

POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: Informatyka

SPECJALNOŚĆ: Inteligentne Systemy Informatyczne

PRACA DYPLOMOWA
MAGISTERSKA

Automatyczne wydobywanie i klasyfikowanie
kolokacji z korpusów języka polskiego.

Automatic Extraction and Classification of
Collocations from Polish Corpora.

AUTOR:

Michał Jan Wendelberger

PROMOTOR:

dr inż. Maciej Piasecki, I-32

OCENA PRACY:

WROCŁAW 2015

Streszczenie

Streszczenie pracy...

SPIS TREŚCI

1. WSTĘP

1.1. WPROWADZENIE

1.2. CEL PRACY

1.3. STRUKTURA PRACY

1.4. WYJAŚNIENIE WAŻNIEJSZYCH TERMINÓW

Zamieszczona poniżej lista przedstawia wykaz ważniejszych terminów wraz z ich definicjami ustalonymi na potrzeby niniejszej pracy:

- **wyraz** - na potrzeby niniejszej pracy termin ten będzie określał dowolny ciąg znaków, w tym znaków diakrytycznych, z wyłączeniem znaków białych; w szczególności za wyraz uznawany będzie także pojedynczy znak; w niniejszej pracy termin ten jest stosowany zamiennie z określeniem *słowo*;
- **zdanie** - zbiór wyrazów w określonej kolejności oddzielonych odstępami, zakończony kropką, powstały poprzez dokonanie segmentacji tekstu ciągłego;
- **segment** - część tekstu ciągłego z zachowaniem kolejności ich składowych (przykładowo: zdanie, ich zbiór lub paragraf);
- **token** - podstawowa jednostka będąca obiektem rozważań lingwistycznych; najmniej element składowy danych tekstowych wykorzystywany w ich przetwarzaniu; najczęściej jest to pojedynczy wyraz lub zdanie, rzadziej całe segmenty;
- **częstość** - liczba, która określa ile razy w rozważanych danych tekstowych wystąpił dany byt, najczęściej słowo lub konkretny zbiór wyrazów;
- **tagset** - zestaw metadanych wykorzystywany do opisu składowych tekstów języka;
- **tager** - narzędzie wykonujące znakowanie morfo-syntaktyczne elementów tekstu z wykorzystaniem określonego tagsetu; ewentualnie spełniać może także funkcję ujednoliczania znaczenia słów w przypadku ich polisemiczności;

- **korpus** - zebrany na potrzeby analiz lingwistycznych zbiór tekstów; dane zawarte w korpusie mogą zawierać różne metadane opisujące zebrane w nim informacje; przykładem popularnych metadanych jest oznakowanie morfo-syntaktyczne słów zawartych w tym zbiorze; pożądaną cechą korpusu jest zrównoważenie polegające na zapewnieniu w nim odpowiednich stosunków ilości danych z wszystkich zakresów języka ¹, co ma zapewnić jego dużą reprezentatywność; część korpusów jest także określona mianem równoległych, co oznacza, że każdemu tekstowi zawartemu w korpusie przyporządkowany jest tekst z co najmniej jednego innego języka, a zestawione mogą być na różnym poziomie - zdania czy akapitu;
- **słowo centralne w korpusie** - termin wprowadzony na potrzeby niniejszej pracy przez jej autora; określenie wyrazu, który podczas przeglądania korpusu jest aktualnie czytany;

¹ Zakres języka rozumiany jako jeden z obszarów jego użycia, na przykład mowa potoczna, słownictwo naukowe, teksty pseudonaukowe, wiadomości, terminologia związana ze sztuką czy innymi dziedzinami wiedzy lub działalności ludzkiej itp.

2. DEFINICJA KOLOKACJI

2.1. PRÓBY DEFINICJI TERMINU “KOLOKACJA” W LITERATURZE

Termin *kolokacja* został pierwszy raz użyty w 1930 roku przez jego autora, angielskiego lingwistę nazywającego się John Rupert Firth. Początkowo zwrot ten miał opisywać charakterystyczne i typowe dla języka kombinacje słów, które łączą się ze sobą w konkretny sposób dostarczając dzięki temu cenną wiedzę i informacje o tym języku [?, str. 15]. Firth sformułował zwrot *You shall know a word by company it keeps!* [?], którym zwrócił uwagę ludzi zajmujących się lingwistyką na fakt, że teksty nie są tylko zbitką losowo występujących słów ograniczonych jedynie przez zasady składniowe języka [?, str. 15]. Zwrot ten w wolnym tłumaczeniu oznacza, że powinniśmy być w stanie określić słowo na podstawie innych wyrazów w jego otoczeniu.

Trudność jednoznacznego i ścisłego zdefiniowania kolokacji została opisana w obrazowy sposób przez Choueka i przytoczona przez Stefana Everta - *Even though any two lexicographers would agree that 'once upon a time', 'hit the road' and similar idioms are collocations, they would most certainly disagree on almost anything else.* [?, str. 15]. Cytat ten jest metaforą obrazującą bardzo niską korelację oceny wyrażen wielowyrazowych jako kolokację nawet pomiędzy lingwistami.

Na przestrzeni ponad osiemdziesięciu ostatnich lat definicja terminu *kolokacja* była wielokrotnie modyfikowana, dostosowywana i poddawana uszczegółowieniu co sprawiło, że pojawiło się wiele jej wersji. Powstałe definicje można jednak w większości przyporządkować do jednej z dwóch grup reprezentujących różne podejścia do zagadnień kolokacji: *distributional* oraz *intensional* [?, str. 15]. Pierwsza z tych grup, zwana także *szkołą Neo-Firthańską* opiera się głównie na wykorzystaniu informacji pozyskanych w sposób empiryczny z przebadanych zbiorów danych [?, str. 15]. Druga grupa to natomiast koncept bardziej teoretyczny, skupia się na tym, że kolokacje są bytami umiejscowionymi pomiędzy wolną kombinacją słów a idiomami, parami składającymi się z jednego słowa wolnego, zwanego bazą i drugiego zdeterminowanego leksykalnie - podejście teoretyczne do tematu wyrażen wielowyrazowych [?, str. 16]. W dalszej części tej pracy na potrzeby realizacji tematu będę starał się określić definicję bliższą pierwszej z tych dwóch grup.

Liczbę wielu różnych definicji jednostki wielowyrazowej odzwierciedla także liczba zwrotów o praktycznie tym samym znaczeniu, które są używane zamiennie jako terminy określające kolokację, przykładowo: *jednostka wielowyrazowa* (ang. *MWU*), *wyrażenie wielowyrazowe* (*MWE*), *n-gramy* dla wieloelementowych kolokacji czy wręcz po prostu *idiomy* [?, str. 16].

2.2. CECHY WYRAŻEŃ WIELOWYRAZOWYCH

2.2.1. Częściowa lub całkowita niekompozycyjność

Specjalną cechą kolokacji jest ich ograniczona kompozycyjność lub całkowity jej brak. Cecha ta jest uwzględniona w większości definicji kolokacji w literaturze. Wyrażenie językowe w pełni kompozycyjne to takie, którego znaczenie może być przewidziane jedynie na podstawie sumy znaczeń jego elementów składowych [?, str. 151, 184].

Jednostki wielowyrazowe mogą być częściowo lub całkowicie niekompozycyjne co oznacza, że ich znaczenie ulega pewnej zmianie, ma pewną wartość dodaną do niego w stosunku do sumy znaczeń składowych. Zmiana ta może być niewielką modyfikacją sensu wyrażenia lub całkowicie zmieniać przekaz kolokacji. Warto wspomnieć, że większość anglojęzycznych kolokacji jest częściowo kompozycyjnych [?, str. 151].

Przykład niewielkiej zmiany sensu wyrażenia zobrazować można na podstawie dwóch wyrażen: *białe włosy* i *białe wino*. Obie frazy zawierają w sobie słowo *białe* określający kolor pewnego obiektu - rzeczownika. Przymiotnik ten określa pewien odcień bieli, ale jednak nieco inny dla każdego z wyrażen. W przypadku białych włosów kolor powinien być lekko szarawy, srebrzysty, a w przypadku wina o lekkim, żółtym zabarwieniu.

Dodatkowo za przykład niewielkiej wartości dodanej do znaczenia kolokacji można podać wyrażenie *czerwona kartka*. Faktycznie obiekt ten jest czerwonym kartonikiem, ale symbolizuje on coś jeszcze - wielokrotne lub poważne przewinienie zawodnika piłki nożnej podczas meczu.

Przykładem całkowitego oderwania znaczenia wyrażenia od sumy znaczeń jego składowych są idiomy takie jak na przykład *nawarzyć piwa* czy *wyjść jak Zabłocki na mydle*. Pierwsze z nich mówi o zaistnieniu pewnej sytuacji z powodu czynów danej osoby, której jednocześnie ta zazwyczaj nieprzyjemna sytuacja dotyczy. Znaczenie tego zwrotu w zasadzie nie ma nic wspólnego z sumą znaczeń jego elementów składowych, zwłaszcza że wspomniany trunek jest zazwyczaj przez wiele osób ceniony i chętnie spożywany.

Idiomy jasno obrazują, że są kolokacje, których znaczenie nie może być nawet przybliżone na podstawie sumy znaczeń składowych tego wyrażenia, ponieważ jest ono zupełnie inne niż znaczenie tego zwrotu rozpatrywanego jako całość. W związku z powyższym w celu poprawnego określenia znaczenia jednostek wielowyrazowych należy rozpatrywać je jako całość, a nie skupiać się jedynie na każdej z jej składowych z osobna.

2.2.2. Niezmiennność szyku

Szyk elementów składowych wyrażenia wielowyrazowego może być zmienny, tego typu kolokacje są trudniejsze do wykrycia niż te o stałym szyku, co wymaga innego podejścia przy ich wyszukiwaniu.

Jeśli rozważymy pojęcia *Unia Europejska* czy *ptasie mleczko* to zmiana szyku poprzez zamianę miejscami składowych kolokacji w obrębie tych wyrażen wielowyrazowych na

odpowiednio *Europejska Unia* oraz *mleczko ptasie* sprawi, że ich znaczenie ulegnie zmianie. Jeśli przyjęta definicja kolokacji uwzględnia tę cechę to według niej dwa powyższe zwroty powinny zostać uznane za kolokację w kontekście tej właśnie cechy.

2.2.3. Nieciągłość

Kolejną cechą niektórych jednostek wielowyrazowych jest ich nieciągłość wyrazowa. Wyrażenie nieciągłe składa się z określonych elementów tworzących kolokację, ale jednocześnie pomiędzy nimi znajdują się wyrazy, które do niej nie należą.

Przykładem wyrażenia nieciągłego może być *druga, straszna wojna światowa*. Kolokacją jest tutaj termin *druga wojna światowa*, a przymiotnik *straszna*, mimo że znajduje się pomiędzy jej elementami, nie powinien być uznany za składową tego wyrażenia wielowyrazowego.

Bardziej problematyczny do wykrycia przykład to analogia do podanego w artykule [?, str. 1] dotyczącego byłej premier Wielkiej Brytanii - *Margaret Thatcher*. Przy rozważaniu kolokacji *Karol Wojtyła* w tekście o byłym papieżu istnieje duże prawdopodobieństwo, że słowo *Wojtyła* wystąpi zaraz po wyrazie *Karol* lub odwrotnie. Można sobie jednak wyobrazić sytuację, w której pierwszy człon tej jednostki wielowyrazowej wystąpi w jednym zdaniu, a drugi w innej jego części lub nawet dopiero w zdaniu kolejnym. Oba słowa tej kolokacji zachowują się wtedy niczym synonimy określające tę samą osobę, ale nie występują w swoim bezpośrednim otoczeniu.

Powyższe przykłady obrazują, że elementy składowe kolokacji nie muszą występować bezpośrednio po sobie, aby rozważany zwrot uznać za jednostkę wielowyrazową.

2.2.4. Niezastępowalność składniowa

Istotną cechą kolokacji jest niezastępowalność jej elementów składowych synonimami tych wyrazów z zachowaniem znaczenia tego wyrażenia wielowyrazowego. Dla przykładu nie można zamienić zwrotu *białe wino* na *żółte wino*, mimo że drugie wyrażenie opisuje obiekt równie dobrze rodzaj trunku co pierwsze, a może i nawet trafniej. Jednak drugi z tych zwrotów nie powinien zostać uznany za kolokację [?, str. 184]. Podobnie ma się przykład z kolokacją *czerwona kartka* - zamiana rzeczownika *kartka* na *kartonik* lub nawet *karta* zmieni znaczenie tego wyrażenia, a tym samym fraza ta może zostać uznana za wyrażenie wielowyrazowe w kontekście tej cechy.

2.2.5. Niemodyfikowalność

Kolokacja posiadająca tę cechę nie może być dowolnie rozszerzana lub modyfikowana poprzez zmianę szyku albo dodawanie do niej dodatkowych słów. Ważne dla zachowania sensu wyrażenia wielowyrazowego jest także pozostawienie liczby, w której występuje rzeczownik [?, str. 184]. Innymi słowy wyrażenia wielowyrazowe posiadające tę cechę muszą pozostać w jednej określonej formie.

Przykładowo idiom *kopnąć w kalendarz* jest kolokacją, ale jego rozszerzenie do postaci *kopnąć butem w kalendarz*, *kopnąć mocno w kalendarz* lub *kopnąć w wielki kalendarz* sprawi, że przekaz jaki niesie ze sobą to wyrażenie mocno się zmieni.

2.2.6. Przynależność domenowa

Wiele kolokacji jest mocno związanych z wiedzą i tematyką dziedzinową. Żargon pewnej grupy osób, np. inżynierów jest często dla nich hermetyczny i zawiera wiele technicznych pojęć przez co bywa niezrozumiały dla laików w danej dziedzinie wiedzy. Ponadto słowa znane osobom niewtajemniczonym bywają używane w inny sposób, do określenia innych rzeczy czy zjawisk wewnątrz określonej grupy [?, str. 4].

Przykładem zwrotu związanego z informatyczną wiedzą dziedzinową może być zwrot *twardy reset* polegający na odłączeniu zasilania komputera w sposób mechaniczny lub inne wyrażenie, które stało się już dość powszechne i przeniknęło do języka codziennego - *program się powiesił* czyli przestał odpowiadać na akcje użytkownika lub obliczenia przez niego wykonane utknęły w martwym punkcie.

2.3. PRZYKŁADOWE DEFINICJE JEDNOSTKI WIELOWYRAZOWEJ STOSOWANE W LITERATURZE

Przykładowe definicje kolokacji z różnych źródeł wraz z krótkim komentarzem do nich zostały zamieszczone na poniższej liście:

1. Manning i Schütze - *collocations correspond to some conventional way of saying things*. [?, str. 151];
kolokacje to sposób w jaki przyjęło się mówić pewne rzeczy.
2. Manning i Schütze - *A collocation is any turn of phrase or accepted usage where somehow the whole is perceived to have an existence beyond the sum of the parts.* [?, str. 29];
kolokacja to wyrażenie, którego znaczenie wykracza poza sumę znaczeń jej elementów składowych - ma pewną wartość dodaną do znaczenia lub ulega ono zmianie.
3. Choueka - *a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components* [?, str. 1];
wyrażenie wielowyrazowe to jednostka syntaktyczna i semantyczna, której dokładne znaczenie nie może być określone bezpośrednio na podstawie znaczeń jego składowych - definicja analogiczna w swym sensie do zamieszczonej przez Manninga i Schütze.

4. Stefan Evert - *A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon.* [?, str. 17];

podobna definicja co poprzednicy jednak z dodaniem części, mówiącej o tym, że kolokacja to związek, który powinien zostać zamieszczony w słowniku. Rozszerzenie tej definicji było inspirowane pracami Choueka, który opracował sugestywną instrukcję do oceny czy dane wyrażenie jest kolokacją, a której esencja wyciągnięta przez Stefana Everta została sprowadzona do postaci pytania: *Does it deserve a special entry in a dictionary or lexical database of the language?* - czy wyrażenie zasługuje na wpis w leksykonie? [?, str. 17].

Mimo ogromnej liczby definicji terminu *kolokacja* zazwyczaj większość z nich jest zgodna co do trzech cech, które wyrażenie powinno spełniać, aby zostać uznane za wyrażenie wielowyrazowe, a są to: semantyczna niekompozycyjność, niezmienność syntaktyczna oraz niemożliwość podmiany słów składowych nawet na ich synonimy [?, str. 16]. Wyjaśnienie tych oraz innych cech jednostek wielowyrazowych zostało zamieszczone w dalszej części tej pracy.

2.4. PRZYJĘTA DEFINICJA KOLOKACJI

Na potrzeby niniejszej pracy przyjęto następującą definicję kolokacji zamieszczoną poniżej:

... definicja kolokacji ...

3. ISTNIEJĄCE SPOSOBY WYDOBYWANIA KOLOKACJI OPISANE W LITERATURZE

Znaczna część dotychczasowych prac nad wydobywaniem kolokacji ogranicza się w dużym stopniu tylko do badania wyrażen dwuelementowych. Jeśli natomiast badane są kolokacje o większej liczbie elementów to często podchodzi się do tego stosując uproszczone modele statystyczne. Dodatkowo sposoby wyszukiwania jednostek wielowyrazowych w tekstach języka polskiego nie były do tej pory często badane, zwłaszcza w kontekście wyrażen o długości większej niż dwóch.

Przykładem pracy która dotyczy innego zagadnienia niż kolokacje, ale traktującej także o ekstrakcji wyrażen dwuelementowych z tekstów języka polskiego jest praca magisterska Aleksandra Buczyńskiego [?].

W tej części niniejszej pracy, na podstawie literatury skupię się na opisie metod wydobywania wyrażen dwuelementowych z niepolskojęzycznych tekstów, głównie anglojęzycznych. Zaznaczyć trzeba jednak, że omówione tutaj metody mogą być także stosowane dla języka polskiego i innych języków słowiańskich i o bogatej fleksji. Ponadto dla lepszego zrozumienia przedstawianych metod podawane będą przykłady ich zastosowania przy zadaniu ekstrakcji bi-gramów, a w dalszej części tego rozdziału zamieszczone zostanie także wprowadzenie do miar mogących posłużyć do wydobywania kolokacji wieloelementowych - dłuższych niż dwie składowe.

3.1. METODA ZLICZANIA

3.1.1. Opis metody

Pierwsza z przytoczonych w tym rozdziale metod ekstrakcji kolokacji jest jednym z najprostszych i dość naiwnych sposobów polegającym na wykorzystaniu podstawowej cechy opisującej słowa w korpusie - ich częstości. Cecha ta określa ile razy w rozważanych danych tekstowych wystąpił konkretny token, gdzie ten może być rozumiany jako pojedyncze słowo, cały n-gram lub inne zestawienie określonej liczby wyrazów.

Technika zliczania korzysta z założenia mówiącego o tym, że jeśli dany zestaw wyrazów występuje w tekście w określonej kolejności bardzo często, częściej niż inne tego typu zbiory słów, to jest to znak, że rozważane zestawienie ma pewne specjalne znaczenie, może nieść ze sobą interesujące informacje i należy zwrócić na nie uwagę [?, str. 153].

Niestety w tym podejściu pojawia się problemem związany z częstościami słów taki, że przy zliczaniu liczby wystąpień danych wyrazów lub ich zbiorów w konkretnym zesta-

wie tekstów będących podzbiorem wszystkich tekstów języka dokonuje się na nim pewnej generalizacji, która jest jedynie przybliżeniem rzeczywistości, ponieważ powstała jedynie na podstawie pewnej próbki losowej tekstów tego języka. Taka generalizacja jest pożądana o ile jest poprawna - reprezentacyjna dla całego języka, ale jednak bardzo trudnym zadaniem jest jej utworzenie na podstawie tylko części tekstów rozważanego języka [?, str. 20]. Sama generalizacja polega na tym, że jeśli w rozważanym tekście pewien zbiór wyrazów został uznany za jednostkę wielowyrazową to z punktu widzenia całego języka także będzie to kolokacja.

Mimo problemu związanego z niepewną generalizacją wydającego się być poważnym warto przedstawić tę metodę i osiągnięte przez badaczy wyniki z jej wykorzystaniem, ponieważ może być to dobry punkt wyjścia do rozważań nad innymi sposobami wydobywania kolokacji. Dzięki przybliżeniu tej metody pozyskać można podstawową wiedzę na temat ekstrakcji kolokacji, a ponadto obserwacje badaczy mogą dostarczyć wielu cennych informacji.

Metoda częstości polega na zliczeniu wszystkich wystąpień każdego z tokenów i wykorzystaniu założenia wspomnianego wcześniej w tym podrozdziale, takiego że:

Jeśli dane słowa występują w tekście razem bardzo często to znaczy, że dany zestaw wyrazów spełnia jakąś ważną funkcję, której nie można w prosty sposób wyjaśnić jako funkcji będącej jedynie wynikiem kombinacji tych wyrazów [?, str. 153].

Korzystając z tego założenia, jeśli jest ono spełnione to dane współwystępowanie wyrazów powinno zostać uznane za kolokację lub przynajmniej za ciekawego kandydata na nią.

3.1.2. Wyniki metody

Przedstawione w [?, str. 154] wyniki badań przeprowadzonych przez Manninga i Schütze obrazują bardzo złe wyniki tej metody i wskazują słaby punkt metody. Powodem złych wyników przedstawionych przez wspomnianych powyżej autorów jest przewaga częstości występowania słów funkcyjnych języka oraz bi-gramów z nich złożonych, głównie anglojęzycznych zwrotów *of the* oraz kilku innych współwystąpień z wyrazem *the*. Spośród dwudziestu najczęściej występujących dwuelementowych zwrotów, aż dziewiętnaście z nich to właśnie wyrażenia funkcyjne języka angielskiego, które nie są kolokacjami. Tylko jedna pozycja z listy może zostać uznana za wyrażenie wielowyrazowe i jest to nazwa miasta - *New York* [?, 154]. Jednak jakość wyników tej metody okazała się zgodna z przewidywaniami autorów [?] - zwykle wybieranie najczęściej współwystępujących wyrazów nie prowadzi do interesujących wyników w dziedzinie badania sposobów wydobywania kolokacji [?, str. 153].

3.1.3. Rozszerzenie algorytmu o filtr części mowy

Manning i Schütze w swojej pracy przedstawili także sposób na znaczną poprawę jakości wyników poprzez zastosowanie prostej heurystyki w postaci kilku filtrów opartych o części mowy. Pomysł na taką technikę poprawy wyników został zaczerpnięty z pracy Justesona i Katza wspomnianej w [?, 154]. Kandydaci na wyrażenia wielowyrazowe, którzy nie spełniali określonych wzorców składniowych części mowy, a konkretnie postaci przymiotnik-rzeczownik lub rzeczownik-rzeczownik [?, 155] byli wykluczani z listy potencjalnych kolokacji.

3.1.4. Wyniki po zastosowaniu filtra

Wyniki po filtrowaniu okazały się znacznie lepsze od metody bez filtrowania, ponieważ tym razem z dwudziestu par wyrazów, które wystąpiły najczęściej w badanych tekstach autorzy nie zakwalifikowaliby jedynie trzech bi-gramów jako niekompozycyjne [?, 155].

Dużą zaletą tej heurystyki jest fakt, że można ją zastosować także w połączeniu z innymi metodami wydobywania kolokacji, także z tymi, które zostaną przez autora tej pracy przytoczone w dalszej części tego rozdziału.

Zastosowanie takiego filtrowania wymaga jednak dziedzinowej wiedzy lingwistycznej lub badań w celu oceny, które połączenia części mowy mogą ewentualnie tworzyć ciekawych kandydatów na wyrażenia wielowyrazowe, a które raczej nie mają ku temu tendencji. Jako przykład badań mogących pomóc w doborze filtrów dla języka czeskiego będącego bardziej skomplikowanym w analizie niż język angielski można podać artykuł autorstwa Pavla Peciny *Reference Data for Czech Collocation Extraction* prezentujący częstość występowania par słów wraz z ich częściami mowy w zbadanych przez niego zestawach danych [?].

Trudnym zadaniem może być też wyznaczenie filtrów, których należy użyć w procesie wydobywania kolokacji, a których nie warto. Zwiększanie liczby filtrów może zaowocować wzrostem kompletności, ale spadkiem precyzji, zwłaszcza kiedy dane zestawienia części mowy w danym filtrze będą rzadko tworzyć kolokacje w stosunku do liczby wszystkich wyrażen przez nie generowanych.

3.1.5. Wnioski

Manning i Schütze zwracają uwagę na istotny wniosek płynący z badań nad tą prostą metodą wydobywania kolokacji. Zauważają, że nawet użycie prostych technik ilościowych wspieranych przez niewielką wiedzę lingwistyczną może dać wręcz niespodziewanie dobre wyniki w zadaniu automatycznej ekstrakcji wyrażen wielowyrazowych [?, str. 155, 157].

3.2. WARIANCJA I ODLEGŁOŚĆ SŁÓW

3.2.1. Wstęp i motywacje

Metoda zliczania wzbogacona o filtrowanie oparte o części mowy słów sprawdza się nad wyraz dobrze dla języka angielskiego przy wydobywaniu kolokacji ciągłych, ale niestety ta technika nie sprawdzi się tak dobrze przy ekstrakcji jednostek wielowyrazowych o zmiennym szyku lub z przerwami pomiędzy składowymi kolokacji - wyrażenia nieciągłe [?, str. 157]. Sposobem mogącym pomóc w radzeniu sobie z tym problemem może być metoda oparta o średnią odległość słów oraz jej wariancję i odchylenie standardowe.

3.2.2. Opis metody

Technika wykorzystuje okno przesuwne o określonej długości oznaczającej rozmiar otoczenia aktualnie rozpatrywanego słowa w korpusie. Innymi słowy okno przesuwne zawiera w sobie słowo centralne i jego kontekst. Długość okna jest liczbą całkowitą określającą liczbę wyrazów po każdej ze stron słowa centralnego, które w danym momencie będą rozważane [?, str. 158]. Dla przykładu zastosowanie okna o długości dwóch wyrazów, przesuwanego za każdym razem o jedno słowo w przykładowym zdaniu *Niestety trzeba przyznać, że pogada dzisiaj nam nie dopisała* oraz przy pominięciu znaków interpunkcyjnych spowoduje rozpatrzenie kolejno takich n-gramów jak *Niestety trzeba przyznać*, *Niestety **trzeba** przyznać że*, *Niestety trzeba **przyznać** że pogada*, *trzeba przyznać **że** pogada dzisiaj* i tak dalej, aż okno zostanie przesunięte do końca aktualnie rozważanego zdania. Słowa pogrubione to słowa, na które ustawione było okno przesuwne w danej iteracji - słowa centralne, aktualnie rozważane w korpusie [?, str. 158]. Przypadek braku wystarczającej liczby słów w oknie może być rozpatrzony dwojako - można wtedy wziąć pod uwagę mniejszą liczbę wyrazów do rozważań lub ominąć ten fragment tekstu i przesunąć okno dalej.

Dla każdego przyłożenia okna tworzeni są kandydaci na kolokacje o konkretnej długości wyrażonej w liczbie słów. Dla uproszczenia ograniczymy rozważania dla wyrażeń dwuelementowych. Kandydaci tworzeni są w taki sposób, że kreowane i zapamiętywane są wszystkie możliwe kombinacje o długości dwóch słów spośród znajdujących się aktualnie w granicach okna. Istotne jest, aby ustalić czy po dodaniu kandydatów z jednego przyłożenia okna powielać ich po przesunięciu tego okna, czy jedynie dodać pary powstałe dzięki nowemu słowu w oknie co wydaje się bardziej rozsądnym posunięciem.

Przykład tworzenia kandydatów na kolokacje w obrębie jednego przyłożenia okna został opisany poniżej. Rozważmy zdanie zamieszczone poniżej z wykorzystaniem okna o długości dwóch wyrazów, gdzie tekstem pogrubionym zostało oznaczone słowo centralne - to do którego w danym momencie zostało przyłożone okno przesuwne. Przyjmijmy też przeskok okna równy jednemu wyrazowi - okno przesuwane jest o jeden wyraz w stosunku do poprzedniej pozycji z każdym nowym przyłożeniem.

*Zdenerwowało mnie **Twoje** wczorajsze zachowanie.*

Na podstawie przytoczonego zdania można utworzyć dwadzieścia różnych kombinacji dwuelementowych, a pięć z nich zostało zamieszczonych w poniższej tabeli 3.1.

Nr.	wyrażenie	słowa zdania
1.	Zdenerwowało mnie	pierwsze, drugie
2.	Zdenerwowało Twoje	pierwsze, trzecie
3.	Zdenerwowało wczorajsze	pierwsze, czwarte
4.	Zdenerwowało zachowanie	pierwsze, piąte
5.	mnie Twoje	drugie, trzecie
6.

Tabela 3.1. Wyrażenia dwuelementowe utworzone z przykładowego zdania za pomocą przykładowego okna przesuwne

Dla każdej kombinacji słów utworzonej we wszystkich przyłożeniach okna przesuwne-
go na przestrzeni całego korpusu zapamiętywane są odległości pomiędzy wyrazami w ten
sposób badając w jakich odległościach od danego słowa występują inne konkretne wyrazy.
Przykładem niech będą trzy wyrażenia, pierwsze *Duże czerwone buty*, drugie *Buty koloru
czerwonego* oraz trzecie *Moje ulubione buty czerwone*. W pierwszym zwrocie odległość
słowa *buty* od wyrazu *czerwone* jest równa jeden, ponieważ słowo *buty* wystąpiło zaraz po
słowie *czerwone*, natomiast w drugim odległość wyniosła minus dwa słowa, gdyż wyraz
buty wystąpił przed przymiotnikiem *czerwonego* dodatkowo z jednym wyrazem pomię-
dzy nimi. Trzeci termin jest przykładem analogicznym do poprzedniego, odległość słowa
buty od wyrazu *czerwony* także jest ujemna i z tego samego powodu, ale tym razem równa
minus jeden - w tym przypadku rzeczownik także wystąpił przed przymiotnikiem, ale tym
razem bezpośrednio przed nim. Średnia odległość słowa *buty* od wyrazu *czerwone* będzie
zatem równa:

$$\bar{d} = \frac{1}{3}(1 + (-2) + (-1)) = -\frac{2}{3}$$

Natomiast wariancja na podstawie [?, str. 159]:

$$s^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1} = \frac{(1 - (-\frac{2}{3}))^2 + (-2 - (-\frac{2}{3}))^2 + (-1 - (-\frac{2}{3}))^2}{3-1} \approx \frac{2.78 + 7.11 + 0.11}{2} = 5$$

Na podstawie zebranych z korpusu informacji o odległościach słów w obrębie każdego
przyłożenia okna przesuwne należy obliczyć średnią odległość występowania wszyst-
kich kombinacji słów będących kandydatami na kolokacje. Dodatkowo wymagane jest
także obliczenie wartości wariancji i odchylenia standardowego ($s = \sqrt{s^2}$) odległości
słów dla każdego z tych kandydatów, a otrzymane w ten sposób wartości posłużą do oce-
ny kandydujących wyrażen wieloelementowych. Niska wartość odchylenia standardowego
oznacza, że słowa składowe wyrażenia zazwyczaj występują względem siebie w podob-
nej odległości. Natomiast odchylenie standardowe równe zero należy interpretować jako

sytuację, gdy wszystkie wyrazy występują zawsze w tej samej odległości oraz kolejności. [?, str. 158, 159]

Średnia odległość słów względem siebie nie niesie informacji pozwalających ocenić czy rozpatrywany kandydat na wyrażenie wielowyrazowe jest interesujący czy nie, jest to prostu wiedza w jakiej średniej odległości dane wyrazy występują względem siebie oraz zazwyczaj w jakiej kolejności, chyba że użyta definicja kolokacji nie pozwala na występowanie wyrażen nieciągłych - w takiej sytuacji wartość ta może okazać się ciekawym dyskryminatorem. Ze względu na zastosowanie średniej arytmetycznej należy mieć na uwadze długość zastosowanego okna, ponieważ przy niewielkich częstościach nawet pojedyncze wystąpienia obserwacji mocno odstających mogą mocno zmodyfikować wartości średniej - miara ta jest na nie wrażliwa. Dlatego z tego powodu podczas oceny wartości średniej odległości międzywyrazowej ważne jest, aby brać pod uwagę także wariancję lub odchylenie standardowe.

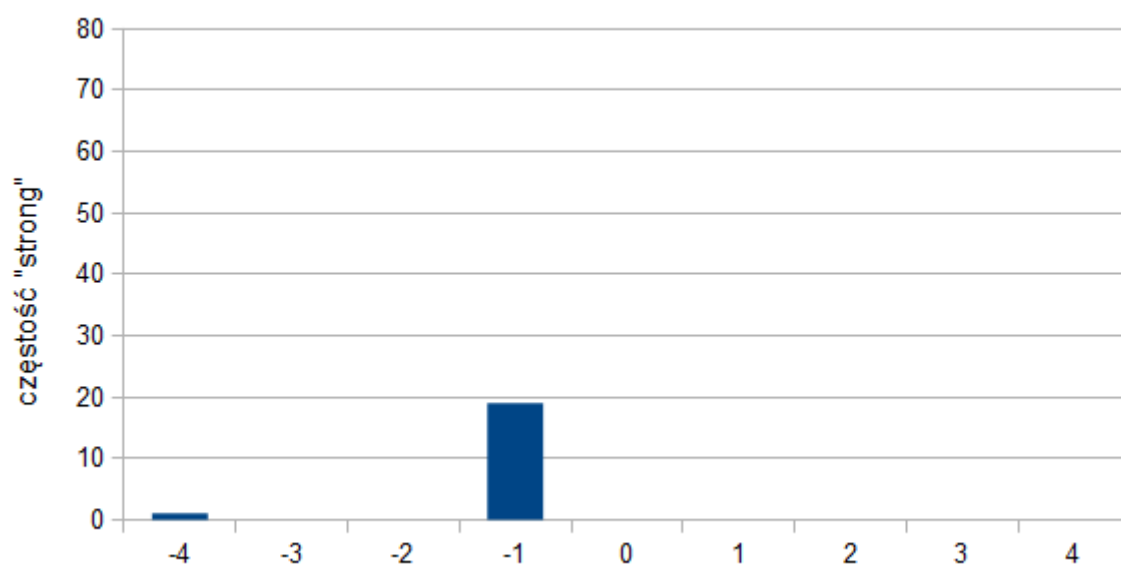
3.2.3. Wyniki działania algorytmu

Wyniki otrzymane tą metodą mogą być łatwiejsze do oceny w przypadku ich wizualizacji za pomocą nawet prostych wykresów. Manning i Schütze w swojej pracy zamieścili trzy wykresy odnośnie trzech różnych anglojęzycznych wyrażen: *strong opposition*, *strong support* oraz *strong for*. Warto wspomnieć, że badania wykonane były z oknem przesuwającym o rozmiarze czterech słów. Poniższe wykresy 3.1, 3.2 oraz 3.3 zostały wykonane na podstawie danych z wykresów zamieszczonych w pracy powyżej wymienionych autorów i są odpowiednikami ich wykresów [?, str. 160].

Pierwszy z wykresów (3.1) składa się w zasadzie z pojedynczego, wysokiego piku, obrazującego, że większość wszystkich wystąpień słowa *strong* razem z *opposition* miało miejsce w odległości minus jeden, czyli słowa te występowały prawie zawsze po sobie w ustalonej kolejności tworząc zwrot *strong opposition*. Dodatkowo według Manninga i Schütze wartość odchylenia standardowego jest niska i równa 0.67, a średnia odległość to -1.15 słowa. Taki wynik mimo dużego skupienia w okolicy argumentu minus jeden został spowodowany szumem w danych - pojedynczym wystąpieniem tej pary słów w odległości równej minus cztery wyrazy [?, str. 159].

Na drugim z wykresów (3.2) zamieszczone zostały wyniki, z których odczytać można, że słowa *strong* i *support* występowały zazwyczaj w odległości minus jeden od siebie, ale wystąpiło także dość dużo obserwacji takich, że ta odległość była większa, jednak prawie zawsze ujemna. Zaobserwować można także prawie monotoniczność wykresu, przy pominięciu mało znaczących, pojedynczych wystąpień w odległościach dwa i trzy. Zwiększenie zakresu argumentów, dla których występują znaczące wartości określające liczbę wystąpień miało wpływ na wzrost wartości wariancji do poziomu 1.07 [?, str. 161].

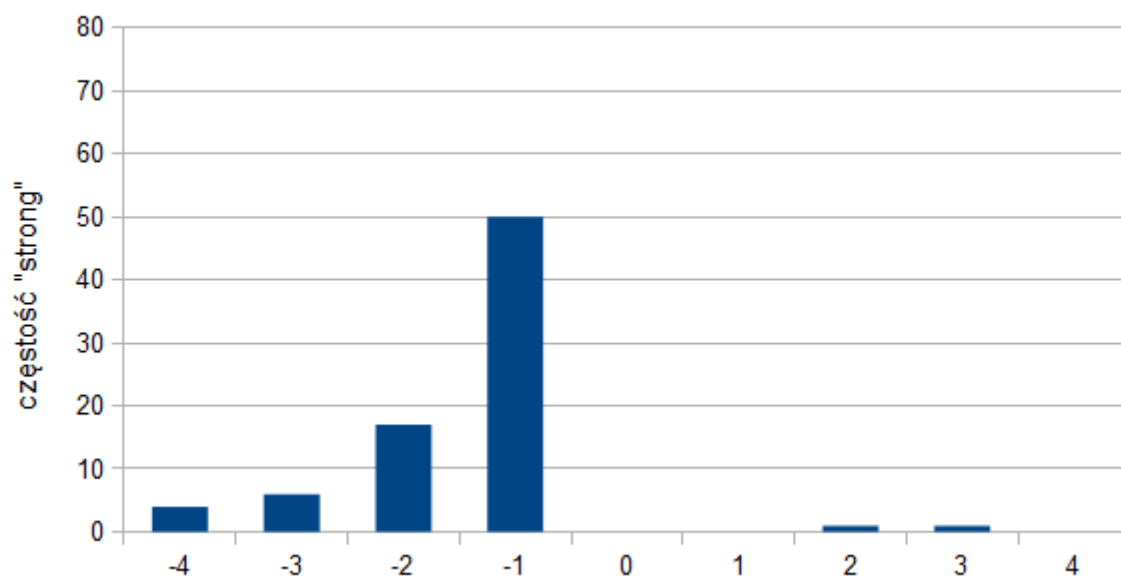
Trzeci wykres (3.3) obrazuje wyniki dla zwrotu niebędącego kolokacją składającego się z pary wyrazów *strong* oraz *for*. Wartości na osi rzędnych dla różnych argumentów -



Rys. 3.1. Słowo *strong* w odniesieniu do słowa *opposition*

$$(\bar{d} = -1.15, s = 0.67)$$

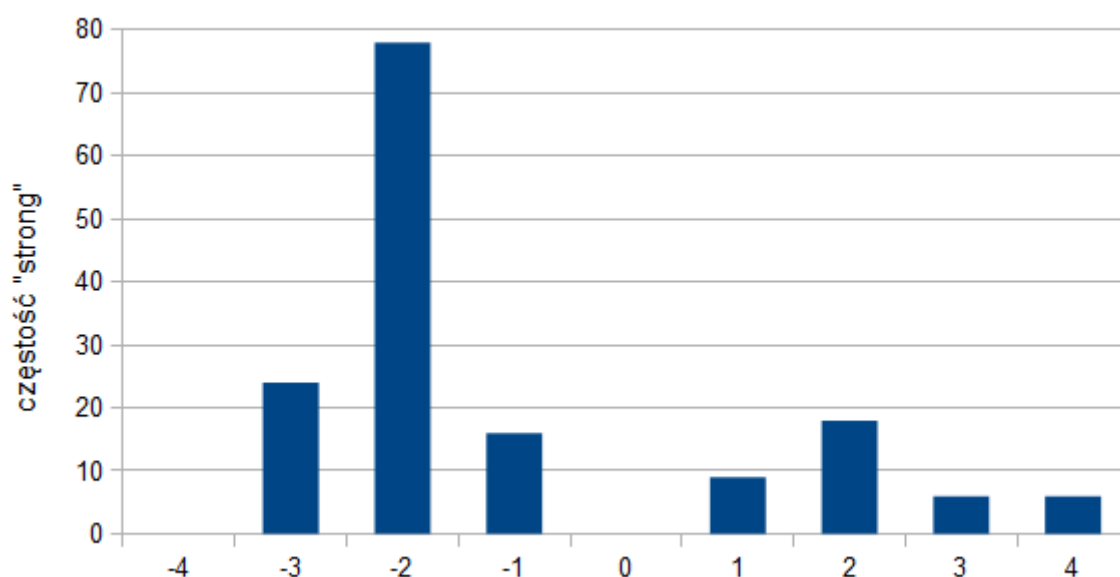
na podstawie wykresów z [?, str. 160]



Rys. 3.2. Słowo *strong* w odniesieniu do słowa *support*

$$(\bar{d} = -1.45, s = 1.07)$$

na podstawie wykresów z [?, str. 160]



Rys. 3.3. Słowo *strong* w odniesieniu do słowa *for*

$$(\bar{d} = -1.12, s = 2.15)$$

na podstawie wykresów z [?, str. 160]

wartości odległości, osiągają znaczące poziomy nie wykazując żadnej monotoniczności czy skoncentrowania w okolicy jednej z wartości odległości jak to miało miejsce w przypadku dwóch poprzednich wykresów. Mimo wysokiej wartości częstości wystąpień dla odległości równej minus dwa, duży pik, taki rozkład wyników spowodował znaczny wzrost wariancji w stosunku do poprzednich wyników, ponieważ odpowiednio ponad trzy- i dwukrotny. Wysoka wartość wariancji i ocena rozkładu odległości rozpatrywanych wyrazów od siebie spowodowała odrzucenie kandydata *strong for* jako interesującego w kontekście wyrażen wielowyrazowych. [?, str. 161]

Interesującym zestawieniem wyników może być poniższa tabela 3.2 utworzona na podstawie danych z tabeli numer 5.5 w pracy [?, str. 161]

Cztery pierwsze pozycje tabeli obrazują dobrych kandydatów na kolokacje, których cechuje niska wartość odchylenia standardowego. Cztery ostatnie wiersze to nieinteresujące wyrażenia w kontekście jednostek wielowyrazowych - ich odchylenie standardowe jest za wysokie, a dodatkowo średnia odległość wyrazów składowych od siebie jest bliska zeru. Świadczy to o tym, że te dwa słowa mogą występować w zasadzie w dowolnej kolejności lub szyku. Natomiast cztery środkowe pozycje to wyrażenia, których wyrazy występowały w kilku różnych odległościach znaczącą liczbę razy [?, 162] i są trudniejsze w jednoznacznej ocenie niż poprzednie przypadki.

s	odległość	częstość	słowo A	słowo B
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by

Tabela 3.2. Wyniki ekstrakcji kolokacji z zastosowaniem metody opartej o wariancję i odległość słów [?, str. 161]

3.2.4. Rozszerzenie metody o filtr pików

Przedstawiona tutaj metoda średniej arytmetycznej i wariancji jest uproszczoną wersją techniki stosowanej przez Smajdę, który opisał ją w 1993 roku [?]. Wspomniany powyżej badacz używał dodatkowo filtru, który odrzucał kandydatów, dla których na wykresie pojawiał się tak zwany płaski pik. Termin ten określa sytuację, gdy istnieje pewna wysoka wartość dla któregoś z argumentów, ale jednocześnie wartości dla argumentów z jego otoczenia także są znaczącymi liczbami. Manning i Schütze jako przykład płaskiego pików podali trzeci z poprzednio omawianych wykresów 3.3 dotyczący pary słów *strong for*.

