz menedżerów wyższych szczebli w firmie Amazon, w analizie opinii i ocen oraz w zidentyfikowaniu głównych przyczyn niezadowolenia pracowników. Ułatwienie analizy poprzez przetłumaczenie opinii (na jeden język) i wykorzystania technik text mining, identyfikacje przyczyn niezadowolenia pracowników oraz stworzenie wizualizacji danych istotnych dla zagadnienia. Ponadto aplikacja wspomogę w opracowaniu racjonalnych tj. możliwych do realizacji celów i działań strategicznych, celem poprawy ocen i zadowolenia pracowników. Praca opierać się będzie o zbiór opinii pracowników Amazon ze świata, z portalu internetowego Indeed.

Projekt aplikacji będzie obejmować pobranie i import danych, przygotowanie danych do analizy. Następnie użycie narzędzi informatycznych, w szczególności do importu/pobrania danych ze strony internetowej, text miningu, analizy statystycznej oraz analizy wizualizacji. Analizę danych, w tym identyfikacja przyczyn, oraz istotnych kwestii dotyczących zarządzania personelem, stworzenie wizualizacji wspomagających ich odczyt. W oparciu o tą analizę zaproponowanie przyszłych celów oraz działań strategicznych.

Wstępna struktura pracy:

Wstęp

- Stworzenie wstępu do pracy inżynierskiej w tym sformułowanie jakich zagadnień dotyczy praca,
- określenie problemu oraz celu pracy i sposobu osiągnięcia celu,
- określenie zawartości pracy w tym rozdziałów i podrozdziałów,
- przy tworzeniu wstępu do pracy równoległe studiowanie i pogłębianie źródeł literaturowych związanych z zagadnieniami takimi jak: analiza danych, zadowoleniem pracy w organizacjach, zarządzanie personelem oraz

metod i narzędzi informatycznych przydatnych w powyższych zagadnieniach.

Rozwinięcie (obejmować będzie część teoretyczną oraz praktyczną)

- Rozdział teoretyczny zawierać ma prezentacje metod i narzędzi, które są wykorzystywane w podobnych analizach, oraz rozwiązujących zbliżone problemu tj. adresowania przyczyn niezadowolenia z pracy, radzenia sobie z trudnościami danych, a także jakie rozwiązania tworzą kadry zarządzające oraz menedżerowie,
- Przedstawienie ważniejszych oraz niezbędnych zagadnień dotyczących zarządzania personelem i jego zadowoleniem oraz analizy danych (teksty opinii oraz ocen),
- Część praktyczna to przedstawienie, organizacji Amazon oraz tego jaki problem dotyczący opinii pracowników występuje w organizacji i jak go rozwiązać. Analiza wstępna organizacji oraz analiza problemów powiązanych w literaturze i w innych organizacjach,
- Przedstawienie metod użytych w projekcie aplikacji. Metod związanych z analizą danych i tworzeniu wizualizacji, text mining i opracowywania działań/celów. Na tym etapie pobrane i zaimportowane dane na temat opinii pracowników wraz z ocenami firmy Amazon są analizowane i tworzone są wizualizacje przy pomocy narzędzi informatycznych,
- zaprezentowanie rozwiązania i omawianie wyników w formie wizualizacji,
- przedstawienie możliwych usprawnienia w dziedzinie zarządzania personelem w firmie Amazon w szczególności działania strategiczne mające wspomóc prace menedżerów, zalet i wad rozwiązania oraz użyteczności proponowanych rozwiązań,
- podsumowanie rozdziału praktycznego.

Zakończenie

- odniesienie do celu pracy i czy cel pracy został osiągnięty,
- podsumowanie wyników analiz/wizualizacji a także zaproponowanych rozwiązań,
- sformułowanie wniosków.