3.2.5. Wyniki dla algorytmu po zastosowaniu filtru

Przy zastosowaniu tego sposobu wydobywania kolokacji i omówionego filtrowania płaskich pików, Smajda osiągnął dokładność na poziomie osiemdziesięciu procent. [?, str. 162]

3.2.6. Wnioski

Wizualizacja danych, tak jak wspomniano wcześniej, może być dodatkowym, pomocnym narzędziem przy ocenie kandydata na kolokację lub próbie wyznaczenia progu wartości wariancji, który będzie używany przy do odrzucania wyrażeń potencjalnie niebędących jednostkami wielowyrazowymi. Mimo małego skomplikowania modelu statystycznego i prostego filtru Smajda [?] [?, str. 162] pokazał, że metoda ta może być skuteczna, a dodatkowo pozwala na ekstrakcję kolokacji nieciągłych co nie było możliwe z wykorzystaniem przedstawionej we wcześniejszej części tej pracy metody zliczania.

3.3. FUNKCJE ASOCJACYJNE

3.3.1. Motywacje i wprowadzenie

Pierwsze dwie omówione metody wykorzystywały częstości słów i konkretnych zestawów wyrazów do oceny wyrażen w kontekście kandydatów na kolokacje. Niestety występują trzy problemy związane z uwzględnianiem tylko surowej częstości tokenów w korpusie, dwa zostały przytoczone przez Everta w [?, str. 20], a trzeci przez Manninga i Schütze [?, str. 162].

Pierwszy problem polega na tym, że częste współwystępowanie słów może być czytym przypadkiem, zwłaszcza jeśli częstość tych wyrazów w tekście jest wysoka. Fakt ten wpływa negatywnie na jakość oceny kandydatów na jednostki wielowyrazowe z wykorzystaniem metod opartych o zliczanie, więc istotne jest wprowadzanie interpretowalności statystycznej częstości [?, str. 20].

Drugi problem to wspomniana już wcześniej generalizacja dla całego języka na podstawie tylko danych wydobytych z pewnego korpusu. Informacje wyekstrahowane z podzbioru tekstów danego języka, przykładowo z tylko pojedynczego korpusu, mogą okazać się nie zrównoważone, niepełne, być jedynie zbliżonymi do tych, które są prawdziwe dla tego języka w ogólności. Problem ten to ewentualna mała reprezentatywność wykorzystanych danych.

Manning i Schütze zwrócili uwagę na trzeci problem związany z metodą zliczania opartą o odległość i wariancję. Zauważyli, że wysoka częstość i niska wariancja współwystępowania słów może być przypadkowa w sytuacji kiedy składowe wyrażenia występują często - wtedy oczekiwać można dużej liczby ich współwystąpień tylko ze względu na szansę na takie zdarzenie, a nie na poziom istotności tego wyrażenia w kontekście wydobywania kolokacji [?, 162]. Problem polega na tym, że oczekiwanym wynikiem jest informacja nie o tym czy dany zestaw słów współwystępuje często, tylko czy współwystępuje częściej niż wynika to jedynie z szansy na takie zdarzenie [?, 162].

Próba dokonania generalizacji języka, nadania interpretowalności statystycznej częstości z wykorzystaniem próbek losowych w postaci korpusów tekstowych dotyczących części języka będącego w centrum zainteresowania oraz problemu zdefiniowanego przez Manninga i Schütze w [?] są różne modele statystyczne. Mają one udzielać informacji czy określone obserwacje współwystępowania bytów są czysto przypadkowe, wynikające po prostu z szansy lub prawdopodobieństwa na ich wystąpienie, czy faktycznie istnieją wystarczająco mocne przesłanki statystyczne do tego, aby stwierdzić, że powiązanie pomiędzy nimi faktycznie istnieje i jest istotne. Takimi szeroko stosowanymi metodami statystycznymi są *funkcje asocjacyjne* zwane także *miarami asocjacji*, za pomocą których można obliczyć *miarę powiązania* pomiędzy argumentami zadanymi dla tej funkcji. Dla przykładu w przypadku bi-gramów argumentami dla tych metod będą pary wyrazów rozważanego kandydata na wyrażenie wielowyrazowe. Otrzymane za pomocą funkcji aso-

racji wyniki będące liczbami rzeczywistymi określają stopień powiązania elementów - jak nieprzypadkowe jest ich współwystąpienie, a dodatkowo wyniki te mogą posłużyć do utworzenia rankingu i dokonania selekcji kandydatów na kolokacje.

Sytuacja, w której elementy współwystępują częściej niż gdyby były od siebie niezależne określana jest mianem *asocjacji pozytywnej*, a w przypadku przeciwnym, kiedy występują rzadziej, mianem *asocjacji negatywnej*. Z powyższym wiąże się ważna cecha funkcji asocjacyjnych w kontekście dostarczanych przez nią wyników - przynależność do jednej z dwóch następujących grup: *miar jednostronnych* albo *miar dwustronnych*. Pierwsza z grup zawiera te miary asocjacji, w przypadku których wysoki wynik oznacza silne powiązanie pozytywne, a niski brak wystarczającego dowodu na powiązanie pozytywne pomiędzy badanymi elementami - może występować powiązanie negatywne lub brak powiązania, ale nie da się tego określić na podstawie wyniku działania tej funkcji. Druga z grup natomiast zawiera miary, których wysoki wynik oznacza silne powiązanie, ale bez informacji o tym jakiego jest ono rodzaju - pozytywne lub negatywne, a wynik o niskiej wartości to według tej funkcji słabe powiązanie pomiędzy elementami lub jego brak. [?, str. 20, 21, 75, 76]

Większość metod statystycznych omawiane w tej części pracy korzysta z uproszczenia polegającego na tym, że tekst to jedynie zbitka losowo występujących wyrazów ograniczonych tylko przez zasady syntaktyczne języka [?, str. 6], a tym samym szansa na wystąpienia danego słowa po dowolnym wyrazie poprzedzającym, nawet tym samym, jest zawsze taka sama na przestrzeni całego tekstu, jednak może być różna dla każdego ze słów. Chociaż już w 1957 roku Firth zwrócił uwagę osób zajmujących się lingwistyką, że stwierdzenie to jest nieprawdziwe [?, str. 15] to i tak jest ono wykorzystywane, niejednokrotnie z powodzeniem, w znacznej części niekontekstowych metod statystycznych jakimi są miary asocjacyjne.

Funkcje asocjacyjne stosowane są od przynajmniej niemalże pół wieku. Już w roku 1965 dostępnych była olbrzymia liczba funkcji asocjacyjnych, a przez kolejne prawie pięćdziesiąt lat powstało i zostało zbadanych wiele kolejnych. Jednak mimo dużej liczby tych miar niewiele z nich zyskało dużą popularność, do najbardziej znanych należą *MI* (*Mutual Information*), *T-score*, *Log-likelihood* oraz H_i^2 . [?, str. 21]

Pavel Pecina w swoich pracach [?][?] przedstawił dziesiątki miar asocjacyjnych [?, str. 3] [?, str. 18]. Wszystkie zostały podzielone na sześć grup [?, str. 2] (sześć pierwszych wpisów na liście), dodatkowo autor tej pracy postanowił wyróżnić kolejne *dwie* (dwie ostatnie pozycje):

1. estymacje współwystąpienia i prawdopodobieństwa warunkowego
2. *Mutal Information* i jej pochodne
3. testy statystyczne niezależności
4. miary z rodziny *likelihood*
5. zestaw heurystycznych miar asocjacyjnych oraz współczynników zależności

6. miary kontekstowe
7. *miary mieszane*
8. *uczenie maszynowe*

Grupy te wraz z przykładami zostały przedstawione w kolejnych częściach tej sekcji. Przed ich omówieniem należy jednak zdefiniować pojęcia używane we wzorach opisujących te miary:

- x - element x ;
- \bar{x} - element inny niż x ;
- (x, y, \dots) - zbiór elementów x, y, \dots ;
- $p(x)$ - prawdopodobieństwo wystąpienia elementu x ;
- $f(x)$ - częstość elementu x , zaobserwowana w zbiorze danych;
- $\hat{f}(x)$ - wartość oczekiwana elementu x ;

3.3.2. Współwystąpienia i prawdopodobieństwo warunkowe

Pierwsza z grup zawiera w sobie trzy podstawowe miary: prawdopodobieństwo współwystąpienia $p(x, y)$, prawdopodobieństwo warunkowe $p(x|y)$ oraz odwrócone prawdopodobieństwo warunkowe $p(y|x)$. Przytoczone miary to podstawa wykorzystywana w kolejnych rodzinach miar asocjacji.

3.3.3. Miary informacji

Jedną z najpopularniejszych funkcji asocjacyjnych będących miarą informacji jest *Mutual Information* wyrażona wzorem[?, str. 2]:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x) * p(y)}$$

Miara ta oblicza ile informacji w jednej zmiennej losowej X jest jednocześnie zawarte w drugiej zmiennej losowej Y [?, str. 2]. Natomiast miara *Pointwise Mutual Information* jest wykorzystana do obliczenia poziomu asocjacji dwóch konkretnych instancji zmiennej losowej X i Y . Funkcja *PMI* wyrażana jest wzorem[?, str. 2] oraz jest szeroko wykorzystywana w badaniach nad wydobywaniem kolokacji:

$$pmi(x, y) = \log_2 \frac{p(x, y)}{p(x) * p(y)}$$

Obie z przytoczonych funkcji zostały wykorzystane do budowy innych miar informacji wykorzystywanych w wydobywaniu jednostek wielowyrazowych.

3.3.4. Testy statystyczne niezależności

Jako przykłady miar należących do tej grupy autor niniejszej pracy wybrał trzy: *T-score*, *Z-score* oraz *Pearson's X^2* .

Pierwsza z nich jest częścią *T-testu* opisanego w pracy [?, str. 166], który składa się trzech głównych punktów - obliczenia wartości dla testu, ustalenia progu wiarygodności

i wykonania testu. *T-score* natomiast składa się z jednego punktu i jest wykorzystana do tworzenia rankingu, wyliczania wartości zamiast klasyfikacji obserwacji. Miara *T-score* w pracach [?, str. 3] i [?, str. 18] została nazwana mianem *T-testu*, jednak ze względu na to, że jest jedynie jego częścią polegającą na wyliczeniu wartości, którą można wykorzystać w *T-teście*, to w niniejszej pracy będzie nazywana *T-score*. Wartość tej miary obliczyć można wykorzystując następujący wzór [?, str. 18]:

$$\frac{f(x,y) - \hat{f}(x,y)}{\sqrt{f(x,y) * (1 - (f(x,y)/N))}}$$

Drugą z miar to *Z-score*, która także służy do wyliczenia wartości rankingowej obserwacji, a nie klasyfikacji. Miara ta w dwóch wcześniej wspomnianych pracach [?, str. 3] i [?, str. 18] nazywana została już *score*, a nie *test*, jak w przypadku *T-Score*. Opisana jest za pomocą poniższego wzorem [?, str. 18]:

$$\frac{f(x,y) - \hat{f}(x,y)}{\sqrt{\hat{f}(x,y) * (1 - (\hat{f}(x,y)/N))}}$$

Różnica pomiędzy miarami *T-score*, a *Z-score* to wykorzystanie wartości oczekiwanej zamiast zaobserwowanej w mianowniku wzoru. Obie te funkcje jednak mają ważną cechę wspólną - zakładają, że rozkład prawdopodobieństwa w zbiorze danych jest rozkładem normalnym co w znacznej części przypadków praktycznych nie jest prawdziwe [?, str. 169].

Trzecia z miar to *Pearson's X²*. Jej zaletą w stosunku do dwóch poprzednich to brak założenia co do rozkładu normalnego prawdopodobieństwa w zbiorze danych. Zamiast tego test zakłada rozkład X^2 (Hi^2), który powinien być bardziej odpowiedni w zadania ekstrakcji kolokacji. Miara korzysta z tabli wielodzzielczej (contingency table) [str. 169] i w celu porównania wartości obserwowanych z oczekiwanymi i na tej podstawie oceny czy hipoteza zerowa może zostać odrzucona na niezerową przyjętą za prawdziwą. *Pearson's X²* w niniejszej pracy oraz w pracach Pavla Peciny został wykorzystany jako funkcja rankingowa. Wykonane zostało to analogicznie jak w przypadku *T-testu*, wyliczana jest jedynie wartość dla testu, ale on sam nie zostaje wykonany, a wartość ta służy jako wartość rankingowa. Wzór funkcji został zamieszczony poniżej [?, str. 18]:

$$\sum_{i,j} \frac{(f_{i,j} - \hat{f}_{i,j})^2}{\hat{f}_{i,j}}$$

3.3.5. Funkcje z rodziny Likelihood

Dwie funkcje z rodziny likelihood zostały zaprezentowanych i zbadane w pracach Pavla Peciny [?, str. 3]. Pierwsza z nich została nazwana *Log likelihood ratio* i została zapisana w formie poniższego wzoru:

$$-2 \sum_{i,j} f_{i,j} \log \frac{f_{i,j}}{\hat{f}_{i,j}}$$

Druga funkcja jest modyfikacją pierwszej, jest pozbawiona jednego członu i wyrażona wzorem zamieszczonym poniżej:

$$-2 \sum_{i,j} \frac{f_{i,j}^2}{\hat{f}_{i,j}}$$

Według [?, str. 172] miary z rodziny likelihood są lepsze od testów H_i^2 dla danych rzadkich jakimi są kolokacje w dużych zbiorach tekstowych. Dodatkowo wyniki są bardziej interpretowalne, ponieważ określają jak jedna z hipotez jest bardziej prawdopodobna od drugiej.

3.3.6. Miary kontekstowe

Pavel Pecina w pracach [?, str. 18] oraz [?, str. 3] prezentuje 24 miary kontekstowe, których jakość także bada. Miary kontekstowe bazują na kontekstach (otoczeniu) słów i kandydatów na kolokacje. Wykorzystywane w przedstawionych przez Pavla Pecine miarach są konteksty dwu- oraz jednostronne i na ich podstawie oraz z wykorzystaniem danych o częstościach obliczane są wartości rankingowe. Podsumowując miary kontekstowe - wykorzystują one także informacje pochodzące z otoczenia kandydatów na wyrażenia wielowyrazowe, a nie tylko dane dotyczące ich bezpośrednio (ich częstości).

Praca [?, str. 4] wskazuje na to, że miara kontekstowa o nazwie *Cosine context similarity in boolean vector space* osiągnęła jeden z dwóch najlepszych wyników spośród wszystkich osiemdziesięciu dwóch sprawdzonych przez Pavla Pecine funkcji [?, str. 4]. Obrazuje to, że kontekst może być bardzo pomocny przy ocenie kandydatów na kolokacje.

3.3.7. Miary mieszane

Podejście miar mieszanych zostało zaczerpnięte z prac [?] oraz [?]. Pomysł przedstawiony w [?] polega na wygenerowaniu zestawu rankingów dla różnych miar asocjacyjnych, a następnie opcjonalne wykonanie ich przepunktowania polegającego na zachowaniu kolejności w rankingu, ale zmianie wartości dla każdej z pozycji. Przykładowymi miarami przepunktowania rankingów są *Rank distance* i *Borda score* [?, str. 2]. Kiedy rankingi są gotowe należy poddać je agregacji w celu wygenerowania pojedynczego, finalnego rankingu, który jest także wynikiem algorytmu.

Pavel Pecina wykorzystuje natomiast miary mieszane w innym celu - opisanym w kolejnym punkcie tej pracy.

3.3.8. Uczenie maszynowe

Praca [?] pokazuje sposób na wykorzystanie miar asocjacyjnych w celu wygenerowania cech dla algorytmów maszynowego uczenia i regresji liniowej.

Proces generowania cech polega na trzech krokach. Pierwszy z nich to dobór miar asocjacyjnych, które zostaną wykorzystane jako składowe generatora cech do wykorzystania w procesie maszynowego uczenia. Pavel Pecina w swojej pracy [?] wskazuje, że jest to krok istotny, ze względu na to, że wiele miar dzieli ze sobą w części te same informacje, a tym samym wykorzystanie niektórych z nich nie powoduje uzyskania dużo lepszych wyni-

ków. Autor pracy [?] pokazał, że można osiągnąć podobne jakościowo wyniki redukując liczbę dobranych miar asocjacyjnych z 82 do 17 przy zachowanie około 95% ich całkowitej wariancji. Natomiast wykorzystanie jedynie 42 miar powoduje utrzymanie poziomu w okolicy 99.9% [?, str. 7]. Dodatkowo Mariusz Paradowski w swojej pracy [?] zbadał część funkcji asocjacyjnych i wykazał, które z nich są ze sobą skorelowane i w efekcie generują takie same rankingi, z zachowaniem kolejności pozycji, ale przy różnej ich punktacji. Informacje te mogą posłużyć do lepszego doboru miar w procesie tworzenia cech dla algorytmów maszynowego uczenia.

Drugi krok generacji cech polega na wykorzystaniu miar asocjacyjnych dla każdego badanego kandydata na kolokacje. Efektem tego działa jest jeden wektor dla każdej krotki, a każda pozycja tego wektora przechowuje wynik jednej z funkcji asocjacyjnych. Dodatkowo jeden z jego elementów przechowuje informacje o tym czy kandydat jest kolokacją czy nie, informacja ta jest wykorzystywana w procesie uczenia nadzorowanego. W pracy [?] klasa jest dwuwartościowa, ale wydaje się, że nic nie stoi na przeszkodzie, żeby była wielowartościowa. Przykład wektora został zamieszczony poniżej:

$$wektor_cech = [x_1, x_2, \dots, x_n, klasa]$$

1. x_i - wynik dla miary i
2. $klasa$ - klasa kandydata na wyrażenie wielowyrazowe

Trzeci krok jest opcjonalny i polega na przetworzeniu otrzymanych cech. Przykładowym algorytmem wykorzystanym w tym celu może być zwykła normalizacja lub jak w pracy [?, str. 6] - *standaryzacja*.

Po wykonaniu dwóch pierwszych i opcjonalnie trzeciego kroku cechy są gotowe do wykorzystania ich w procesie uczenia maszynowego i klasyfikacji. Pavel Pecina wykorzystał tak wygenerowane cechy do nauczania klasyfikatorów takich jak jednowarstwowa sztuczna sieć neuronowa czy *Support Vector Machine*. Wyniki przedstawione w [?, str. 7] są obiecujące i wskazują, że metody maszynowego uczenia mogą okazać się znacząco lepsze od samych funkcji asocjacyjnych.

3.4. EKSTRAKCJA KOLOKACJI TRI-GRAMOWYCH I DŁUŻSZYCH

3.4.1. Wstęp

Ekstrakcja wyrażen wieloelementowych składających się z trzech lub więcej wyrazów jest procesem bardziej złożonym obliczeniowo i trudniejszym koncepcyjnie niż wyszukiwanie kolokacji dwuelementowych. Wiele podejść do wydobywania wyrażen trzy- oraz więcej elementowych polegało na dokonaniu generalizacji metod wykorzystywanych do ekstrakcji par słów lub wykorzystaniu specjalnie do tego przygotowanych miar heurystycznych.

3.4.2. Rozszerzenie miary *Mutual Information*

Praca Tima Van de Cruys [?] jest poświęcona rozważaniom nad dwoma alternatywnymi wersjami generalizacji funkcji *Mutual Information*. Pierwsza z nich bazuje na mierze *Interaction Information* pojawiającej się w pracach McGilla z 1954 roku oraz Bella z 2003 wyrażanej wzorem[?, str. 2]:

$$I(X; Y|Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}$$

Na podstawie tej miary Tim Van de Cruys zaproponował rozszerzenie funkcji *Mutual Information* do postaci uwzględniającej trzy zmiennych losowe zamiast dwóch. Zaproponowana przez niego miara opisana jest następującym wzorem[?, str. 2]:

$$I(X; Y; Z) = I(X; Y|Z) - I(X; Y) = I(X; Z|Y) - I(X; Z) = I(Y; Z|X) - I(Y; Z)$$

Rozszerzona wersja pierwszego z powyższych przypadków ma zatem postać[?, str. 2]:

$$I(X; Y; Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)} - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Odpowiednikiem miary *Pointwise Mutual Information* dla *Interaction Information* jest funkcja *Specific Interaction Information*[?, str. 2], a jej wzór jest następujący[?, str. 3]:

$$SI(x, y, z) = \log \frac{p(x, y)}{p(x)p(y)} - \log \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)} = \log \frac{p(x, y)p(y, z)p(x, z)}{p(x)p(y)p(z)p(x, y, z)}$$

Autor artykułu [?] przedstawił także jeszcze jedną wersję generalizacji *Mutual Information*. Miara zwana *Total Correlation* lub także *Multi-Information* bada ilość informacji współdzielonej pomiędzy zmiennymi losowymi z pewnego ich zbioru[?, str. 3]. Jej zaletą jest prosta generalizacja do dowolnej liczby zmiennych losowych, stosunkowo niska złożoność wzoru w porównaniu z poprzednią propozycją, a do tego była używana w zadaniach przetwarzania języka naturalnego w zadaniu ekstrakcji kolokacji[?, str. 3]. Funkcja została przedstawiona w pracy Watanabe w roku 1960 i opisana za pomocą poniższego wzoru[?, str. 3]:

$$I(X_1, X_2, \dots, X_n) = \sum_{x_1 \in X_1, x_2 \in X_2, \dots, x_n \in X_n} p(x_1, x_2, \dots, x_n) \log \frac{p(x_1, x_2, \dots, x_n)}{\prod_{i=1}^n p(x_i)}$$

Wersja miary *Total Correlation* analogiczna do dwuargumentowego *Pointwise Mutual Information* nazwana *Specific Correlation* wyraża się następującym wzorem[?, str. 3]:

$$SI(x_1, x_2, \dots, x_n) = \log \frac{p(x_1, x_2, \dots, x_n)}{\prod_{i=1}^n p(x_i)}$$

Dwie przedstawione miary będące rozszerzeniem *Mutual Information* są przykładami podejścia wykorzystującego generalizację miar działających na parach argumentów w celu wykorzystania ich do obliczeń na większej liczbie parametrów.

3.4.3. Wyniki dla miar *Interaction Information* oraz *Total Correlation*

Autor artykułu [?] badał wydobywanie jednostek trójelementowych postaci *subject*, *verb*, *object* z wykorzystaniem obu przedstawionych tutaj generalizacji *Mutual Informa-*

tion oraz za pomocą zwykłego rankingu częstości wystąpień[?, str. 4]. Według jego wyników zamieszczonych w [?, str. 4], dla jego zbiorów danych lepsza okazała się prostsza wersja generalizacji miary *Mutual Information - Specific Correlation*. Wyniki przytoczone z tej pracy zostały zamieszczone w poniższej tabeli 3.3¹:

miara	precyzja
częstości	0.00
<i>Specific Interaction Information</i>	0.24
<i>Specific Correlation</i>	0.31

Tabela 3.3. Wyniki badań dwóch generalizacji funkcji *Mutual Information* zaczerpnięte z [?, str. 4].

Różnica w precyzji pomiędzy dwoma funkcjami w wynikach przedstawionych przez autora [?] jest według niego wartością znaczącą[?, str. 5]. Ponadto obie z tych miar spisały się nieporównywalnie lepiej niż zwykle wybieranie wyrażeń o najwyższej częstości[?, str. 5] co może świadczyć o zasadności użycia opisanych funkcji i ich przydatności. Zamieszczone wyniki zdają się być dobrymi przesłankami do wykorzystania opisanych miar asocjacyjnych w badaniach.

3.4.4. Heurystyki generalizujące miary dwuelementowych

Autorzy artykułu [?] w swojej pracy zebrali, a także zaproponowali nowe wzorce budowy funkcji heurystycznych do wydobywania kolokacji dłuższych niż dwuelementowe. Opisane przez nich wzorce podchodzą do problemu w podobny sposób - rozbijają kolokacje N-elementowe w zestaw kolokacji 2-elementowych, ale robią to na inne sposoby, wyjątkiem jest pierwsza z miar oznaczona G_0 . Każda z sześciu kolejnych metod po podzieleniu N-elementowej krotki ocenia wszystkie powstałe bi-gramy wykorzystując do tego dwuelementową funkcję asocjacyjną, przykładowo jedną z zaprezentowanych wcześniej - *Pointwise Mutual Information*. Następnie na podstawie zestawu wyników dokonuje ich przetworzenia w pojedynczą wartość. Praca zawiera w sumie opis siedmiu wzorców generalizacji miar dwuelementowych.

Pierwszym przykładem jest wzorec G_0 , który wyróżnia się podejściem w stosunku do pozostałych metod. Zakłada on intuicyjną generalizację już istniejących metod. Autorzy pracy podali dwa przykłady dla funkcji *Pointwise Mutual Information* i *Dice* [?, str. 4]. Wzory zostały zamieszczone poniżej:

$$G_0(PMI, x_1 \dots x_n) = \log_2 \frac{P(x_1 \dots x_n)}{\prod_{i=1}^n P(x_i)}$$

$$G_0(Dice, x_1 \dots x_n) = \frac{nf(x_1 \dots x_n)}{\sum_{i=1}^n f(x_i)}$$

1. x_i - i-ta składowa N-elementowej krotki;

¹ Dane na podstawie wyników pracy zamieszczonych w [?, str. 4].

2. $x_1...x_n$ - krotka N-elementowa;
3. n - liczba elementów krotki.

Drugim przykładem prostszej z miar dokonujących podziału jest *Average Bi-gram*, oznaczona G_3 w pracy [?, str. 5] i opisana wzorem:

$$G_3(x_1...x_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} AF(x_i, x_{i+1})$$

1. $AF(x_i, x_{i+1})$ - wynik funkcji asocjacyjnej dla elementów x_i oraz x_{i+1} .

Trzeci przykład jest bardziej skomplikowany i został oznaczony G_6 w artykule [?, str. 5], a przez autora niniejszej pracy nazwany *Smoothed Bi-gram*. Wzorzec dzieli N-elementową krotkę na zestaw $N - 1$ dwuelementowych krotek taki, że:

$$Krotka(x_1, x_2, \dots, x_n) \rightarrow \{Krotka(x_1, x_2), Krotka(x_2, x_3), \dots, Krotka(x_{n-1}, x_n)\}$$

Po takim podziale wykorzystuje się inny wzorzec generalizacji, gdzie argumentami są wszystkie krotki powstałe poprzez podział opisany w poprzednim kroku, a ich częstości są równe liczbie wystąpień tych właśnie krotek 2-elementowych w zbiorze danych. Dodatkowo według autorów wewnętrzną funkcją asocjacyjną dla wzorca wewnętrznego G_0 jest miara wykorzystana we wzorcu zewnętrznym G_6 [?, str. 5]. Wydaje się jednak, że nic nie stoi na przeszkodzie, aby wykorzystać inny wzorzec wewnętrzny niż G_0 oraz różne miary asocjacji w obu wzorcach - wewnętrznym i zewnętrznym. Formalnie wzór dla tego wzorca przyjmuje następującą postać:

$$G_6(AF, x_1...x_n) = G_0(AF, (x_1x_2, x_2x_3, \dots, x_{n-1}x_n))$$

Dodatkowo Mariusz Paradowski opracował miarę *Minimalnego Bi-gramu*, która także rozбивa kolokacje na zestaw krotek 2-elementowych. Ocena kolokacji jest równa minimum z ocen wszystkich kolejnych, ciągłych i zachodzących na siebie bi-gramów. Wzór miary prezentuje się następująco:

$$ocena = \min(s \forall s \in S : S = AF(b) \forall b \in t)$$

1. t - krotka N-elementowa;
2. b - bi-gram utworzony z fragmentu N-elementowej krotki;
3. $AF(b)$ - wynik funkcji asocjacyjnej dla bi-gramu.

3.4.5. Podsumowanie

Literatura okazuje się być bogata w opisy, prezentacje, badania i porównania wielu różnych funkcji asocjacyjnych oraz innych sposobów wydobywania kolokacji dwuelementowych z tekstów. Znaleźć można też różne sposoby i podejścia do ekstrakcji kolokacji dłuższych niż dwuelementowe. Skrupulatny i szeroki przegląd literatury może być dobrą bazą do wykonywania kolejnych badań i testów metod ekstrakcji wyrażeń wieloelementowych z korpusów tekstowych.

4. NARZĘDZIE I OPROGRAMOWANIE WYKORZYSTANE NA POTRZEBY REALIZACJI TEMATU

4.1. BIBLIOTEKA CORPUS2

Biblioteka *Corpus2* to zestaw struktur danych i funkcji do przetwarzania korpusów oznaczonych morfo-syntaktycznie, z tagsetem pozycyjnym. Umożliwia ona wczytywanie i zapisywanie danych do różnych formatów danych takich jak *XCES*, jego rozszerzeniem *CCL* czy formatów tekstowo-kolumnowych takich jak przykładowo *IOB-Chan*. Opcjonalnie biblioteka wspiera także pracę, wczytywanie danych, z binarnym i zindeksowanym formatem *Poliqarp* autorstwa pracowników instytutu *Podstaw Informatyki Polskiej Akademii Nauk*, a opracowanym do przechowywania korpusów tekstowych. Warto nadmienić, że biblioteka *Corpus2* wspiera także wiele innych formatów plików.

Biblioteka została zaimplementowana w *Instytucie Informatyki Politechniki Wrocławskiej* w języku C++, ale powstały także skrypty w języku Python, opakowania, umożliwiające pracę z tą biblioteką z wykorzystaniem tego właśnie języka skryptowego. Dodatkowo narzędzie jest dostępne nieodpłatnie na licencji *GNU LGPL 3.0* lub w przypadku wykorzystania formatu *Poliqarp* na licencji *GNU GPL 3.0*.

Corpus2 został wykorzystany w badaniach na potrzeby niniejszej pracy do wczytywania oznaczonych korpusów tekstowych z formatów *IOB-Chan*, *CCL*, a także *Poliqarp*, a także do konwersji pomiędzy nimi. Ponadto inne narzędzia wykorzystane w tej pracy także wykorzystują funkcjonalność tej biblioteki.

4.2. TAGERY WCRFT I WCRFT2

Nazwa *WCRFT* jest akronimem powstałym od słów *Wrocław Conditional Random Fields Tagger* - wrocławskie narzędzie tagujące oparte o model matematyczny *Conditional Random Field*. *WCRFT* został zaimplementowany z myślą o oznaczaniu tekstów języków fleksyjnych, zwłaszcza słowiańskich, a szczególnie języka polskiego.

Tager napisano w języku Python, z wykorzystaniem gotowej implementacji *CRF* o nazwie *CRF++*, która to napisana została w języku C++ [?]. Narzędzie *WCRFT* było implementowane głównie przez Adama Radziszewskiego.

WCRFT2 natomiast jest następcą tagera *WCRFT*. Został on napisany całkowicie w języku C++ – zostało dokonane portowanie kodu z języka Python do języka C++. Takie zabieg pozwolił na usunięcie zależności związanych z językiem skryptowym, znaczne przyspieszenie tagera – trzykrotne, a także usprawniło i ułatwiło proces budowy narzędzia.

Zaznaczyć trzeba, że po wykonaniu konwersji pomiędzy językami zachowano kompatybilność modeli z poprzednią wersją tagera, a wyniki dla obu wersji są takie same. Ta wersja jest teraz zalecaną do użycia, a poprzednia nie jest już wspierana.

Narzędzie to, w celach badawczych na potrzeby tej pracy, zostało wykorzystane do oznaczenia morfo-syntaktycznego i ujednoznaczniania semantycznego słów z zebranych tekstów języka polskiego, które następnie zostały wykorzystane w procesie badania sposobów ekstrakcji kolokacji w tej pracy. Do wykonania tych prac wykorzystany został nowy, gotowy model dla tagera udostępniony przez grupę G4.19.

4.3. FORMALIZM WCCL

4.3.1. Wyjaśnienie terminu

WCCL (Wrocław Corpus Constraint Language) jest nowym formalizmem, językiem ograniczeń i narzędziem pozwalającym tworzyć wyrażenia funkcyjne, które można wykorzystać jako cechy, kluczowe informacje, dla wielu algorytmów przetwarzania języka naturalnego i maszynowego uczenia. Wyrażenia *WCCL* działają na tekście uprzednio oznakowanego morfo-syntaktycznie na przykład za pomocą narzędzia wspomnianego we wcześniejszej części tej pracy - tagera *WCRFT2*. Chociaż formalizm ten był konstruowany z myślą o pracy z językiem polskim to według autorów powinien on także móc być użyty do pracy z innymi językami fleksyjnymi. Ograniczeniami mogą być jednak przyjęta tekstowa reprezentacja tagów oraz formaty oznakowanych korpusów, na których miałyby pracować wyrażenia *WCCL*. [?, str. 1]

Omawiany język ograniczeń został zaimplementowany w języku *C++* w postaci bibliotek wykorzystujących *Corpus2* oraz system *MACA (Morphological Analysis Converter and Aggregator)*. Dodatkowo napisane zostały także skrypty w języku Python opakowujące funkcjonalność *WCCL* i umożliwiające prace z poziomu języka skryptowego, bez konieczności zaznajomienia się z językiem natywnym, w jakim formalizm ten został zaimplementowany. Dzięki temu tworzenie narzędzi i praca z *WCCL* może być szybsza i prostsza. [?, str. 3]

Formalizm *WCCL* został wykorzystany w badaniach z poziomu języka *C++*, bez użycia opakowań, i posłuży głównie do filtrowania kandydatów na jednostki wielowyrazowe na podstawie części mowy składowych wyrażenia wielowyrazowego oraz innych cech takich jak przykładowo uzgodnienie rzeczownika z przymiotnikiem. Autor niniejszej pracy zajął się badaniami nad kolokacjami w języku polskim co rozwiązało ewentualne problemy mogące się pojawić przy pracy z innymi językami fleksyjnymi.

4.3.2. Przykładowe wyrażenie

Przykładowy, prosty operator *WCCL* zbliżony swoją składnią do jednego z operatorów stosowanych w badaniach przeprowadzonych przez autora tej pracy zamieszczono poniżej.

```

@b: "SubstAdjOrSubstSubst"
(
  or
  (
    and
    (
      inter(class[0], {subst}),
      inter(class[1], {adj, subst}),

      setvar($Case, 0)
    ),
    and
    (
      inter(class[1], {subst}),
      inter(class[0], {adj, subst}),

      setvar($Case, 1)
    )
  )
)

```

Listing 4.1. Przykładowe wyrażenie w języku *WCCL*

Powyższy listing prezentuje kod wyrażenia w języku *WCCL*. Składa się on z dwóch części - nagłówka i ciała. Istotną informacją jest fakt, że operator *WCCL* jest kontekstowy, wywoływany jest dla konkretnego miejsca w korpusie, jakiegoś wyrazu w nim zawartego. Wyraz ten nazwijmy *początkiem kontekstu operatora*.

Nagłówek zawiera dwie informacje o wyrażeniu, pierwsza z nich, *@b*, to typ zwracany przez ten operator na podstawie wykonania instrukcji zawartych w jego ciele. Druga informacja to nazwa wyrażenia znajdująca się po dwukropku, tutaj będzie to *SubstAdjOrSubstSubst*. Powyższy kod zwróci jedną z dwóch wartości logicznych - *True* lub *False*. Inne typy, które mogą zwracać operatory *WCCL* to *Position*, *Set of strings* oraz *Tagset symbol set*. Pierwszy z nich to wartość całkowita - odpowiednik typu *int* w języku *C++*, drugi typ jest rozumiany jako zbiór ciągów znaków tekstowych, a trzeci to zestaw symboli używanego tagsetu.

Przed opisem ciała powyższego, przykładowego operatora należy wyjaśnić funkcje w nim użyte. Pierwsze dwie z nich są intuicyjne - funkcja *or* i *and*. Obie z nich zachowują się jak funkcje logiczne o tych samych nazwach angielskich, a argumenty dla nich są oddzielone przecinkiem tak samo jak w każdej funkcji języka *WCCL*. Funkcja *setvar* ustawia wartość zmiennej o nazwie podanej w pierwszym argumencie na wartość zadaną parametrem drugim. Natomiast funkcja *inter* sprawdza przecięcie zbiorów zadanych jej argumentami i zwraca *True* jeśli moc przecięcia jest niezerowa. Kluczowa w tym wyrażeniu jest także funkcja *class* przyjmująca tylko jeden argument typu całkowitego. Funkcja ta zwróci klasę gramatyczną słowa oddalonego o zadaną argumentem liczbę wyrazów od początku kontekstu tego wyrażenia. Przykładowo przyjmijmy, że jakiś operator wywołuje

kolejno funkcje *class[-1]*, *class[0]* i *class[1]* na poniższym zdaniu, gdzie słowo będące początkiem kontekstu operatora zostało oznaczone pogrubioną czcionką:

Długie, ciężkie **spodnie** jeszcze się suszą.

Pierwsza funkcja z argumentem równym minus jeden zwróci klasę gramatyczną słowa *ciężkie* czyli przymiotnik, druga klasę wyrazu będącego początkiem kontekstu operatora - *spodnie*, czyli rzeczownik, natomiast wynikiem działania trzeciej dla wyrazu *jeszcze* będzie partykuła.

Ciało powyższego wyrażenia *WCCL* składa się z pojedynczego bloku - funkcji *or*, sprawdzającej czy którykolwiek z dwóch bloków wewnętrznych *and* zwróci wartość *True* i jeśli tak to *or* także zwróci *True* jako wartość testu wykonanego przez ten operator. Oba wewnętrzne bloki *and* wykonują podobny test na dwóch kolejnych wyrazach z korpusu poczynając od *początku kontekstu* wywołania wyrażenia. Pierwszy blok *and* sprawdza czy klasa gramatyczna słowa, dla którego operator został wywołany jest rzeczownikiem (*subst*), a wyraz kolejny odpowiednio przymiotnikiem (*adj*) lub też rzeczownikiem. Jeśli obie funkcje *inter* zwrócą *True* to ustawiana jest zmienna, która potem może zostać odczytana w kodzie programu, który ten operator wywołał. W przypadku omawianego operatora zmienna *Case* będzie równa zero lub jeden w zależności od tego, która z funkcji *and* zwróci wartość *prawda* lub też może być niezdefiniowana jeśli obie funkcje *and* zwrócą wartość *fałsz*. Zabieg taki pozwalać może przykładowo na określenie kolejności wyrazów w jakiej wystąpiły w korpusie w celu odczytania tej informacji w programie. Drugi blok *and* zachowuje się analogicznie do pierwszego - sprawdza ten sam warunek, ale dla wyrazów ułożonych w odwrotnej kolejności.

Więcej informacji o wyrażeniach języka *WCCL* można pozyskać zapoznając się z pracą [?].

4.4. SŁOWOSIEĆ

Słowsieć jest polskim, podczas opisywania tego narzędzia drugim co wielkości na świecie *Wordnetem* utworzonym i nieustannie rozwijanym przez Grupę Technologii Językowych Politechniki Wrocławskiej G4.19 w ramach projektów *Clarin*, *Synat* oraz *Nekst* przy wsparciu uczelni oraz Ministerstwa Nauki i Szkolnictwa Wyższego.

Aktualna wersja *Słowsieci* jest dostępna na darmowej licencji i zawiera 140 tysięcy słów, 200 tysięcy znaczeń oraz pół miliona relacji, jednak jest także szybko rozwijana przez jej pracowników i powyższe wartości mogą ulec znacznej zmianie w najbliższym wydaniu nowej wersji tego polskiego *Wordnetu*. Zawiera ona także 110 tysięcy haseł polsko-angielskich pozyskanych z *Princeton Wordnet 3.1*. Dostępnych jest także wiele publikacji traktujących o *Słowsieci*, których spis dostępny jest na stronie internetowej [?]. Znaleźć tam można także dokładne statystyki, interfejs webowy oraz inne informacji tym o projekcie.

Baza danych polskiego *Wordnetu* wykorzystywana była w celu pozyskania wzorcowych wyrażen wielowyrazowych.

4.5. SUPERMATRIX

SuperMatrix to pakiet wielu narzędzi zaimplementowanych w języku C++, z założenia przystosowany do wykonywania różnych operacji na dwuelementowych obiektach takich jak przykładowo bi-gramy. Narzędzie to oferuje szereg funkcji, programów i gotowych algorytmów takich jak obliczanie wartości asocjacji z wykorzystaniem wielu miar - ponad osiemdziesięciu, wyliczanie wartości funkcji podobieństwa wierszy względem siebie, filtrowanie danych, transformacje, dzielenie macierzy, ich łączenie i inne. Zawiera programy oraz skrypty umożliwiające budowanie krotek dwuelementowych i macierzy na kilka sposobów z korpusów tekstowych oznaczonych morfo-syntaktycznie. Dla wygody i przyspieszenia pracy napisane zostały także opakowania w języku Python umożliwiające używanie bibliotek i narzędzi z pakietu *SuperMatrix* z poziomu tego języka skryptowego.

Dużą zaletą omawianego narzędzia jest fakt, że było ono z powodzeniem wielokrotnie wykorzystywane, oraz że mimo dużej funkcjonalności jest ono dostępne na darmowej licencji i można je nieodpłatnie pobrać z sieci internet.

Istotne jest, że dane składowane są w postaci macierzy dwuwymiarowej (podstawowego formatu danych dla narzędzi w tym pakiecie) co jest ograniczeniem i problemem w przypadku chęci pracy z kolokacjami dłuższymi niż dwuelementowe. Jednak zarówno to ograniczenie jak i fakt, że *SuperMatrix* nie był projektowany do pracy z kolokacjami, nie jest problemem dla prowadzenia badań nad wyrażeniami dwuelementowymi. Wystarczy jedynie przygotować dane w odpowiedni sposób i wedle potrzeby, a następnie wykorzystać możliwości zawartych w pakiecie narzędzi.

Pierwsze badania na potrzeby tej pracy wykonywane były z wykorzystaniem właśnie tego pakietu narzędzi ze względu na licznosc zaimplementowanych miar asocjacji, które można było zbadać bez konieczności przystosowywania narzędzia, implementacji nowych modułów pakietu, kolejnych programów czy całej aplikacji specjalnie do tego celu. Badania te pozwoliły przetestować dziesiątki miar i ocenić, którymi z nich warto zająć się przy dalszej pracy z jednostkami wielowyrazowymi dłuższymi niż dwuelementowe. Pierwsze wyniki nadały pewien kierunek dla prowadzenia dalszych prac w temacie wyrażen wielowyrazowych.

4.6. MWEXTRACTOR

Pakiet narzędzi *SuperMatrix* z założenia miał być wykorzystywany do pracy z parami elementów, z których jeden jest reprezentowany danymi zawartymi w wierszu, a drugi w kolumnie macierzy. Cecha ta jest swoistego rodzaju ograniczeniem i problemem uniemożliwiającym, a przynajmniej bardzo utrudniającym prowadzenie badań obiektów składają-

cych się z trzech lub większej liczby elementów. Sposób przechowywania danych i ich wykorzystania musiałby zostać przygotowany do składowania w formacie pakietu *SuperMatrix* czyli dwuwymiarowej macierzy co znacząco utrudniałoby wykonanie zadania. Dodatkowo zabieg ten wydaje się zły koncepcyjnie - to format danych powinien być dopasowany do składowania określonych danych, a nie dane upakowane na siłę w taki sposób, aby zmieściły się w narzuconym formacie niebędącym przygotowanym do przechowywania informacji tego rodzaju. Sam sposób przechowywania danych był jednak tylko wierzchołkiem góry lodowej. Praktycznie wszystkie narzędzia pakietu *SuperMatrix* zostały przygotowane do pracy z obiektami dwuelementowymi. Fakt ten oznacza, że wykorzystanie omawianego narzędzia byłoby niewielką wartością dodaną w prowadzeniu badań, ponieważ dużą część funkcjonalności trzeba byłoby zaimplementować od nowa mając na uwadze ograniczenia i koncepcje wprowadzone na potrzeby realizacji projektu *Supermatrix*. Reimplementacja metod wymagana byłaby także ze względu na wspomnianą wcześniej konieczność upakowania danych N-elementowych do postaci dwuelementowej. Mogłoby to doprowadzić do poświęcenia dużej ilości czasu i uwagi na zaplanowanie, dostosowanie, implementację i integrację nowych fragmentów oprogramowania *SuperMatrix* zamiast wykorzystania tych zasobów do realizacji faktycznego zadania, czyli celu niniejszej pracy.

Omówione problemy były jednymi z głównych przyczyn opracowania nowego narzędzia, niezależnego od pakietu *SuperMatrix*, specjalizowanego pod kątem prowadzenia badań dotyczących wyrażeń wielowyrazowych, między innymi na potrzeby niniejszej pracy. Narzędzie *MWeXtractor* utworzone przez autora tej pracy jest biblioteką programistyczną oraz pakietem struktur danych, programów, skryptów i niewielkiej ilości danych. Posłużyło ono do prowadzenia badań nad kolokacjami, których wyniki zostały zamieszczone w dalszej części niniejszej pracy.

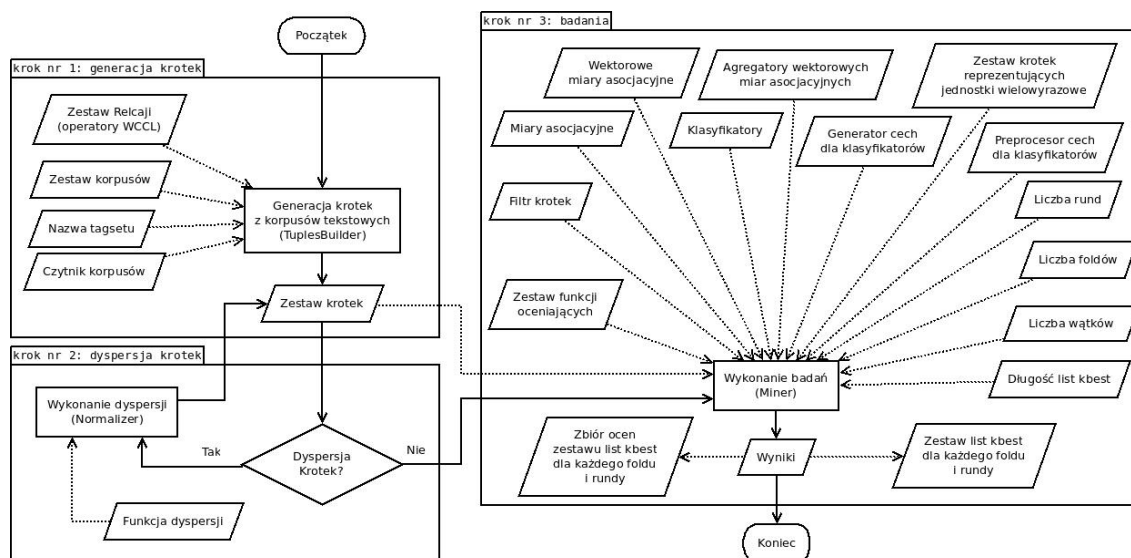
Większość oprogramowania, a w szczególności jego główne funkcjonalności zostały zaimplementowane w języku C++ w celu osiągnięcia jak największej wydajności pamięciowej odnośnie składowania danych oraz obliczeniowej narzędzi i algorytmów. Wykorzystany standard tego języka to C++11. Język skryptowy *Python* został wykorzystany do utworzenia części oprogramowania pomocniczego - skryptów. Oprogramowanie związane z algorytmem genetycznym pochodzi z kodów zaimplementowanych przez Łukasza Kłyka [?], a zostało dostosowane przez autora niniejszej pracy do współpracy z omawianym tutaj pakietem narzędzi *MWeXtractor*.

4.6.1. Przykładowe schematy użytkowania narzędzia

Dwie typowe ścieżki przetwarzania umożliwiające przez oprogramowanie *MWeXtractor* zostały opisane w dalszej części niniejszej pracy. Pierwsza z nich jest związana z badaniem jakości funkcji asocjacyjnych i klasyfikatorów, a druga służy do ekstrakcji kolokacji z korpusów tekstowych.

4.6.2. Schemat procesu badania metod ekstrakcji kolokacji

Schemat 4.1 prezentuje pierwszą z dwóch typowych ścieżek przetwarzania, odpowiedzialną za badania metod ekstrakcji wyrażen wielowyrazowych.



Rys. 4.1. Powyższy schemat prezentuje jeden z typowych procesów użytkowania oprogramowania *MWeXtractor* – badania jakości generowanych przez rankery kolokacji

Krok numer 1: generacja krotek

Najpierw należy przygotować listę korpusów, z których mają zostać wydobyte dane do przeprowadzenia badań nad metodami ekstrakcji wyrażen wielowyrazowych. Należy podać typ czytnika korpusów oraz tagset jakim dane tekstowe zostały opisane. Następnie zdefiniować zestaw relacji – operatorów zapisanych w języku *WCCL*, które specyfikują jakich kandydatów na wyrażenia wielowyrazowe należy wyszukiwać w badanych korpusach tekstowych. Dzięki nim można ograniczyć zestaw kandydatów, nad którymi chce się prowadzić badania. Jeśli użytkownik jest zainteresowany wszystkimi możliwymi do utworzenia *n*-gramami może przykładowo wykorzystać do tego celu operator *WCCL* generujący wszystkie możliwe kombinacje *n*-elementowe. Zaznaczyć trzeba, że wydobywani kandydaci na kolokacje nie muszą być ciągli, a ich szyk może być zmienny, wszystko zależy od tego jak zostaną przygotowane relacje. Tak przygotowany zestaw danych i parametrów należy następnie podać jako argumenty dla programu *TuplesBuilder*, który na ich podstawie przygotuje skład krotek – podstawową strukturę danych dla oprogramowania *MWeXtractor* składującą krotki.

Krok numer 2: dyspersja krotek

Drugi krok tego schematu przetwarzania jest opcjonalny i dotyczy modyfikacji informacji o zebranych krotkach – ich częstości będących kluczowym elementem związanym z późniejszą oceną kandydatów na kolokacje. Do tego celu wykorzystana może zostać jedna z funkcji dyspersji zaimplementowana w tym celu, zostaną one opisane w dalszej

części tej pracy. Zależnie od wybranej metody dyspersji pod uwagę są brane różne statystyki związane z krotkami, przykładowo będą to częstości danych krotek, liczba krotek w danym korpusie czy informacja o tym w ile korpusach dana krotka została odnaleziona. W skrócie, miara dyspersji ma za zadanie zmienić rozkład częstości krotek w składzie promując instancje ciekawe – mniej typowe.

Jednak z zaimplementowanych funkcji dyspersji – *Lynes D3*, bazuje na mierze Hi^2 , a tym samym potrzebne są jej pewne dane statystyczne. Jeśli ta miara ma zostać wykorzystana do dyspersji krotek to trzeba jej te dane statystyczne przygotować. Sposób generacji tych informacji został opisany w trzecim kroku omawiania kolejnego z przykładowych schematów przetwarzania. Zaznaczyć jednak trzeba, że ta miara to wyjątek.

Krok numer 3: badania

Niniejszy krok jest krokiem finalnym tej ścieżki przetwarzania. Polega on na wykonaniu r-rundowej, f-foldowej walidacji krzyżowej dla danego zbioru danych, z wykorzystaniem określonych miar asocjacyjnych i klasyfikatorów. Oprogramowanie przeznaczone do tego celu zostało nazwane *Miner*. Wyniki wygenerowane przez ten program zostają poddane ocenie przez użytkownika narzędzi *MWeXtractor*. Do łączenia dużych zestawów wyników przygotowany został skrypt w języku *Python* uśredniający wyniki dla wszystkich z foldów z każdej z rund, dla poszczególnych funkcji rankingowych z osobna. Pozwala to na szybkie wygenerowanie zbiorczych wyników i przykładowo utworzenie wykresów prezentujących jakości rezultatów zastosowanych metod wydobywania według określonych miar.

Parametry programu *Miner* zostały wymienione i opisane w poniższej tabeli 4.1:

Wspomniany zbiór wyników jest obszerny, podzielony na pliki, których są dwa rodzaje - pliki z listami k-najlepszych kandydatów na wyrażenia wielowyrazowe, oraz pliki z ocenami tych list. Liczba wygenerowanych plików rankingowych jest równa $((A + V + C) * R * F)$, gdzie A , V oraz C oznaczają kolejno liczbę wykorzystanych funkcji asocjacyjnych, wektorowych miar asocjacyjnych oraz klasyfikatorów, natomiast R i F to kolejno liczba rund i foldów walidacji krzyżowej. Dodatkowo dla każdego pliku z rankingiem wygenerowanych zostaje Q plików oceny tego rankingu, gdzie Q jest liczbą wykorzystanych funkcji oceny list k-najlepszych. Wzorzec nazwy pliku z rankingiem jest generowany w sposób następujący: *kbest.nr_rankera.nr_rundy.nr_foldu.csv*, natomiast wzorzec dla plików z wynikami funkcji oceny list k-najlepszych to: *kbest.nr_rankera.nr_rundy.nr_foldu.nr_funkcji_oceny.csv*. Numer rankera jest z przedziału:

1. $[0 : A - 1]$ - dla funkcji asocjacyjnych;
2. $[A : A + V - 1]$ - dla wektorowych miar asocjacyjnych;
3. $[A + V : A + V + C - 1]$ - dla klasyfikatorów.

Ze względu na dużą liczbę plików wynikowych zaimplementowany został skrypt wspomniany wcześniej, łączący wyniki poprzez uśrednienie wyników dla każdego foldu z każdej rundy dla poszczególnych funkcji rankingujących z osobna i generuje pojedynczy plik

Nazwa	Typ	Opis
Skład krotek	nazwa folderu	ścieżka do folderu ze składem krotek
Wyjście programu	nazwa folderu	ścieżka do folderu, w którym zostaną zamieszczone wyniki
Miary asocjacji	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcje asocjacyjną, którą program ma wykorzystać
Miary wektorowe	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący wektorową miarę asocjacyjną, którą program ma wykorzystać
Agregatory	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcję agregującą wyniki miar wektorowych, po jednej funkcji dla każdej z nich
Klasyfikatory	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący klasyfikator, który program ma wykorzystać
Generator cech	ciąg tekstowy	wektorowa miara asocjacyjna, która zostanie wykorzystana jako generator cech dla klasyfikatorów
Preprocesor cech	ciąg tekstowy	parametr opcjonalny, tekst reprezentujący funkcję, która ma zostać wykorzystana do normalizacji cech
Zestaw JW	nazwa pliku	plik zawierający w każdej linii ciąg wyrazów oddzielonych spacjami, reprezentujący wyrażenie wielowyrazowe
Funkcje oceny	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcję oceny, którą program ma wykorzystać
Filtr krotek	ciąg tekstowy	tekst reprezentujący filtr, który program ma wykorzystać do wyznaczenia zestawu krotek, z których ma korzystać i, dla których wygenerować dane
Liczba wątków	liczba całkowita	maksymalna liczba wątków do wykorzystania przez program
Liczba rund	liczba całkowita	liczba rund walidacji krzyżowej
Liczba foldów	liczba całkowita	liczba foldów dla każdej walidacji krzyżowej

Tabela 4.1. Tabela zawiera nazwy, typy oraz opisy parametrów wykorzystywanych przez program *Miner*

z jedną kolumną dla każdego rankera. Skrypt nazwany został *cv_quality_merger.py*, został napisany w języku *Python* i jest częścią pakietu *MWeXtractor*.

4.6.3. Schemat procesu ekstrakcji wyrażen wielowyrazowych

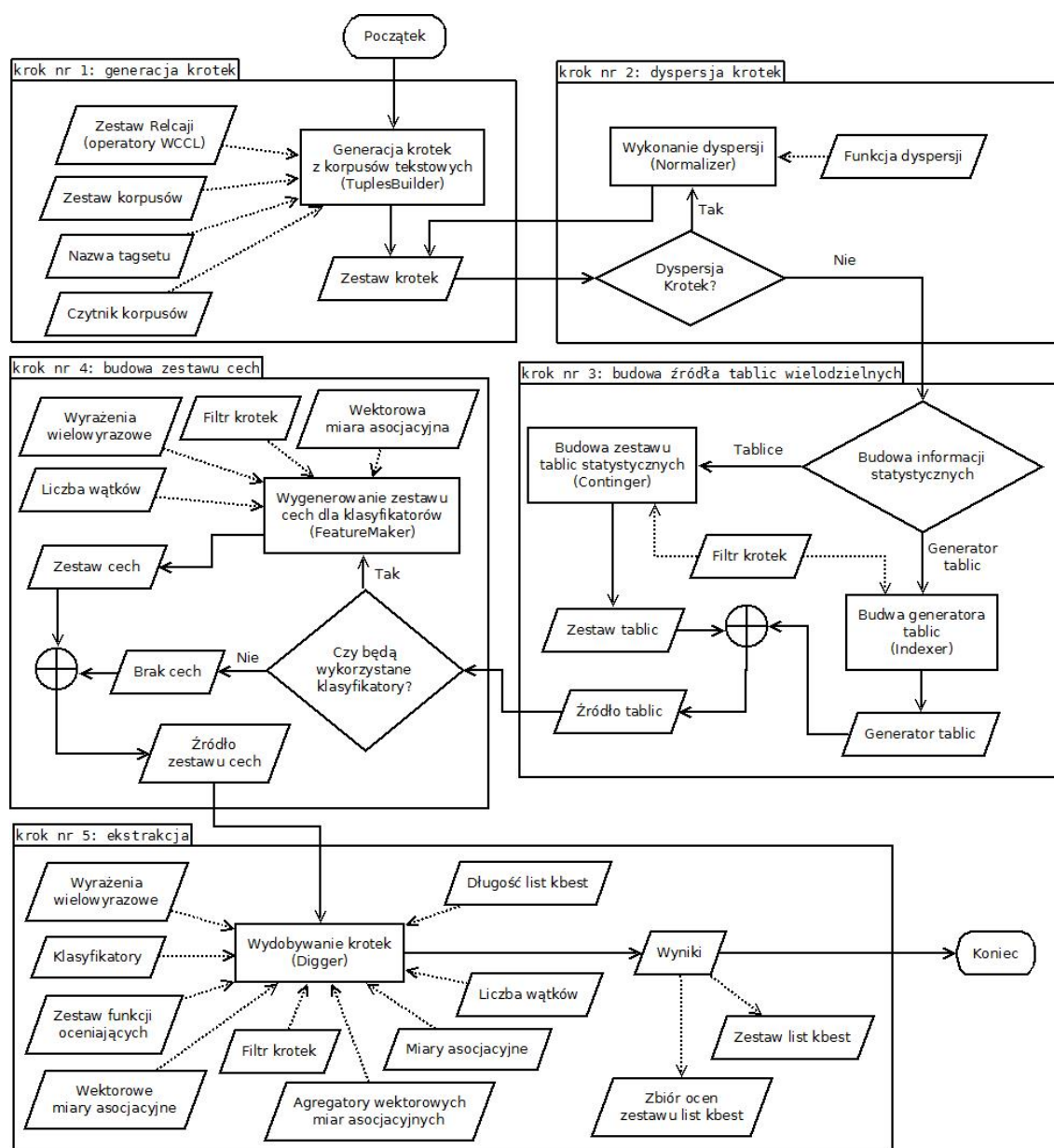
Schemat 4.2 prezentuje pierwszą z dwóch typowych ścieżek przetwarzania, odpowiedzialną za badanie metod ekstrakcji wyrażen wielowyrazowych.

Kroki numer 1 i 2

Oba kroki są takie same jak w przypadku poprzedniego przykładowego schematu przetwarzania – badania miar asocjacyjnych i klasyfikatorów.

Krok numer 3: budowa źródła tablic wielodzielných

Prawie wszystkie zaimplementowane miary i klasyfikatory korzystają z informacji za-



Rys. 4.2. Powyższy schemat prezentuje jeden z typowych procesów wydobywania wyrażeń wielowyrazowych za pomocą oprogramowania *MWEExtractor*

wartych w tablicach wielodzielnych dla krotek. Miary korzystają z tych informacji bezpośrednio do obliczania wartości dla kolokacji, a w przypadku klasyfikatorów miary te wykorzystywane są do generowania cech instancji. Wyjątkiem są miary oparte na szyku, ponieważ nie korzystają takich informacji. Krok trzeci polega na utworzeniu jednego z dwóch dostępnych źródeł tablic – ich generatora lub składu. Programy przygotowane do wykonania tego zadania to odpowiednio *Indexer* oraz *Continger*. Zazwyczaj generator tablic zajmuje znacznie mniej pamięci operacyjnej – kilkakrotnie mniej, ale odwołanie do tablic wielodzielnych jest znacznie wolniejsze niż w przypadku składu tablic, ponieważ generator tworzy je na bieżąco, a skład przechowuje już gotowe tablice. Zestaw krotek, na podstawie których źródło tablic ma zostać wygenerowane można ograniczyć za pomocą

filtrów. Zabieg taki pozwala na używanie określonego składu krotek do różnych celów, bez konieczności wydobywania kandydatów z korpusów wielokrotnie. Dodatkowo umożliwia to zbudowanie informacji statystycznych na podstawie innego zestawu krotek niż ten, z którego wyrażenia wielowyrazowe mają być wydobywane. Jako praktyczny przykład takiego zadania można podać ten, w którym informacja statystyczna budowana jest na podstawie krotek zebranych za pomocą relacji okna, a wydobywanie wyrażen wielowyrazowych jedynie na podstawie innych relacji, przykładowo o określonych wzorcach strukturalnych.

Krok numer 4: budowanie zestawu cech

Niniejszy krok jest opcjonalny i może zostać pominięty jeśli użytkownik nie ma zamiaru korzystać z klasyfikatorów. Jeśli jednak użytkownik będzie chciał skorzystać z klasyfikatorów lub wygenerować zestaw cech, z który będzie można wykorzystać w oprogramowaniu *WEKA* to może do tego celu użyć programu o nazwie *FeatureMaker*. Zadaniem tego modułu jest wygenerowanie cech dla kandydatów na kolokacje, spośród których mają być w przyszłości wydobyte wyrażenia wielowyrazowe. Do wykonania tego zadania generator cech wykorzystuje miary asocjacyjne, których wyniki traktowane są jako cechy. Dopuszczalne jest także wykorzystanie uprzednio wyuczonych klasyfikatorów, przykładowo sieci neuronowych, do generowania cech opisujących kandydatów. Zestaw kandydatów można ograniczyć za pomocą filtru jeśli nie ma potrzeby wyznaczenia cech dla wszystkich krotek.

Aktualnie jedynym wspieranym formatem pliku wykorzystywanym do zapisu zestawu cech w pamięci nieulotnej jest *ARFF*.

Zamieszczona poniżej tabela 4.2 zawiera opisy parametrów programu *FeatureMaker*:

Nazwa	Typ	Opis
Skład krotek	nazwa folderu	ścieżka do folderu ze składem krotek
Wyjście programu	nazwa pliku	ścieżka do pliku, w którym zostaną zapisane wyniki
Zestaw JW	nazwa pliku	plik zawierający w każdej linii ciąg wyrazów oddzielonych spacjami, reprezentujący wyrażenie wielowyrazowe, wykorzystany w przypisywaniu klas krotkom
Generator cech	ciągi tekstowe	tekst reprezentujący wektorową miarę asocjacyjną, którą program ma wykorzystać do generowania cech dla krotek
Filtr krotek	ciąg tekstowy	tekst reprezentujący filtr, który program ma wykorzystać do wyznaczenia zestawu krotek, dla których wygenerować cechy
Liczba wątków	liczba całkowita	maksymalna liczba wątków do wykorzystania przez program

Tabela 4.2. Tabela zawiera nazwy, typy oraz opisy parametrów wykorzystywanych przez program *FeatureMaker*

Krok numer 5: ekstrakcja

Ostatnim krokiem tej ścieżki przetwarzania jest wykorzystanie wcześniej utworzonych danych do ekstrakcji wyrażeń wielowyrazowych z korpusów tekstowych. Zadanie to może zostać wykonane za pomocą programu *Digger*. Zestaw parametrów programu jest zbliżony do tych dla narzędzie *Miner*, w poniższej tabeli 4.3 opisane zostały argumenty dla oprogramowania *Digger*:

Nazwa	Typ	Opis
Skład krotek	nazwa folderu	ścieżka do folderu ze składem krotek
Źródło tablic	nazwa pliku	ścieżka do pliku ze składem lub generatorem krotek
Wyjście programu	nazwa folderu	ścieżka do folderu, w którym zostaną zamieszczone wyniki
Miary asocjacji	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcje asocjacyjną, którą program ma wykorzystać
Miary wektorowe	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący wektorową miarę asocjacyjną, którą program ma wykorzystać
Agregatory	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcję agregującą wyniki miar wektorowych, po jednej funkcji dla każdej z nich
Klasyfikatory	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący klasyfikator, który program ma wykorzystać
Źródło cech	nazwa pliku	Źródło zestawu cech dla krotek
Zestaw JW	nazwa pliku	plik zawierający w każdej linii ciąg wyrazów oddzielonych spacjami, reprezentujący wyrażenie wielowyrazowe
Funkcje oceny	ciągi tekstowe	parametr wielokrotny, tekst reprezentujący funkcję oceny, którą program ma wykorzystać
Filtr krotek	ciąg tekstowy	tekst reprezentujący filtr, który program ma wykorzystać do wyznaczenia zestawu krotek, z których ma korzystać i, dla których wygenerować dane
Liczba wątków	liczba całkowita	maksymalna liczba wątków do wykorzystania przez program

Tabela 4.3. Tabela zawiera nazwy, typy oraz opisy parametrów wykorzystywanych przez program *Digger*

Zestaw generowanych wyników jest podobny do tego generowanego przez narzędzie *Miner*, ale jest mniejszy, ponieważ nie ma podziału na rundy i foldy ze względu na brak walidacji krzyżowej. Wzorce nazw dla plików k-najlepszych krotek i plików z ocenami dla tych list prezentują się teraz następująco: *kbest.nr_rankera.csv* oraz *quality.nr_rankera.nr_funkcji_oceny.csv*. Zasada generowania numeru rankera jest taka sama jak w przypadku programu *Miner*.

4.6.4. Funkcjonalności dodatkowe

Pakiet narzędzi *MWeXtractor* posiada też dodatkowe funkcjonalności, z których część została opisana w tej części niniejszej pracy.

Program *Cover*

Program nazwany *Cover* służy do weryfikacji relacji oraz badania przecięcia zbioru wydobytych krotek pomiędzy relacjami – sprawdzenie jak dużo z nich zostało przyporządkowanych do więcej niż jednej relacji. Praktyczne wykorzystanie tego oprogramowania to sprawdzenie pokrycia wzajemnego operatorów *WCCL* zastosowanych do wydobywania krotek z tekstów. Wynikiem jego działania jest macierz kwadratowa o wymiarze równym liczbie relacji wykorzystanych w procesie tworzenia *składu krotek*, gdzie każdy z wierszy i każda z kolumn mają przypisaną do siebie określoną nazwę operatora *WCCL*. Liczby całkowite zapisane na przecięciu wierszy z kolumnami informują ile różnych krotek zostało przypisanych zarówno do relacji będącej etykietą wiersza jak i kolumny. Interpretacją liczby na przecięciu wiersza i kolumny – diagonalu, odpowiadającym tej samej relacji jest informacja ile krotek zostało zakwalifikowanych do tej relacji. Jeśli w danym wierszu lub kolumnie wystąpi więcej niż jedna wartość większa od zera poza polem na diagonalu macierzy oznacza to, że relacje są nierozłączne – część krotek przyporządkowano do więcej niż jednej relacji. Należy jednak pamiętać, że narzędzie to nie jest wyrocznią, ponieważ jeśli dla danego zbioru danych okaże się, że relacje są rozłączne to nie oznacza to, że dla innego zbioru, większego czy lepiej oddającego rzeczywistość, sytuacja się powtórzy. Zbudowana w ten sposób macierz jest symetryczna.

Opisywany tutaj program *Cover* ma także drugie zastosowanie wspomniane we wcześniejszej części tego fragmentu niniejszej pracy – badanie przecięcia zbioru krotek ze składu i tego zadanego parametrem programu. Na podstawie zadanego pliku z ciągami wyrazów program bada ile z nich, oraz które konkretnie ciągi zostały odnalezione w *składzie krotek*. Funkcjonalność ta może posłużyć przykładowo do wyznaczania zbioru krotek pozytywnych, czyli wyrażeń wielowyrazowych, które zostały w tym tekście odnalezione podczas wydobywania kandydatów na kolokacje. Dodatkowo zakres poszukiwań wśród krotek w składzie można zawęzić za pomocą filtru. Zabieg taki pozwala na lepsze poznanie zbioru danych.

Wydobywanie form napotkanych

Program *Digger* umożliwia dodatkowo wydobycie form napotkanych dla krotek z wygenerowanych list *k*-najlepszych kandydatów na kolokacje. Jeśli użytkownik jest zainteresowany ekstrakcją form napotkanych to powinien podać cztery dodatkowe parametry dla tego programu, które zostały opisane w poniższej tabeli 4.4:

Wyekstrahowany zestaw form napotkanych dla krotek zostanie zapisany w pojedynczym pliku w formacie *csv*. Format pliku jest następujący: najpierw zapisana zostaje reprezentacja krotki z jednej z list *k*-najlepszych, a część kolejnych linii rozpoczyna się tabulacją i zawiera ciąg wyrazów napotkanych dla tej krotki, wraz z jego częstością. Następnie zapis taki się powtarza dla wszystkich wydobytych krotek jeśli jest ich więcej, ale bez powtórzeń jeśli krotki wystąpiły na różnych listach *k*-najlepszych. Dodatkowo napisany został skrypt łączący listy *k*-najlepszych z zestawem wydobytych dla nich form napotkanych i tworzy

Nazwa	Typ	Opis
Zestaw korpusów	nazwa pliku	ścieżka do pliku zawierającego listę ścieżek do korpusów, po jednej ścieżce w każdej linii pliku
Operatory WCCL	nazwa pliku	ścieżka do pliku z operatorami języka <i>WCCL</i> , które były użyte przy wydobywaniu kandydatów na kolokacje w kroku numer jeden
Tagset	ciąg znaków	nazwa wykorzystanego tagsetu w korpusach
Czytnik	ciąg znaków	nazwa czytnika korpusów

Tabela 4.4. Tabela zawiera nazwy, typy oraz opisy parametrów dodatkowych wykorzystywanych przez program *Digger*

nowe listy k-najlepszych zachowując kolejność rankingu i przypisując do każdej pozycji zestaw form napotkanych tej krotki wraz częstościami tych form napotkanych. Kolejność form napotkanych po łączeniu będzie posortowana malejąco po ich częstościach. Skrypt ten napisany w języku *Python* został nazwany *kbest_orth_merger.py* i jest częścią pakietu *MWeXtractor*.

4.6.5. Format składowania krotek zebranych z korpusów

Jako format przechowywania danych wykorzystano krotki o różnych długościach¹ co umożliwia składowanie informacji o wyrażeniach wielowyrazowych dowolnej długości w intuicyjny sposób. Krotki są dość powszechnym, elastycznym i prostym formatem stosowanym do przechowywania danych, a dodatkowo zapisywane są w postaci tekstu czytelnego dla człowieka co czyni format przejrzystym, łatwym do ewentualnej edycji i ułatwia analizę zapisanych informacji.

Krotka reprezentująca wyrażenie wielowyrazowe zawiera w sobie słowa w ustalonej kolejności wraz z ich częściami mowy, które wchodzi w skład tego wyrażenia. Każda z krotek zawiera także informację w jakiej relacji wystąpiły zawarte w tej krotce wyrazy. Relacja jest jednym z operatorów *WCCL* wykorzystanym w procesie tworzenia krotek na podstawie korpusów - proces ten został opisany w dalszej części tej sekcji. Jeśli dane zestawienie wyrazów w tekście spełnia wymagania dla kilku różnych wyrażen języka ograniczeń to dla każdego z nich powstanie osobna krotka zawierająca te same elementy, ale różniącą się relacją. Dodatkowo w krotce zawarte są także metadane o niej takie jak jej częstość² w przetworzonych korpusach danych. Relacja nie jest uwzględniana w rozmiarze krotki. Format składowania krotek w pliku tekstowym wraz z przykładem został zamieszczony w poniższej tabeli 4.5:

¹ Długość krotki rozumiana jest jako liczba jej elementów składowych, przykładowo wyrazów.

² Częstość krotki określa liczbę wystąpień konkretnego zestawienia słów w danej relacji

nazwa relacji	arność relacji	częstość krotki	cz.m.:s1	cz.m.:s2	...
AdjSubst	2	17	adj:nowy	subst:but	

Tabela 4.5. Format składowania krotek w pliku tekstowym wraz z przykładem. Elementy (ciągi znakowe) składowane w pliku tekstowym oddzielone są od siebie tabulatorami. Arność relacji jest tożsama z rozmiarem krotki. Skrótowiec *cz.m.:sN* pochodzi od: część mowy, dwukropek, *N*-te słowo krotki.

Istotnym elementem przetwarzania tak dużych zbiorów danych jest wydajny sposób składowania ich w pamięci operacyjnej maszyny przetwarzającej. Do tego celu opracowana, zaimplementowana i wykorzystana została struktura starająca się minimalizować zużycie pamięci przy zachowaniu szybkiego dostępu do danych poprzez wykorzystanie implementacji zbiorów opartych o drzewa czerwono-czarne i funkcje skrótu³. Struktura składa się z pięciu modułów odpowiedzialnych za przechowywanie innych informacji.

Moduł numer jeden zawiera w sobie informacje o wykorzystanych korpusach. Zapisana jest nazwa korpusu, ścieżka do pliku z tym korpusem, a także częstości słów w każdym z korpusów z osobna oraz suma wystąpień tych wyrazów.

Drugi z modułów zawiera w sobie podzbiór części mowy występujących w tagsecie wykorzystanym do opisu morfo-syntaktycznego korpusów danych użytych do ekstrakcji kandydatów na wyrażenia wielowyrazowe. Podzbiór ten zawiera tylko te części mowy, które wystąpiły przynajmniej raz w przetwarzanych tekstach, ale nie stoi na przeszkodzie by dodać także części mowy niewchodzące w skład wykorzystanego tagsetu.

Trzeci moduł wykorzystuje informacje z moduły związanego z częściami mowy wyrazów. Jego zadaniem jest składowanie wszystkich słów wraz z przyporządkowanymi im częściami mowy oraz częstością tych właśnie wyrazów wygenerowaną na podstawie korpusów tekstowych. Moduł musi przechowywać wszystkie słowa, które będą składowymi krotek, ale może przechowywać także wyrazy nadmiarowe.

Moduł numer cztery składa się dane o relacjach jakie zostały użyte do tworzenia krotek na podstawie korpusów tekstowych. Przechowywane są w nim informacje takie jak nazwy relacji, ich arności określające długość generowanych krotek oraz liczba krotek, które zostały utworzone z wykorzystaniem tej właśnie relacji. Relacje, które nie wykreowały żadnego kandydata na kolokacje także są w tym składzie uwzględnione wraz z informacjami o sobie. Ważną informacją jest także ta, że w tej strukturze nie są przechowywane ciała wyrażen *WCCL* definiujących daną relację.

Piąty z modułów jest odpowiedzialny za składowanie krotek oraz metadanych ich opisujących. Struktura składowania tych danych może być postrzegana jako *N*-hipermacierz[?, rozdział 15], której liczba wymiarów *N* jest równa długości najdłuższej z krotek, powiększonej o jeden, formalnie:

³ Popularna nazwa anglojęzyczna to *hash function* - funkcja haszująca.

$$N = 1 + maks(rozmiar(k) \forall k \in K : K = \{zestaw\ wszystkich\ krotek\})$$

Każdy wymiar odpowiada jednemu ze słów krotki lub jej relacji. Pierwsze K wymiarów odpowiada kolejno wyrazom kolokacji wraz z ich częściami mowy, a wymiar $K + 1$ jest interpretowany jako opisujący relację tej krotki. Prostymi wnioskami są $(K+1) \leq N$ oraz to, że nie wszystkie wymiary w macierzy dla danej krotki są zawsze wykorzystane. Przykładowo rozważmy skład zawierający w sobie tylko dwie poniższe krotki:

relacja	arność	cz.m:s1	cz.m:s2	cz.m:s3	częstości...
AdjSubstAdj	3	adj:czerwony	subst:samochód	adj:sportowy	...
AdjSubst	2	adj:czerwony	subst:kartka	[brak]	...

Dla podanego zestawu krotek wymiar hipermacierzy będzie równy cztery - trzy wyrazy najdłuższej z krotek plus jeden. Pierwszy wymiar tej hipermacierzy dla obu krotek opisuje pierwszą składową krotki - dla obu z nich będzie to *adj:czerwony*. Drugi wymiar zawiera informacje o drugiej składowej kandydatów na kolokację i są to odpowiednio słowa *subst:samochód* oraz *subst:kartka*. Trzeci wymiar w przypadku pierwszej krotki odpowiada jej trzeciej składowej, a w przypadku krotki numer dwa - jej relacji. Ostatni z wymiarów, czwarty, jest zdefiniowany tylko dla dłuższej z obu krotek i odnosi się do jej relacji. *Hipermacierz* krotek jest indeksowana słowami wraz z ich częściami mowy i opcjonalnie za pomocą relacji. Podczas odnoszenia się do tej macierzy nie jest wymagane podanie wszystkich składowych indeksu⁴ - wszystkich słów i relacji, ale wymagane jest określanie wartości po kolei - nie mogą powstać dziury pomiędzy zdefiniowanymi wartościami dla wymiarów, zachowana musi zostać ciągłość w procesie tworzenia indeksu. Zapis formalny ciągłości indeksu:

$$\forall D_i : D_i \in D = \{wymiar\ hipermacierzy\}, i \geq 1 \exists D_{i-1} : \\ warty(D_{i-1}) \neg niezdefiniowany$$

Poniższa tabela 4.6 korzysta z przytoczonego wcześniej przykładu i zawiera przykłady poprawnych oraz błędnych - nieciągłych, indeksów dla 4-hipermacierzy krotek.

Niech pojęcie *I-indeks* odnosi się do indeksu *I*-elementowego, czyli struktury ze zdefiniowanymi *I* elementami odpowiedzialnymi za indeksowanie danych w *hipermacierzy* krotek. Efektem wykorzystania *I-indeksu* do odwołania się do danych w *N-hipermacierzy* krotek jest *K-hipermacierz* danych⁵, gdzie spełniona jest nierówność $K \leq N$, a liczba wymiarów K jest zależna od N oraz I i równa:

$$K = N - I$$

⁴ Pojęcie indeksu odnosi się do struktury wykorzystywanej do odwoływania się do danych w *hipermacierzy* krotek.

⁵ W przypadku indeksu o rozmiarze zero tej samej *hipermacierzy*.

cz.m:s1	cz.m:s2	cz.m:s3	relacja	poprawny
adj:ładny,	subst:samochód,	adj:sportowy	AdjSubstAdj	tak
adj:ładny,	subst:samochód,	- ,	-	tak
adj:ładny,	- ,	- ,	-	tak
- ,	subst:samochód,	adj:sportowy	-	nie
- ,	adj:ładny,	subst:samochód	AdjSubstAdj	nie
adj:ładny,	adj:ładny,	- ,	AdjSubstAdj	nie

Tabela 4.6. Przykładowe indeksy dla 4-hipermacierzy krotek. Wszystkie przykłady niepoprawnych indeksów złamały tę samą zasadę dotyczącą ciągłości indeksu - nie można definiować wartości dla wymiaru D jeśli nie została ona zdefiniowana dla $D - 1$ ($D = 0$ jest wyjątkiem).

Wartość $K = 0$ jest całkowicie poprawna i oznacza skalar, konkretną krotkę, $K = 1$ oznacza wektor, dla $K = 2$ jest macierzą dwuwymiarową, $K = 3$ to 3-hipermacierz i tak dalej. Dla wyjaśnienia i lepszego zrozumienia indeksowania rozważmy poniższą *hipermacierz* z przyporządkowaniem wierszom, kolumną itd., liczb zamiast słów i relacji.

$$M = \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 4 & 5 & 6 \\ 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \end{array} \right]$$

Powyższy zapis symbolizuje 3-hipermacierz liczb naturalnych o wymiarach $4 \times 3 \times 2$ - cztery wiersze, trzy kolumny oraz dwie warstwy, indeksowane od zera. Liczby 1, 11, 21, 31 oraz 4, 14, 24, 34 są etykietami wierszy, 1, 2, 3 oraz 4, 5, 6 etykietami kolumn, a liczby 1 oraz 4 są także etykietami warstw tej macierzy. Część po lewej stronie pionowej linii odzwierciedla pierwszą warstwę *hipermacierzy*, a po prawej drugą. Utworzenie i wykorzystanie 1-indeksu postaci $[4, -, -]$ zwróci zredukowaną do dwóch wymiarów macierz będącą po prostu drugą warstwą 3-hipermacierzy wyjściowej, postaci:

$$M[4] = \left[\begin{array}{ccc} 4 & 5 & 6 \\ 14 & 15 & 16 \\ 24 & 25 & 26 \\ 34 & 35 & 36 \end{array} \right]$$

Dodanie wartości 5 do 1-indeksu spowoduje tym razem utworzenie 2-indeksu postaci $[4, 5, -]$, którego wykorzystanie zaowocuje odwołaniem się do konkretnego wektora - kolumny, zawierającego się w macierzy i równego:

$$M[4, 5] = \begin{bmatrix} 5 \\ 15 \\ 25 \\ 35 \end{bmatrix}$$

Ostatecznie rozwinięcie 2-indeksu o jeszcze jedną wartość, przykładowo 25, i utworzenie 3-indeksu $[4, 5, 25]$, a następnie odwołanie się za jego pomocą do rozważanej 3-hipermacierzy liczb spowoduje zwrócenie konkretnej krotki zawierającej kolejne słowa oznaczone numerami 4 i 5 oraz przyporządkowanej do relacji oznaczonej liczbą 25. Dodać należy, że kolejność interpretacji wymiarów jest jednak dowolna i nie musi przebiegać w tej kolejności.