Spis treści

1. Wstęp i cel pracy.....	1
2. Rozdział teoretyczny.....	2
2.1. Eksploracja danych.....	2
2.1.1. Modele eksploracji danych.....	2
2.1.2. Narzędzia informatyczne eksploracji danych	4
2.2. Eksploracja tekstu.....	6
2.2.1. Zastosowanie eksploracji tekstu.....	6
2.2.2. Analiza danych niestrukturyzowanych	7
Bibliografia	7
Spis rysunków	7

1. Wstęp i cel pracy

Pracownicy czy to byli czy obecni mają prawo i często też chęć wyrazić opinię na temat swojego pracodawcy, stanowiska czy też warunków w pracy. Celem takiej opinii może m.in. być chęć okazania swojego zadowolenia/niezadowolenia lub poinformowania potencjalnych nowych pracowników. Należałoby zadać pytanie, czy jest więc potrzeba i chęć odczytania takiej opinii. Powodem potrzeby odczytania opinii związanej z pracą mogą się różnić. Osoba szukająca zatrudnienia w tej firmie może chcieć dowiedzieć się o niej lub też pracodawca sprawdza dlaczego pracownicy są niezadowoleni z pracy. Analiza opinii w zależności od ich formy, ilości, celu tej analizy czy wielu innych możliwych uwarunkowań może być mniej lub bardziej utrudniona. Pracownicy danej firmy lub byli pracownicy tej firmy chętnie dzielą się opiniami na temat swojej pracy (np. na forum publicznym). Można przeanalizować opinie aby wydobyć z nich potrzebne informacje. Projekt aplikacji wspomagającej analizę opinii pracowników firmy Amazon. To właśnie przedsięwzięcie, o którym głośno tak i ze względu na silną reklamę ofert pracy jak i liczne afery związane z warunkami pracy będzie obiektem badania. Problemem jest duża liczba opinii i ocen negatywnych pracowników Amazon (około 25% ocen negatywnych na portalu Indeed) oraz ich wielojęzyczność. Przekłada się to na trudności w analizie danych, część danych jest w formie słownej, a część jako ocena w skali 1-5. Utrudnienia to identyfikacje negatywnych przyczyn opinii i ocen, w konsekwencji, potencjalne podejmowanie decyzji, przez menedżerów zarządzających kapitałem ludzkim oraz menedżerów wyższych szczebli. Spada też średnia wartość ocen pracy w Amazon co może oznaczać konieczność podjęcia działań mogących zmienić ten trend. Celem pracy jest projekt aplikacji, która wspomogę menedżerów zarządzających zasobami ludzkimi lub też menedżerów wyższych szczebli w firmie Amazon, w analizie opinii i ocen oraz w zidentyfikowaniu głównych przyczyn niezadowolenia pracowników, stworzenia wizualizacji danych istotnych dla zagadnienia. Ponadto opracowania racjonalnych tj. możliwych do realizacji celów i działań strategicznych, celem poprawy ocen i zadowolenia pracowników. Aby osiągnąć cel zostanie przeprowadzony proces analityczny oraz zastosuję narzędzia informatyczne w tym wybrane środowiska analityczne. Projekt aplikacji będzie obejmować pobranie i import danych, przygotowanie danych do analizy. Następnie użycie narzędzi informatycznych, w szczególności do importu/pobrania danych ze strony internetowej, text miningu, analizy statystycznej oraz analizy wizualizacji. Analizę danych, w tym identyfikacja przyczyn, oraz istotnych kwestii dotyczących zarządzania personelem, stworzenie wizualizacji wspomagających ich odczyt. W oparciu o tę analizę zaproponowanie przyszłych celów oraz działań strategicznych.

W Rozdziale Teoretycznym zostanie przedstawione podejście do procesu analitycznego tj. cztery modele eksploracji danych następnie pokażę metody i narzędzia wykorzystywane w analizie danych, analizie opinii, tworzenia wizualizacji w tym platformy przedstawione przez raport Gartnera oraz wybrane przeze mnie platformy. Następnie postaram się przedstawić w krótki sposób eksplorację tekstu oraz wybrane podejścia, metody i techniki związane z text mining. W następnym rozdziale znajdzie się przedstawienie organizacji Amazon oraz problemu, który potencjalnie jest związany z organizacją. Pokażę również dane przedstawiające opinie internetowe pracowników firmy Amazon, które zostały wykorzystane do stworzenia analizy. Zostaną przedstawione również problemy oraz wątpliwości, które pojawiły się podczas wykorzystywania danych empirycznych. W kolejnym rozdziale przedstawię podejście, metody i narzędzia użyte w projekcie aplikacji oraz ukażę przeprowadzenie procesu analitycznego, rozwiązanie oraz wizualizacje stworzone w środowisku analitycznym. Następnie zaproponuje możliwe usprawnienia w dziedzinie zarządzania personelem w firmie Amazon w szczególności

działania strategiczne mające wspomóc prace menedżerów, zalet i wad rozwiązania oraz użyteczności proponowanych rozwiązań. Na koniec podsumujemy wnioski i efekty analizy.