Wykorzystanie indeksów dłuższych niż N dla N -hipermacierzy nie jest też błędem, ale spowoduje niezwrócenie żadnej wartości. Podanie indeksu niepełnego, tak jak było napisane we wcześniejszej części tego fragmentu niniejszej pracy, zwróci w ogólności pewną hipermacierz, a tym samym zestaw krotek. Innymi słowy można się odwoływać do konkretnego zestawu krotek spośród całego ich zbioru. Każda z krotek zwróconych poprzez wykorzystanie I -indeksu będzie miała I pierwszych elementów takich jak w indeksie, a reszta będzie dowolnymi elementami. Także relacja może być konkretna jeśli zostanie podana jako składowa indeksu. Możliwe jest utworzenie indeksu nie posiadającego zdefiniowanego żadnego z elementów, a jego wykorzystanie zaowocuje zwróceniem pełnej hipermacierzy.

4.6.6. Przedstawienie sposobu wydobywania i przechowywania informacji statystycznych wykorzystywanych w obliczeniach

Przed omówieniem szczegółów dotyczących składowania i pozyskiwania danych wprowadzone zostaną następujące oznaczenia, które będą także wykorzystywane w dalszej części tej pracy. Oznaczenia zostały zamieszczone na poniższym spisie:

1. x lub x_i - element krotki x lub i -ty element krotki;
2. \bar{x} - element inny niż x ;
3. x_1, x_2, \dots, x_n - zestaw n -elementów krotki;
4. n - długość krotki;
5. N - liczba wszystkich krotek w zbiorze danych;
6. $f(x)$ - częstość x , wartość zaobserwowana, liczba wystąpień;
7. $p(x)$ - prawdopodobieństwo x ;
8. $\hat{f}(x)$ - wartość oczekiwana x ;
9. $FA(x, y)$ - wartość funkcji asocjacyjnej dla elementów x oraz y .

Podstawowymi danymi stosowanymi w procesie wydobywania wyrażeń wielowyrazowych z grona kandydatów są tablice wielodzielne dla krotek skonstruowane przy wykorzystaniu częstości kolokacji zebranych z korpusów tekstowych. Tablica wielodzielna jest terminem stosowanym w niniejszej pracy rozumianym jako n -wymiarowa tablica, której rozmiar każdego z wymiarów jest równy dwa. Liczba wymiarów tej tablicy jest równy 2^n , a każde jej pole jest numerowane w implementacji od 0 do $n - 1$, ale dla w dalszych częściach tej pracy dla uproszczenia przyjęta zostanie numeracja od 1 do n . Omówiony tutaj zostanie sposób generowania tablicy dla wartości zaobserwowanych z danych, natomiast dla drugiej z tablic w dalszej części tego rozdziału. Pierwszy indeks tablicy zawiera informacje o częstości konkretnej krotki, gdzie każdy z elementów jest znany, a kolejne pola tej *hiperkoci* zawierają informacje o liczbie wystąpień innych krotek, które mają tylko część składowych, czyli wyrazów i relację, takich samych. Ostatnie pole w tej tablicy mówi o tym ile jest krotek takich, że wszystkie ich elementy składowe są inne niż te, dla której tablica została utworzona. Kolejne pola tablicy zawiera informacje o częstościach kolejnych krotek: $x_1, x_2, x_3, \dots, x_n, x_1, \bar{x}_2, x_3, \dots, x_n, \bar{x}_1, x_2, x_3, \dots, x_n, \bar{x}_1, \bar{x}_2, x_3, \dots, x_n$, a pełna tablica zawiera informacje o częstościach dla wszystkich możliwych kombinacji. Przykładowa tablica dla trójelementowej krotki została zaprezentowana w tabeli 4.7 zamieszczonej poniżej:

Indeks.	częstość krotki zawierającej elementy
1.	x_1, x_2, x_3
2.	x_1, \bar{x}_2, x_3
3.	\bar{x}_1, x_2, x_3
4.	$\bar{x}_1, \bar{x}_2, x_3$
5.	x_1, x_2, \bar{x}_3
6.	$x_1, \bar{x}_2, \bar{x}_3$
7.	$\bar{x}_1, x_2, \bar{x}_3$
8.	$\bar{x}_1, \bar{x}_2, \bar{x}_3$

Tabela 4.7. Przykładowa tablica wielodzielna obrazująca informacje o częstościach w niej składowanych

Indeksowanie tablicy dla innej liczby argumentów przebiega analogicznie. Wzory zamieszczone w dalszej części tego rozdziału mogą wykorzystywać informacje z takiej tablicy poprzez zapis $t(i)$, co oznacza odwołanie do i -tego elementu tablicy utworzonej dla konkretnego kandydata na kolokację.

Oprogramowanie *MWeXtractor* umożliwia tworzenie omówionych tablic wielodzielnych na dwa sposoby tak jak wspomniano wcześniej przy okazji opisu schematu przetwarzania – utworzenie gotowego składu tablic wielodzielnych dla każdego kandydata na wyrażenie wielowyrazowe lub poprzez przygotowanie generatora, który tworząc odpowiedni indeks będzie w stanie generować tablice wielodzielne dla kolokacji w trakcie działania programu. Obie przedstawione tutaj metody w efekcie generują takie same wyniki i tego

samego typu – tablice wielodzielne. Bez względu na wykorzystaną metodę tworzenia takich tablic, ich efektem będzie struktura nazywana *Źródłem tablic wielodzielnych*. Proces tworzenia tablic z wartościami zaobserwowanymi składa się z kilku kroków. Pierwszym z nich jest wyselekcjonowanie zestawu krotek, które mają zostać wykorzystane do budowy tablic wielodzielnych, za pomocą zaimplementowanego mechanizmu filtrów. Kolejnym zadaniem jest już właściwy proces budowania tablic i polega on na utworzeniu generatora poprzez zbudowanie indeksu zawierającego informacje o częstościach danych krotek, a także sumy częstości krotek utworzonych poprzez kombinacje elementów składowych tych kolokacji – kombinacje te są rozumiane tak jak te przedstawione w opisie tablic wielodzielnych z tym, że zamiast elementu innego niż stosowany jest zapis mówiący o tym, że element jest dowolny. Następnie gotowy generator może być już wykorzystany do tworzenia tablic wielodzielnych lub można na jego podstawie wygenerować pełny skład tablic dla konkretnych krotek.

Generowanie tablic z wartościami oczekiwanymi jest procesem szybszym od zbierania danych o wartościach zaobserwowanych i polegającym na obliczeniu iloczynu prawdopodobieństw odpowiadających konkretnym elementom krotki, a następnie wynik ten jest przemnożony przez liczbę wszystkich krotek o zadanej długości, przykładowo dwuelementowych. Dane do tych obliczeń pozyskiwane są z tablicy wielodzielnej wartości zaobserwowanych. Jeśli element krotki jest określony i konkretny to jego prawdopodobieństwo jest równe $\frac{f(element)}{E}$, gdzie E jest sumą częstości wszystkich krotek dwuelementowych w generatorze (uwzględnionych w obliczeniach po zastosowaniu filtru). Natomiast jeśli element krotki jest zanegowany, czyli *inny niż*, to jego prawdopodobieństwo wykorzystane we wzorze jest równe $1 - p(element)$. Dla lepszego zrozumienia tego zagadnienia rozważmy przykład obliczenia wartości oczekiwanej dla krotki 3-elementowej ($A\bar{B}C$), gdzie $E = 10$, $A = 1$, $B = 3$ i $C = 2$:

$$\bar{f}((A\bar{B}C)) = \frac{A}{E} * (1 - \frac{B}{E}) * \frac{C}{E} * E = \frac{1}{10} * (1 - \frac{3}{10}) * \frac{2}{10} * 10 = 0,14$$

Otrzymany w ten sposób wynik jest wartością oczekiwaną dla krotki ($A\bar{B}C$).

Obie omówione metody generowania tablic mają wady i zalety, porównanie efektów działania obu metod zostało zamieszczone w tabeli 4.8 zamieszczonej poniżej:

Odwoływanie się do *Źródła tablic wielodzielnych* jest możliwe poprzez podanie poprawnego identyfikatora krotki w składzie krotek lub podanie konkretnej krotki, dla której tablica ma zostać utworzona. Jeśli odwołanie następuje przez krotkę, a nie identyfikator, to musi ona istnieć w składzie krotek tylko w przypadku korzystania ze składu tablic, generatora nie ma tego ograniczenia.

Cecha	porównanie
Zajętość pamięciowa	generator zajmuje kilkukrotnie mniej pamięci operacyjnej niż skład;
Szybkość generacji źródła	stworzenie generatora jest szybsze od utworzenia składu;
Szybkość generacji tablic	skład działa znacznie szybciej od generator, ponieważ musi jedynie znaleźć gotową tablicę, a generator musi ją konstruować na bieżąco za każdym razem;
Generalizacja	skład nie posiada żadnych możliwości generalizacji, ponieważ każda z tablic jest utworzona dla konkretnej krotki, generator natomiast na podstawie zebranych danych jest w stanie utworzyć tablicę, dla krotki, z którą wcześniej nie miał styczności.

Tabela 4.8. Przykładowa tablica wielodzielna obrazująca informacje o częstościach w niej składowanych

4.6.7. Mechanizm filtrów i zaimplementowane metody filtracji

Mechanizm filtrów w oprogramowaniu *MWeXtractor* został zaimplementowany jako ciągi funkcji filtrujących wspierane przez logikę dwuwartościową. Dodatkowo napisana została funkcjonalność tworzenia złożonych filtrów z ciągów tekstowych co pozwala na ich łatwą reprezentację, składowanie na dysku twardym w postaci czytelnej formy i proste, intuicyjne modyfikacje. Ze względu na zastosowane poziomy abstrakcji możliwe jest tworzenie nowych filtrów i łączenie ich z już istniejącymi. Przykładowy filtr w reprezentacji tekstowej został zamieszczony poniżej:

or(relation(\wedge , file = relacje.csv, SubstAdj), and(frequency($>$, 5), corpora_frequency($<$, 17)), not

Wykorzystanie powyższego ciągu tekstowego spowoduje utworzenie filtru, który zwróci identyfikatory wszystkich krotek znajdujących się w zadanym składzie krotek, dla których spełniony zostanie przynajmniej jeden z następujących warunków:

1. *relation(\wedge , file = relacje.csv, SubstAdj)* - relacja krotki musi być zawarta w zbiorze zamieszczonym w pliku lub będzie dokładnie relacją o nazwie *SubstAdj*;
2. *and(frequency($>$, 5), corpora_frequency($<$, 17))* - częstość krotki po dyspersji musi być większa od 5, ale jednocześnie jej częstość sprzed dyspersji musi być mniejsza niż 17;
3. *not(tuple(\wedge , file = zestaw_krotek.csv))* - wyrazy krotki nie tworzą żadnego z ciągów słów zdefiniowanych w pliku *zestaw_krotek*.

Filtry dzielą się na dwa typy: filtry logiczne (łańcuchowe) i filtry cech. Zadaniem filtrów logicznych jest budowanie ciągów złożonych filtrów. Nie wykonują one same w sobie żadnych konkretnych operacji filtrujących, a jedynie łączą działania innych filtrów, w tym filtrów cech. Argumentami dla tego typu filtru mogą być inne filtry logiczne lub filtry cech.

Wspierane przez oprogramowanie są następujące operatory logiczne, które można w sobie zagnieżdżać i budować ciągi filtrów:

1. *or* - n-elementowa suma logiczna;
2. *and* - iloczyn logiczny n-elementów;
3. *xor* - n-elementowa suma wykluczająca;
4. *not* - negacja tylko pojedynczego argumentu.

Postać ogólna funkcji logicznej wykorzystywanej w filtracji jest następująca:

$$\textit{nazwa_funkcji_logicznej}([, \textit{arg1}][, \textit{arg2}]...[, \textit{argN}])$$

Opisane w dalszej tej pracy filtry są tak zwanymi *filtrami cech*, z których każdy wspierać może wszystkie lub tylko część z wybranych operatorów filtrujących przedstawionych poniżej:

1. *=* - jest równy;
2. *>=* - jest większy równy;
3. *<=* - jest mniejszy równy;
4. *>* - jest większy;
5. *<* - jest mniejszy;
6. *^* - argumenty zawarte są w zbiorze;

Filtry cech są filtrami końcowymi co oznacza, że w ogólności nie można w nich zagnieżdżać kolejnych metod filtrujących, chyba że określona funkcja filtrująca zostanie w odpowiedni sposób zaimplementowana. Postać ogólna filtru cech jest następująca:

$$\textit{nazwa_filtru}(\textit{typ_operatora}[, \textit{file} = \textit{sciezka_do_pliku.ext}][, \textit{arg1}][, \textit{arg2}]...[, \textit{argN}])$$

W dalszej części tego rozdziału zamieszczono opisy filtrów cech aktualnie zaimplementowanych w oprogramowaniu *MWeXtractor*.

Częstość kolokacji

Zaimplementowane operatory tego filtru to: *=*, *>=*, *<=*, *>*, *<*. Filtr sprawdzający czy częstość danej krotki spełnia odpowiedni warunek. Argumentem dla tego filtru jest pojedyncza wartość będąca drugim argumentem dla operatora. Podanie większej liczby argumentów spowoduje zignorowanie parametrów nadmiarowych. Parametr *file* jest dla tego filtru ignorowany. Przykładowy filtr częstości:

$$\textit{frequency}(>, 1337)$$

Częstość kolokacji w korpusach

Różnica pomiędzy tym filtrem a poprzednim jest taka, że ten bierze pod uwagę częstość krotki z korpusów tekstowych, a nie jej globalną częstość. Praktyczna różnica pomiędzy nimi jest tylko w sytuacji kiedy skład krotek poddano dyspersji kolokacji, jeśli tego nie zrobiono to wynik będzie taki sam dla obu filtrów. Ta wersja filtru bierze pod uwagę częstość krotki sprzed wykonania dyspersji. Nic nie stoi na przeszkodzie, aby wykorzystywać oba filtry naraz jeśli pracuje się ze składem poddanym uprzednio dyspersji krotek. Przykład tego filtru zamieszczono poniżej:

$$corpora_{frequency}(>, 1337)$$

Zawieranie się krotki w podzbiorze ciągów słów

Zaimplementowane operatory tego filtru to: \wedge . Filtr zwróci wartość *prawda* tylko w sytuacji, gdy ciąg utworzony z wyrazów w krotce, bez relacji, zostanie odnaleziony w zadanym zbiorze. Argumentami dla tego filtru są wartość parametru *file* będąca ścieżką do pliku z ciągami słów oraz informacja o tym czy ciągi słów powinny być rozważane jako o szyku wolnym czy ustalonym. Przykład filtru:

$$tuple(\wedge, file = ciagi_wyrazow.csv)$$

Zawartość słów w zadanym podzbiorze

Zaimplementowane operatory tego filtru to: \wedge . Zadaniem tego filtru cech jest sprawdzenie czy wszystkie słowa zawarte w krotce z pominięciem ich części mowy zawierają się w zbiorze słów. Filtr przyjmuje na wejście dowolną liczbę parametrów będących słowami oraz parametr *file* – ścieżkę do pliku z zestawem wyrazów. Kolejność podawania argumentów nie ma znaczenia. Końcowy zbiór słów jest połączeniem wyrazów wczytanych z pliku z tymi zadanymi argumentami. Przykład filtru:

$$every_word(\wedge, wyraz2, file = wyrazy.csv, slowo1, magister)$$

Relacja kolokacji należąca do zbioru

Zaimplementowane operatory tego filtru to: \wedge . Celem tego filtru jest odnalezienie wszystkich krotek, których nazwa relacji znajduje się w zadanym zbiorze. Można o nim myśleć jak o uproszczonym filtrze badającym zawartość słów w krotce i biorącym pod uwagę relacje, a nie wyrazy. Kolejność definiowania argumentów dla tego filtru nie ma znaczenia, a sposób budowania zbioru i parametry są takie same jak w filtrze badającym zawartość słów w określonym zbiorze. Jedyna różnica jest taka, że argumenty to nazwy relacji. Przykład filtru:

$relation(\wedge, file = relacje.csv, relacja1, friendzone)$

4.6.8. Dostępne funkcje dyspersji

Zadaniem zaimplementowanych funkcji dyspersji jest zmiana częstości krotek na podstawie danych zawartych w krotkach i informacji o zbiorach danych – korpusach. Zmiana ta ma za zadanie wyróżnić z zestawu krotek te, które wydają się być bardziej nietypowe ze względu na swoją częstość i swój rozkład w dokumentach przykładowo rozumiany jako liczba dokumentów, w których krotka wystąpiła. Im dana kolokacja jest częstsza w obrębie jak najmniejszej liczby dokumentów to funkcja ta powinna ją wyróżnić, aby funkcje asocjacji przywiązywały do niej większą wagę podczas jej oceny. Natomiast jeśli krotka jest pospolita w rozumieniu jej występowania w prawie każdym korpusie to funkcja prawdopodobnie zmniejszy jej wagę dla funkcji asocjacyjnych – zależy to także od częstości samej krotki, ponieważ krotka pospolita, ale o dużej częstości także może być interesująca z punktu widzenia tej funkcji dyspersji, a w efekcie także dla miary asocjacji.

Na potrzeby wzorów opisujących te funkcje wprowadzono następujące oznaczenia:

1. c - korpus c ;
2. C - liczba korpusów;
3. t - krotka t ;
4. $f(t_c)$ - częstość krotki t w korpusie c .
5. $S(t)$ - zbiór wszystkich krotek opisujących tego samego kandydata t , ale w różnych szykach;
6. $S(t)_i$ - kandydat t w i -tym szyku;

Dalsza część tego fragmentu niniejszej pracy opisuje funkcje dyspersji zaimplementowane w oprogramowaniu *MWeXtractor*:

Distributional consistency

$$dc = \frac{\sum_{c=1}^C \sqrt{f(t_c)}}{C}$$

Źródło: [?, str. 7]

Odchylenie standardowe

$$\sigma = \sqrt{\frac{\sum_{c=1}^C (f(t_c) - \frac{\sum_{c=1}^C f(t_c)}{C})^2}{C}}$$

Źródło: [?, str. 6]

Variation coefficient – współczynnik wariancji

$$vc = \frac{\sigma C}{\sum_{c=1}^C f(t_c)}$$

Źródło: [?, str. 6]

Julliaand

$$jd = 1 - \frac{vc}{\sqrt{C - 1}}$$

Źródło: [?, str. 6]

Lynes D3

$$ld3 = \frac{1 - X^2}{0.25 * \sum_{c=1}^C f(t_c)}$$

Term frequency - Inverse document frequency

Opisywana tutaj funkcja dyspersji jest połączeniem dwóch elementów, jednego badającego unikalność na przestrzeni dokumentów – *IDF*, oraz drugiego mówiącego o tym jak bardzo dany obiekt, tutaj krotka, jest częsta w badanych danych. Definicje dla *Term frequency* oraz *Inverse document frequency* mogą przybierać różne formy i badać inne cechy. *TFIDF* służy do zbadania jak bardzo określona krotka jest unikalna dla zadanego dokumentu, a wzór tej funkcji jest następujący:

$$tfidf_{t,c} = tf_{t,c} * idf_c$$

Składowa *Term frequency* badająca częstość termu, zaimplementowana w niniejszym oprogramowaniu ma prostą postać i jest po prostu równa częstości krotki w zbadanym dokumencie:

$$tf_{t,c} = f(t_c)$$

Składowa *Inverse document frequency* określa jak bardzo unikalna jest badana krotka na przestrzeni danego zestawu korpusów. Wzór *IDF* przyjęty w implementacji jest popularny i opisany następującym wzorem wykorzystującym indykator:

$$idf_t = \log_{10} \frac{C}{\sum_{c=1}^C (f(t_c) > 0 ? 1 : 0)}$$

Ze względu na fakt, że *TFIDF* sprawdza jak bardzo krotka jest interesująca w danym korpusie, a nie całym ich zestawie, należało dokonać połączenia wyników tej miary dla

każdego z korpusów z osobna w jedną wartość. Zadanie to zostało wykonane z wykorzystaniem następującego wzoru:

$$TFIDF_t = \sum_{c=1}^C tfidf_{t,c}$$

4.6.9. Zaimplementowane miary asocjacji

Zestaw zaimplementowanych w oprogramowaniu *MWeXtractor* miar asocjacyjnych został opisany w tym rozdziale.

Największa częstość

$$y = f(x, y)$$

Jedna z pierwszych i najprostszych funkcji wykorzystanych do wydobywania kolokacji. Generalizacja tej funkcji jest trywialna, jej wzór zamieszczono poniżej:

$$y = f(x_1, x_2, \dots, x_n)$$

Wartość oczekiwana

$$y = \hat{f}(x, y)$$

Pomysł na funkcję wprowadzony przez autora niniejszej pracy. Generalizacja wzoru jest analogiczna do *Największej częstości*.

Odwrotna wartość oczekiwana

$$y = \frac{1}{\hat{f}(x, y)}$$

Pomysł na funkcję wprowadzony przez autora niniejszej pracy. Generalizacja wzoru jest analogiczna do *Największej częstości*.

Jaccard

$$y = \frac{f(x, y)}{f(x, y) + f(x, \hat{y}) + f(\hat{x}, y)}$$

Źródło: [?, str. 18].

Zaproponowana generalizacja funkcji opisana jest poniższym wzorem:

$$y = \frac{f(x_1, x_2, \dots, x_n)}{N - f(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)}$$

Dice

$$y = \frac{2f(x, y)}{f(x) + f(y)}$$

Źródło: [?, str. 18].

Generalizacja funkcji została zaczerpnięta z pracy [?, str. 2] i reprezentowana jest wzorem:

$$y = \frac{nf(x_1, x_2, \dots, x_n)}{\sum_{i=1}^n f(x_i)}$$

Sorgenfrei

$$y = \frac{f(x, y)^2}{(f(x, y) + f(x, \hat{y}))(f(x, y) + f(\hat{x}, y))}$$

Źródło: [?, str. 4]

Zaproponowana generalizacja funkcji wyrażona jest wzorem:

$$y = \frac{f(x_1, x_2, \dots, x_n)^n}{\prod_{i=2}^{n-1} (f(x_1, x_2, \dots, x_n) + table(i))}$$

Odds ratio

$$y = \frac{f(x, y)f(\hat{x}, \hat{y})}{f(x, \hat{y})f(\hat{x}, y)}$$

Źródło: [?, str. 18]

Zaproponowana generalizacja tej funkcji została zapisana w postaci poniższego wzoru, zmiana dotyczy także sposobu obliczania mianownika – dodawana jest jedynka ze względu na problem częstości zerowych:

$$y = \frac{f(x_1, x_2, \dots, x_n)f(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)}{\prod_{i=2}^{n-1} (t(i) + 1)}$$

Unigram subtuples

$$y = \log\left(\frac{f(x, y)f(\hat{x}, \hat{y})}{f(x, \hat{y})f(\hat{x}, y)}\right) - 3.29\sqrt{\frac{1}{f(x, y)} + \frac{1}{f(x, \hat{y})} + \frac{1}{f(\hat{x}, y)} + \frac{1}{f(\hat{x}, \hat{y})}}$$

Źródła: [?, str. 3], [?, str. 19]

Jest to miara, która w badaniach Pavla Peciny opisanych w [?, str. 5] okazała się jedną z dwóch najlepszych spośród zestawu 82 funkcji. Zaproponowana w tej pracy generalizacja i implementacja miary *Unigram subtuples* została dodatkowo zmodyfikowana przez wygładzenie części wartości - dodanie wartości jeden do częstości wystąpień wszystkich obserwacji w celu rozwiązania problemu częstości zerowych. Wzór po generalizacji i modyfikacji został zapisany poniżej:

$$y = \log(OddsRatio) - 3.29\sqrt{\sum_{i=1}^n \frac{1}{t(i)+1}}$$

Consonni T1

$$y = \frac{\log(1+f(x, y)+f(\hat{x}, \hat{y}))}{\log(1+N)}$$

Źródło: [?, str. 4]

Zaproponowana prosta generalizacja tej funkcji opisana została wzorem zamieszczonym poniżej:

$$y = \frac{\log(1+f(x_1, x_2, \dots, x_n) + f(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n))}{\log(1+N)}$$

Consonni T2

$$y = \frac{\log(1+N) - \log(1+f(x, \hat{y}) + f(\hat{x}, y))}{\log(1+N)}$$

Źródło: [?, str. 4]

Wprowadzona na potrzeby programu prosta generalizacja zapisana została za pomocą poniższego wzoru:

$$y = \frac{\log(1+N) - \log(1 + \sum_{i=2}^{n-1} t(i))}{\log(1+N)}$$

Mutual Expectation

$$y = p(x, y) \frac{2f(x, y)}{f(x) + f(y)}$$

Źródła: [?, str. 18], [?]

Zastosowana generalizacja miary *Mutual Expectation* zapisana za pomocą poniższego wzoru:

$$y = p(x_1, x_2, \dots, x_n) \frac{n f(x_1, x_2, \dots, x_n)}{\sum_{i=1}^n f(x_i)}$$

Pointwise Mutual Information

$$y = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Źródła: [?, str. 18], [?, str. 2]

Autor artykułu [?] prezentuje dwie generalizacje miary *MI* oraz wyprowadza z nich dwie generalizacje funkcji *PMI*. Nazwy generalizacji *Mutual Information* to *Total Correlation* i *Interaction Information*, a *Pointwise Mutual Information* to odpowiednio *Specific Correlation* oraz *Specific Information*. Wzory i opis obu miar został zamieszczony we wcześniejszej części tej pracy. Autor pracy przedstawił generalizację dowolnego poziomu dla miary *Specific Correlation* oraz generalizację dla rozmiaru trzy dla miary *Specific Information* z komentarzem, że można ją analogicznie rozwinąć dla obserwacji z większą liczbą elementów [?, str. 3].

Funkcja *Pointwise Mutual Information* w opisywanym tutaj programie została zastąpiona dwoma generalizacjami opisanymi we wspomnianym artykule [?] i noszą one takie same nazwy jak w artykule – *Specific Correlation* oraz *Specific Information*.

W Specific Correlation

$$y = p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Źródła: pomysł autora pracy oraz chyba jakieś źródło...

Autor niniejszej pracy do stworzenia tej miary opierał się mocno na funkcji *Specific Correlation* omówionej we wcześniejszej części tej pracy. Generalizacja omawianej miary opisana została poniższym wzorem:

$$y = p(x_1, x_2, \dots, x_n) \log_2 \frac{p(x_1, x_2, \dots, x_n)}{\prod_{i=1}^n p(x_i)}$$

Mutual Dependency

$$y = \log_2 \frac{p(x, y)^2}{p(x)p(y)}$$

Źródło: [?, str. 2]

Miara jest rozwinięciem funkcji *PMI* i według autorów pracy [?] sprawdza się lepiej niż pierwowzór. Zastosowana w omawianym tutaj programie generalizacja tej miary opisana została wzorem:

$$y = \log_2 \frac{p(x_1, x_2, \dots, x_n)^2}{\prod_{i=1}^n p(x_i)}$$

Funkcja w omawianym programie nosi nazwę *Specific Mutual Dependency*, ponieważ jest zaimplementowana w wersji uogólnionej, a sam pomysł na jej generalizację autor niniejszej pracy zaczerpnął z artykułu [?].

Frequency biased Mutual Dependency

$$y = \log_2 \frac{p(x, y)^3}{p(x)p(y)}$$

Źródło: [?, str. 2]

Omawiana miara także jest rozwinięciem funkcji *PMI* lub *MD* i według autorów pracy [?] sprawdza się lepiej niż obie z nich. Zaproponowana generalizacja tej miary została opisana za pomocą poniższego wzoru:

$$y = \log_2 \frac{p(x_1, x_2, \dots, x_n)^3}{\prod_{i=1}^n p(x_i)}$$

Funkcja w omawianym programie nosi nazwę *Specific Frequency biased Mutual Dependency*, ponieważ jest zaimplementowana w wersji uogólnionej, a sam pomysł na jej generalizację autor niniejszej pracy zaczerpnął z pracy [?].

Specific Exponential Correlation

$$y = \log_2 \frac{p(x, y)^e}{p(x)p(y)}$$

Źródło: autor niniejszej pracy oraz *buczynski*

Pomysł na funkcję wziął się z obserwacji trzech uprzednio opisanych funkcji - *PMI*, *MD* oraz *FbMD* i jest ich parametryczną generalizacją, gdzie parametrem jest wykładnik e . Funkcja jest swoim wzorem zbliżona do metody opisanej przez autora pracy [?], który zastosował miarę opisaną poniższym wzorem:

$$y = \log_2 \frac{p(x, y)^{2+e}}{p(x)p(y)}$$

Widać, że różnica polega tylko na wykładniku, dodaniu do niego stałej o wartości dwa. Generalizacja miary jest analogiczna do generalizacji *Frequency biased Mutual Dependency* i opisana wzorem:

$$y = \log_2 \frac{p(x_1, x_2, \dots, x_n)^e}{\prod_{i=1}^n p(x_i)}$$

W Specific Exponential Correlation

$$y = p(x, y) \log_2 \frac{p(x, y)^e}{p(x)p(y)}$$

Źródło: autor niniejszej pracy

Pomysł na funkcję wziął się z obserwacji kilku poprzednich i jest ich połączeniem. Generalizacja jest zbliżona do poprzednich generalizacji miar podobnych do tej funkcji i przedstawiona jest za pomocą wzoru:

$$y = p(x_1, x_2, \dots, x_n) \log_2 \frac{p(x_1, x_2, \dots, x_n)^e}{\prod_{i=1}^n p(x_i)}$$

T-score

$$y = \frac{f(x, y) - \hat{f}(x, y)}{\sqrt{f(x, y)(1 - \frac{\hat{f}(x, y)}{N})}}$$

Źródło: [?, str. 18]

Miara opiera się o test statystyczny hipotez o nazwie *T-test* zakładający rozkład normalny prawdopodobieństwa w zbiorze danych za co też jest krytykowany, ponieważ ten rozkład w praktyce często nie opisuje rzeczywistych danych [?, str. 169]. Ze względu omówionego w poprzedniej części pracy, dotyczącego faktu, że miara ta jest jedynie częścią *T-testu*, została ona nazwana *T-score*, a nie tak jak w źródle *T-test* [?, str. 18]. Miara została opisana także we wcześniejszej części tej pracy. Generalizacja jest w zasadzie taka sama jak wersja dwuelementowa i opisana poniższym wzorem:

$$y = \frac{f(x_1, x_2, \dots, x_n) - \hat{f}(x_1, x_2, \dots, x_n)}{\sqrt{f(x_1, x_2, \dots, x_n)(1 - \frac{\hat{f}(x_1, x_2, \dots, x_n)}{N})}}$$

Z-score

$$y = \frac{f(x, y) - \hat{f}(x, y)}{\sqrt{\hat{f}(x, y)(1 - \frac{\hat{f}(x, y)}{N})}}$$

Źródło: [?, str. 18]

Miara jest podobnie do *T-score* i podobnie jak ona, została opisana we wcześniejszej części tej pracy. Generalizacja *Z-score* zapisana została za pomocą poniższego wzoru:

$$y = \frac{f(x, y) - \hat{f}(x_1, x_2, \dots, x_n)}{\sqrt{\hat{f}(x_1, x_2, \dots, x_n)(1 - \frac{\hat{f}(x_1, x_2, \dots, x_n)}{N})}}$$

Pearson's χ^2

$$X^2 = \sum_i \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

Źródło: [?, str. 18]

Ta miara będąca częścią statystyczną jak dwie poprzednie, także została omówiona we wcześniejszej części tej pracy. Jej przewagą w stosunku do *Z-score* i *T-score* ma być niezakładanie rozkładu normalnego prawdopodobieństwa. Funkcja nie wymaga generalizacji.

Loglikelihood

$$G^2 = -2 \sum_i f_i \log \frac{f_i}{f_i}$$

Źródło: [?, str. 18]

Funkcja nie wymaga generalizacji.

Fair dispersion point normalization

$$y = \frac{1}{n-1} \sum_{i=1}^{n-1} AF(w_1 \dots w_i, w_{i+1} \dots w_n)$$

Źródła: [?], [?, str. 5]

Funkcja nie wymaga generalizacji, ponieważ została zaprojektowana jako wieloelementowa. Wykorzystuje wewnętrznie jedną z miar dwuelementowych do obliczania poziomów asocjacji bi-gramów, z których każdy został utworzony poprzez podział n-gramu w jednym z punktów dyspersji - miejscu między wyrazami. Przykładowo 4-elementowa krotka $x_1 x_2 x_3 x_4$ posiada trzy punkty dyspersji: pomiędzy x_1 i x_2 , x_2 i x_3 oraz x_3 i x_4 . Funkcja oblicza średni poziom asocjacji każdego z trzech utworzonych bi-gramów, gdzie każdy z nich składa się z pewnej liczby elementów, a jego częstość to liczba wystąpień danego zestawu słów w zbiorze danych. Dla przykładowej krotki utworzone zostać mogą następujące bi-gramy: $(x_1, x_2 x_3 x_4)$, $(x_1 x_2, x_3 x_4)$ i $(x_1 x_2, x_3 x_4)$. Znacznie szerszy opis metody został zaprezentowany w artykule źródłowym [?].

Average bigram

$$y = \frac{1}{n-1} \sum_{i=1}^{n-1} FA(w_i, w_{i+1})$$

Źródło: [?, str. 5]

Podobnie jak metoda poprzednia ta także jest uogólniona do obliczeń wieloelementowych. Funkcja została opisana we wcześniejszej części tej pracy i zakłada podział n-elementowej krotki na $n - 1$ kolejnych, ciągłych i zachodzących na siebie bi-gramów, następnie obliczenia wartości asocjacji za pomocą jednej z dwuelementowych funkcji asocjacji i uśrednienia wyniku, który zostanie zwrócony jako rezultat obliczeń tej funkcji.

Smoothed bigram

$$y = FA(x_1 x_2, x_2 x_3, \dots, x_{n-1} x_n)$$

[?, str. 5]

Miara dzieli n-elementową krotkę na bi-gramy tak jak *Average bigram*, a następnie traktuje każdy z nich jako pojedynczą składową krotki o liczbie elementów $n - 1$, gdzie

częstość tego elementu to liczba wystąpień danego bi-gramu w zbiorze danych. Natomiast częstość krotki pozostaje taka sama. Następnie poziom asocjacji tak zmodyfikowanej krotki $n - 1$ -elementowej jest obliczany za pomocą jednej z wieloelementowych funkcji asocjacyjnych.

Minimal bigram

$$y = \min_{i=1}^{n-1} FA(w_i, w_{i+1})$$

Źródło: Mariusz Paradowski

Funkcja działa analogicznie do *Average bigram*, ale zamiast obliczać średnią zwraca najmniejszą z wartości.

W order

$$y = \frac{1}{\prod_{i=1}^n (1 + \frac{f(S(t)_i)}{\max(f(S(t))) + 1})}$$

Źródło: autor niniejszej pracy

Motywacją do utworzenia tej funkcji jest założenie, że im kandydat na kolokację częściej występuje w mniejszej liczbie różnych szyków tym jest on ciekawszy, pewniejszy. Iloczyn został wybrany jako składowa tej miary ze względu, że jego wartość przy zachowaniu tej samej sumy składowych jest największa, gdy te składowe są sobie równe, a najmniejsza, gdy jedna z nich jest równa tej sumie lub wystąpią zera. Jest to cecha, która idealnie odwzorowuje zapotrzebowanie założenia jeśli odwróci się wynik takiego iloczynu. Funkcja abstrahuje od interpretacji kolejności szyku badając jedynie ich liczbę i rozkład częstości w tych szykach. Miara też bezpośrednio nie korzysta z częstości krotek dla danego kandydata, a jedynie bada ich stosunek – normalizacja w obrębie krotek dla pojedynczego kandydata. Dodanie jedynki do każdej częstości spełnia kilka funkcji. Pierwsza z nich to rozwiązanie problemu częstości zerowych, dodanie jedynki umożliwia otrzymywanie wartości innych niż nieskończoność w przypadku wystąpienia nawet pojedynczego zera, w którejkolwiek krotce dla danego kandydata. Druga funkcja dodania jedynki jest związana z pierwszą i umożliwia utworzenie rankingu, w którym krotki kandydata w danym szyku mają częstość równą zero. Konkretnie umożliwia przypisanie większej punktacji kandydatowi, którego mniejsza liczba krotek w danych szykach ma częstość zero. Trzecia funkcja dodania jedynki sprawia, że miara zaczyna brać pod uwagę ilość informacji statystycznych o danym kandydacie, ponieważ jeśli iloczyn częstości krotek w obrębie pojedynczego kandydata dla różnych kandydatów będzie taki sam to promowany będzie ten, o którym udało się zdobyć więcej informacji statystycznych – wystąpił więcej razy. Dodatkowo jedynka zmienia zakres możliwych do osiągnięcia wartości eliminując możliwość wystąpienia nieskończoności dodatniej.

W term frequency order

$$y = f(t)WOrder(t)$$

Źródło: autor niniejszej pracy

Motywacją tej funkcji było to, że wiele różnych miar asocjacyjnych ocenianych jako dobre korzysta bezpośrednio z częstości kandydata.

4.6.10. Inne metody ekstrakcji

Oprócz miar asocjacyjnych w oprogramowaniu zaimplementowane zostały także inne metody ekstrakcji jednostek wielowyrazowych.

Kombinacja liniowa

Metoda opiera się o wykorzystanie zestawu funkcji asocjacyjnych, z których każda generuje osobny ranking kolokacji. Następnie opcjonalnie każdy z tych rankingów jest normalizowany w określony sposób, dokonywane jest ich przepunktowanie za pomocą odpowiedniej funkcji. Każdemu z rankingów przypisywana jest waga, która określa jak bardzo dany ranking jest istotny w stosunku do pozostałych. Kolejnym krokiem jest agregacja rankingów, ich kombinacja liniowa, za pomocą pewnej funkcji agregującej. Zadaniem takiej funkcji jest wykonanie agregacji zbioru ocen dla każdej krotki z osobna, gdzie ocenami są wartości asocjacji wygenerowane przez miary asocjacji dla tej kolokacji. Przykładem takiej funkcji agregującej może być suma, a wzór na tę funkcję agregacyjną został zamieszczony poniżej:

$$ocena(t) = \sum_{i=1}^m w * r_i(t)$$

1. t - kolokacja;
2. m - liczba miar asocjacji, rankingów;
3. $r_i(t)$ - ocena krotki t w rankingu i .

Po wykonaniu agregacji posortowanie otrzymanych wartości spowoduje utworzenie nowego, pojedynczego rankingu będącego wynikiem kombinacji liniowej.

Wagi rankingów to parametry algorytmu kombinacji, które powinny zostać wyznaczone eksperymentalnie, przykładowo poprzez ich optymalizację. Prezentowane oprogramowanie umożliwia wykorzystanie do tego celu pięciu różnych algorytmów heurystycznych i metaheurystycznych implementacji Łukasza Kłyka [?]. Utworzony przez Niego *Optimizer* został przystosowany przez autora niniejszej pracy do działania z opisywanym w tym rozdziale *MWeXtractorem*. Łukasza Kłyk zaimplementował następujące algorytmu w swoim oprogramowaniu:

1. Algorytm ewolucyjny;
2. Hill climbing;
3. Random Search;
4. Tabu Search;
5. Symulowane wyżarzanie.

Nazwy wspomnianych algorytmów heurystycznych i metaheurystycznych dokładnie określają jakie są to algorytmy, jednak poza dwoma wyjątkami - symulowanym wyżarzaniem i algorytmem ewolucyjnym. Pierwszy z nich nie jest ścisły co do schematu chłodzenia, ale domyślnie w oprogramowaniu zaimplementowanym przez Łukasz Kłyka stosowana jest funkcja $T(k) = 0.3^k$ [?, str. 36]. Przypadek algorytmu ewolucyjnego wymaga dłuższego opisu, ponieważ pojęcie to jest znacznie szersze od nazw pozostałych metod.

Opisanie wykorzystanego algorytmu ewolucyjnego wymaga wyjaśnienia schematu przetwarzania w algorytmie, operatora selekcji, mutacji oraz krzyżowania, a także sposobu kodowania danych w genotypie osobników.

Zaimplementowany algorytm ewolucyjny jest algorytmem genetycznym o standardowym schemacie przetwarzania. Pierwszy krok to inicjalizacja populacji początkowej i ocena jej osobników. Po tym kroku następuje rozpoczęcie algorytmu, którego kolejne kroki przetwarzania to selekcja osobników, krzyżowanie wybranych przedstawicieli populacji oraz mutacja ich informacji genetycznych zapisanych w genotypie. Po tych krokach ustalona zostaje nowa populacja, która także podlega ocenie, a następnie algorytm wykonuje kolejną iterację zaczynając od operatora selekcji. Cykl jest powtarzany przez określoną liczbę iteracji lub do przerwania obliczeń.

Funkcja przystosowania osobników F' to przeskalowana funkcja liniowa oceny F , wyrażona za pomocą wzoru [?, str. 28]:

$$F' = aF + b$$

Natomiast współczynniki a i b to odpowiednio: $a = -\frac{F_{av}}{F_w - F_{av}}$ $b = \frac{F_{av}F_w}{F_w - F_{av}}$

1. F - funkcja oceny;
2. F_{av} - średnia wartość przystosowania osobników w populacji;
3. F_w - wartość przystosowania najgorszego z osobników populacji.

Algorytm dopuszcza kodowanie trzech typów wartości: liczby całkowite, zmiennie-przecinkowe oraz logiczne. Implementacja genotypu dopuszcza możliwość mieszania typów genów, mogą także występować zależności pomiędzy kodowanymi wartościami. Genotyp każdego osobnika składa się z pojedynczego chromosomu zawierającego zestaw genów, z których każdy przechowuje pojedynczą wartość jednego z optymalizowanych parametrów [?, str. 30].

Operator selekcji to połączenie metody turniejowej z ruletkową. Pierwszym krokiem jest obliczenie przystosowania osobników w populacji. Następnie metodą ruletki losowanych jest k osobników do turnieju, gdzie k jest rozmiarem turnieju. Metoda ruletki polega na zwiększaniu szans na wylosowanie osobników lepiej przystosowanych, zwiększaniem prawdopodobieństwa na wybranie ich do drugiego etapu selekcji - turnieju. Drugi krok selekcji to już wspomniany wcześniej turniej, który rozgrywany jest pomiędzy k wylosowanymi osobnikami z populacji. Turniej wyłania jednego zwycięzcę, który trafi do populacji w kolejnej iteracji, a gdy wyłonionych zostanie odpowiednia liczba zwycięzców ze

wszystkich turniejów następuje krzyżowanie ich genotypów i wymiana informacji [?, str. 29].

Autor pracy [?] zbadał dwa operatory krzyżowania osobników - jednopunktowy i równomierny. Pierwszy polega na wybraniu punktu podziału genotypu rodziców i utworzeniu nowego na podstawie ich dwóch ciągłych fragmentów genotypów - lewego i prawego, po jednym od każdego rodzica. Przykładowo dla dwóch rodziców i ich przykładowych genotypów: $[x_1, x_2, x_3, x_4]$ i $[y_1, y_2, y_3, y_4]$, oraz punktu podziału pomiędzy genem pierwszym i drugim utworzone mogą zostać dwa genotypy: $[y_1, x_2, x_3, x_4]$ oraz $[x_1, y_2, y_3, y_4]$. Drugi typ krzyżowania, równomierny, polega na utworzeniu maski krzyżowania, mówiącej o tym, który gen pochodzi, od którego rodzica. Sposób ten jest bardziej elastyczny, ponieważ kolejność genów w genotypie nie ma znaczenia co było ograniczeniem pierwszego typu krzyżowania. Wynik zastosowania operatora równomiernego dla dwóch powyższych genotypów rodziców i przykładowej maski 0101 byłoby dwuwartościowy i następujący: $[x_1, y_2, x_3, y_4]$ oraz $[y_1, x_2, y_3, x_4]$ [?, str. 32].

Operator mutacji składa się z dwóch różnych funkcji ze względu na zastosowanie różnych typów wartości kodowanych w genach osobnika. Pierwszy z nich dotyczy wartości logicznych i polega na zmianie wartości na przeciwną - wartość *prawda* na *falsz* lub odwrotnie. Drugi z nich dotyczy wartości całkowitych oraz rzeczywistych i polega na losowym odjęciu lub dodaniu pewnej wartości do tej składowanej w danym genie. Minimalna i maksymalna wartość mutowanego genu jest ograniczona za pomocą parametrów. Proces mutacji polega na wylosowaniu prawdopodobieństwa i sprawdzeniu czy jest ono mniejsze od szansy na mutację i jeśli tak to wykonanie mutacji osobnika. Sama mutacja polega na obliczeniu prawdopodobieństwa mutacji poszczególnych genów i w zależności od wyniku wykonanie samej modyfikacji wartości w odpowiednim genie [?, str. 31].

Perceptron wielowarstwowy

Motywacją do zaimplementowania i zbadania perceptronu wielowarstwowego były dobre wyniki sztucznej sieci neuronowej z jedną warstwą ukrytą w pracy [?], które według rezultatów w pracy Pavla Peciny okazały się lepsze nawet od *Support Vector Machine*, który uważany jest za bardzo dobry klasyfikator dwuklasowy, chociaż relatywnie trudny do nauczania ze względu na wrażliwość na część parametrów. Zaznaczyć trzeba także, że autorzy tej pracy skorzystali z kernela liniowego.

Zaimplementowany perceptron jest wielowarstwowy, umożliwia zdefiniowanie dowolnej liczby warstw zaczynając od trzech, a każda z dwóch sąsiednich warstw jest ze sobą w pełni połączona. Do każdej z tych warstw dodany został neurony typu *bias*. Zaimplementowano także *momentum*, którego wartość można ustalić argumentem klasyfikatora tak samo jak parametru określającego szybkość uczenia sieci.

Umożliwiony został także zapis struktury perceptronu wielowarstwowego w celu pozwolenia na przechowywanie go w pamięci nieulotnej i możliwości jego późniejszego

odtworzenia. Zapisywane są zarówno parametry klasyfikujące - wagi oraz struktura, jak i dane związane z uczeniem co pozwala na kontynuację uczenia sieci lub jej naukę z przerwami.

Funkcja klasyfikująca została pominięta przez co wartość zwracana przez ten klasyfikator jest liczbą rzeczywistą co z kolei umożliwia wykorzystanie jej jako mechanizmu generującego wartości rankingowe dla kolokacji, a nie jako klasyfikatora. Wykorzystana funkcja aktywacji ϕ została zaczerpnięta z pracy [?], a jej wzór został zamieszczony poniżej:

$$\phi(z) = \frac{\exp(z)}{1+\exp(z)}$$

Wzór zastosowanej pochodnej d funkcji ϕ został zapisany poniżej:

$$d(z) = \frac{\exp(z)}{(1+\exp(z))^2}$$

Parametry optymalizowalne w sieci to liczba neuronów w warstwie ukrytej [?], a także liczba warstw ukrytych, *momentum* oraz wartość określająca szybkość uczenia sieci.

4.6.11. Miary punktuje rankingi i normalizatory cech

Funkcje te początkowo zostały zaimplementowane w celu wykonywania przepunktowania rankingów będących wynikami miar asocjacyjnych. Dodana została jednak następnie możliwość wykorzystania ich do wykonania modyfikacji i przygotowania cech dla klasyfikatorów.

Normalizacja

Klasyczna normalizacja zakresu wartości. Każda z wartości z danej serii liczb, przykładowo wartości konkretnej cechy, zostaje obliczona z wykorzystaniem poniższego zbioru:

$$nowa_warto = \frac{stara_warto - \min(S)}{\max(S) - \min(S)}$$

Borda score

Zestaw wartości zostaje posortowany malejąco, a następnie każda z wartości, poczynając od największej, zostaje zastąpiona liczbą całkowitą równą mocy tego zbioru wartości powiększonej o jeden i zmniejszonej o odległość tej wartości od największej z liczb, gdzie odległość jest rozumiana jako liczba liczb. Przykładowo poniższy zestaw wartości:

90, 17, 7

zostanie zamieniony na:

3, 2, 1,

Borda score w oparciu o prawo Zipfa

Niniejsza funkcja jest rozszerzeniem metody *Borda score*, modyfikacja polega na pomnożeniu każdej z nowych wartości przez odwrotność odległości tej liczby od liczby maksymalnej, gdzie odległość jest rozumiana tak samo jak w *Borda score* – liczba liczb od największej liczby ze zbioru. Przykładowo poniższy zestaw wartości:

90, 17, 7, 3, 2, 1

zostanie zamieniony na:

$\frac{6}{1}, \frac{5}{2}, \frac{4}{3}, \frac{3}{4}, \frac{2}{5}, \frac{1}{6}$

Centering

Centering polega na wyzerowaniu średniej rozkładu danego zestawu wartości. Metoda polega na odjęciu od każdej liczby ze zbioru średniej wartości tego zbioru liczb.

$$\text{nowa_warto} = \text{stara_warto} - \text{avg}(S)$$

Standaryzacja

Standaryzacja jest rozszerzeniem metody centrowania. Modyfikacja polega na dodatkowym podzieleniu każdej liczby przez odchylenie standardowe dla tego zbioru wartości.

$$\text{nowa_warto} = \frac{\text{stara_warto} - \text{avg}(S)}{\text{sd}(S)}$$

4.6.12. Dostępne agregatory rankingów

Przedstawione tutaj funkcje są wykorzystywane do łączenia kilku zestawów wyników w jeden ranking k-najlepszych pozycji. We wzorach wykorzystane zostały następujące oznaczenia:

1. $s(t)$ - ocena krotki;
2. $s_i(t)$ - ocena krotki t w rankingu i ;
3. w_i - waga dla rankingu i ;
4. R - liczba rankingów;

Maksymalna suma

Miara obliczająca średnią ważoną dla ocen konkretnej krotki ze wszystkich rankingów.

$$s(t) = \sum_{i=1}^R s_i(t) * w_i$$

Kandydat z większą oceną po wykonaniu sumy jest uznawany za lepszego.

Maksymalny iloczyn

Funkcja jest analogiczna do maksymalnej sumy, ale zamiast sumy obliczany jest iloczyn. Należy jednak zauważyć, że wagi dla rankingów nie mają w tym wypadku znaczenia. Dodatkowo miara ta jest bardzo wrażliwa na wartości ujemne, a konkretnie ich liczbę. Powyższe powody sprawiły, że funkcja ta nie została wykorzystana przy dostrajaniu wag algorytmem genetycznym.

$$s(t) = \sum_{i=1}^R s_i(t) * w_i$$

4.6.13. Zaimplementowane miary oceny wyników

Każda miara oceny została zaimplementowana w taki sposób, że generuje pewną liczbę wartości dla każdej długości rankingu zaczynając od pozycji numer jeden do ostatniej pozycji rankingu lub do pozycji zdefiniowanej argumentem danej funkcji jakości. Generowanie wyników dla każdej pozycji rankingu jest istotne przykładowo w przypadku chęci utworzenia wykresu obrazującego przebieg poziomu danej wartości oceny wraz ze wzrostem długości rankingu.

Precyzja

Jedna z podstawowych miar oceny jakości rozwiązania. Każda z wartości określa ile elementów w wyznaczonym zbiorze zostało poprawnie przyporządkowanych do danej klasy. W przypadku zadania polegającego na utworzeniu rankingu precyzja jest wartością określającą stosunek liczby elementów interesujących – wyrażeń wielowyrazowych, do określonej długości rankingu.

Kompletność

Kompletność jest drugą z podstawowych miar oceny jakości rozwiązań. Jej wartość określa jak dużo elementów interesujących zostało wykrytych spośród wszystkich elementów interesujących znajdujących się w zbiorze. W przypadku kompletności ranking jest po prostu pewnym zbiorem.

Para precyzja i kompletność

Miara oblicza zarówno precyzję jak i kompletność, czyli generuje parę liczb dla każdej pozycji rankingu.

4.6.14. Precyzja uśredniona

Średnia precyzja przy trafieniu

Miara oblicza średnią precyzję liczoną tylko z tych wartości przypisanych do pozycji rankingu, dla których na tej pozycji ulokowany został element poprawny – wyrażenie wielowyrazowe.

Miara F

Popularna miara stosowana w literaturze, opisana następującym wzorem:

$$F_{\beta} = (1 + \beta^2) \frac{\textit{precyzja} * \textit{kompletno}}{\beta^2 * \textit{precyzja} + \textit{kompletno}}$$

5. BADANIA

Badania na potrzeby niniejszej pracy obejmują sprawdzenie jakości miar asocjacyjnych w procesie wydobywania wyrażen wielowyrazowych z dużych korpusów języka polskiego. Do badanych miar zaliczają się wszystkie funkcje asocjacyjne zaimplementowane w oprogramowaniu *MWeXtractor*, kombinacja liniowa rankingów wygenerowanych za pomocą tych funkcji, a także klasyfikator będący perceptronem wielowarstwowym. Metodologia, proces badań, wyniki ich omówienie oraz wnioski zostały zawarte w dalszej części tego rozdziału niniejszej pracy.

5.1. OPIS WYKORZYSTANYCH ZBIORÓW DANYCH

Podczas badań wykorzystanych w różnych celach zostało kilka zbiorów danych, które zostały dokładniej omówione w dalszej części tego rozdziału. Pozyskanie wszystkich z nich przebiegło bardzo szybko ze względu na współpracę z grupą technologiczną G4.19, która udostępniła te zbiory danych na potrzeby badań prowadzonych przez autora niniejszej pracy.

5.1.1. Korpus KIPI

Właściwa nazwa korpusu to *Korpus IPI PAN*, a został on stworzony przez Zespół Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki Polskiej Akademii Nauk. Według autorów korpus posiada 250 milionów segmentów anotowanych morfosyntaktycznie. Korpus jest dostępny publicznie i składowany w formacie *Poliqarp*. Te i więcej informacji można pozyskać ze strony internetowej autorów korpusu [?].

5.1.2. Korpus KGR7

Korpus utworzony przez grupę naukową G4.19 składowany także w plikach o formacie *Poliqarp*, otagowane tagsetem *KIPI*. Jedną z jego części jest korpus *KIPI*. Korpus *KGR7* jest około siedmiokrotnie większy niż opisany we wcześniejszej części tej pracy *Korpus IPI PAN*.

Poniższa tabel 5.1 zawiera statystyki podkorpusów składowych korpus *KGR7*.

5.1.3. Słowność i praca lingwistów

Wspomniany wcześniej twór - *Słowność*, w procesie badań został wykorzystany jako baza wiedzy, z której pozyskane zostały wyrażenia wielowyrazowe uznane za popraw-

nazwa korpusu	liczba tokenów
1002	19 512 317
1003	10 006 539
blogi	9 613 618
interia	611 402
kipi	255 516 328
knigi_joined	1 010 676 150
naukawe	2 594 225
ornitologia	544 937
plwiki20120428	275 578 635
pogoda	593 538
poig_biznes_data_sub_0	35 439 099
poig_biznes_data_sub_1	30 676 362
polityka	82 480 654
prace	12 665 419
pryzmat	2 183 403
rzepa	116 317 357
sjp	2 177 299
wordpress	439 304
zwiazki	820 991
suma	1 868 447 577

Tabela 5.1. Podkorpusy i statystyki dotyczące korpusu *KGR7*

ne. Dodatkowo w trakcie prac opisanych w niniejszym dokumencie grupa lingwistów na bieżąco oceniała kolejne zestawy kandydatów na jednostki wielowyrazowe. Oba zbiory kolokacji zostały ze sobą połączone w jeden, który następnie wykorzystywany był jako dane niezbędne do oceny wyników generowanych przez rankery, a także do generacji cech dla sprawdzonych klasyfikatorów.

5.2. PRZYGOTOWANIE DANYCH

Tak jak wspomniano wcześniej, wszystkie dane zostały udostępnione przez grupę naukową G4.19. Otrzymane korpusy były składowane w formacie *Poliqarp* i opisane za pomocą tagsetu *KIPI*. Korpusy *KGR7* oraz *KIPI* zostały przygotowane do badań w taki sam sposób, którego opis znajduje się w dalszej części tej sekcji. Dodać jednak trzeba, że z korpusu *KGR7* wyłączony został podkorpus *sjp*. Motywacją do tego były problemy związane z błędami w pliku z danymi oraz tym, że jest to zestaw definicji słownikowych, a nie spójny tekst traktujący na jakieś konkretne tematy.

5.2.1. Tagowanie

Autor niniejszej pracy postanowił dokonać rozłożenia danych otagowanych do nieotagowanego tekstu ciągłego, a następnie ponownego otagowania tych danych. Do wykonania

tego zadania użyte zostało narzędzi omówione we wcześniejszej części tej pracy, tager *WCRFT2*. Tekst został otagowany tagsetem *NKJP*, a tekst już przetworzony jest składowany w formacie *IOB-chan*. Motywacje do tego działu były dwie, pierwsza to chęć otagowania danych za pomocą znanego tagera i modelu wyuczonego przez grupę G4.19, a nie korzystanie z już gotowych danych otagowanych. Drugą motywacją były pewne problemy związane z czytaniem korpusów w formacie *Poliqarp*. Jako format zapisu na nowo otagowanego tekstu wybrano wspomniany *IOB-chan* ze względu na małą objętość pamięciową i szybkość jego wczytywania, a także ze względu na jego przejrzystość.

5.2.2. Ekstrakcja kandydatów na kolokacje

Gotowe, otagowane korpusy należało wczytywać i wydobyć z nich potrzebne informacje o kandydatach na kolokacje takie jak składowe krotek i ich częstości czy licznosc konkretnych słów. Częstości kandydatów na kolokacje składowane są z podziałem na konkretne podkorpusy co umożliwia zastosowanie funkcji dyspersji w przypadku korpusów składających się z podkorpusów. Wszystkie zebrane informacje zapisane są w omówionym uprzednio *składzie krotek*. Do wydobywania kandydatów na kolokacje użyto omawianego wcześniej języka ograniczeń *WCCL*. Skorzystano z 80 relacji w celu ekstrakcji 2-elementowych kandydatów na kolokacje oraz 12 w przypadku ekstrakcji relacji 3-elementowych. Większość wykorzystanych operatorów sprawdzała części mowy wyrazów składowych kandydata na kolokacje, ale część także pewne ich cechy i powiązania takie jak przykładowo uzgodnienie rzeczownika z przymiotnikiem. Część użytych wyrażeń *WCCL* miała na celu zebranie wszystkich kandydatów na kolokacje, których można było utworzyć w oknie o danej długości. Spisy relacji użytych w konkretnych badaniach zostały zamieszczone przy ich opisach.

5.3. OPIS PLANOWANEGO PRZEBIEGU BADAŃ

Zaplanowany proces badania metod ekstrakcji został podzielony na kilka kroków. Wszystkie zostały opisane w dalszej części tego rozdziału, a wykonane zostały z wykorzystaniem narzędzi z pakietu *MWeXtractor*. Skrótowy opis przebiegu badania został zawarty poniżej ??:

5.4. MIARY DWUELEMENTOWE, KORPUS KIPI

Celem tego badania było zdobycie informacji o jakości miar dwuelementowych wykorzystanych do ekstrakcji kolokacji z korpusu *KIPI*. Na podstawie wyników ocenić można było, które funkcje dwuelementowe generują najlepsze wyniki co jest jednocześnie wstępem do dalszych badań opisanych w tym rozdziale – informacja ta jest pomocna w doborze miary dla specjalnych funkcji N-elementowych bazujących na rozbijaniu krotek

nr.	zadanie	cel i motywacje
1	zbadanie dwuelementowych miar asocjacyjnych na korpusie kipi	zdobycie informacji pozwalających na ocenę jakości miar w zadaniu ekstrakcji kolokacji dwuelementowych; zdobycie wiedzy pozwalającej na ocenę, które funkcje asocjacyjne powinny zostać wybrane jako składowe generatora cech dla klasyfikatorów; punkt wyjścia do dalszych badań;
2	zbadanie dwuelementowych miar asocjacyjnych na korpusie kipi poddanym dyspersji za pomocą miary TF-IDF	sprawdzenie wpływu zastosowania miary dyspersji na zmianę jakości wyników;
3	zbadanie dwuelementowych miar asocjacyjnych na korpusie kipi po wykonaniu podpróbkowania klasy negatywnej	korpus zawiera w znacznej przewadze krotki niebędące kolokacjami; stężenie wyrażen wielowyrazowych wśród kandydatów na kolokacje jest mniejszy niż pół procenta – dokładne statystyki zostały zawarte w dalszej części tej pracy; ze względu na to autor pracy postanowił sprawdzić jakość miar po wykonaniu podpróbkowania klasy negatywnej;
4	zbadanie wariantów perceptronu wielowarstwowego dla kolokacji dwuelementowych	sprawdzenie jakości rozwiązań generowanych przez klasyfikatory; przetestowanie wpływu liczby neuronów i warstw ukrytych na jakość rozwiązań dostarczanych przez ten klasyfikator;
5	zbadanie miar trójelementowych	sprawdzenie jakości generalizacji miar zaproponowanych w literaturze oraz przez autora niniejszej pracy;
6	dyspersja na większym korpusie (fragmenty KGR7) ?	;
7	SVM?	;

Tabela 5.2. Planowany przebieg badań

N-elementowych na 2-elementowe. Dodatkowo wiedza ta posłuży do doboru miar będących generatorami cech dla badanych klasyfikatorów – sieci neuronowych.

5.4.1. Przygotowanie i zbadanie podkorpusów KIPI

Pierwszym etapem tego badania było sprawdzenie podzbiorów danych korpusu *KIPI*, czyli zebranie i zestawienie statystyk o nich w celu ich lepszego poznania oraz oceny zastosowanych relacji. Dane zostały pozyskane na sześć różnych sposobów – sposoby różniły się zestawem wykorzystanych operatorów *WCCL*. Możliwym było wydobycie ich tylko w jeden sposób ze względu na możliwość wykorzystania funkcjonalności pakietu *MWeXtractor*, takich jak między innymi filtrowanie, ale podejście z wieloma może przyspieszyć kolejne etapy badań ze względu na szybkość wczytywania danych, a także zmniejszyć ilość

pamięci RAM potrzebnej do obliczeń. Wydobyte informacje zostały zapisane w strukturach współpracujących ze wspomnianym oprogramowaniem – składach krotek. Gotowe składy krotek poddano analizie z wykorzystaniem programu *Cover*, który generuje dwa zestawy informacji opisane we wcześniejszej części tej pracy. Dla przypomnienia, pierwsza z nich to macierz liczb z etykietami na wierszach i kolumnach zawierająca dane o tym, w jakim stopniu relacje zachodzą na siebie – ile krotek zostało przyporządkowanych do konkretnych relacji. Drugi wynik to liczba wyrażeń wielowyrazowych odnalezionych w każdej z relacji z osobna. Dane te posłużyły do oceny wykorzystanych operatorów i wyborze wzorców *częstych*, czyli takich wyrażeń *WCCL*, których wśród odnalezionych przez nie instancji w tekście jest stosunkowo dużo wyrażeń pozytywnych. Do oceny wyrażeń pod kątem tego czy są *częste* wykorzystano informacje o liczbie jednostek wielowyrazowych należących do danego wzorca jak i o stosunku tej liczby do liczby wszystkich krotek należących do rozpatrywanej relacji. Wzorzec został uznany za częsty jeśli jego stosunek liczby jednostek wielowyrazowych w nim zawartych do wszystkich krotek należących do tego wzorca był większy od jednego procenta. W dalszej części tej pracy podane zostaną szczegóły dotyczące przygotowanych i sprawdzonych zbiorów danych.

Podzbiór 2W

Statystyki dotyczące podzbioru 2W zostały zamieszczone w tabeli ??.

relacje	liczba krotek	liczba JW	procent JW	częsta?
Window2P0	19752289	41221	0,20869	nie
Window2P1	19752289	25560	0,12940	nie
Suma:	39504578	66781	0,16904	

Tabela 5.3. Statystyki dotyczące podzbioru danych 2W, pozyskanego z korpusu *KIPI*.

Ważną obserwacją jest stężenie jednostek wielowyrazowych wśród wszystkich kandydatów na kolokacje – tylko niecałe 0,17%. Tak niska ich zawartość obrazować może poziom trudności zadania ekstrakcji kolokacji z korpusów tekstowych.

Podzbiór 2R

Statystyki dotyczące podzbioru 2R zostały zamieszczone w tabeli ??.

Autor pracy postanowił wykluczyć relacje *AllSubstSubstP0* i *AllSubstSubstP1* z grona *częstych* ze względu na niski procent jednostek wielowyrazowych w gronie kandydatów przez nią wybranych. Co prawda relacje odnalazły w sumie 4577 poprawnych wyrażeń wielowyrazowych co stanowi znaczącą liczbę, prawie 10% wszystkich poprawnych kolokacji, ale zauważyć trzeba, że relacje te będąc dwiema z 38 wysunęły prawie 62% wszystkich kandydatów na kolokacje. Podsumowując, dwa wspomniane operatory wygenerowały znaczącą liczbę jednostek wielowyrazowych, ale jednocześnie także nieporównywalnie

relacje	liczba krotek	liczba JW	procent JW	częsta?
AdjGenCosP0	128	0	0,00000	nie
AdjGenCosP1	840	0	0,00000	nie
AgrAdjSubstP0	1336463	1230	0,09203	nie
AgrAdjSubstP1	706331	645	0,09132	nie
AgrSubstAdjP0	706331	18162	2,57132	<i>TAK</i>
AgrSubstAdjP1	1336463	11774	0,88098	<i>TAK</i>
AllAdvPartP0	122560	6	0,00490	nie
AllAdvPartP1	74187	1	0,00135	nie
AllBurkSubstP0	2544	30	1,17925	<i>TAK</i>
AllBurkSubstP1	7126	2	0,02807	nie
AllGerQubP0	11518	2012	17,46831	<i>TAK</i>
AllGerQubP1	20772	1016	4,89120	<i>TAK</i>
AllNumSubstP0	69889	8	0,01145	nie
AllNumSubstP1	50929	4	0,00785	nie
AllPartAdvP0	74187	1171	1,57844	<i>TAK</i>
AllPartAdvP1	122560	1064	0,86815	<i>TAK</i>
AllQubGerP0	20772	1	0,00481	nie
AllQubGerP1	11518	1	0,00868	nie
AllSiebieSubstP0	8149	0	0,00000	nie
AllSiebieSubstP1	2017	0	0,00000	nie
AllSubstBurkP0	7126	9	0,12630	nie
AllSubstBurkP1	2544	0	0,00000	nie
AllSubstNumP0	50929	3	0,00589	nie
AllSubstNumP1	69889	1	0,00143	nie
AllSubstSiebieP0	2017	12	0,59494	nie
AllSubstSiebieP1	8149	3	0,03681	nie
AllSubstSubstP0	4124657	3847	0,09327	nie
AllSubstSubstP1	4124657	730	0,01770	nie
CosAdjGenP0	840	47	5,59524	<i>TAK</i>
CosAdjGenP1	128	5	3,90625	<i>TAK</i>
GndAdjSubstP0	42964	67	0,15594	nie
GndAdjSubstP1	82688	63	0,07619	nie
GndSubstAdjP0	82676	3234	3,91166	<i>TAK</i>
GndSubstAdjP1	42920	779	1,81500	<i>TAK</i>
Ppron3GenSubstP0	18323	6	0,03275	nie
Ppron3GenSubstP1	10350	5	0,04831	nie
SubstPpron3GenP0	10350	5	0,04831	nie
SubstPpron3GenP1	18323	10	0,05458	nie
Suma	13384814	45953	0,34332	
Suma częstych	2400939	39294	1,63661	

Tabela 5.4. Statystyki dotyczące podzbioru danych 2R, pozyskanego z korpusu *KIPI*.

większą ilość szumu jakim są błędni kandydaci. Możliwe jest, że metodom wydobywającym udałoby się odnaleźć część z poprawnych wyrażen wielowyrzowych należących do omawianych relacji, ale jednak procent jednostek wielowyrzowych w niej wydaje się na tyle mały, że ich wydobywanie mogłoby znacząco pogorszyć ogólny wynik pod kątem precyzji.

Wśród wszystkich krotek pozyskanych za pomocą przedstawionych w 2R operatorów *WCCL* znajduje się około 68,8% wszystkich jednostek wielowyrzowych wykrytych w tekście za pomocą operatora okna. Wynik ten otrzymano przy zmniejszeniu grona kan-

dydatów na kolokacje do 33,9%. Natomiast procent wyrażający stosunek jednostek wielowyrzowych do wszystkich kandydatów wzrósł z niecałych 0,17% do ponad 0,34%. Ponadto w przypadku wzięcia pod uwagę tylko relacji częstych wartości te będą wynosić odpowiednio ponad 58,8%, prawie 6,1% oraz niecałe 1.64.

Podsumowując powyższe obserwacje zauważyć można, że poprzez zastosowanie filtrów opartych o części mowy, dla korpusu *KIPI* i zestawu jednostek wielowyrzowych pozyskanych ze Słownosieci, da się zachować około 58,8% jednostek wielowyrzowych przy zmniejszeniu liczby kandydatów do jedynie niecałych 6,1%. Skutkuje to także podniesieniem stężenia jednostek wielowyrzowych wśród kandydatów około dziesięciokrotnie – skutkiem tego może być znaczny wzrost precyzji systemu wykrywającego kolokacje.

Zauważyć można także kilkukrotne różnice w procencie jednostek wielowyrzowych w całym zestawie kandydatów z danej relacji w zależności od wybranego szyku, przykładowo w przypadku relacji *AllSubstSubstP0* i *AllSubstSubstP1* różnica ta jest około czterokrotna, a w przypadku *CosAdjGenP0* i *CosAdjGenP1* ponad dziewięciokrotna.

Podzbiór 2RW

Omawiany tutaj zbiór powstał poprzez połączenie dwóch zestawów relacji, jednego wykorzystanego do utworzenia podzbioru 2W oraz drugiego przygotowanego na potrzeby generacji podzbioru 2R. Statystyki niniejszego podkorpusu są takie same jak korpusów 2R i 2W. Zmianie uległa jedynie statystyka dotycząca sumy.

Podzbiór 2W1H

Zbiór został rozszerzony o operatory akceptujące wszystkie pary wyrazów oddzielone dowolnym słowem znajdującym się pomiędzy nimi. Operatory te generują wszystkie możliwe do utworzenia pary wyrazów tworząc zestaw kandydatów na kolokacje nieciągłe. Nazwy nowych operatorów zostały zmodyfikowane poprzez dodanie do ich nazw fragmentu *H1* reprezentującego nieciągłość wielkości pojedynczego wyrazu.

relacje	liczba krotek	liczba JW	procent JW	częsta?
Window2P0	19752289	41221	0,20869	nie
Window2P1	19752289	25560	0,12940	nie
Window2H1P0	29740688	20176	0,06784	nie
Window2H1P1	29740688	24998	0,08405	nie
Suma nieciągłych	59481376	45174	0,0759465	
Suma wszystkich	98985954	111955	0,11310	

Tabela 5.5. Statystyki dotyczące podzbioru danych 2W1H, pozyskanego z korpusu *KIPI*.

Zastanawiające mogłoby być dlaczego kolokacji nieciągłych jest więcej niż ciągłych skoro dla zdania *N*-elementowego zawsze można wygenerować *N* – 1 bi-gramów oraz *N* – 2 tri-gramów. Zadać sobie można pytanie dlaczego więc tri-gramów jest więcej.

Odpowiedź jest następująca: jest ich mniej. Statystyki pokazują ile różnych kandydatów udało się utworzyć za pomocą danych relacji, ale trzeba pamiętać, że każdy z kandydatów mógł wystąpić wielokrotnie. Omawiany tutaj zbiór zawiera w sumie 98985954 różnych krotek, których suma częstości jest równa 942472698, gdzie tylko 457972144 par zostało wygenerowanych przez relacje wyszukujące kolokacje nieciągłe, a tym samym 484500554 przez relacje wyszukujące kandydatów ciągłych. Podsumowując kandydaci wyszukiwani przez relacje ciągłe byli mniej zróżnicowani, ale było ich więcej niż w zestawie kandydatów pozyskanych dzięki operatorom ciągłym. Sytuacja może pojawiać się także przy innych relacjach niż oknowe, ale powód takiej sytuacji może być taki sam, a dodatkowo pojawia się także inne wyjaśnienie – po prostu wyrazy tak się ułożyły, że konkretne pary częściej występowały oddzielone jakimś wyrazem.

Wyniki z tabeli obrazują duży potencjał operatorów nieciągłych w zwiększeniu liczby możliwych do wykrycia wyrażeń wielowyrazowych, a ponadto wpływają na dane statystyczne wykorzystywane przez miary powiązania i klasyfikatory w ich pracy.

Podzbiór 2R1H

Niniejszy podzbiór jest nadzbiorem 2R. Zawarte w nim są te same relacje co w zbiorze 2R, ale rozszerzone zostały o akceptowanie także nieciągłych kandydatów na kolokacje, których odległość pomiędzy wyrazami składowymi krotki była równa dwa – jeden dowolny wyraz pomiędzy składowymi. Liczba relacji zwiększyła się dwukrotnie do 76, a dodatkowo dzięki zastosowaniu tego mechanizmu informacja statystyczna uległa zmianie. Szczegółowe informacje na temat tego podzbioru zawarte zostały w tabeli ???. Zauważyć jednak należy, że podano w niej tylko informacje, które zostały pozyskane poprzez zastosowanie nowych relacji – nieciągłych, ponieważ dane wydobyte za pomocą relacji ciągłych są takie same jak w przypadku podzbioru 2R.

Ciekawą obserwacją jest, że relacje *AgrAdjSubstH1P0* i *AgrAdjSubstH1P1* generują prawie 48,7% wszystkich nieciągłych wyrażeń wielowyrazowych na podstawie badanego tekstu, z pośród wszystkich przedstawionych relacji. Warto dodać, że zbiór relacji oznaczonych jako częste prawie nie uległ zmianie z wyjątkiem *AllBurkSubstP0*. Nie uległ zmianie w takim rozumieniu, że jeśli relacje ciągłe A i B weszły w skład relacji częstych to ich wersje wykrywające relacje nieciągłe (z dodatkiem *H1* w nazwie) także znalazły się w tym zbiorze.

Przedstawione statystyki obrazują, że ogólna jakość rozwiązań na podstawie tak zebranych danych teoretycznie powinna spaść w przypadku zastosowania wyboru wyrażeń wielowyrazowych w procesie losowania, wniosek taki wysunąć można na podstawie kolumny *procent JW* tabeli ??, która obrazuje spadek stężenia jednostek wielowyrazowych w stosunku do podzbioru 2R. Zaznaczyć jednak trzeba, że także tutaj usunięcie relacji *AllSubstSubstH1P0* i *AllSubstSubstH1P1* powinna zaowocować znaczną poprawą wyniku, ponieważ według obliczeń tylko prawie 0.041% jednostek w niej zawartych stanowią

relacje	liczba krotek	liczba JW	procent JW	częsta?
AdjGenCosH1P0	405	0	0	nie
AdjGenCosH1P1	564	0	0	nie
AgrAdjSubstH1P0	441381	470	0,1064839674	nie
AgrAdjSubstH1P1	435442	410	0,0941572012	nie
AgrSubstAdjH1P0	435442	5097	1,170534767	TAK
AgrSubstAdjH1P1	441381	6691	1,5159238844	TAK
AllAdvPartH1P0	92077	3	0,0032581426	nie
AllAdvPartH1P1	70914	3	0,0042304764	nie
AllBurkSubstH1P0	7411	6	0,080960734	nie
AllBurkSubstH1P1	8195	6	0,0732153752	nie
AllGerQubH1P0	14164	614	4,3349336346	TAK
AllGerQubH1P1	30860	1491	4,8314970836	TAK
AllNumSubstH1P0	79359	6	0,0075605791	nie
AllNumSubstH1P1	105045	5	0,0047598648	nie
AllPartAdvH1P0	70914	604	0,8517359055	TAK
AllPartAdvH1P1	92077	1555	1,6888039358	TAK
AllQubGerH1P0	30860	1	0,0032404407	nie
AllQubGerH1P1	14164	1	0,0070601525	nie
AllSiebieSubstH1P0	10628	0	0	nie
AllSiebieSubstH1P1	7743	0	0	nie
AllSubstBurkH1P0	8195	2	0,0244051251	nie
AllSubstBurkH1P1	7411	2	0,0269869113	nie
AllSubstNumH1P0	105045	2	0,0019039459	nie
AllSubstNumH1P1	79359	3	0,0037802896	nie
AllSubstSiebieH1P0	7743	6	0,0774893452	nie
AllSubstSiebieH1P1	10628	7	0,0658637561	nie
AllSubstSubstH1P0	7888228	1790	0,0226920419	nie
AllSubstSubstH1P1	7888228	1415	0,0179381225	nie
CosAdjGenH1P0	564	46	8,1560283688	TAK
CosAdjGenH1P1	405	25	6,1728395062	TAK
GndAdjSubstH1P0	149805	95	0,0634157738	nie
GndAdjSubstH1P1	156749	99	0,0631582977	nie
GndSubstAdjH1P0	156749	1780	1,1355734327	TAK
GndSubstAdjH1P1	149805	1931	1,2890090451	TAK
Ppron3GenSubstH1P0	17518	9	0,0513757278	nie
Ppron3GenSubstH1P1	19023	6	0,0315407664	nie
SubstPpron3GenH1P0	19023	4	0,0210271776	nie
SubstPpron3GenH1P1	17518	12	0,0685009704	nie
Suma ciągłych	13384814	45953	0,34332	
Suma nieciągłych	19071022	24197	0,1268783603	
Suma wszystkich	32455836	70150	0,2161398646	
Suma ciągłych częstych	2400939	39294	1,63661	
Suma nieciągłych częstych	1392361	19834	1,4244868967	
Suma wszystkich częstych	3793300	59128	1,5587483194	

Tabela 5.6. Statystyki dotyczące podzbioru danych 2R1H, pozyskanego z korpusu *KIPI*.

jednostki wielowyrazowe, a do tych relacji należy ponad 82, 7% wszystkich kandydatów na kolokacje. Nie poprawia to jednak jakości rozwiązań losowych generowanych na podstawie danych zebranych przez inne relacje w ich obrębie. Trzeba jednak mieć na uwadze to o czym autor niniejszej pracy napisał przy okazji omawiania zbioru *2W1H* – zebranie informacji statystycznych w ten sposób może zmienić w znacznym stopniu wyniki

generowane przez funkcje asocjacyjne i klasyfikatory dzięki pozyskaniu nowych danych statystycznych z korpusu tekstowego. Wyniki mogą zostać poprawione dlatego, że sieci neuronowe i miary powiązania są bardziej skomplikowane niż wybór losowy jednostek wielowyrazowych z grona kandydatów.

Statystyki z tabeli pozwalają na wywnioskowanie, że zastosowanie filtrów opartych o dane lingwistyczne – części mowy i inne, może jeszcze bardziej zmniejszyć grono kandydatów na wyrażenia wielowyrazowe w przypadku wyszukiwania zarówno kandydatów ciągłych jak i nieciągłych. W przypadku wszystkich relacji zbiór kandydatów na kolokacje został zmniejszony do niecałych 32,8%, a maksymalna możliwa do osiągnięcia kompletność to niecałe 62,7%. Jeśli natomiast rozważymy tylko relacje uznane za częste to zbiór kandydatów zostanie zmniejszony do zaledwie nieco ponad 3,83%, a maksymalna kompletność to ponad 52,8%. Tak duże ograniczenie kandydatów na kolokacje skutkować może także kilkudziesięciokrotnym przyspieszeniem wydobywania wyrażen wielowyrazowych.

Tak duże zmiany w tych wartościach mogą być godne uwagi i dowodzą, że filtry części mowy są w stanie znacznie zmniejszyć grono kandydatów na kolokacji, a w konsekwencji tego zmienić wyniki pracy metod wykrywających wyrażenia wielowyrazowe poprzez zwiększenie ich szans na osiągnięcie większych poziomów precyzji. Trzeba pamiętać jednak, że nie jest to złote rozwiązanie, ponieważ kosztem może być spadek kompletności rozwiązania.

Podzbiór 2RW1H

Omawiany tutaj zbiór powstał w sposób analogiczny do 2W i 2R, ale poprzez połączenie dwóch innych zestawów relacji, jednego wykorzystanego do utworzenia podzbioru 2W1H oraz drugiego przygotowanego na potrzeby generacji podzbioru 2R1H. Statystyki niniejszego podkorpusu są takie same jak korpusów 2R1H i 2W1H. Zmianie uległy jedynie statystyki dotyczące sum.

Wersje podkorpusów poddane dyspersji

Oprócz sześciu opisanych w poprzedniej części tej sekcji podkorpusów korpusu *KIPI* utworzone zostały też ich odpowiedniki powstałe po podzieleniu korpusu na pewną liczbę mniej więcej równych części. Sposób podziału korpusu *KIPI* polegał na podzieleniu go na 10 podobnych rozmiarem części. Rozmiar był determinowany przez liczbę tokenów, a sam korpus był dzielony z dokładnością do zdania, żeby uniknąć cięcia w środku zdania.

Po wykonaniu podziału zostały przygotowane odpowiedniki podkorpusów 2R, 2W, 2RW, 2R1H, 2W1H, 2RW1H poprzez wykorzystanie tych samych relacji. Z racji, że żaden token czy zdanie nie zostało pominięte, a cięcia następowały z dokładnością do zdania to statystyki podkorpusów nie uległy zmianie.

Przygotowane podkorpusy następnie poddano dyspersji z wykorzystaniem miary TF-IDF,

ponieważ w literaturze uchodzi ona za dobrą miarę w zadaniach ekstrakcji informacji. Specyfiką tej miary dyspersji jest to, że części kandydatów przyporządkowuje ona częstość równą zero – tym krotkom, które wystąpiły we wszystkich podkorpusach. Kandydaci, których częstość po wykonaniu dyspersji była równa zero zostali usunięci z grona kandydatów.