2. Rozdział teoretyczny

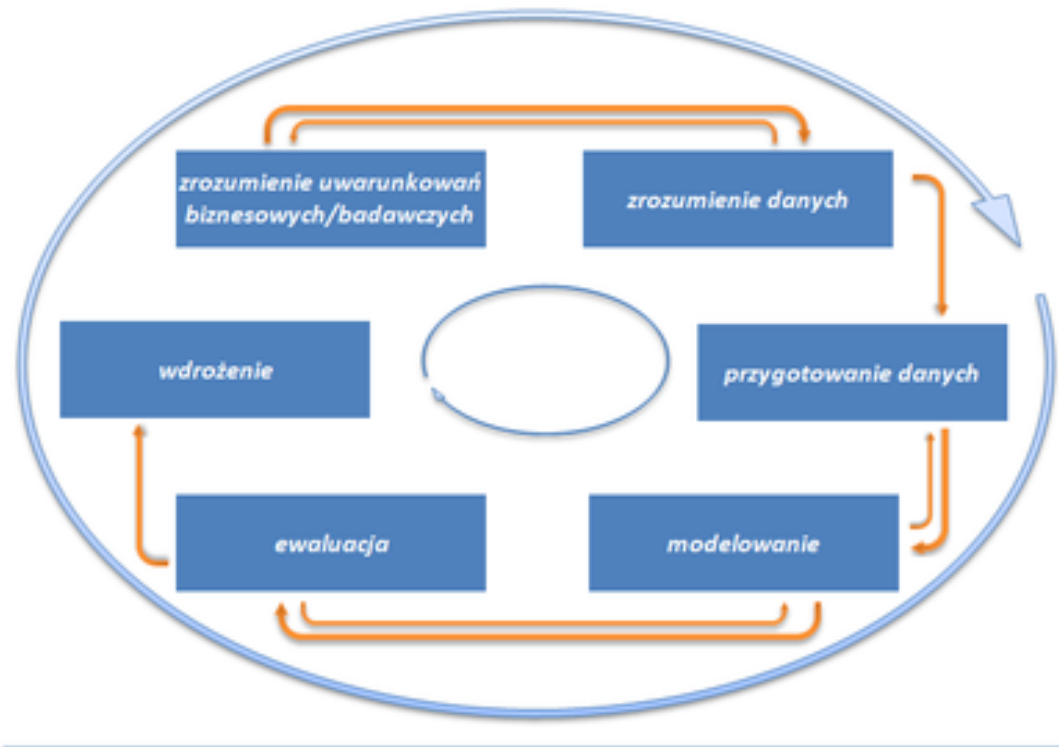
2.1. Eksploracja danych

W dzisiejszych czasach generowane oraz gromadzone są ogromne ilości danych, wiąże się to z rozwojem technologicznym. Większość tych danych przechowywana jest cyfrowo. Eksploracja danych (ang. data mining) jest procesem znajdowania istotnych zależności czy wzorców w dużych ilości danych przy użyciu metod oraz technik matematycznych i statystycznych (Bartuś i Olszak, 2009). Zadaniem metod eksploracji danych, jest automatyczne odkrywanie nietrywialnych, dotychczas nieznanych, zależności, związków, podobieństw lub trendów-wzorców w dużych zbiorach danych (T. Morzy, 2013). K. Bartuś oraz M. Olszak zaznaczają czym eksploracja danych nie jest. Według analityków eksploracja danych nie jest odkrywaniem wiedzy (ang. Knowledge Discovery in Database) i nie powinno się traktować tych dwóch pojęć jako synonimicznych, nie jest klasycznym narzędziem generującym analizy i sprawozdania, nie jest procesem automatycznym tzn. potrzebny jest człowiek, analityk uczestniczący w procesie. Autorzy podkreślają też, że eksploracja danych nie jest łatwym i szybkim procesem. Na złożoność procesu wpływa niezbędność posiadania przez ekspertów wiedzy, często w wielu dziedzinach (2009).