Wersje poddane podpróbkowaniu klasy negatywnej

Ze względu na fakt, że klasa pozytywna reprezentująca wyrażenia wielowyrazowe jest słabo reprezentowana (poniżej 0,5% wszystkich kandydatów) autor niniejszej pracy dokonał podpróbkowania klasy negatywnej w celu utworzenia zbioru zawierającego znacznie bardziej reprezentatywną liczbę instancji klasy pozytywnej.

Podpróbkowanie polegało na wybraniu z zestawu kandydatów wszystkich wyrażen wielowyrazowych, a następnie z wykorzystaniem rozkładu jednorodnego usuwanie kolokacji z klasy negatywnej aż do momentu uzyskania zadanego stężenia jednostek wielowyrzowych. Pożądany stosunek jednostek pozytywnych do negatywnych został ustalony na poziomie około i przynajmniej 20%. Motywacją do wyboru takiego progu był artykuł [?], gdzie autorzy badali zbiór kolokacji, gdzie liczba instancji pozytywnych została ustalona na poziomie prawie 21%.

5.4.2. Szczegółowy opis przebiegu tej części badań

Tabela ?? przedstawia zestaw 30 różnych wariantów badań przeprowadzonych dla funkcji dwuelementowych na korpusie *KIPI*.

Tabela ?? przedstawia zestaw zbadanych funkcji na korpusie *KIPI*.

5.4.3. Wyniki

Podczas przedstawiania jakości osiągniętych wyników przez miary dwuelementowe dla korpusu *KIPI* używane będą odpowiednie numery badań zamiast ich pełnych nazw w celu identyfikacji rodzaju badania. Analogicznie stosowane będą identyfikatory miar zamiast ich pełnych nazw. Cały zestaw funkcji został zbadany dla każdego z wariantów badań w przypadku korpusu *KIPI*.

Wyniki badań zostały w dużym stopniu skompresowane do tabel zawierających wartości średniej uśrednionej precyzji ze względu na ilość otrzymanych wyników. Średnia uśredniona precyzja jest rozumiana jako miara uśrednionej precyzji (Average Precision) obliczana na każdej pozycji rankingu, na której znajdowało się wyrażenie wielowyrzowe, na pewnym zakresie tego rankingu – miara została opisana we wcześniejszej części tej pracy. Następnie wartości te zostały skompresowane do pojedynczej wartości poprzez obliczenie ich średniej. Autor niniejszej pracy postanowił ograniczyć zakres obliczania średniej uśrednionej precyzji z wartości z rankingów od ich 10% do 90%. Decyzja taka została podjęta ze względu na inspiracje artykułem [?].

nr	źródło danych statystycznych	źródło kandydatów	filtry
1	okno ciągłe	okno ciągłe	
2	okno ciągłe	okno ciągłe	morfeusz
3	okno ciągłe	okno ciągłe	morfeusz, częstość > 5
4	okna ciągłe i nieciągłe	okno ciągłe i nieciągłe	
5	okna ciągłe i nieciągłe	okno ciągłe i nieciągłe	morfeusz
6	okna ciągłe i nieciągłe	okno ciągłe i nieciągłe	morfeusz, częstość > 5
7	relacje ciągłe	relacje ciągłe	
8	relacje ciągłe	relacje ciągłe	morfeusz
9	relacje ciągłe	relacje ciągłe	morfeusz, częstość > 5
10	relacje ciągłe	częste relacje ciągłe	
11	relacje ciągłe	częste relacje ciągłe	morfeusz
12	relacje ciągłe	częste relacje ciągłe	morfeusz, częstość > 5
13	relacje i okno, ciągłe	relacje ciągłe	
14	relacje i okno, ciągłe	relacje ciągłe	morfeusz
15	relacje i okno, ciągłe	relacje ciągłe	morfeusz, częstość > 5
16	relacje i okno, ciągłe	częste relacje ciągłe	
17	relacje i okno, ciągłe	częste relacje ciągłe	morfeusz
18	relacje i okno, ciągłe	częste relacje ciągłe	morfeusz, częstość > 5
19	relacje ciągłe i nieciągłe	relacje ciągłe i nieciągłe	
20	relacje ciągłe i nieciągłe	relacje ciągłe i nieciągłe	morfeusz
21	relacje ciągłe i nieciągłe	relacje ciągłe i nieciągłe	morfeusz, częstość > 5
22	relacje ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	
23	relacje ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	morfeusz
24	relacje ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	morfeusz, częstość > 5
25	relacje i okno, ciągłe i nieciągłe	relacje ciągłe i nieciągłe	
26	relacje i okno, ciągłe i nieciągłe	relacje ciągłe i nieciągłe	morfeusz
27	relacje i okno, ciągłe i nieciągłe	relacje ciągłe i nieciągłe	morfeusz, częstość > 5
28	relacje i okno, ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	
29	relacje i okno, ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	morfeusz
30	relacje i okno, ciągłe i nieciągłe	częste relacje ciągłe i nieciągłe	morfeusz, częstość > 5

Tabela 5.7. Zestaw przeprowadzonych badań dla funkcji dwuelementowych na korpusie *KIPI*.

Najlepsza funkcja oraz zestaw funkcji, które osiągnęły wyniki na poziomie przynajmniej 95% jakości najlepszej z funkcji został oznaczony w tabelach za pomocą czcionki pogrubionej – wartości w odpowiednich komórkach.

Wyniki badań miar dwuelementowych

Cztery tabele ??, ??, ?? oraz ?? prezentują jakość wyników osiągniętych przez 72 funkcje w 30 różnych badaniach – 30 zestawów danych pozyskanych z korpusu *KIPI*. Ze względu na ilość wyników autor pracy postanowił zmniejszyć rozmiary tabel przedstawiających jakość wyników poprzez używanie liczb zamiast pełnych nazw badań i funkcji. Etykiety kolumn odpowiadają numerowi badania a wiersze numerom funkcji. Zestaw funkcji i typów badań (sposobu przygotowania korpusu *KIPI*) został opisane we wcześniejszej części tego rozdziału w tabelach odpowiednio ?? oraz ??.

Na podstawie tych czterech tabel zauważyć można, że najlepsze jakościowo wyniki zostały osiągnięte dla następujących miar asocjacyjnych: *W Specific Correlation*, *Specific Frequency Biased Mutual Dependency*, *Loglikelihood*, *W Order*, *W Term Frequency Order*

nr	nazwa	nr	nazwa
1	Frequency()	37	Specific Exponential Correlation(e=4.7)
2	Expected Frequency()	38	Specific Exponential Correlation(e=4.8)
3	Inversed Expected Frequency()	39	Specific Exponential Correlation(e=4.9)
4	Jaccard()	40	Specific Exponential Correlation(e=5)
5	Dice()	41	Specific Exponential Correlation(e=5.1)
6	Sorgenfrei()	42	Specific Exponential Correlation(e=5.2)
7	Odds Ratio()	43	Specific Exponential Correlation(e=5.3)
8	Unigram Subtuples()	44	Specific Exponential Correlation(e=5.4)
9	Consonni T1()	45	Specific Exponential Correlation(e=5.5)
10	Consonni T2()	46	Specific Exponential Correlation(e=5.6)
11	Mutual Expectation()	47	Specific Exponential Correlation(e=5.7)
12	Specific Correlation()	48	Specific Exponential Correlation(e=5.8)
13	W Specific Correlation()	49	Specific Exponential Correlation(e=5.9)
14	Specific Mutual Dependency()	50	Specific Exponential Correlation(e=6)
15	Specific Frequency Biased Mutual Dependency()	51	W Specific Exponential Correlation(e=1.05)
16	Tscore()	52	W Specific Exponential Correlation(e=1.1)
17	Zscore()	53	W Specific Exponential Correlation(e=1.15)
18	Pearsons Chi Square()	54	W Specific Exponential Correlation(e=1.2)
19	W Chi Square()	55	W Specific Exponential Correlation(e=1.25)
20	Loglikelihood()	56	W Specific Exponential Correlation(e=1.3)
21	Specific Exponential Correlation(e=3.1)	57	W Specific Exponential Correlation(e=1.35)
22	Specific Exponential Correlation(e=3.2)	58	W Specific Exponential Correlation(e=1.4)
23	Specific Exponential Correlation(e=3.3)	59	W Specific Exponential Correlation(e=1.45)
24	Specific Exponential Correlation(e=3.4)	60	W Specific Exponential Correlation(e=1.5)
25	Specific Exponential Correlation(e=3.5)	61	W Specific Exponential Correlation(e=1.55)
26	Specific Exponential Correlation(e=3.6)	62	W Specific Exponential Correlation(e=1.6)
27	Specific Exponential Correlation(e=3.7)	63	W Specific Exponential Correlation(e=1.65)
28	Specific Exponential Correlation(e=3.8)	64	W Specific Exponential Correlation(e=1.7)
29	Specific Exponential Correlation(e=3.9)	65	W Specific Exponential Correlation(e=1.75)
30	Specific Exponential Correlation(e=4)	66	W Specific Exponential Correlation(e=1.8)
31	Specific Exponential Correlation(e=4.1)	67	W Specific Exponential Correlation(e=1.85)
32	Specific Exponential Correlation(e=4.2)	68	W Specific Exponential Correlation(e=1.9)
33	Specific Exponential Correlation(e=4.3)	69	W Specific Exponential Correlation(e=1.95)
34	Specific Exponential Correlation(e=4.4)	70	W Specific Exponential Correlation(e=2)
35	Specific Exponential Correlation(e=4.5)	71	W Order()
36	Specific Exponential Correlation(e=4.6)	72	W Term Frequency Order()

Tabela 5.8. Zestaw zbadanych funkcji dwuelementowych na korpusie *KIPI*.

oraz zestaw miar *Specific Exponential Correlation* i *W Specific Exponential Correlation* dla pewnych wartości ich jedyne go parametru – wykładnika.

Analizując jakość wyników dla miar *Specific Exponential Correlation* i *W Specific Exponential Correlation* należy skupić się także na sposobie przygotowania danych dla nich, ponieważ wartości ich parametru, dla którego osiągnięto najlepsze wyniki, różnią się z tego względu.

Zauważyć można, że im dane są bardziej przefiltrowana tym lepsze wyniki są osiągnięte dla mniejszych wartości wykładnika. Przykładowo dla miary *Specific Exponential Correlation* w zadaniu ekstrakcji kolokacji z grona kandydatów pozyskanego operatorami oknowymi wartości parametru w okolicach liczby 3,0 są preferowane dla bardziej przefiltrowanych danych, a do ekstrakcji kolokacji z mniej przefiltrowanych danych lepszą war-

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0184	0,0261	0,0370	0,0100	0,0136	0,0219	0,0522	0,0612	0,0906	0,2095	0,2117	0,3000	0,0523	0,0614	0,0910
2	0,0027	0,0038	0,0067	0,0024	0,0033	0,0053	0,0071	0,0085	0,0189	0,0372	0,0376	0,0792	0,0078	0,0094	0,0198
3	0,0014	0,0022	0,0244	0,0007	0,0011	0,0141	0,0024	0,0042	0,0324	0,0109	0,0110	0,0930	0,0024	0,0045	0,0362
4	0,0078	0,0171	0,0584	0,0032	0,0058	0,0300	0,0117	0,0313	0,0810	0,0836	0,0849	0,2490	0,0122	0,0359	0,0892
5	0,0078	0,0171	0,0584	0,0032	0,0058	0,0300	0,0117	0,0313	0,0810	0,0836	0,0849	0,2490	0,0122	0,0359	0,0892
6	0,0081	0,0231	0,0714	0,0034	0,0079	0,0392	0,0107	0,0368	0,0868	0,0979	0,0993	0,2698	0,0106	0,0413	0,0958
7	0,0030	0,0057	0,0471	0,0014	0,0023	0,0252	0,0047	0,0109	0,0576	0,0274	0,0277	0,1717	0,0045	0,0111	0,0598
8	0,0048	0,0103	0,0522	0,0020	0,0038	0,0278	0,0077	0,0191	0,0651	0,0486	0,0492	0,1966	0,0067	0,0195	0,0684
9	0,0015	0,0020	0,0215	0,0008	0,0010	0,0115	0,0029	0,0041	0,0331	0,0107	0,0108	0,0949	0,0030	0,0045	0,0359
10	0,0015	0,0020	0,0215	0,0008	0,0010	0,0115	0,0029	0,0041	0,0332	0,0107	0,0108	0,0949	0,0030	0,0045	0,0359
11	0,0264	0,0368	0,0646	0,0115	0,0164	0,0374	0,0500	0,0705	0,1116	0,2242	0,2278	0,3584	0,0486	0,0711	0,1140
12	0,0028	0,0053	0,0418	0,0013	0,0023	0,0231	0,0043	0,0094	0,0507	0,0238	0,0239	0,1483	0,0044	0,0107	0,0578
13	0,0344	0,0440	0,0698	0,0169	0,0211	0,0466	0,0621	0,0744	0,1155	0,2430	0,2463	0,3753	0,0610	0,0740	0,1186
14	0,0081	0,0231	0,0714	0,0034	0,0079	0,0392	0,0107	0,0368	0,0868	0,0979	0,0993	0,2698	0,0106	0,0413	0,0958
15	0,0223	0,0466	0,0815	0,0090	0,0193	0,0483	0,0336	0,0745	0,1172	0,2345	0,2377	0,3794	0,0310	0,0759	0,1219
16	0,0292	0,0376	0,0566	0,0160	0,0195	0,0400	0,0557	0,0656	0,1003	0,2172	0,2205	0,3286	0,0563	0,0665	0,1049
17	0,0079	0,0224	0,0711	0,0031	0,0072	0,0385	0,0106	0,0362	0,0864	0,0965	0,0978	0,2685	0,0103	0,0401	0,0950
18	0,0080	0,0227	0,0711	0,0033	0,0078	0,0382	0,0107	0,0367	0,0867	0,0977	0,0991	0,2692	0,0107	0,0417	0,0958
19	0,0079	0,0224	0,0711	0,0031	0,0071	0,0385	0,0106	0,0362	0,0864	0,0964	0,0977	0,2684	0,0103	0,0401	0,0950
20	0,0287	0,0384	0,0649	0,0097	0,0136	0,0300	0,0628	0,0751	0,1181	0,2509	0,2530	0,3869	0,0569	0,0678	0,1143
21	0,0238	0,0469	0,0808	0,0097	0,0198	0,0481	0,0367	0,0758	0,1184	0,2397	0,2430	0,3835	0,0338	0,0769	0,1228
22	0,0251	0,0470	0,0800	0,0104	0,0202	0,0479	0,0398	0,0767	0,1193	0,2438	0,2471	0,3866	0,0367	0,0777	0,1234
23	0,0263	0,0469	0,0790	0,0110	0,0205	0,0475	0,0428	0,0775	0,1199	0,2470	0,2503	0,3889	0,0395	0,0783	0,1238
24	0,0273	0,0467	0,0780	0,0115	0,0207	0,0470	0,0455	0,0780	0,1204	0,2494	0,2526	0,3906	0,0423	0,0787	0,1241
25	0,0281	0,0464	0,0770	0,0120	0,0208	0,0465	0,0479	0,0783	0,1207	0,2511	0,2544	0,3917	0,0448	0,0789	0,1242
26	0,0287	0,0460	0,0759	0,0124	0,0209	0,0459	0,0501	0,0786	0,1209	0,2524	0,2557	0,3922	0,0471	0,0791	0,1242
27	0,0292	0,0456	0,0748	0,0128	0,0209	0,0454	0,0519	0,0787	0,1209	0,2533	0,2566	0,3924	0,0491	0,0792	0,1241
28	0,0296	0,0452	0,0737	0,0131	0,0208	0,0448	0,0535	0,0788	0,1209	0,2540	0,2572	0,3923	0,0508	0,0792	0,1239
29	0,0298	0,0448	0,0727	0,0134	0,0208	0,0442	0,0549	0,0788	0,1208	0,2543	0,2576	0,3919	0,0524	0,0791	0,1237
30	0,0299	0,0444	0,0717	0,0136	0,0207	0,0436	0,0560	0,0787	0,1206	0,2545	0,2577	0,3913	0,0537	0,0790	0,1234
31	0,0300	0,0439	0,0707	0,0137	0,0206	0,0430	0,0569	0,0786	0,1204	0,2545	0,2578	0,3906	0,0548	0,0789	0,1231
32	0,0299	0,0435	0,0697	0,0138	0,0205	0,0424	0,0577	0,0785	0,1201	0,2544	0,2577	0,3897	0,0557	0,0788	0,1227
33	0,0299	0,0431	0,0688	0,0139	0,0204	0,0418	0,0584	0,0783	0,1198	0,2542	0,2575	0,3887	0,0565	0,0786	0,1223
34	0,0298	0,0427	0,0679	0,0140	0,0203	0,0413	0,0590	0,0782	0,1195	0,2540	0,2572	0,3877	0,0572	0,0784	0,1219
35	0,0296	0,0423	0,0670	0,0140	0,0202	0,0407	0,0594	0,0780	0,1191	0,2536	0,2568	0,3866	0,0578	0,0782	0,1215
36	0,0295	0,0419	0,0662	0,0140	0,0201	0,0402	0,0598	0,0778	0,1187	0,2533	0,2564	0,3854	0,0582	0,0780	0,1211

Tabela 5.9. Wyniki badań miar dwuelementowych dla korpusu *KIPI*, część 1.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0,2097	0,2097	0,3011	0,0273	0,0340	0,0625	0,1577	0,1585	0,2688	0,0273	0,0337	0,0628	0,1589	0,1588	0,2691
2	0,0396	0,0400	0,0804	0,0047	0,0062	0,0127	0,0383	0,0388	0,0850	0,0067	0,0085	0,0155	0,0423	0,0432	0,0870
3	0,0121	0,0123	0,1105	0,0014	0,0025	0,0223	0,0095	0,0095	0,0863	0,0013	0,0024	0,0232	0,0103	0,0104	0,1013
4	0,1057	0,1028	0,2917	0,0054	0,0139	0,0517	0,0524	0,0531	0,2080	0,0052	0,0141	0,0538	0,0660	0,0645	0,2417
5	0,1057	0,1028	0,2917	0,0054	0,0139	0,0517	0,0524	0,0531	0,2080	0,0052	0,0141	0,0538	0,0660	0,0645	0,2417
6	0,1202	0,1178	0,3175	0,0053	0,0182	0,0610	0,0653	0,0659	0,2356	0,0047	0,0171	0,0624	0,0762	0,0748	0,2670
7	0,0294	0,0297	0,1879	0,0025	0,0058	0,0403	0,0208	0,0209	0,1530	0,0022	0,0052	0,0379	0,0222	0,0222	0,1633
8	0,0520	0,0521	0,2186	0,0037	0,0095	0,0445	0,0341	0,0343	0,1725	0,0030	0,0084	0,0432	0,0362	0,0360	0,1876
9	0,0121	0,0124	0,1092	0,0015	0,0023	0,0203	0,0087	0,0088	0,0826	0,0015	0,0023	0,0210	0,0103	0,0104	0,0969
10	0,0121	0,0124	0,1092	0,0015	0,0023	0,0203	0,0087	0,0088	0,0826	0,0015	0,0023	0,0210	0,0103	0,0104	0,0969
11	0,2315	0,2331	0,3763	0,0232	0,0371	0,0791	0,1449	0,1464	0,3033	0,0211	0,0354	0,0784	0,1537	0,1536	0,3210
12	0,0285	0,0286	0,1813	0,0023	0,0053	0,0349	0,0192	0,0193	0,1352	0,0021	0,0051	0,0367	0,0219	0,0219	0,1598
13	0,2440	0,2461	0,3932	0,0331	0,0422	0,0875	0,1629	0,1633	0,3285	0,0286	0,0373	0,0858	0,1571	0,1569	0,3422
14	0,1202	0,1178	0,3175	0,0053	0,0182	0,0610	0,0653	0,0659	0,2356	0,0047	0,0171	0,0624	0,0762	0,0748	0,2670
15	0,2482	0,2489	0,4117	0,0154	0,0408	0,0871	0,1512	0,1523	0,3277	0,0129	0,0370	0,0855	0,1596	0,1587	0,3504
16	0,2193	0,2216	0,3458	0,0302	0,0374	0,0748	0,1502	0,1508	0,2908	0,0273	0,0342	0,0761	0,1457	0,1460	0,3071
17	0,1168	0,1142	0,3150	0,0052	0,0177	0,0605	0,0637	0,0640	0,2338	0,0044	0,0160	0,0613	0,0715	0,0699	0,2624
18	0,1214	0,1194	0,3179	0,0054	0,0186	0,0613	0,0666	0,0674	0,2371	0,0051	0,0190	0,0635	0,0840	0,0835	0,2731
19	0,1167	0,1142	0,3149	0,0052	0,0177	0,0605	0,0636	0,0640	0,2338	0,0044	0,0160	0,0613	0,0714	0,0698	0,2624
20	0,2371	0,2345	0,3895	0,0324	0,0413	0,0869	0,1711	0,1688	0,3354	0,0248	0,0307	0,0696	0,1304	0,1270	0,2848
21	0,2521	0,2532	0,4141	0,0169	0,0418	0,0883	0,1556	0,1566	0,3316	0,0141	0,0380	0,0866	0,1632	0,1625	0,3534
22	0,2550	0,2563	0,4157	0,0184	0,0426	0,0893	0,1591	0,1601	0,3347	0,0154	0,0388	0,0874	0,1661	0,1656	0,3557
23	0,2571	0,2586	0,4165	0,0199	0,0433	0,0900	0,1620	0,1631	0,3372	0,0167	0,0395	0,0881	0,1685	0,1681	0,3574
24	0,2586	0,2603	0,4168	0,0214	0,0438	0,0905	0,1644	0,1655	0,3390	0,0179	0,0400	0,0886	0,1704	0,1701	0,3585
25	0,2596	0,2614	0,4167	0,0228	0,0442	0,0909	0,1664	0,1674	0,3404	0,0192	0,0404	0,0889	0,1720	0,1717	0,3591
26	0,2603	0,2621	0,4162	0,0242	0,0444	0,0911	0,1680	0,1690	0,3413	0,0204	0,0407	0,0891	0,1732	0,1730	0,3593
27	0,2606	0,2625	0,4153	0,0255	0,0446	0,0913	0,1693	0,1703	0,3419	0,0216	0,0410	0,0892	0,1743	0,1741	0,3592
28	0,2607	0,2627	0,4142	0,0266	0,0448	0,0913	0,1703	0,1713	0,3421	0,0227	0,0412	0,0892	0,1751	0,1750	0,3588
29	0,2606	0,2626	0,4129	0,0276	0,0448	0,0912	0,1712	0,1722	0,3422	0,0238	0,0413	0,0892	0,1757	0,1756	0,3583
30	0,2604	0,2624	0,4115	0,0284	0,0449	0,0911	0,1719	0,1729	0,3420	0,0246	0,0414	0,0891	0,1762	0,1762	0,3575
31	0,2601	0,2621	0,4099	0,0291	0,0449	0,0909	0,1725	0,1735	0,3416	0,0254	0,0415	0,0889	0,1766	0,1766	0,3567
32	0,2597	0,2617	0,4083	0,0297	0,0448	0,0907	0,1729	0,1739	0,3411	0,0261	0,0415	0,0887	0,1769	0,1769	0,3557
33	0,2592	0,2612	0,4067	0,0302	0,0448	0,0904	0,1733	0,1743	0,3405	0,0267	0,0415	0,0885	0,1771	0,1772	0,3546
34	0,2587	0,2607	0,4050	0,0307	0,0447	0,0901	0,1736	0,1746	0,3399	0,0272	0,0415	0,0883	0,1773	0,1774	0,3535
35	0,2581	0,2601	0,4033	0,0310	0,0446	0,0898	0,1738	0,1748	0,3391	0,0277	0,0415	0,0880	0,1774	0,1775	0,3524
36	0,2575	0,2595	0,4016	0,0313	0,0445	0,0895	0,1740	0,1750	0,3383	0,0280	0,0415	0,0877	0,1775	0,1776	0,3512

Tabela 5.10. Wyniki badań miar dwuelementowych dla korpusu *KIPI*, część 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
37	0,0293	0,0415	0,0654	0,0140	0,0199	0,0397	0,0601	0,0776	0,1184	0,2529	0,2560	0,3842	0,0586	0,0778	0,1206
38	0,0292	0,0412	0,0646	0,0140	0,0198	0,0393	0,0604	0,0774	0,1180	0,2524	0,2556	0,3830	0,0590	0,0776	0,1202
39	0,0290	0,0408	0,0639	0,0140	0,0197	0,0388	0,0606	0,0772	0,1176	0,2520	0,2551	0,3818	0,0593	0,0774	0,1198
40	0,0288	0,0405	0,0632	0,0139	0,0196	0,0384	0,0607	0,0770	0,1173	0,2515	0,2546	0,3806	0,0595	0,0772	0,1193
41	0,0286	0,0402	0,0625	0,0139	0,0195	0,0379	0,0609	0,0768	0,1169	0,2510	0,2541	0,3794	0,0597	0,0769	0,1189
42	0,0284	0,0398	0,0619	0,0138	0,0194	0,0375	0,0610	0,0766	0,1165	0,2505	0,2536	0,3782	0,0598	0,0767	0,1185
43	0,0282	0,0395	0,0613	0,0138	0,0192	0,0371	0,0611	0,0764	0,1161	0,2500	0,2531	0,3770	0,0600	0,0765	0,1180
44	0,0281	0,0392	0,0607	0,0138	0,0191	0,0368	0,0611	0,0762	0,1157	0,2495	0,2526	0,3758	0,0601	0,0763	0,1176
45	0,0279	0,0390	0,0602	0,0137	0,0190	0,0364	0,0611	0,0760	0,1154	0,2490	0,2520	0,3747	0,0601	0,0761	0,1172
46	0,0277	0,0387	0,0596	0,0137	0,0189	0,0361	0,0612	0,0758	0,1150	0,2485	0,2515	0,3735	0,0602	0,0759	0,1168
47	0,0275	0,0384	0,0591	0,0136	0,0188	0,0357	0,0612	0,0756	0,1147	0,2480	0,2510	0,3724	0,0603	0,0757	0,1164
48	0,0274	0,0382	0,0586	0,0136	0,0187	0,0354	0,0612	0,0754	0,1143	0,2475	0,2505	0,3713	0,0603	0,0755	0,1161
49	0,0272	0,0379	0,0581	0,0135	0,0186	0,0351	0,0612	0,0752	0,1140	0,2471	0,2500	0,3702	0,0603	0,0753	0,1157
50	0,0270	0,0377	0,0577	0,0135	0,0186	0,0348	0,0611	0,0750	0,1136	0,2466	0,2495	0,3692	0,0603	0,0751	0,1153
51	0,0364	0,0455	0,0744	0,0161	0,0201	0,0484	0,0621	0,0747	0,1174	0,2432	0,2464	0,3814	0,0603	0,0738	0,1206
52	0,0369	0,0455	0,0783	0,0143	0,0181	0,0491	0,0617	0,0746	0,1191	0,2420	0,2451	0,3869	0,0584	0,0723	0,1216
53	0,0352	0,0436	0,0806	0,0116	0,0150	0,0480	0,0605	0,0737	0,1202	0,2384	0,2410	0,3904	0,0546	0,0688	0,1212
54	0,0312	0,0391	0,0804	0,0081	0,0110	0,0446	0,0581	0,0714	0,1204	0,2308	0,2326	0,3903	0,0483	0,0624	0,1182
55	0,0253	0,0330	0,0761	0,0053	0,0078	0,0394	0,0538	0,0671	0,1189	0,2162	0,2172	0,3849	0,0401	0,0544	0,1122
56	0,0199	0,0279	0,0702	0,0029	0,0047	0,0348	0,0475	0,0604	0,1152	0,1936	0,1940	0,3722	0,0308	0,0448	0,1037
57	0,0141	0,0219	0,0645	0,0017	0,0025	0,0294	0,0394	0,0516	0,1088	0,1639	0,1636	0,3510	0,0207	0,0335	0,0927
58	0,0081	0,0148	0,0570	0,0012	0,0018	0,0235	0,0297	0,0405	0,0992	0,1263	0,1253	0,3185	0,0119	0,0221	0,0791
59	0,0039	0,0079	0,0476	0,0010	0,0014	0,0179	0,0200	0,0293	0,0871	0,0874	0,0872	0,2761	0,0060	0,0118	0,0644
60	0,0023	0,0037	0,0377	0,0008	0,0011	0,0136	0,0119	0,0188	0,0735	0,0527	0,0524	0,2303	0,0038	0,0063	0,0507
61	0,0016	0,0025	0,0282	0,0007	0,0009	0,0107	0,0063	0,0102	0,0599	0,0256	0,0257	0,1857	0,0027	0,0044	0,0390
62	0,0013	0,0019	0,0210	0,0006	0,0008	0,0089	0,0041	0,0061	0,0477	0,0168	0,0169	0,1453	0,0022	0,0035	0,0301
63	0,0011	0,0016	0,0162	0,0005	0,0007	0,0076	0,0030	0,0045	0,0376	0,0127	0,0128	0,1125	0,0018	0,0029	0,0241
64	0,0009	0,0013	0,0132	0,0005	0,0007	0,0067	0,0024	0,0035	0,0298	0,0101	0,0102	0,0890	0,0016	0,0025	0,0201
65	0,0008	0,0012	0,0111	0,0005	0,0006	0,0060	0,0020	0,0029	0,0244	0,0084	0,0085	0,0731	0,0014	0,0022	0,0173
66	0,0008	0,0011	0,0096	0,0004	0,0006	0,0055	0,0017	0,0025	0,0205	0,0073	0,0074	0,0617	0,0013	0,0020	0,0154
67	0,0007	0,0010	0,0087	0,0004	0,0006	0,0051	0,0016	0,0023	0,0178	0,0065	0,0065	0,0532	0,0013	0,0019	0,0141
68	0,0007	0,0009	0,0079	0,0004	0,0006	0,0048	0,0014	0,0021	0,0157	0,0059	0,0060	0,0470	0,0012	0,0018	0,0131
69	0,0007	0,0009	0,0074	0,0004	0,0006	0,0046	0,0013	0,0019	0,0142	0,0056	0,0057	0,0425	0,0012	0,0018	0,0124
70	0,0006	0,0009	0,0070	0,0004	0,0006	0,0045	0,0013	0,0018	0,0131	0,0055	0,0055	0,0393	0,0012	0,0018	0,0119
71	0,0150	0,0191	0,0226	0,0280	0,0364	0,0430	0,0356	0,0422	0,0498	0,1569	0,1587	0,1873	0,0357	0,0423	0,0500
72	0,0182	0,0239	0,0310	0,0250	0,0327	0,0425	0,0418	0,0493	0,0647	0,1826	0,1843	0,2364	0,0418	0,0494	0,0649

Tabela 5.11. Wyniki badań miar dwuelementowych dla korpusu *KIPI*, część 3.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
37	0,2569	0,2589	0,3999	0,0316	0,0444	0,0891	0,1741	0,1751	0,3375	0,0284	0,0414	0,0874	0,1775	0,1776	0,3500
38	0,2563	0,2582	0,3982	0,0318	0,0443	0,0888	0,1742	0,1752	0,3366	0,0287	0,0414	0,0871	0,1775	0,1776	0,3488
39	0,2557	0,2576	0,3965	0,0320	0,0442	0,0884	0,1743	0,1752	0,3357	0,0289	0,0413	0,0867	0,1775	0,1776	0,3476
40	0,2551	0,2569	0,3949	0,0321	0,0441	0,0880	0,1743	0,1753	0,3348	0,0291	0,0412	0,0864	0,1775	0,1775	0,3464
41	0,2544	0,2563	0,3933	0,0322	0,0440	0,0877	0,1743	0,1753	0,3339	0,0293	0,0412	0,0861	0,1774	0,1775	0,3451
42	0,2538	0,2556	0,3917	0,0323	0,0439	0,0873	0,1743	0,1753	0,3330	0,0295	0,0411	0,0857	0,1773	0,1774	0,3440
43	0,2532	0,2550	0,3902	0,0324	0,0437	0,0869	0,1743	0,1752	0,3321	0,0296	0,0410	0,0854	0,1772	0,1773	0,3428
44	0,2526	0,2543	0,3886	0,0324	0,0436	0,0866	0,1742	0,1752	0,3312	0,0297	0,0410	0,0851	0,1771	0,1772	0,3416
45	0,2519	0,2537	0,3871	0,0325	0,0435	0,0862	0,1742	0,1751	0,3303	0,0298	0,0409	0,0848	0,1770	0,1771	0,3405
46	0,2513	0,2531	0,3857	0,0325	0,0434	0,0859	0,1741	0,1750	0,3294	0,0299	0,0408	0,0845	0,1769	0,1770	0,3394
47	0,2508	0,2525	0,3842	0,0326	0,0433	0,0855	0,1740	0,1750	0,3286	0,0300	0,0407	0,0841	0,1768	0,1768	0,3383
48	0,2502	0,2518	0,3829	0,0326	0,0431	0,0852	0,1739	0,1749	0,3277	0,0301	0,0407	0,0838	0,1766	0,1767	0,3372
49	0,2496	0,2513	0,3815	0,0326	0,0430	0,0849	0,1738	0,1748	0,3269	0,0301	0,0406	0,0835	0,1765	0,1766	0,3362
50	0,2491	0,2507	0,3802	0,0326	0,0429	0,0845	0,1737	0,1747	0,3261	0,0302	0,0405	0,0832	0,1763	0,1764	0,3351
51	0,2437	0,2455	0,4011	0,0328	0,0419	0,0893	0,1600	0,1600	0,3325	0,0274	0,0363	0,0869	0,1526	0,1520	0,3464
52	0,2394	0,2405	0,4061	0,0319	0,0411	0,0906	0,1552	0,1549	0,3353	0,0252	0,0341	0,0869	0,1436	0,1427	0,3466
53	0,2279	0,2283	0,4060	0,0303	0,0396	0,0912	0,1484	0,1478	0,3358	0,0216	0,0302	0,0848	0,1272	0,1260	0,3396
54	0,2054	0,2060	0,3966	0,0280	0,0372	0,0907	0,1388	0,1378	0,3330	0,0169	0,0250	0,0804	0,1045	0,1043	0,3236
55	0,1752	0,1781	0,3775	0,0245	0,0333	0,0884	0,1246	0,1231	0,3245	0,0117	0,0191	0,0733	0,0791	0,0799	0,2982
56	0,1402	0,1448	0,3488	0,0198	0,0281	0,0840	0,1056	0,1034	0,3099	0,0067	0,0124	0,0644	0,0499	0,0514	0,2654
57	0,0990	0,1049	0,3104	0,0141	0,0212	0,0771	0,0807	0,0779	0,2872	0,0035	0,0066	0,0537	0,0254	0,0270	0,2254
58	0,0586	0,0652	0,2642	0,0085	0,0138	0,0677	0,0531	0,0506	0,2545	0,0023	0,0039	0,0427	0,0163	0,0170	0,1827
59	0,0262	0,0306	0,2129	0,0044	0,0076	0,0562	0,0287	0,0278	0,2140	0,0017	0,0029	0,0326	0,0122	0,0128	0,1408
60	0,0162	0,0177	0,1637	0,0028	0,0045	0,0451	0,0180	0,0178	0,1742	0,0014	0,0023	0,0246	0,0097	0,0101	0,1075
61	0,0120	0,0129	0,1220	0,0021	0,0033	0,0349	0,0137	0,0136	0,1367	0,0011	0,0019	0,0192	0,0080	0,0084	0,0852
62	0,0095	0,0102	0,0935	0,0017	0,0026	0,0268	0,0109	0,0109	0,1074	0,0010	0,0017	0,0156	0,0070	0,0073	0,0699
63	0,0079	0,0084	0,0748	0,0014	0,0022	0,0211	0,0089	0,0090	0,0860	0,0009	0,0015	0,0132	0,0063	0,0065	0,0596
64	0,0068	0,0072	0,0623	0,0012	0,0019	0,0172	0,0077	0,0077	0,0713	0,0008	0,0014	0,0115	0,0058	0,0060	0,0519
65	0,0061	0,0065	0,0535	0,0010	0,0017	0,0146	0,0068	0,0068	0,0610	0,0008	0,0013	0,0103	0,0056	0,0057	0,0465
66	0,0058	0,0060	0,0472	0,0010	0,0015	0,0126	0,0062	0,0062	0,0532	0,0008	0,0012	0,0094	0,0055	0,0056	0,0426
67	0,0056	0,0057	0,0431	0,0009	0,0014	0,0111	0,0058	0,0058	0,0472	0,0007	0,0012	0,0088	0,0054	0,0055	0,0401
68	0,0055	0,0056	0,0402	0,0008	0,0013	0,0100	0,0056	0,0056	0,0429	0,0007	0,0012	0,0083	0,0054	0,0054	0,0383
69	0,0054	0,0055	0,0383	0,0008	0,0013	0,0092	0,0055	0,0055	0,0398	0,0007	0,0012	0,0080	0,0053	0,0054	0,0371
70	0,0053	0,0054	0,0370	0,0008	0,0013	0,0087	0,0054	0,0054	0,0376	0,0007	0,0012	0,0077	0,0053	0,0054	0,0362
71	0,1571	0,1572	0,1879	0,0643	0,0792	0,0987	0,2668	0,2680	0,3298	0,0643	0,0786	0,0993	0,2688	0,2686	0,3302
72	0,1827	0,1828	0,2371	0,0572	0,0706	0,0962	0,2519	0,2531	0,3278	0,0572	0,0701	0,0967	0,2537	0,2536	0,3281

Tabela 5.12. Wyniki badań miar dwuelementowych dla korpusu *KIPI*, część 4.

tością parametru wydają się być wartości bliższe 4,0. Dodatkowo dla zadań związanych z ekstrakcją wyrażeń wielowyrazowych preferowane dla obu funkcji są w większości raczej niskie wartości ich parametru – niskie jak na przyjęty przedział optymalizacji.

Na podstawie wyników można spróbować wysunąć wniosek, że w okolicach wartości optymalnej parametru funkcji *Specific Exponential Correlation* i *W Specific Exponential Correlation* dopuszczalny margines błędu podczas jego dostrajania jest dość duży. Innymi słowy jeśli zostanie znaleziona wartość bliska optymalnej to osiągnęte przez funkcje wyniki też będą bliskie optimum. Nie występują duże skoki w jakości rozwiązania co może uprościć dostrajanie tego parametru dla tej funkcji. Tym samym funkcje wydają się być odporne na niewielkie zmiany wartości parametrów w okolicach ich optimum. Przykładowo dla wartości parametru 1,05 i 1,1 dla badania ze składem oznaczonym numerem 2 różnica w wyniku jest minimalna lub wręcz niezauważalna, ale z drugiej strony zmiana z 1,5 na 1,45 zaowocowała poprawieniem wyniku ponad dwukrotnie. Na uwadze trzeba jednak mieć to, że w pierwszym przypadku dla obu wartości otrzymywane są stosunkowo dobre wyniki w porównaniu z innymi funkcjami, a w przypadku drugiej zmiany obie wartości parametru były znacząco bardziej oddalone od wartości optymalnej.

Dodatkowo w przypadku operatorów oknowych dodanie kandydatów nieciągłych pogarsza wyniki dla każdej z miar z wyjątkiem dwóch – *W Order* oraz *W Term Frequency Order*, dla których jakość wyników została poprawiona o około 37% dla wszystkich trzech przypadków filtrowania – braku, filtru opartego o morfeusza oraz filtru morfeusza i częstości. Powodem takiego wzrostu może być to, że obie miary badają szyk kandydata kolokacji, a tym samym przy dodaniu kolejnych relacji i kandydatów nieciągłych funkcje te miały około dwa razy więcej informacji o kandydatach niż przy braku kandydatów nieciągłych. Dodać także należy, że dzięki poszerzeniu grona kandydatów miary te wysunęły się na prowadzenie jeśli chodzi o jakość wyników dla zbioru bez filtracji i filtracji z użyciem morfeusza. Natomiast w przypadku zbioru poddanemu filtrowaniu zbiorem słów morfeusza i częstości funkcje te przestały być tak dobre, ale ich wyniki w dalszym ciągu mogą być godne uwagi. Natomiast w przypadku operatorów relacyjnych sytuacja jest prawie taka sama, czyli ogólne pogorszenie wyników z wyjątkiem miar opartych o szyk – *W Order* oraz *W Term Frequency Order*. Istnieją jednak pewne różnice. Tym razem miary oparte o szyk wysunęły się na prowadzenie we wszystkich trzech przypadkach związanych z filtrowaniem danych, a nie tylko dwóch pierwszych. Dodatkowo w tym przypadku wzrost jakości rozwiązań generowanych przez miary oparte o szyk wyniósł około 81%, 88% oraz 98% dla odpowiednio trzech kolejnych sposobach filtracji – brak, morfeusz, morfeusz i częstość większa od pięciu.

Istotną obserwacją może być też ta, że funkcje z rodziny *Specific Exponential Correlation* i *W Specific Exponential Correlation* zdają się mieć jedynie pojedyncze optimum dla swojego parametru, a jeśli tak jest (wyniki wydają się to potwierdzać) to proces optymalizacji tego parametru – w dodatku jednego, powinien być zadaniem dość prostym.

Niestety optymalna wartość parametru dla obu funkcji jest różna, a dodatkowo zmienna dla różnych zestawów danych. Mimo to wydaje się, że można wydzielić pewne zakresy wartości dla tych parametrów, w obrębie, których można próbować optymalizować wartość wykładnika. Dla *W Specific Exponential Correlation* zakres taki można spróbować ustalić na od 0,0+ do 1,3. Sytuacja jest trudniejsza dla *Specific Exponential Correlation*, ponieważ w większości przypadków zakres taki można byłoby ograniczyć do wartości od mniej niż 3,0 do 4,1, ale w kilku sytuacjach zakres ten należałoby rozszerzyć do nawet 6,0. Zaznaczyć jednak trzeba, że w sytuacji, w której *Specific Exponential Correlation* osiąga dobre wyniki dla parametru o wartości 6,0 nie znajduje się ona w czołówce najlepszych funkcji (ich 5%). Dodatkowo istotnym jest, że są to tylko propozycje zakresów wartości parametrów dla tych miar ustalone przez autora tej pracy na podstawie konkretnego zbioru danych, ale z drugiej strony mogą być pomocne do wyznaczenia wartości początkowej optymalizowanego parametru.

Ciekawym jest też to, że przyjęta przez autora niniejszej pracy ziarnistość optymalizacji tych parametrów pozwala na dokładność optymalizacji umożliwiającą dostrojenie parametrów do wartości bliskiej optymalnej i dla której ich ewentualna zmiana o tę właśnie ziarnistość powoduje jedynie niewielkie zmiany jakości – na poziomie czwartej cyfry licząc od pierwszej cyfry niezerowej wyniku. Jednak powodem tego może być też to o czym autor niniejszej pracy napisał wcześniej – funkcje wydają być się odporne na niewielkie zmiany wartości parametrów w okolicach ich optimów dla danego zbioru danych.

Należy jedna mieć na uwadze, że wszystkie zamieszczone tutaj wnioski pochodzą z pewnego określonego zbioru danych, a tym samym co do części z nich nie ma pewności w przypadku pracy z innym zbiorem tekstów.

Wyniki badań miar dwuelementowych na korpusie podzielonym na 10 fragmentów i poddanym dyspersji

Cztery tabele ??, ??, ?? oraz ?? prezentują jakość wyników osiągniętych przez 72 funkcje w 30 różnych badaniach – 30 zestawów danych pozyskanych z korpusu *KIPI*. Różnica pomiędzy tym a poprzednim badaniem jest taka, że w tym przypadku korpus *KIPI* został podzielony na 10 części i poddany dyspersji funkcją *TF-IDF*. Indeksy miar i typów badań pozostały takie same jak w poprzednim badaniu.

Najlepsze jakościowo wyniki zostały osiągnięte dla następujących miar asocjacyjnych: *Sorgenfrei*, *Specific Mutual Dependency*, *Specific Frequency Biased Mutual Dependency*, *W Specific Correlation*, *T-Score*, *Z-Score*, *Pearson's Chi²*, *Loglikelihood*, *W Order*, *W Term Frequency Order* oraz zestaw miar *Specific Exponential Correlation* i *W Specific Exponential Correlation* dla pewnych wartości ich jedyne parametru – wykładnika. Zestaw najlepszych funkcji wyznaczony na podstawie wyników niniejszego badania zawiera ich znaczną liczbę. Jednak część z nich okazała się pomocna jedynie przy ekstrakcji kolekacji z kandydatów wyznaczonych za pomocą operatorów oknowych.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0072	0,0116	0,0196	0,0042	0,0062	0,0111	0,0176	0,0246	0,0397	0,1098	0,1093	0,1869	0,0176	0,0248	0,0393
2	0,0011	0,0015	0,0028	0,0012	0,0016	0,0027	0,0031	0,0042	0,0095	0,0206	0,0206	0,0479	0,0037	0,0050	0,0102
3	0,0015	0,0026	0,0279	0,0008	0,0013	0,0158	0,0027	0,0052	0,0378	0,0130	0,0132	0,0963	0,0025	0,0051	0,0400
4	0,0034	0,0054	0,0321	0,0015	0,0021	0,0157	0,0058	0,0118	0,0487	0,0301	0,0311	0,1496	0,0060	0,0130	0,0514
5	0,0034	0,0054	0,0321	0,0015	0,0021	0,0157	0,0058	0,0118	0,0487	0,0301	0,0311	0,1496	0,0060	0,0130	0,0514
6	0,0033	0,0074	0,0400	0,0014	0,0028	0,0206	0,0052	0,0142	0,0531	0,0366	0,0373	0,1641	0,0049	0,0147	0,0557
7	0,0022	0,0042	0,0363	0,0010	0,0018	0,0192	0,0037	0,0084	0,0467	0,0209	0,0212	0,1327	0,0034	0,0082	0,0484
8	0,0029	0,0056	0,0363	0,0012	0,0023	0,0189	0,0050	0,0115	0,0495	0,0290	0,0295	0,1492	0,0043	0,0114	0,0519
9	0,0015	0,0019	0,0207	0,0008	0,0010	0,0110	0,0030	0,0044	0,0337	0,0108	0,0110	0,0881	0,0030	0,0046	0,0359
10	0,0015	0,0019	0,0207	0,0008	0,0010	0,0110	0,0030	0,0044	0,0337	0,0108	0,0110	0,0881	0,0030	0,0046	0,0359
11	0,0067	0,0118	0,0335	0,0027	0,0043	0,0170	0,0118	0,0247	0,0545	0,0815	0,0830	0,2053	0,0116	0,0256	0,0549
12	0,0022	0,0040	0,0339	0,0010	0,0018	0,0182	0,0035	0,0079	0,0445	0,0196	0,0199	0,1210	0,0033	0,0080	0,0476
13	0,0108	0,0175	0,0337	0,0040	0,0067	0,0192	0,0196	0,0303	0,0529	0,1204	0,1197	0,2260	0,0178	0,0292	0,0535
14	0,0033	0,0074	0,0400	0,0014	0,0028	0,0206	0,0052	0,0142	0,0531	0,0366	0,0373	0,1642	0,0049	0,0147	0,0558
15	0,0052	0,0125	0,0407	0,0021	0,0044	0,0212	0,0080	0,0232	0,0578	0,0700	0,0711	0,2081	0,0075	0,0235	0,0586
16	0,0100	0,0153	0,0249	0,0050	0,0073	0,0154	0,0195	0,0268	0,0429	0,1127	0,1126	0,1955	0,0190	0,0268	0,0442
17	0,0033	0,0073	0,0400	0,0014	0,0027	0,0207	0,0052	0,0141	0,0531	0,0364	0,0370	0,1640	0,0049	0,0145	0,0557
18	0,0033	0,0073	0,0400	0,0014	0,0028	0,0206	0,0052	0,0141	0,0531	0,0364	0,0371	0,1640	0,0049	0,0146	0,0557
19	0,0033	0,0073	0,0400	0,0014	0,0027	0,0207	0,0052	0,0141	0,0531	0,0364	0,0370	0,1639	0,0049	0,0145	0,0557
20	0,0100	0,0171	0,0357	0,0033	0,0058	0,0191	0,0188	0,0307	0,0549	0,1202	0,1195	0,2311	0,0171	0,0297	0,0553
21	0,0054	0,0129	0,0405	0,0021	0,0045	0,0211	0,0084	0,0240	0,0579	0,0736	0,0747	0,2110	0,0078	0,0242	0,0586
22	0,0056	0,0134	0,0403	0,0022	0,0047	0,0210	0,0088	0,0247	0,0580	0,0771	0,0781	0,2135	0,0081	0,0249	0,0586
23	0,0058	0,0138	0,0400	0,0023	0,0049	0,0209	0,0091	0,0254	0,0580	0,0805	0,0815	0,2157	0,0085	0,0255	0,0585
24	0,0060	0,0141	0,0397	0,0024	0,0050	0,0208	0,0095	0,0260	0,0580	0,0838	0,0847	0,2177	0,0088	0,0260	0,0584
25	0,0062	0,0145	0,0395	0,0024	0,0052	0,0207	0,0099	0,0266	0,0579	0,0869	0,0878	0,2195	0,0092	0,0265	0,0582
26	0,0064	0,0148	0,0392	0,0025	0,0053	0,0206	0,0103	0,0271	0,0579	0,0898	0,0907	0,2210	0,0095	0,0270	0,0581
27	0,0066	0,0151	0,0389	0,0026	0,0054	0,0205	0,0107	0,0275	0,0578	0,0926	0,0934	0,2223	0,0099	0,0274	0,0579
28	0,0068	0,0153	0,0385	0,0027	0,0056	0,0204	0,0110	0,0279	0,0576	0,0952	0,0959	0,2234	0,0102	0,0277	0,0577
29	0,0070	0,0155	0,0382	0,0028	0,0057	0,0202	0,0114	0,0283	0,0575	0,0976	0,0982	0,2244	0,0106	0,0280	0,0575
30	0,0072	0,0157	0,0379	0,0028	0,0058	0,0201	0,0118	0,0286	0,0573	0,0998	0,1004	0,2252	0,0109	0,0283	0,0573
31	0,0074	0,0159	0,0376	0,0029	0,0059	0,0200	0,0121	0,0289	0,0571	0,1018	0,1024	0,2259	0,0113	0,0286	0,0570
32	0,0076	0,0161	0,0373	0,0030	0,0060	0,0198	0,0125	0,0291	0,0570	0,1037	0,1042	0,2265	0,0116	0,0288	0,0568
33	0,0077	0,0162	0,0370	0,0030	0,0061	0,0197	0,0128	0,0294	0,0568	0,1055	0,1059	0,2269	0,0119	0,0290	0,0566
34	0,0079	0,0163	0,0367	0,0031	0,0062	0,0195	0,0132	0,0295	0,0566	0,1071	0,1074	0,2273	0,0122	0,0291	0,0563
35	0,0080	0,0164	0,0363	0,0032	0,0063	0,0194	0,0135	0,0297	0,0564	0,1085	0,1089	0,2275	0,0125	0,0293	0,0561
36	0,0081	0,0165	0,0360	0,0032	0,0064	0,0193	0,0138	0,0299	0,0562	0,1098	0,1101	0,2277	0,0128	0,0294	0,0559

Tabela 5.13. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 10 części, i poddanemu dyspersji miarą TF-IDF, część 1.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0,1097	0,1114	0,1873	0,0091	0,0129	0,0239	0,0811	0,0810	0,1593	0,0091	0,0129	0,0238	0,0809	0,0812	0,1599
2	0,0223	0,0225	0,0488	0,0022	0,0031	0,0061	0,0244	0,0246	0,0563	0,0037	0,0049	0,0082	0,0280	0,0280	0,0587
3	0,0131	0,0135	0,1084	0,0015	0,0031	0,0264	0,0109	0,0110	0,0875	0,0013	0,0026	0,0255	0,0107	0,0109	0,0974
4	0,0360	0,0367	0,1702	0,0027	0,0052	0,0294	0,0192	0,0198	0,1192	0,0026	0,0052	0,0294	0,0235	0,0237	0,1384
5	0,0360	0,0367	0,1702	0,0027	0,0052	0,0294	0,0192	0,0198	0,1192	0,0026	0,0052	0,0294	0,0235	0,0237	0,1384
6	0,0403	0,0411	0,1853	0,0027	0,0070	0,0357	0,0250	0,0254	0,1386	0,0022	0,0060	0,0342	0,0259	0,0262	0,1526
7	0,0210	0,0216	0,1428	0,0020	0,0046	0,0325	0,0158	0,0160	0,1154	0,0016	0,0038	0,0300	0,0155	0,0158	0,1221
8	0,0299	0,0307	0,1623	0,0025	0,0058	0,0331	0,0204	0,0207	0,1256	0,0019	0,0049	0,0317	0,0204	0,0207	0,1358
9	0,0119	0,0122	0,0993	0,0015	0,0024	0,0211	0,0085	0,0087	0,0749	0,0015	0,0023	0,0212	0,0098	0,0099	0,0872
10	0,0119	0,0122	0,0993	0,0015	0,0024	0,0211	0,0085	0,0087	0,0749	0,0015	0,0023	0,0212	0,0098	0,0099	0,0872
11	0,0908	0,0927	0,2179	0,0050	0,0104	0,0336	0,0444	0,0456	0,1611	0,0047	0,0101	0,0327	0,0520	0,0524	0,1753
12	0,0207	0,0212	0,1389	0,0019	0,0044	0,0304	0,0153	0,0155	0,1073	0,0016	0,0037	0,0295	0,0154	0,0157	0,1203
13	0,1175	0,1196	0,2341	0,0090	0,0154	0,0361	0,0743	0,0734	0,1882	0,0066	0,0122	0,0347	0,0651	0,0650	0,1943
14	0,0403	0,0411	0,1854	0,0027	0,0070	0,0357	0,0250	0,0254	0,1385	0,0022	0,0060	0,0342	0,0259	0,0262	0,1526
15	0,0763	0,0775	0,2244	0,0038	0,0108	0,0390	0,0418	0,0423	0,1703	0,0032	0,0092	0,0365	0,0433	0,0435	0,1812
16	0,1126	0,1150	0,2029	0,0103	0,0146	0,0286	0,0769	0,0761	0,1665	0,0083	0,0124	0,0294	0,0690	0,0691	0,1750
17	0,0398	0,0405	0,1850	0,0027	0,0070	0,0357	0,0248	0,0251	0,1383	0,0022	0,0058	0,0342	0,0250	0,0253	0,1520
18	0,0401	0,0409	0,1851	0,0027	0,0070	0,0357	0,0249	0,0253	0,1385	0,0023	0,0063	0,0345	0,0269	0,0275	0,1536
19	0,0398	0,0405	0,1850	0,0027	0,0070	0,0357	0,0248	0,0251	0,1383	0,0022	0,0058	0,0342	0,0250	0,0253	0,1520
20	0,1184	0,1207	0,2395	0,0084	0,0156	0,0376	0,0753	0,0754	0,1923	0,0071	0,0128	0,0349	0,0657	0,0651	0,1962
21	0,0799	0,0812	0,2266	0,0040	0,0112	0,0391	0,0436	0,0442	0,1725	0,0033	0,0095	0,0365	0,0451	0,0453	0,1831
22	0,0833	0,0847	0,2285	0,0041	0,0115	0,0392	0,0454	0,0460	0,1746	0,0034	0,0098	0,0365	0,0469	0,0471	0,1848
23	0,0866	0,0880	0,2302	0,0043	0,0119	0,0392	0,0472	0,0478	0,1764	0,0035	0,0101	0,0365	0,0486	0,0489	0,1863
24	0,0896	0,0912	0,2316	0,0044	0,0122	0,0392	0,0489	0,0495	0,1781	0,0036	0,0103	0,0365	0,0503	0,0506	0,1876
25	0,0925	0,0941	0,2328	0,0046	0,0125	0,0391	0,0507	0,0512	0,1796	0,0038	0,0106	0,0364	0,0519	0,0523	0,1888
26	0,0951	0,0969	0,2338	0,0047	0,0128	0,0391	0,0523	0,0529	0,1809	0,0039	0,0108	0,0364	0,0535	0,0539	0,1898
27	0,0976	0,0995	0,2346	0,0049	0,0130	0,0390	0,0539	0,0545	0,1821	0,0040	0,0111	0,0363	0,0551	0,0554	0,1907
28	0,0999	0,1018	0,2353	0,0051	0,0133	0,0389	0,0555	0,0561	0,1831	0,0042	0,0113	0,0362	0,0565	0,0569	0,1914
29	0,1020	0,1040	0,2358	0,0052	0,0135	0,0387	0,0570	0,0575	0,1840	0,0043	0,0115	0,0360	0,0579	0,0583	0,1921
30	0,1039	0,1060	0,2362	0,0054	0,0137	0,0386	0,0585	0,0590	0,1848	0,0044	0,0117	0,0359	0,0593	0,0597	0,1926
31	0,1057	0,1078	0,2365	0,0055	0,0139	0,0385	0,0598	0,0603	0,1855	0,0045	0,0119	0,0358	0,0605	0,0609	0,1930
32	0,1073	0,1095	0,2366	0,0057	0,0141	0,0383	0,0612	0,0617	0,1861	0,0047	0,0120	0,0357	0,0617	0,0622	0,1934
33	0,1087	0,1110	0,2367	0,0058	0,0142	0,0381	0,0624	0,0629	0,1866	0,0048	0,0122	0,0355	0,0629	0,0633	0,1936
34	0,1100	0,1124	0,2367	0,0060	0,0144	0,0380	0,0636	0,0641	0,1870	0,0049	0,0123	0,0354	0,0640	0,0644	0,1939
35	0,1112	0,1136	0,2367	0,0061	0,0145	0,0378	0,0647	0,0652	0,1874	0,0051	0,0125	0,0352	0,0650	0,0655	0,1940
36	0,1123	0,1147	0,2366	0,0063	0,0146	0,0376	0,0658	0,0662	0,1877	0,0052	0,0126	0,0351	0,0659	0,0664	0,1941

Tabela 5.14. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 10 części, i poddanemu dyspersji miarą TF-IDF, część 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
37	0,0083	0,0166	0,0357	0,0033	0,0065	0,0191	0,0141	0,0300	0,0559	0,1110	0,1113	0,2278	0,0131	0,0295	0,0556
38	0,0084	0,0166	0,0355	0,0034	0,0066	0,0190	0,0144	0,0301	0,0557	0,1121	0,1123	0,2279	0,0134	0,0296	0,0554
39	0,0085	0,0167	0,0352	0,0034	0,0066	0,0189	0,0147	0,0302	0,0555	0,1131	0,1133	0,2279	0,0137	0,0297	0,0552
40	0,0086	0,0167	0,0349	0,0035	0,0067	0,0187	0,0150	0,0303	0,0553	0,1140	0,1142	0,2279	0,0139	0,0297	0,0549
41	0,0087	0,0167	0,0346	0,0035	0,0068	0,0186	0,0152	0,0303	0,0551	0,1149	0,1150	0,2278	0,0142	0,0298	0,0547
42	0,0087	0,0167	0,0344	0,0036	0,0068	0,0185	0,0155	0,0304	0,0549	0,1156	0,1157	0,2277	0,0144	0,0298	0,0545
43	0,0088	0,0167	0,0341	0,0036	0,0069	0,0184	0,0157	0,0304	0,0547	0,1163	0,1163	0,2276	0,0146	0,0299	0,0543
44	0,0089	0,0167	0,0338	0,0037	0,0069	0,0182	0,0159	0,0304	0,0545	0,1169	0,1169	0,2274	0,0149	0,0299	0,0540
45	0,0089	0,0167	0,0336	0,0037	0,0070	0,0181	0,0161	0,0305	0,0543	0,1175	0,1174	0,2272	0,0151	0,0299	0,0538
46	0,0090	0,0167	0,0334	0,0038	0,0070	0,0180	0,0163	0,0305	0,0541	0,1180	0,1179	0,2270	0,0153	0,0299	0,0536
47	0,0091	0,0167	0,0331	0,0038	0,0070	0,0179	0,0165	0,0305	0,0539	0,1184	0,1183	0,2267	0,0155	0,0299	0,0534
48	0,0091	0,0167	0,0329	0,0039	0,0071	0,0178	0,0167	0,0305	0,0537	0,1188	0,1187	0,2265	0,0156	0,0299	0,0532
49	0,0091	0,0167	0,0327	0,0039	0,0071	0,0177	0,0169	0,0305	0,0535	0,1192	0,1191	0,2262	0,0158	0,0299	0,0530
50	0,0092	0,0167	0,0325	0,0039	0,0071	0,0176	0,0171	0,0305	0,0533	0,1195	0,1194	0,2259	0,0160	0,0299	0,0528
51	0,0107	0,0174	0,0352	0,0036	0,0061	0,0200	0,0192	0,0302	0,0540	0,1187	0,1178	0,2282	0,0170	0,0287	0,0548
52	0,0103	0,0169	0,0369	0,0030	0,0052	0,0208	0,0186	0,0298	0,0552	0,1157	0,1146	0,2303	0,0158	0,0276	0,0562
53	0,0095	0,0158	0,0385	0,0024	0,0041	0,0214	0,0176	0,0291	0,0565	0,1110	0,1095	0,2320	0,0140	0,0258	0,0576
54	0,0081	0,0139	0,0400	0,0018	0,0031	0,0217	0,0161	0,0277	0,0577	0,1035	0,1016	0,2326	0,0118	0,0230	0,0586
55	0,0064	0,0115	0,0411	0,0014	0,0024	0,0216	0,0142	0,0255	0,0588	0,0927	0,0906	0,2317	0,0093	0,0193	0,0590
56	0,0048	0,0087	0,0415	0,0012	0,0019	0,0211	0,0120	0,0226	0,0595	0,0787	0,0764	0,2285	0,0069	0,0149	0,0585
57	0,0035	0,0062	0,0411	0,0009	0,0015	0,0200	0,0096	0,0188	0,0594	0,0619	0,0597	0,2219	0,0051	0,0106	0,0568
58	0,0026	0,0044	0,0396	0,0008	0,0013	0,0186	0,0073	0,0145	0,0586	0,0442	0,0425	0,2116	0,0038	0,0075	0,0538
59	0,0020	0,0033	0,0372	0,0007	0,0011	0,0170	0,0055	0,0106	0,0567	0,0300	0,0291	0,1974	0,0030	0,0056	0,0500
60	0,0016	0,0026	0,0341	0,0006	0,0010	0,0153	0,0042	0,0077	0,0538	0,0214	0,0211	0,1796	0,0024	0,0045	0,0454
61	0,0013	0,0021	0,0306	0,0006	0,0009	0,0138	0,0033	0,0059	0,0498	0,0163	0,0162	0,1591	0,0021	0,0037	0,0407
62	0,0012	0,0018	0,0270	0,0006	0,0008	0,0123	0,0028	0,0048	0,0454	0,0130	0,0130	0,1386	0,0018	0,0032	0,0365
63	0,0010	0,0016	0,0236	0,0005	0,0008	0,0111	0,0024	0,0040	0,0410	0,0108	0,0109	0,1195	0,0017	0,0029	0,0326
64	0,0010	0,0015	0,0205	0,0005	0,0008	0,0101	0,0021	0,0035	0,0368	0,0094	0,0095	0,1036	0,0016	0,0027	0,0294
65	0,0009	0,0014	0,0180	0,0005	0,0007	0,0093	0,0019	0,0031	0,0333	0,0084	0,0085	0,0909	0,0015	0,0025	0,0268
66	0,0009	0,0013	0,0160	0,0005	0,0007	0,0086	0,0018	0,0029	0,0301	0,0078	0,0079	0,0805	0,0015	0,0024	0,0247
67	0,0008	0,0013	0,0144	0,0005	0,0007	0,0080	0,0017	0,0027	0,0275	0,0074	0,0074	0,0725	0,0014	0,0023	0,0232
68	0,0008	0,0012	0,0132	0,0005	0,0007	0,0076	0,0016	0,0026	0,0255	0,0071	0,0071	0,0667	0,0014	0,0023	0,0218
69	0,0008	0,0012	0,0122	0,0005	0,0007	0,0072	0,0016	0,0025	0,0238	0,0068	0,0069	0,0622	0,0014	0,0022	0,0207
70	0,0008	0,0012	0,0114	0,0005	0,0007	0,0069	0,0015	0,0024	0,0226	0,0067	0,0067	0,0589	0,0013	0,0022	0,0197
71	0,0057	0,0083	0,0151	0,0058	0,0086	0,0139	0,0121	0,0165	0,0265	0,0773	0,0770	0,1315	0,0120	0,0167	0,0263
72	0,0069	0,0104	0,0177	0,0058	0,0087	0,0138	0,0148	0,0205	0,0316	0,0960	0,0956	0,1577	0,0147	0,0207	0,0314

Tabela 5.15. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 10 części, i poddanemu dyspersji miarą TF-IDF, część 3.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
37	0,1133	0,1158	0,2364	0,0064	0,0147	0,0374	0,0668	0,0672	0,1879	0,0053	0,0127	0,0349	0,0668	0,0673	0,1942
38	0,1142	0,1167	0,2362	0,0065	0,0148	0,0373	0,0678	0,0681	0,1881	0,0054	0,0128	0,0348	0,0677	0,0682	0,1942
39	0,1150	0,1175	0,2359	0,0067	0,0149	0,0371	0,0687	0,0690	0,1882	0,0055	0,0129	0,0346	0,0685	0,0690	0,1941
40	0,1157	0,1182	0,2357	0,0068	0,0150	0,0369	0,0695	0,0699	0,1883	0,0056	0,0130	0,0345	0,0692	0,0698	0,1941
41	0,1164	0,1189	0,2354	0,0069	0,0151	0,0367	0,0703	0,0706	0,1884	0,0058	0,0131	0,0344	0,0699	0,0705	0,1940
42	0,1169	0,1195	0,2350	0,0070	0,0152	0,0365	0,0711	0,0714	0,1884	0,0059	0,0131	0,0342	0,0706	0,0712	0,1939
43	0,1175	0,1201	0,2347	0,0071	0,0152	0,0364	0,0718	0,0721	0,1884	0,0060	0,0132	0,0341	0,0712	0,0718	0,1937
44	0,1179	0,1205	0,2343	0,0072	0,0153	0,0362	0,0724	0,0727	0,1884	0,0061	0,0133	0,0339	0,0718	0,0724	0,1936
45	0,1184	0,1210	0,2339	0,0074	0,0153	0,0360	0,0731	0,0733	0,1883	0,0062	0,0133	0,0338	0,0723	0,0729	0,1934
46	0,1187	0,1214	0,2335	0,0075	0,0154	0,0358	0,0737	0,0739	0,1883	0,0063	0,0134	0,0336	0,0728	0,0735	0,1932
47	0,1191	0,1217	0,2331	0,0076	0,0154	0,0357	0,0742	0,0744	0,1882	0,0064	0,0135	0,0335	0,0733	0,0739	0,1930
48	0,1194	0,1220	0,2327	0,0077	0,0154	0,0355	0,0747	0,0749	0,1881	0,0065	0,0135	0,0334	0,0738	0,0744	0,1928
49	0,1197	0,1223	0,2323	0,0078	0,0154	0,0354	0,0752	0,0754	0,1880	0,0065	0,0135	0,0332	0,0742	0,0748	0,1926
50	0,1199	0,1226	0,2319	0,0078	0,0155	0,0352	0,0757	0,0758	0,1878	0,0066	0,0136	0,0331	0,0746	0,0752	0,1923
51	0,1141	0,1158	0,2371	0,0086	0,0150	0,0373	0,0710	0,0699	0,1895	0,0058	0,0113	0,0357	0,0591	0,0589	0,1953
52	0,1082	0,1094	0,2395	0,0079	0,0144	0,0385	0,0666	0,0653	0,1905	0,0049	0,0100	0,0366	0,0508	0,0505	0,1952
53	0,0984	0,0991	0,2407	0,0071	0,0134	0,0397	0,0606	0,0592	0,1908	0,0040	0,0083	0,0372	0,0406	0,0402	0,1930
54	0,0843	0,0847	0,2395	0,0061	0,0120	0,0407	0,0529	0,0513	0,1899	0,0031	0,0065	0,0372	0,0300	0,0299	0,1874
55	0,0665	0,0669	0,2348	0,0050	0,0102	0,0413	0,0436	0,0420	0,1870	0,0024	0,0049	0,0366	0,0216	0,0217	0,1785
56	0,0465	0,0472	0,2251	0,0040	0,0082	0,0414	0,0337	0,0324	0,1820	0,0019	0,0037	0,0352	0,0161	0,0163	0,1663
57	0,0300	0,0308	0,2106	0,0032	0,0063	0,0408	0,0252	0,0243	0,1741	0,0016	0,0030	0,0330	0,0126	0,0129	0,1509
58	0,0204	0,0211	0,1914	0,0025	0,0049	0,0393	0,0192	0,0187	0,1630	0,0013	0,0025	0,0305	0,0104	0,0106	0,1346
59	0,0152	0,0158	0,1691	0,0021	0,0039	0,0373	0,0151	0,0149	0,1501	0,0012	0,0021	0,0279	0,0088	0,0091	0,1185
60	0,0120	0,0125	0,1462	0,0017	0,0032	0,0345	0,0123	0,0122	0,1347	0,0010	0,0019	0,0253	0,0078	0,0081	0,1040
61	0,0099	0,0104	0,1247	0,0015	0,0027	0,0316	0,0104	0,0103	0,1198	0,0010	0,0017	0,0231	0,0072	0,0074	0,0918
62	0,0086	0,0090	0,1068	0,0013	0,0024	0,0287	0,0090	0,0090	0,1057	0,0009	0,0016	0,0209	0,0068	0,0070	0,0816
63	0,0078	0,0082	0,0922	0,0012	0,0021	0,0260	0,0081	0,0081	0,0934	0,0009	0,0016	0,0192	0,0065	0,0067	0,0740
64	0,0073	0,0076	0,0813	0,0011	0,0020	0,0237	0,0075	0,0075	0,0837	0,0009	0,0015	0,0178	0,0063	0,0065	0,0679
65	0,0069	0,0072	0,0731	0,0011	0,0019	0,0216	0,0071	0,0071	0,0754	0,0008	0,0015	0,0166	0,0062	0,0063	0,0633
66	0,0067	0,0069	0,0670	0,0010	0,0018	0,0198	0,0068	0,0069	0,0689	0,0008	0,0014	0,0157	0,0061	0,0062	0,0598
67	0,0065	0,0067	0,0626	0,0010	0,0017	0,0184	0,0066	0,0067	0,0639	0,0008	0,0014	0,0149	0,0060	0,0062	0,0573
68	0,0064	0,0065	0,0592	0,0010	0,0017	0,0173	0,0065	0,0065	0,0602	0,0008	0,0014	0,0143	0,0060	0,0061	0,0552
69	0,0063	0,0064	0,0565	0,0010	0,0016	0,0164	0,0064	0,0064	0,0574	0,0008	0,0014	0,0137	0,0059	0,0060	0,0535
70	0,0062	0,0063	0,0545	0,0009	0,0016	0,0156	0,0063	0,0063	0,0553	0,0008	0,0014	0,0132	0,0059	0,0060	0,0522
71	0,0773	0,0785	0,1317	0,0122	0,0175	0,0280	0,0995	0,0994	0,1602	0,0122	0,0175	0,0280	0,0992	0,0996	0,1608
72	0,0959	0,0974	0,1580	0,0120	0,0173	0,0279	0,1008	0,1007	0,1644	0,0120	0,0174	0,0278	0,1005	0,1009	0,1650

Tabela 5.16. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 10 części, i poddanemu dyspersji miarą TF-IDF, część 4.

Wykonanie podziału korpusu na 10 części i wykonanie dyspersji kandydatów na kolokacje za pomocą miary *TF-IDF* pogorszyło jakość wyników generowanych przez wszystkie funkcje. Efekt taki nie był spodziewany, a powodem takiej sytuacji może być przykładowo to, że teksty źródłowe w korpusie *KIPI* nie zostały pogrupowane tematycznie. Brak grupowania może zaowocować rozrzuceniem zwrotów i sformułowań z danej dziedziny tematycznej na przestrzeń całego korpusu, a to mogłoby tłumaczyć zaistniały problem pogorszenia się wyników. Strona internetowa [?] nie zawierała informacji o zawartości tematycznej tekstów składowych korpusu (jedynie jego próbki) ani informacji o ich zgrupowaniu lub jego braku. Autor niniejszej pracy nie znalazł także takich informacji w publikacji [?].

Istnieje też możliwość wystąpienia innego problemu, a mianowicie zbyt mała ziarnistość podziału korpusu na części, która została wybrana arbitralnie. Na potrzeby tego badania korpus *KIPI* został podzielony na 10 części, czyli każda z nich zawierała około 10% tekstu z korpusu. Źródła [?] oraz [?] nie podają informacji o zawartości tematycznej artykułów ani o procentowej zawartości każdego działu tematycznego w pełnym korpusie *KIPI*. Takie informacje dostępne są jedynie dla próbki korpusu, a znajdują się na [?].

Wyniki badań miar dwuelementowych na korpusie *KIPI* podzielonym na 20 fragmentów i poddanym dyspersji

Autor niniejszej pracy postanowił sprawdzić jak zmieniają się wyniki po zwiększeniu ziarnistości podziału do 20, czyli podziału na 20 części, gdzie każda z nich będzie zawierać 5% z całego korpusu. Procent ten został wybrany po zasugerowaniu się danymi dotyczącymi próbki z korpusu *KIPI* [?]. Zauważyć można, że najmniejszy dział tematyczny stanowi właśnie około 5% całego korpusu *KIPI*. Jednak zaznaczyć trzeba, że autor niniejszej pracy zdaje sobie sprawę, że próbka nie odzwierciedla statystyk całego korpusu, ale taka informacja wydaje się lepsza niż brak jakiejkolwiek informacji. Innymi słowy ze względu na brak istotnych informacji o składzie korpusu *KIPI* podział na 10 czy 20 części jest tak samo dobry jak na każdą inną ich liczbę.

Cztery tabele ??, ??, ?? oraz ?? prezentują jakość wyników osiągniętych przez 72 funkcje w 30 różnych badaniach – 30 zestawów danych pozyskanych z korpusu *KIPI*. Różnica pomiędzy tym a poprzednim badaniem jest taka, że w tym przypadku korpus *KIPI* został podzielony na 20 części. Wykorzystana została natomiast ta sama miara dyspersji – *TF-IDF*. Indeksy miar i typów badań pozostały takie same jak w poprzednich badaniach.

Dwukrotne zwiększenie ziarnistości spowodowało poprawę jakości generowanych rozwiązań w stosunku do wyników z poprzedniego badania w każdym przypadku. Wzrost jakości sięgnął nawet około 50% w stosunku do poprzedniego badania, ale taki wzrost nie spowodował nawet wyrównania wyników osiągniętych przy badaniu korpusu bez podziału i dyspersji. Wykonanie dyspersji tą metodą spowodowało osiągnięcie wyników na

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0098	0,0158	0,0235	0,0056	0,0083	0,0133	0,0236	0,0323	0,0469	0,1409	0,1405	0,2139	0,0236	0,0328	0,0468
2	0,0014	0,0020	0,0035	0,0015	0,0020	0,0032	0,0037	0,0050	0,0114	0,0243	0,0244	0,0546	0,0046	0,0060	0,0124
3	0,0016	0,0026	0,0288	0,0008	0,0013	0,0161	0,0027	0,0052	0,0392	0,0131	0,0132	0,1036	0,0025	0,0051	0,0416
4	0,0039	0,0065	0,0341	0,0017	0,0025	0,0167	0,0067	0,0137	0,0515	0,0351	0,0361	0,1562	0,0068	0,0153	0,0549
5	0,0039	0,0065	0,0341	0,0017	0,0025	0,0167	0,0067	0,0137	0,0516	0,0351	0,0361	0,1562	0,0068	0,0153	0,0549
6	0,0038	0,0087	0,0425	0,0016	0,0033	0,0220	0,0058	0,0162	0,0563	0,0418	0,0426	0,1719	0,0056	0,0171	0,0593
7	0,0023	0,0045	0,0387	0,0011	0,0019	0,0203	0,0039	0,0090	0,0493	0,0225	0,0228	0,1415	0,0036	0,0088	0,0507
8	0,0032	0,0064	0,0385	0,0014	0,0025	0,0201	0,0054	0,0127	0,0520	0,0319	0,0324	0,1537	0,0047	0,0126	0,0544
9	0,0016	0,0020	0,0222	0,0008	0,0010	0,0116	0,0031	0,0045	0,0357	0,0112	0,0114	0,0957	0,0031	0,0047	0,0381
10	0,0016	0,0020	0,0222	0,0008	0,0010	0,0116	0,0031	0,0045	0,0357	0,0112	0,0114	0,0957	0,0031	0,0047	0,0381
11	0,0087	0,0162	0,0380	0,0035	0,0059	0,0192	0,0146	0,0320	0,0609	0,1049	0,1067	0,2209	0,0145	0,0334	0,0620
12	0,0023	0,0043	0,0352	0,0011	0,0019	0,0189	0,0038	0,0084	0,0466	0,0210	0,0213	0,1292	0,0036	0,0086	0,0498
13	0,0152	0,0243	0,0410	0,0056	0,0093	0,0230	0,0263	0,0397	0,0622	0,1564	0,1554	0,2575	0,0238	0,0382	0,0628
14	0,0038	0,0087	0,0425	0,0016	0,0033	0,0220	0,0058	0,0162	0,0563	0,0418	0,0426	0,1718	0,0056	0,0171	0,0593
15	0,0064	0,0162	0,0453	0,0025	0,0056	0,0236	0,0095	0,0287	0,0631	0,0858	0,0870	0,2193	0,0089	0,0294	0,0645
16	0,0142	0,0215	0,0311	0,0069	0,0100	0,0192	0,0262	0,0355	0,0513	0,1464	0,1463	0,2263	0,0255	0,0355	0,0531
17	0,0038	0,0086	0,0425	0,0016	0,0031	0,0220	0,0058	0,0161	0,0562	0,0415	0,0422	0,1715	0,0055	0,0167	0,0592
18	0,0038	0,0086	0,0425	0,0016	0,0032	0,0220	0,0058	0,0161	0,0562	0,0416	0,0423	0,1716	0,0055	0,0169	0,0593
19	0,0038	0,0086	0,0425	0,0016	0,0031	0,0220	0,0058	0,0161	0,0562	0,0415	0,0422	0,1715	0,0055	0,0167	0,0592
20	0,0133	0,0228	0,0425	0,0042	0,0072	0,0217	0,0250	0,0400	0,0641	0,1553	0,1543	0,2612	0,0224	0,0381	0,0644
21	0,0067	0,0169	0,0453	0,0026	0,0059	0,0236	0,0100	0,0298	0,0634	0,0908	0,0920	0,2229	0,0094	0,0305	0,0647
22	0,0070	0,0176	0,0453	0,0028	0,0061	0,0237	0,0105	0,0309	0,0637	0,0957	0,0969	0,2263	0,0098	0,0314	0,0649
23	0,0073	0,0182	0,0452	0,0029	0,0064	0,0237	0,0110	0,0319	0,0640	0,1005	0,1016	0,2294	0,0103	0,0324	0,0650
24	0,0076	0,0189	0,0451	0,0030	0,0066	0,0237	0,0115	0,0328	0,0642	0,1051	0,1062	0,2323	0,0107	0,0332	0,0651
25	0,0079	0,0194	0,0450	0,0031	0,0069	0,0236	0,0120	0,0337	0,0644	0,1095	0,1105	0,2350	0,0112	0,0340	0,0652
26	0,0083	0,0199	0,0448	0,0032	0,0071	0,0236	0,0126	0,0344	0,0645	0,1137	0,1147	0,2374	0,0117	0,0347	0,0652
27	0,0086	0,0204	0,0447	0,0034	0,0074	0,0236	0,0131	0,0351	0,0646	0,1176	0,1186	0,2396	0,0122	0,0353	0,0652
28	0,0089	0,0209	0,0445	0,0035	0,0076	0,0235	0,0137	0,0358	0,0646	0,1213	0,1222	0,2415	0,0127	0,0359	0,0652
29	0,0092	0,0212	0,0443	0,0036	0,0078	0,0234	0,0142	0,0364	0,0646	0,1248	0,1256	0,2433	0,0132	0,0364	0,0651
30	0,0095	0,0216	0,0441	0,0037	0,0080	0,0234	0,0147	0,0369	0,0646	0,1280	0,1288	0,2449	0,0137	0,0368	0,0650
31	0,0097	0,0219	0,0438	0,0038	0,0082	0,0233	0,0153	0,0374	0,0646	0,1310	0,1317	0,2464	0,0142	0,0372	0,0649
32	0,0100	0,0222	0,0436	0,0040	0,0084	0,0232	0,0158	0,0378	0,0646	0,1338	0,1344	0,2476	0,0147	0,0376	0,0648
33	0,0103	0,0224	0,0433	0,0041	0,0086	0,0231	0,0163	0,0381	0,0645	0,1363	0,1368	0,2488	0,0152	0,0379	0,0647
34	0,0105	0,0226	0,0431	0,0042	0,0087	0,0230	0,0168	0,0385	0,0644	0,1386	0,1391	0,2498	0,0157	0,0382	0,0646
35	0,0108	0,0228	0,0429	0,0043	0,0089	0,0229	0,0173	0,0387	0,0643	0,1407	0,1411	0,2506	0,0161	0,0384	0,0645
36	0,0110	0,0229	0,0426	0,0044	0,0090	0,0228	0,0178	0,0390	0,0642	0,1426	0,1430	0,2514	0,0166	0,0387	0,0643