2.1.1. Modele eksploracji danych

Można wyróżnić kilka podejść, które w większości opracowywały przedsiębiorstwa w potrzebie ustandaryzowania metodyk. Choć ich rdzeń działania jest podobny to możemy jednak znaleźć różnice w etapach eksploracji danych. Obecnie najpopularniejszymi podejściami do eksploracji plików są: CRISP-DM, SEMMA, DMAIC, VcofDM.

CRISP-DM jest to model przedstawiający sześć faz cyklu życia procesu eksploracji danych. Jest najpopularniejszą metodologią analityki i eksploracji danych o czym świadczą np. prowadzone co kilka lat sondaże na stronie KDnuggets (G. Piatetsky-Shapiro, 2014). Twórcy metodyki łączyli ją z SPSS jednym z najbardziej znanych i najczęściej używanym oprogramowaniem do analizy statystycznej (P. Chapman i in., 2000). Należy zaznaczyć, że wstępem do procesu są dane a raczej wcześniej problem który analityk pragnie rozwiązać lub zrozumieć. Kolejność faz nie jest ściśle określona, kolejność można zmieniać w stosunku do potrzeb. Może wystąpić konieczność powrotu do poprzednich etapów lub pominięcia niektórych. Nie jest też określone czy każdy etap musi wystąpić. Na przykład jeśli jesteśmy w stanie rozwiązać nasz problem biznesowy bez etapu modelowania to ta faza nie jest niezbędna. Analityk Meta S. Brown podkreśla rolę zespołów IT w niektórych fazach takich jak zdobyciu dostępu do danych, implementacji czy modelowaniu oraz jak ważna jest współpraca analityki z technicznymi aspektami procesu (Meta S. Brown, 2015). sześć faz cyklu życia w metodyce CRISP-DM przedstawia poniższy rysunek.



Rysunek 1 Schemat ilustrujący proces eksploracji danych wg. CRISP-DM, Źródło: https://pl.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

SEMMA (ang. Sample, Explore, Modify, Model, Assess) jest to model, który został stworzony przez SAS Institute czyli producenta rozwiązań statystycznych i oprogramowania BI. Twórcy twierdzą, że SEMMA jest podejściem określającym działania w jednym z ich oprogramowań SAS Enterprise Miner. Składa się z pięciu etapów:

- próbkowania czyli użycia najpierw mniejszej części danych do modelowania. Celem próbkowania lub też próby jest poznanie widocznych wzorców i trendów,
- eksplorowanie jest fazą w której przeglądamy dane. Obejmuje zrozumienie danych w tym poznaniu oczekiwanych i nieoczekiwanych wzorców lub też związków.
- manipulacji czyli modyfikowanie lub manipulowanie danymi nie jest koniecznym etapem ale bardzo często potrzebnym. „Twórcy modelu zaznaczają, iż dynamiczny charakter eksploracji danych lub większe uaktualnienie baz danych na potrzeby eksploracji powinny być możliwe na wszystkich etapach tego procesu”(Olszak i Bartuś, 2009). Oznacza to, że tak jak w przypadku metodyki CRISP-DM możemy zmieniać kolejność niektórych faz zgodnie z potrzebą, oraz to, że eksploracja danych nie jest procesem liniowym ani stałym tylko dynamicznie się zmieniającym,
- modelowania tzn. wyboru i użyciu takich modeli czy technik, które zapewnią oczekiwany rezultat. Przykładem takiego modelu może być np. wybór drzewa decyzyjnego,
- oceny w której analizujemy wyniki pod kątem przydatności modeli (Bartuś i Olszak, 2009).

DMAIC (Define, Measure, Analyze, Improve, Control) został stworzony przez Instytut Motorili. Jest modelem, którego fundamentem jest metodyka sześć sigma. Wywodzi się z podejścia do problemów związanych z jakością. Model wyszczególnia etapy takie jak (Bartuś i Olszak, 2009).

- definiowanie - w którym określa się cele oraz definiujemy problemy biznesowe, które w późniejszych etapach będą podstawą do analizy i tworzenia usprawnień,
- pomiar – czyli zgromadzenia informacji celem poznania skali problemu oraz uzyskania odniesienia do danej sytuacji,
- analiza – gdzie poznaje się najważniejsze przyczyny wpływające na występowanie problemu oraz ich relacji w danym procesie ,
- usprawnienie – po etapie analizy proponuje się w oparciu o znalezione problemy i ich przyczyny, które będą ich rozwiązaniem,
- kontrola – w ostatnim etapie sprawdzamy zaobserwowane przez nas elementy oraz wyniki.

VCofDM (Virtuous Cycle of Data Mining) jest ostatnim z wymienionych przeze mnie modeli. Nacisk, oprócz identyfikacji problemów biznesowych, jest położony na cykliczność etapów. Michael J. A. Berry i Gordon S. Linoff opracowali model, który definiuje sukces przy eksploracji danych jako zastosowanie czterech etapów:

- identyfikacji problemów biznesowych
- wydobywania z danych informacji
- działania na podstawie wyników
- mierzenia i oceny wyników

2.1.2. Narzędzia informatyczne eksploracji danych

Użycie narzędzi IT jest nie tylko pomocne ale często i niezbędne w związku z rosnącą liczbą danych, zbiorów danych i zbiorów Big Data w dziedzinie analizy danych czy BI (ang. Business Intelligence). Rośnie również liczba możliwych narzędzi czy też systemów BI, dostępnych dla rozwijających się firm. Istnieje wiele narzędzi stosowanych w analityce i inteligencji biznesowej, które firma Gartner co jakiś czas porównuje ze sobą tworząc tak zwany kwadrant, raport który przedstawia niektóre narzędzia pod względem czterech cech. Cech takich jak bycia liderem, pretendentem, wizjonerem czy graczem niszowym.



Source: Gartner (February 2020)

Rysunek 2 Kwadrant przedstawiający platformy analityczne i BI, Źródło: Gartner, luty 2020

Raport stworzony w styczniu 2020 roku a opublikowany miesiąc później, opiera się na ocenach analityków i ekspertów. Firma zastrzega sobie, że są to właśnie opinie a praktyce występuje wiele zmiennych i ciężko stwierdzić, która platforma jest najlepsza. Niektóre narzędzia będą lepsze dla innych projektów, zadań czy potrzeb. Omówię niektóre z nich.

Tableau jest narzędziem, który pomaga zobaczyć i przeanalizować dane oraz jest jednym z liderów i wizjonerów w raporcie. Producent oferuje pięć głównych produktów: Tableau Desktop, Tableau Server, Tableau Online, Tableau Reader i Tableau Public. Mocną stroną platformy są szeroko rozwinięte możliwości wizualizacji danych a także daje względnie łatwą możliwość korzystania z języków skryptowych np. Python lub R. Do słabych stron możemy zaliczyć utrudnione aktualizowanie danych.

Power BI platforma do analityki biznesowej oraz analizy i wizualizacji danych stworzona przez Microsoft. Podobnie jak Tableau umożliwia tworzenie pulpitów (ang. Dashboardów), a dużą zaletą jest intuicyjność i możliwość korzystania z wielu zintegrowanych aplikacji takich jak np. SQL Server czy Github.

RapidMiner to platforma oprogramowania służąca do tworzenia procesów eksploracji danych czy stosowania uczenia maszynowego. Mocną stroną tego oprogramowania są duże możliwości integracji danych, transformacji danych, modelowania i wizualizacji tzn. ponad 1500 metod (H. Chahal i P. Gulia, 2016).

SAS(Statistical Analysis System) jest środowiskiem analitycznym używanym w business Intelligence, który oferuje szeroką funkcjonalność w przetwarzaniu danych, zarządzaniu danymi czy analizę statystyczną. Mocną stroną platformy SAS jest jej szerokie możliwości zastosowania oraz duża liczba aplikacji tworzona przez instytut SAS. Przykładem takiej aplikacji może być SAS text mining.

Analizując raporty z poprzednich lat możemy zaobserwować spadek IBM. Jednym z najbardziej znanych i korzystanych narzędzi od IBM jest SPSS czyli oprogramowanie do statystycznej analizy danych rozwijanej już od 1968r. Choć SPSS nadal jest ważnym graczem to spadek w pozycji w raporcie można przypisać m.in. mniejszym możliwościom wizualizacji danych w programie.

2.2. Eksploracja tekstu

Szczególnym rodzajem danych są dane tekstowe, ponieważ są to dane niestrukturyzowane. Eksploracja tekstu (ang. Text mining) przez R. Feldmana i J. Sangera zdefiniowana jest jako proces przy którym za pośrednictwem narzędzi analitycznych, użytkownik oddziałuje ze zbiorem dokumentów (R. Feldman i J. Sanger, 2006). Text mining zajmuje się analizą tekstu, eksploracją tekstu, sporządzaniem streszczeń, klasteryzacją i kategoryzacją dokumentów, znajdowaniem grup słów o podobnym znaczeniu, czy też automatycznym rozpoznawaniem wyrażen. Text mining, innymi słowami jest eksploatacją tekstu w języku naturalnym i postępuje w oparciu o słownik w danym języku (D. Zwierzchowski, 2016). Z tego powodu łatwiejszymi oraz częstszymi analizami tego typu dotyczą w języku angielskim m.in. z powodu zaimplementowanych w wielu narzędziach zestawy słów ze słownika języka angielskiego. Eksploracja tekstu w języku polskim jest utrudniona m.in. ze względu na dużo większą ilość słów niż w języku angielskim oraz dużo bogatsze odmiany słów.

2.2.1. Zastosowanie eksploracji tekstu

Aby wydobyć potrzebne nam informacje, można zastosować wiele metod. Niektóre metody mimo tego, że są to dane niestrukturyzowane są takie same jak w przypadku data mining. Feldman i Sanger wskazują na podobieństwa w data mining i text mining jako, że oba systemy zawierają w sobie etap preprocesingu, odkrywania wzorców poprzez algorytmy matematyczne czy statystyczne czy prezentacja i wizualizacja jako forma ułatwiania analizy danych (R. Feldman i J. Sanger, 2006). Możemy zastosować poniższe metody do analizy tekstu (M. Lubicz, 2020):

- Klasyfikacja dokumentów
- Grupowanie dokumentów
- Grupowanie pojęć
- Analiza powiązań
- Automatyczne rozpoznawanie języka
- Wyszukiwanie informacji
- Analiza wydźwięku

2.2.2. Analiza danych niestrukturyzowanych

Przy procesie przetwarzania dokumentów można posłużyć się szeregiem technik czy operacji na danych tekstowych, które komputer czy też dane narzędzie analityczne widzi jako szereg znaków (D. Zwierzchowski, 2016). Przedstawię poniżej przykłady takich operacji:

- Podział strumienia znaków na słowa, frazy, symbole lub żetony. Podział taki nazywamy tokenizacją.
- Pominięcie niektórych wyrazów. Na przykład tych często występujących, gdzie ich wkład do analizy jest mały czy też wyrazów niecenzuralnych.
- Łączenie słów o podobnym znaczeniu. Zdaniem Zwierzynieckiego może to poprawić w jakość danych.
- Stemming pomaga nam odnaleźć i przeanalizować rdzenie słowa dzięki czemu można uzyskać słowa w postaci bazowej. Różne słowa, które odnoszą się do tej samej rzeczy można sprowadzić do jednego terminu.
- Lematyzacja również odnosi się do szukania rdzenia słowa i sprowadzania słów do jednego opierając się na słowniku danego języka oraz kontekstu słów stojących obok siebie.
- Wyszukiwanie fraz i poprzez znalezienie słów łączenie ich w pary jeśli nie należy traktować ich osobno.

Bibliografia

- Bartuś K. i Olszak C., Analiza i ocena wybranych modeli eksploracji danych, 2009.
Chahal H. i Gulia P., Comprehensive Study of Open-Source Big Data Mining Tools, 2016.
Chapman P. i in. CRISP-DM-1.0, 2000.
Lubicz M., Wprowadzenie do analityki danych nieustrukturalizowanych, 2020.
Morzy T., Esploracja danych, 2013.
Feldman R. i Sanger J., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.
Piatetsky-Shapiro G., KDnuggets Methodology Poll, 2014. [Online]. Available: www.kdnuggets.com.
Zwierzchowski D., Text mining i narzędzia eksploracji tekstu, 2016

Spis rysunków

- Rysunek 1 Schemat ilustrujący proces eksploracji danych wg. CRISP-DM, Źródło: https://pl.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining..... 3
Rysunek 2 Kwadrant przedstawiający platformy BI, Źródło: Gartner, luty 2020..... 5