Tabela 5.17. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 20 części, i poddanemu dyspersji miarą TF-IDF, część 1.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0,1429	0,1421	0,2136	0,0123	0,0173	0,0286	0,1048	0,1051	0,1846	0,0123	0,0172	0,0286	0,1055	0,1044	0,1839
2	0,0266	0,0266	0,0562	0,0025	0,0036	0,0073	0,0276	0,0280	0,0634	0,0043	0,0056	0,0099	0,0318	0,0317	0,0665
3	0,0134	0,0138	0,1163	0,0015	0,0031	0,0271	0,0111	0,0111	0,0943	0,0013	0,0027	0,0261	0,0110	0,0112	0,1044
4	0,0428	0,0426	0,1775	0,0031	0,0061	0,0311	0,0223	0,0231	0,1261	0,0029	0,0060	0,0311	0,0278	0,0277	0,1458
5	0,0428	0,0426	0,1775	0,0031	0,0061	0,0311	0,0223	0,0231	0,1261	0,0029	0,0060	0,0311	0,0278	0,0277	0,1458
6	0,0474	0,0472	0,1935	0,0030	0,0082	0,0378	0,0288	0,0292	0,1471	0,0025	0,0070	0,0362	0,0305	0,0304	0,1609
7	0,0229	0,0234	0,1511	0,0021	0,0049	0,0342	0,0170	0,0172	0,1240	0,0017	0,0040	0,0312	0,0169	0,0171	0,1296
8	0,0336	0,0339	0,1681	0,0027	0,0065	0,0348	0,0226	0,0229	0,1318	0,0021	0,0054	0,0331	0,0230	0,0231	0,1420
9	0,0125	0,0127	0,1074	0,0016	0,0025	0,0222	0,0089	0,0091	0,0817	0,0015	0,0024	0,0221	0,0102	0,0103	0,0942
10	0,0125	0,0127	0,1074	0,0016	0,0025	0,0222	0,0089	0,0091	0,0817	0,0015	0,0024	0,0221	0,0102	0,0103	0,0942
11	0,1194	0,1182	0,2359	0,0063	0,0140	0,0376	0,0590	0,0607	0,1755	0,0059	0,0134	0,0368	0,0702	0,0692	0,1915
12	0,0226	0,0230	0,1471	0,0021	0,0047	0,0316	0,0164	0,0166	0,1149	0,0017	0,0040	0,0306	0,0168	0,0170	0,1277
13	0,1545	0,1529	0,2646	0,0127	0,0211	0,0430	0,0993	0,0984	0,2176	0,0090	0,0162	0,0407	0,0873	0,0852	0,2208
14	0,0474	0,0472	0,1935	0,0030	0,0082	0,0378	0,0288	0,0292	0,1471	0,0025	0,0070	0,0362	0,0305	0,0304	0,1609
15	0,0964	0,0949	0,2376	0,0046	0,0138	0,0427	0,0522	0,0531	0,1828	0,0038	0,0116	0,0400	0,0555	0,0546	0,1938
16	0,1487	0,1479	0,2345	0,0143	0,0200	0,0349	0,1011	0,1007	0,1953	0,0113	0,0165	0,0355	0,0919	0,0902	0,2037
17	0,0466	0,0463	0,1930	0,0030	0,0080	0,0377	0,0284	0,0288	0,1468	0,0024	0,0066	0,0361	0,0291	0,0289	0,1601
18	0,0470	0,0469	0,1933	0,0030	0,0081	0,0378	0,0287	0,0292	0,1471	0,0026	0,0073	0,0365	0,0318	0,0320	0,1624
19	0,0465	0,0463	0,1930	0,0030	0,0080	0,0377	0,0284	0,0288	0,1468	0,0024	0,0066	0,0361	0,0291	0,0289	0,1601
20	0,1534	0,1513	0,2688	0,0116	0,0208	0,0441	0,1004	0,1005	0,2212	0,0091	0,0152	0,0402	0,0781	0,0749	0,2177
21	0,1015	0,1000	0,2406	0,0048	0,0143	0,0429	0,0550	0,0559	0,1857	0,0039	0,0121	0,0402	0,0583	0,0573	0,1964
22	0,1064	0,1048	0,2434	0,0050	0,0149	0,0432	0,0577	0,0586	0,1885	0,0041	0,0126	0,0404	0,0611	0,0600	0,1987
23	0,1111	0,1095	0,2459	0,0052	0,0154	0,0433	0,0604	0,0614	0,1910	0,0043	0,0130	0,0405	0,0638	0,0627	0,2009
24	0,1156	0,1140	0,2482	0,0054	0,0159	0,0435	0,0631	0,0641	0,1934	0,0045	0,0134	0,0406	0,0665	0,0653	0,2029
25	0,1198	0,1182	0,2503	0,0057	0,0164	0,0436	0,0657	0,0667	0,1956	0,0046	0,0138	0,0407	0,0690	0,0678	0,2048
26	0,1237	0,1221	0,2521	0,0059	0,0169	0,0437	0,0683	0,0693	0,1977	0,0048	0,0142	0,0408	0,0715	0,0703	0,2065
27	0,1273	0,1258	0,2538	0,0061	0,0173	0,0438	0,0708	0,0718	0,1995	0,0050	0,0146	0,0408	0,0739	0,0727	0,2080
28	0,1307	0,1292	0,2552	0,0064	0,0177	0,0438	0,0732	0,0742	0,2013	0,0052	0,0150	0,0408	0,0762	0,0749	0,2094
29	0,1338	0,1323	0,2565	0,0066	0,0181	0,0438	0,0755	0,0765	0,2028	0,0054	0,0153	0,0408	0,0784	0,0771	0,2106
30	0,1366	0,1352	0,2576	0,0069	0,0185	0,0438	0,0778	0,0787	0,2043	0,0056	0,0156	0,0408	0,0805	0,0792	0,2118
31	0,1392	0,1379	0,2586	0,0071	0,0188	0,0437	0,0799	0,0808	0,2056	0,0058	0,0159	0,0408	0,0824	0,0811	0,2128
32	0,1415	0,1403	0,2594	0,0073	0,0191	0,0437	0,0819	0,0828	0,2067	0,0060	0,0162	0,0407	0,0843	0,0830	0,2137
33	0,1437	0,1425	0,2601	0,0076	0,0194	0,0436	0,0838	0,0847	0,2078	0,0062	0,0164	0,0407	0,0860	0,0847	0,2145
34	0,1456	0,1444	0,2607	0,0078	0,0196	0,0436	0,0856	0,0865	0,2088	0,0064	0,0167	0,0406	0,0876	0,0863	0,2152
35	0,1473	0,1463	0,2612	0,0081	0,0198	0,0435	0,0873	0,0882	0,2096	0,0066	0,0169	0,0406	0,0892	0,0879	0,2158
36	0,1489	0,1479	0,2616	0,0083	0,0201	0,0434	0,0889	0,0897	0,2104	0,0068	0,0171	0,0405	0,0906	0,0893	0,2163

Tabela 5.18. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 20 części, i poddanemu dyspersji miarą TF-IDF, część 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
37	0,0112	0,0230	0,0423	0,0045	0,0092	0,0227	0,0183	0,0392	0,0641	0,1444	0,1447	0,2521	0,0170	0,0388	0,0642
38	0,0114	0,0231	0,0421	0,0046	0,0093	0,0226	0,0187	0,0394	0,0640	0,1460	0,1462	0,2527	0,0175	0,0390	0,0640
39	0,0116	0,0232	0,0418	0,0047	0,0094	0,0225	0,0192	0,0396	0,0639	0,1474	0,1476	0,2532	0,0179	0,0391	0,0638
40	0,0117	0,0233	0,0416	0,0048	0,0095	0,0224	0,0196	0,0397	0,0637	0,1487	0,1488	0,2536	0,0183	0,0393	0,0637
41	0,0119	0,0233	0,0413	0,0049	0,0096	0,0223	0,0200	0,0398	0,0636	0,1499	0,1500	0,2540	0,0187	0,0394	0,0635
42	0,0120	0,0234	0,0411	0,0050	0,0097	0,0222	0,0204	0,0399	0,0634	0,1509	0,1510	0,2543	0,0190	0,0394	0,0633
43	0,0122	0,0234	0,0408	0,0051	0,0098	0,0220	0,0207	0,0400	0,0633	0,1519	0,1519	0,2545	0,0194	0,0395	0,0632
44	0,0123	0,0234	0,0406	0,0052	0,0099	0,0219	0,0211	0,0401	0,0631	0,1528	0,1528	0,2547	0,0198	0,0396	0,0630
45	0,0124	0,0234	0,0404	0,0053	0,0099	0,0218	0,0214	0,0401	0,0630	0,1536	0,1535	0,2548	0,0201	0,0396	0,0628
46	0,0125	0,0234	0,0401	0,0053	0,0100	0,0217	0,0218	0,0401	0,0628	0,1543	0,1542	0,2549	0,0204	0,0397	0,0627
47	0,0126	0,0234	0,0399	0,0054	0,0100	0,0216	0,0221	0,0402	0,0627	0,1549	0,1548	0,2550	0,0207	0,0397	0,0625
48	0,0127	0,0234	0,0397	0,0055	0,0101	0,0215	0,0223	0,0402	0,0625	0,1555	0,1553	0,2550	0,0210	0,0397	0,0623
49	0,0128	0,0233	0,0394	0,0055	0,0101	0,0214	0,0226	0,0402	0,0624	0,1560	0,1558	0,2550	0,0213	0,0397	0,0621
50	0,0128	0,0233	0,0392	0,0056	0,0102	0,0213	0,0229	0,0402	0,0622	0,1565	0,1563	0,2549	0,0215	0,0397	0,0620
51	0,0149	0,0239	0,0424	0,0048	0,0081	0,0237	0,0256	0,0394	0,0632	0,1533	0,1520	0,2588	0,0224	0,0371	0,0639
52	0,0140	0,0228	0,0439	0,0038	0,0067	0,0242	0,0244	0,0385	0,0643	0,1482	0,1465	0,2596	0,0203	0,0351	0,0650
53	0,0123	0,0207	0,0452	0,0028	0,0050	0,0244	0,0227	0,0370	0,0652	0,1403	0,1381	0,2594	0,0176	0,0321	0,0658
54	0,0101	0,0176	0,0461	0,0021	0,0035	0,0242	0,0204	0,0346	0,0660	0,1285	0,1260	0,2575	0,0143	0,0278	0,0661
55	0,0076	0,0138	0,0464	0,0015	0,0025	0,0235	0,0176	0,0312	0,0665	0,1128	0,1100	0,2536	0,0108	0,0225	0,0655
56	0,0054	0,0100	0,0458	0,0012	0,0019	0,0224	0,0143	0,0268	0,0663	0,0931	0,0903	0,2464	0,0076	0,0165	0,0636
57	0,0036	0,0065	0,0442	0,0009	0,0015	0,0210	0,0109	0,0215	0,0651	0,0705	0,0677	0,2353	0,0053	0,0111	0,0606
58	0,0025	0,0043	0,0418	0,0008	0,0012	0,0193	0,0079	0,0158	0,0630	0,0477	0,0456	0,2202	0,0038	0,0074	0,0566
59	0,0019	0,0031	0,0388	0,0007	0,0010	0,0176	0,0056	0,0108	0,0600	0,0304	0,0294	0,2024	0,0029	0,0053	0,0517
60	0,0015	0,0024	0,0352	0,0006	0,0009	0,0161	0,0042	0,0075	0,0560	0,0208	0,0204	0,1814	0,0023	0,0042	0,0469
61	0,0012	0,0020	0,0316	0,0006	0,0008	0,0148	0,0032	0,0056	0,0514	0,0155	0,0154	0,1597	0,0020	0,0035	0,0424
62	0,0011	0,0017	0,0283	0,0005	0,0008	0,0136	0,0026	0,0044	0,0466	0,0122	0,0122	0,1397	0,0018	0,0030	0,0386
63	0,0010	0,0015	0,0255	0,0005	0,0007	0,0126	0,0022	0,0037	0,0424	0,0102	0,0102	0,1229	0,0016	0,0027	0,0352
64	0,0009	0,0013	0,0230	0,0005	0,0007	0,0118	0,0020	0,0032	0,0387	0,0088	0,0089	0,1092	0,0015	0,0025	0,0325
65	0,0009	0,0013	0,0210	0,0005	0,0007	0,0111	0,0018	0,0028	0,0356	0,0079	0,0079	0,0982	0,0014	0,0023	0,0304
66	0,0008	0,0012	0,0193	0,0005	0,0007	0,0105	0,0017	0,0026	0,0329	0,0073	0,0073	0,0893	0,0014	0,0022	0,0288
67	0,0008	0,0011	0,0180	0,0005	0,0007	0,0100	0,0016	0,0024	0,0307	0,0069	0,0069	0,0823	0,0013	0,0021	0,0275
68	0,0008	0,0011	0,0169	0,0005	0,0007	0,0096	0,0015	0,0023	0,0291	0,0066	0,0066	0,0774	0,0013	0,0021	0,0264
69	0,0008	0,0011	0,0160	0,0005	0,0007	0,0092	0,0015	0,0022	0,0278	0,0064	0,0064	0,0735	0,0013	0,0020	0,0255
70	0,0007	0,0011	0,0152	0,0005	0,0007	0,0089	0,0014	0,0022	0,0268	0,0063	0,0063	0,0707	0,0013	0,0020	0,0246
71	0,0061	0,0087	0,0150	0,0068	0,0100	0,0147	0,0129	0,0174	0,0267	0,0786	0,0784	0,1281	0,0129	0,0176	0,0267
72	0,0083	0,0123	0,0188	0,0071	0,0106	0,0154	0,0173	0,0234	0,0333	0,1075	0,1073	0,1624	0,0173	0,0237	0,0333

Tabela 5.19. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 20 części, i poddanemu dyspersji miarą TF-IDF, część 3.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
37	0,1504	0,1494	0,2619	0,0085	0,0202	0,0433	0,0904	0,0912	0,2111	0,0070	0,0173	0,0404	0,0919	0,0907	0,2168
38	0,1516	0,1507	0,2621	0,0087	0,0204	0,0432	0,0918	0,0926	0,2118	0,0072	0,0174	0,0403	0,0932	0,0919	0,2172
39	0,1528	0,1519	0,2623	0,0090	0,0206	0,0430	0,0931	0,0939	0,2123	0,0074	0,0176	0,0402	0,0943	0,0931	0,2176
40	0,1538	0,1530	0,2624	0,0092	0,0207	0,0429	0,0944	0,0951	0,2128	0,0076	0,0178	0,0401	0,0954	0,0942	0,2179
41	0,1548	0,1540	0,2625	0,0094	0,0208	0,0428	0,0955	0,0962	0,2132	0,0078	0,0179	0,0400	0,0964	0,0952	0,2181
42	0,1556	0,1548	0,2625	0,0096	0,0209	0,0427	0,0966	0,0973	0,2136	0,0080	0,0180	0,0399	0,0973	0,0962	0,2183
43	0,1564	0,1556	0,2624	0,0098	0,0210	0,0425	0,0976	0,0982	0,2139	0,0081	0,0181	0,0398	0,0982	0,0970	0,2184
44	0,1570	0,1563	0,2624	0,0100	0,0211	0,0424	0,0985	0,0992	0,2142	0,0083	0,0182	0,0397	0,0990	0,0979	0,2186
45	0,1576	0,1569	0,2622	0,0102	0,0212	0,0423	0,0994	0,1000	0,2144	0,0085	0,0183	0,0396	0,0998	0,0986	0,2186
46	0,1582	0,1575	0,2621	0,0103	0,0213	0,0421	0,1002	0,1008	0,2146	0,0086	0,0184	0,0395	0,1005	0,0993	0,2187
47	0,1587	0,1580	0,2620	0,0105	0,0213	0,0420	0,1009	0,1015	0,2148	0,0088	0,0185	0,0394	0,1011	0,1000	0,2187
48	0,1591	0,1584	0,2618	0,0107	0,0214	0,0418	0,1016	0,1022	0,2149	0,0089	0,0185	0,0393	0,1017	0,1006	0,2187
49	0,1595	0,1588	0,2615	0,0108	0,0214	0,0417	0,1023	0,1028	0,2150	0,0091	0,0186	0,0392	0,1023	0,1012	0,2187
50	0,1598	0,1592	0,2613	0,0110	0,0214	0,0416	0,1029	0,1034	0,2150	0,0092	0,0187	0,0391	0,1028	0,1018	0,2186
51	0,1485	0,1464	0,2661	0,0119	0,0204	0,0440	0,0943	0,0931	0,2181	0,0078	0,0147	0,0415	0,0781	0,0759	0,2203
52	0,1386	0,1360	0,2665	0,0108	0,0192	0,0451	0,0875	0,0860	0,2178	0,0063	0,0126	0,0420	0,0656	0,0635	0,2179
53	0,1236	0,1206	0,2648	0,0095	0,0176	0,0460	0,0785	0,0768	0,2163	0,0048	0,0101	0,0420	0,0505	0,0487	0,2124
54	0,1029	0,1001	0,2598	0,0079	0,0154	0,0465	0,0671	0,0653	0,2131	0,0035	0,0074	0,0413	0,0350	0,0339	0,2028
55	0,0783	0,0762	0,2502	0,0062	0,0127	0,0465	0,0537	0,0518	0,2073	0,0026	0,0052	0,0398	0,0233	0,0230	0,1899
56	0,0521	0,0511	0,2353	0,0047	0,0096	0,0458	0,0395	0,0378	0,1985	0,0020	0,0038	0,0375	0,0166	0,0166	0,1742
57	0,0314	0,0315	0,2163	0,0035	0,0069	0,0442	0,0274	0,0264	0,1867	0,0016	0,0029	0,0346	0,0126	0,0128	0,1566
58	0,0202	0,0206	0,1940	0,0026	0,0050	0,0419	0,0198	0,0193	0,1726	0,0013	0,0024	0,0316	0,0102	0,0105	0,1390
59	0,0147	0,0151	0,1701	0,0021	0,0038	0,0389	0,0151	0,0149	0,1565	0,0011	0,0021	0,0287	0,0086	0,0088	0,1227
60	0,0114	0,0119	0,1478	0,0017	0,0031	0,0357	0,0120	0,0119	0,1396	0,0010	0,0018	0,0262	0,0076	0,0078	0,1093
61	0,0094	0,0098	0,1287	0,0014	0,0026	0,0324	0,0100	0,0100	0,1240	0,0009	0,0017	0,0240	0,0069	0,0072	0,0983
62	0,0081	0,0085	0,1133	0,0013	0,0022	0,0294	0,0087	0,0087	0,1102	0,0009	0,0016	0,0222	0,0065	0,0067	0,0894
63	0,0074	0,0077	0,1008	0,0012	0,0020	0,0270	0,0077	0,0078	0,0994	0,0009	0,0015	0,0208	0,0063	0,0065	0,0826
64	0,0069	0,0071	0,0913	0,0011	0,0018	0,0249	0,0071	0,0072	0,0906	0,0008	0,0014	0,0196	0,0061	0,0063	0,0773
65	0,0065	0,0068	0,0843	0,0010	0,0017	0,0231	0,0068	0,0068	0,0832	0,0008	0,0014	0,0187	0,0060	0,0061	0,0731
66	0,0063	0,0065	0,0790	0,0010	0,0017	0,0217	0,0065	0,0065	0,0773	0,0008	0,0014	0,0180	0,0059	0,0060	0,0701
67	0,0062	0,0063	0,0750	0,0009	0,0016	0,0206	0,0063	0,0064	0,0730	0,0008	0,0013	0,0174	0,0058	0,0060	0,0679
68	0,0060	0,0062	0,0721	0,0009	0,0016	0,0196	0,0062	0,0062	0,0696	0,0008	0,0013	0,0168	0,0058	0,0059	0,0662
69	0,0060	0,0061	0,0699	0,0009	0,0015	0,0189	0,0061	0,0061	0,0671	0,0008	0,0013	0,0164	0,0058	0,0059	0,0648
70	0,0059	0,0060	0,0681	0,0009	0,0015	0,0183	0,0061	0,0061	0,0653	0,0008	0,0013	0,0160	0,0057	0,0058	0,0636
71	0,0799	0,0794	0,1279	0,0157	0,0226	0,0319	0,1190	0,1194	0,1707	0,0157	0,0225	0,0320	0,1199	0,1186	0,1700
72	0,1091	0,1084	0,1622	0,0157	0,0226	0,0325	0,1218	0,1222	0,1799	0,0157	0,0225	0,0326	0,1227	0,1214	0,1792

Tabela 5.20. Wyniki badań miar dwuelementowych dla korpusu *KIPI* podzielonemu na 20 części, i poddanemu dyspersji miarą TF-IDF, część 4.

poziomie od około 30% do 60% jakości wyników badań przeprowadzonych na korpusie *KIPI* bez dyspersji.

Najlepsze jakościowo wyniki zostały osiągnięte dla następujących miar asocjacyjnych: *W Specific Correlation*, *Specific Frequency Biased Mutual Dependency*, *T-Score*, *Loglikelihood*, *W Order*, *W Term Frequency Order* oraz zestaw miar *Specific Exponential Correlation* i *W Specific Exponential Correlation* dla pewnych wartości ich parametru. Opisany zestaw funkcji uległ zmianie w stosunku do poprzedniego badania wpływu dyspersji na jakość wyników, ale w konkretny sposób, ponieważ nie pojawiła się żadna nowa funkcja w tym zestawie, a został on jedynie okrojony. Następującą obserwacją jest to, że może dobre wyniki pewnych funkcji w poprzednim badaniu były spowodowane jedynie ogólnie wynikami niskiej jakości. Potwierdzeniem poprawności takiej obserwacji może być przykład miary *Sorgenfrei*, która okazała się być jedną z najlepszych w poprzednim badaniu, a mianowicie jej wynik w poprzednim badaniu jest zbliżony do wyniku w tym badaniu, natomiast przykładowo wynik funkcji *Loglikelihood* jest widocznie lepszy. Innymi słowy pojawienie się miary *Sorgenfrei* w gronie najlepszych mogło być spowodowane ogólnym obniżeniem jakości wszystkich funkcji z wyjątkiem właśnie tej, a tym samym było jej łatwiej osiągnąć wynik pozwalający na wpisanie jej do czołówki miar.

Niniejsze i poprzednie badanie pokazują, że taka metoda dyspersji nie spisała się w przypadku tutaj badanego korpusu *KIPI*. Nie można jednak jednoznacznie stwierdzić, że winna jest sama metoda. Powodem takiej niemożności są informacje zamieszczone przez autora niniejszej pracy przy omawianiu poprzedniego badania, dyspersji korpusu podzielonego na 10 części, a mianowicie brak informacji o grupowaniu tematycznym tekstów składowych korpusu *KIPI*.

Wyniki badań jakości wyników dla miar asocjacyjnych po podpróbkowaniu klasy negatywnej do 80%

Różnica pomiędzy tym, a poprzednimi badaniami jest taka, że w tym przypadku krotki z korpusu zostały poddane podpróbkowaniu klasy negatywnej. Proces ten został wykonany dla każdego z 6 zestawów danych opisanych w poprzedniej części pracy – *2R*, *2W*, *2RIH*, *2WIH*, *2RW* oraz *2RWIH*. Składy krotek zostały najpierw poddane ewentualnemu filtrowaniu, a dopiero potem podpróbkowaniu w celu zapewnienia pożądanego stosunku liczby wyrażen wielowyrazowych do wszystkich kandydatów. Pożądany stosunek wyrażen wielowyrazowych do wszystkich kandydatów został ustalony na poziomie 20%, a idea dobrania takiej ich liczby pochodzi z artykułu [?]. Tak przygotowane zestawy krotek zostały wykorzystane w tej części badań.

Cztery tabele ??, ??, ?? oraz ?? prezentują jakość wyników osiągniętych przez 72 funkcje w 30 różnych badaniach – 30 zestawów danych pozyskanych z korpusu *KIPI*. Indeksy miar i typów badań pozostały takie same jak w poprzednich badaniach.

Zgodnie z oczekiwaniami jakość wyników wzrosła i to znacząco, ponieważ kilkukrot-

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,7085	0,7200	0,4872	0,6583	0,6700	0,4657	0,7588	0,7432	0,5387	0,9410	0,9218	0,8379	0,8312	0,8243	0,6595
2	0,8174	0,8035	0,4562	0,8428	0,8389	0,5147	0,7262	0,6764	0,3780	0,8473	0,8043	0,6263	0,8218	0,7971	0,5405
3	0,0625	0,0601	0,1541	0,0541	0,0547	0,1465	0,0613	0,0655	0,1383	0,2473	0,2040	0,3796	0,0907	0,0992	0,1980
4	0,1993	0,2128	0,3357	0,1624	0,1764	0,3092	0,1874	0,2258	0,2952	0,5560	0,5080	0,6511	0,2604	0,3145	0,3925
5	0,1995	0,2128	0,3358	0,1624	0,1764	0,3090	0,1873	0,2259	0,2957	0,5556	0,5080	0,6513	0,2606	0,3145	0,3923
6	0,1870	0,2079	0,3305	0,1468	0,1681	0,3018	0,1759	0,2201	0,2887	0,5393	0,4948	0,6437	0,2287	0,2893	0,3615
7	0,1207	0,1250	0,2409	0,0910	0,0975	0,2201	0,1218	0,1390	0,2186	0,3992	0,3494	0,5334	0,1326	0,1561	0,2392
8	0,1769	0,1808	0,2765	0,1313	0,1385	0,2491	0,1826	0,2005	0,2450	0,5175	0,4659	0,5793	0,1984	0,2273	0,2876
9	0,0887	0,0817	0,1919	0,0751	0,0722	0,1786	0,0836	0,0862	0,1679	0,3037	0,2521	0,4380	0,1188	0,1245	0,2331
10	0,0888	0,0816	0,1917	0,0751	0,0722	0,1785	0,0836	0,0863	0,1682	0,3037	0,2522	0,4380	0,1188	0,1245	0,2331
11	0,4833	0,4908	0,4496	0,3933	0,4058	0,4190	0,5416	0,5548	0,4587	0,8656	0,8364	0,8116	0,6242	0,6491	0,5675
12	0,0896	0,0902	0,2157	0,0750	0,0792	0,2022	0,0840	0,0950	0,1848	0,3174	0,2719	0,4749	0,1214	0,1412	0,2534
13	0,5738	0,5841	0,4586	0,4659	0,4732	0,4170	0,6570	0,6518	0,5091	0,9073	0,8825	0,8399	0,7150	0,7165	0,6123
14	0,1869	0,2080	0,3304	0,1468	0,1681	0,3018	0,1760	0,2203	0,2888	0,5394	0,4948	0,6438	0,2287	0,2894	0,3614
15	0,4143	0,4389	0,4297	0,3113	0,3409	0,3957	0,4449	0,4806	0,4195	0,8168	0,7858	0,7865	0,4942	0,5542	0,5005
16	0,6358	0,6454	0,4630	0,5293	0,5335	0,4154	0,7109	0,6988	0,5283	0,9226	0,8995	0,8382	0,7668	0,7613	0,6403
17	0,1843	0,2054	0,3268	0,1448	0,1654	0,2975	0,1739	0,2183	0,2862	0,5364	0,4927	0,6402	0,2274	0,2874	0,3601
18	0,1865	0,2073	0,3290	0,1462	0,1672	0,2999	0,1765	0,2201	0,2881	0,5390	0,4945	0,6428	0,2283	0,2887	0,3606
19	0,1838	0,2049	0,3263	0,1445	0,1650	0,2968	0,1738	0,2182	0,2860	0,5362	0,4922	0,6403	0,2274	0,2873	0,3601
20	0,5695	0,5833	0,4583	0,4768	0,4918	0,4247	0,6562	0,6539	0,5044	0,9107	0,8875	0,8405	0,7233	0,7293	0,6052
21	0,4307	0,4541	0,4348	0,3262	0,3550	0,4011	0,4645	0,4966	0,4276	0,8279	0,7976	0,7931	0,5163	0,5711	0,5102
22	0,4456	0,4677	0,4393	0,3402	0,3681	0,4060	0,4819	0,5108	0,4350	0,8373	0,8077	0,7987	0,5361	0,5861	0,5192
23	0,4590	0,4800	0,4433	0,3534	0,3803	0,4103	0,4975	0,5235	0,4417	0,8453	0,8163	0,8036	0,5541	0,5995	0,5274
24	0,4709	0,4911	0,4468	0,3655	0,3915	0,4142	0,5115	0,5348	0,4477	0,8521	0,8237	0,8079	0,5701	0,6115	0,5349
25	0,4818	0,5012	0,4500	0,3768	0,4019	0,4177	0,5241	0,5450	0,4532	0,8581	0,8300	0,8116	0,5844	0,6223	0,5418
26	0,4917	0,5104	0,4528	0,3872	0,4115	0,4208	0,5353	0,5542	0,4582	0,8633	0,8356	0,8149	0,5972	0,6320	0,5481
27	0,5007	0,5188	0,4552	0,3968	0,4205	0,4236	0,5454	0,5626	0,4628	0,8678	0,8405	0,8177	0,6087	0,6407	0,5539
28	0,5089	0,5266	0,4574	0,4058	0,4290	0,4261	0,5547	0,5702	0,4669	0,8719	0,8449	0,8202	0,6192	0,6487	0,5592
29	0,5165	0,5337	0,4594	0,4142	0,4369	0,4284	0,5631	0,5772	0,4708	0,8755	0,8488	0,8225	0,6285	0,6559	0,5640
30	0,5235	0,5404	0,4612	0,4221	0,4443	0,4305	0,5708	0,5836	0,4743	0,8787	0,8524	0,8244	0,6371	0,6626	0,5685
31	0,5300	0,5465	0,4628	0,4295	0,4513	0,4323	0,5779	0,5895	0,4775	0,8816	0,8555	0,8262	0,6449	0,6687	0,5727
32	0,5361	0,5523	0,4643	0,4365	0,4579	0,4340	0,5845	0,5949	0,4805	0,8843	0,8584	0,8278	0,6520	0,6744	0,5766
33	0,5417	0,5577	0,4656	0,4430	0,4641	0,4355	0,5905	0,6000	0,4832	0,8867	0,8611	0,8292	0,6586	0,6797	0,5801
34	0,5470	0,5628	0,4668	0,4492	0,4700	0,4369	0,5961	0,6048	0,4857	0,8890	0,8635	0,8304	0,6647	0,6846	0,5834
35	0,5519	0,5676	0,4679	0,4551	0,4756	0,4382	0,6013	0,6092	0,4881	0,8910	0,8658	0,8315	0,6704	0,6892	0,5865
36	0,5566	0,5721	0,4690	0,4606	0,4809	0,4394	0,6062	0,6134	0,4902	0,8929	0,8679	0,8325	0,6756	0,6935	0,5894

Tabela 5.21. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkowaniu klasy negatywnej do 80%, część 1.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0,9612	0,9510	0,8879	0,6980	0,6840	0,5425	0,9444	0,9221	0,8631	0,7779	0,7737	0,6615	0,9643	0,9516	0,9146
2	0,9005	0,8839	0,7540	0,7896	0,7388	0,4258	0,9055	0,8671	0,7376	0,8699	0,8451	0,5981	0,9434	0,9280	0,8446
3	0,3500	0,3001	0,5118	0,0528	0,0588	0,1320	0,3110	0,2559	0,4457	0,0781	0,0929	0,2049	0,4336	0,3771	0,6141
4	0,6917	0,6500	0,7743	0,1522	0,1887	0,2753	0,6039	0,5569	0,6947	0,2154	0,2747	0,3907	0,7466	0,7012	0,8279
5	0,6917	0,6499	0,7744	0,1522	0,1888	0,2751	0,6041	0,5568	0,6952	0,2153	0,2746	0,3907	0,7465	0,7012	0,8280
6	0,6566	0,6152	0,7488	0,1421	0,1854	0,2729	0,5867	0,5404	0,6894	0,1897	0,2578	0,3679	0,7157	0,6703	0,8085
7	0,4624	0,4123	0,5959	0,0951	0,1129	0,2085	0,4477	0,3901	0,5866	0,1119	0,1424	0,2555	0,5435	0,4884	0,6893
8	0,5829	0,5330	0,6645	0,1399	0,1602	0,2312	0,5525	0,4949	0,6289	0,1606	0,1989	0,2970	0,6455	0,5925	0,7441
9	0,4173	0,3622	0,5748	0,0702	0,0743	0,1549	0,3686	0,3083	0,5021	0,1013	0,1140	0,2330	0,5017	0,4424	0,6717
10	0,4174	0,3623	0,5748	0,0701	0,0743	0,1548	0,3688	0,3083	0,5022	0,1013	0,1140	0,2331	0,5017	0,4423	0,6717
11	0,9126	0,8934	0,8766	0,4295	0,4531	0,4476	0,8477	0,8110	0,8263	0,5112	0,5516	0,5653	0,9029	0,8785	0,8998
12	0,4381	0,3893	0,6097	0,0719	0,0849	0,1760	0,3830	0,3282	0,5392	0,1046	0,1320	0,2625	0,5247	0,4696	0,7017
13	0,9306	0,9117	0,8859	0,5376	0,5388	0,5036	0,8836	0,8486	0,8510	0,5884	0,5989	0,6046	0,9074	0,8817	0,9045
14	0,6565	0,6153	0,7489	0,1421	0,1853	0,2730	0,5868	0,5404	0,6893	0,1896	0,2578	0,3678	0,7157	0,6703	0,8085
15	0,8688	0,8442	0,8505	0,3369	0,3835	0,4073	0,7989	0,7592	0,8039	0,3882	0,4648	0,5032	0,8624	0,8331	0,8795
16	0,9423	0,9255	0,8859	0,6000	0,5914	0,5211	0,9068	0,8747	0,8551	0,6430	0,6423	0,6239	0,9219	0,8980	0,9050
17	0,6541	0,6124	0,7474	0,1409	0,1839	0,2705	0,5839	0,5376	0,6870	0,1882	0,2550	0,3657	0,7119	0,6657	0,8066
18	0,6555	0,6144	0,7481	0,1423	0,1856	0,2730	0,5868	0,5407	0,6889	0,1900	0,2594	0,3683	0,7179	0,6734	0,8088
19	0,6540	0,6124	0,7472	0,1408	0,1838	0,2704	0,5838	0,5374	0,6869	0,1882	0,2549	0,3656	0,7117	0,6655	0,8065
20	0,9367	0,9208	0,8863	0,5626	0,5678	0,5078	0,9006	0,8694	0,8537	0,6500	0,6694	0,6197	0,9362	0,9190	0,9110
21	0,8774	0,8537	0,8550	0,3542	0,3976	0,4162	0,8084	0,7693	0,8094	0,4066	0,4794	0,5129	0,8689	0,8406	0,8829
22	0,8846	0,8618	0,8590	0,3700	0,4103	0,4244	0,8167	0,7780	0,8142	0,4239	0,4927	0,5218	0,8747	0,8471	0,8859
23	0,8908	0,8687	0,8624	0,3843	0,4219	0,4320	0,8240	0,7858	0,8185	0,4402	0,5048	0,5300	0,8797	0,8529	0,8885
24	0,8961	0,8747	0,8655	0,3975	0,4324	0,4388	0,8304	0,7926	0,8223	0,4553	0,5158	0,5375	0,8842	0,8580	0,8908
25	0,9007	0,8799	0,8681	0,4094	0,4420	0,4450	0,8360	0,7987	0,8256	0,4690	0,5259	0,5444	0,8883	0,8626	0,8929
26	0,9047	0,8844	0,8705	0,4202	0,4508	0,4507	0,8411	0,8042	0,8286	0,4816	0,5351	0,5507	0,8919	0,8667	0,8947
27	0,9082	0,8883	0,8725	0,4302	0,4589	0,4558	0,8457	0,8092	0,8312	0,4930	0,5435	0,5565	0,8951	0,8705	0,8964
28	0,9113	0,8918	0,8743	0,4393	0,4664	0,4605	0,8498	0,8137	0,8336	0,5034	0,5513	0,5619	0,8981	0,8739	0,8978
29	0,9140	0,8950	0,8759	0,4477	0,4733	0,4648	0,8536	0,8179	0,8357	0,5130	0,5586	0,5668	0,9008	0,8770	0,8991
30	0,9165	0,8978	0,8772	0,4556	0,4798	0,4688	0,8571	0,8217	0,8376	0,5218	0,5653	0,5713	0,9033	0,8798	0,9003
31	0,9187	0,9003	0,8785	0,4628	0,4859	0,4725	0,8603	0,8253	0,8393	0,5300	0,5716	0,5754	0,9057	0,8825	0,9014
32	0,9207	0,9026	0,8796	0,4696	0,4915	0,4758	0,8633	0,8286	0,8409	0,5375	0,5775	0,5793	0,9078	0,8850	0,9023
33	0,9225	0,9047	0,8805	0,4760	0,4969	0,4789	0,8661	0,8317	0,8423	0,5446	0,5831	0,5828	0,9098	0,8873	0,9032
34	0,9242	0,9066	0,8814	0,4819	0,5020	0,4818	0,8687	0,8346	0,8436	0,5512	0,5883	0,5861	0,9117	0,8894	0,9040
35	0,9257	0,9084	0,8823	0,4875	0,5067	0,4844	0,8711	0,8373	0,8447	0,5575	0,5932	0,5892	0,9134	0,8914	0,9047
36	0,9271	0,9100	0,8830	0,4928	0,5113	0,4869	0,8734	0,8399	0,8458	0,5633	0,5979	0,5921	0,9151	0,8933	0,9053

Tabela 5.22. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkowaniu klasy negatywnej do 80%, część 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
37	0,5610	0,5764	0,4699	0,4659	0,4860	0,4404	0,6108	0,6173	0,4923	0,8947	0,8699	0,8334	0,6805	0,6975	0,5921
38	0,5652	0,5804	0,4707	0,4709	0,4908	0,4414	0,6151	0,6210	0,4941	0,8964	0,8717	0,8342	0,6851	0,7013	0,5946
39	0,5691	0,5843	0,4715	0,4757	0,4954	0,4424	0,6192	0,6245	0,4959	0,8979	0,8734	0,8350	0,6894	0,7048	0,5969
40	0,5729	0,5880	0,4722	0,4803	0,4998	0,4432	0,6231	0,6278	0,4975	0,8994	0,8750	0,8356	0,6934	0,7082	0,5991
41	0,5765	0,5914	0,4729	0,4847	0,5040	0,4440	0,6267	0,6310	0,4991	0,9008	0,8765	0,8362	0,6973	0,7114	0,6011
42	0,5799	0,5948	0,4735	0,4889	0,5080	0,4448	0,6302	0,6340	0,5005	0,9021	0,8780	0,8368	0,7009	0,7145	0,6031
43	0,5831	0,5979	0,4741	0,4929	0,5118	0,4455	0,6335	0,6368	0,5019	0,9033	0,8793	0,8373	0,7043	0,7174	0,6049
44	0,5862	0,6010	0,4746	0,4967	0,5155	0,4461	0,6366	0,6395	0,5031	0,9045	0,8806	0,8377	0,7076	0,7201	0,6066
45	0,5892	0,6038	0,4751	0,5004	0,5190	0,4467	0,6396	0,6421	0,5043	0,9056	0,8818	0,8381	0,7107	0,7227	0,6083
46	0,5920	0,6066	0,4756	0,5039	0,5224	0,4473	0,6425	0,6446	0,5055	0,9066	0,8830	0,8385	0,7137	0,7252	0,6098
47	0,5948	0,6093	0,4760	0,5073	0,5257	0,4479	0,6452	0,6469	0,5065	0,9076	0,8841	0,8388	0,7165	0,7276	0,6112
48	0,5974	0,6118	0,4765	0,5106	0,5288	0,4484	0,6479	0,6492	0,5075	0,9085	0,8851	0,8391	0,7192	0,7299	0,6126
49	0,5999	0,6143	0,4768	0,5137	0,5318	0,4489	0,6504	0,6514	0,5085	0,9094	0,8861	0,8394	0,7218	0,7321	0,6139
50	0,6022	0,6166	0,4772	0,5167	0,5346	0,4493	0,6528	0,6534	0,5094	0,9103	0,8871	0,8396	0,7242	0,7342	0,6152
51	0,5482	0,5574	0,4514	0,4344	0,4404	0,4071	0,6440	0,6394	0,5045	0,9012	0,8753	0,8383	0,6961	0,6980	0,6040
52	0,5140	0,5218	0,4415	0,3981	0,4023	0,3951	0,6285	0,6243	0,4985	0,8932	0,8659	0,8359	0,6720	0,6742	0,5934
53	0,4741	0,4803	0,4280	0,3608	0,3647	0,3805	0,6099	0,6062	0,4910	0,8831	0,8540	0,8324	0,6466	0,6493	0,5798
54	0,4509	0,4570	0,4134	0,3360	0,3402	0,3674	0,5878	0,5845	0,4815	0,8700	0,8391	0,8273	0,6210	0,6243	0,5644
55	0,4342	0,4396	0,4065	0,3163	0,3201	0,3621	0,5601	0,5575	0,4696	0,8520	0,8183	0,8200	0,5913	0,5951	0,5467
56	0,4165	0,4209	0,4029	0,2970	0,2999	0,3589	0,5272	0,5252	0,4543	0,8277	0,7909	0,8091	0,5557	0,5601	0,5255
57	0,3958	0,3989	0,3981	0,2762	0,2784	0,3534	0,4913	0,4897	0,4354	0,7996	0,7593	0,7929	0,5126	0,5174	0,4990
58	0,3714	0,3729	0,3898	0,2537	0,2546	0,3434	0,4531	0,4523	0,4132	0,7686	0,7250	0,7716	0,4637	0,4689	0,4673
59	0,3417	0,3420	0,3753	0,2280	0,2279	0,3271	0,4117	0,4112	0,3885	0,7307	0,6838	0,7453	0,4095	0,4157	0,4309
60	0,3076	0,3069	0,3548	0,1998	0,1990	0,3036	0,3670	0,3667	0,3597	0,6870	0,6361	0,7122	0,3534	0,3592	0,3905
61	0,2698	0,2679	0,3269	0,1706	0,1695	0,2765	0,3198	0,3201	0,3289	0,6349	0,5814	0,6734	0,2959	0,3008	0,3478
62	0,2294	0,2261	0,2954	0,1441	0,1426	0,2473	0,2719	0,2715	0,2964	0,5774	0,5198	0,6276	0,2431	0,2467	0,3059
63	0,1889	0,1841	0,2613	0,1238	0,1223	0,2175	0,2239	0,2233	0,2624	0,5114	0,4518	0,5763	0,1995	0,2023	0,2689
64	0,1527	0,1476	0,2262	0,1077	0,1064	0,1899	0,1819	0,1807	0,2302	0,4462	0,3865	0,5233	0,1677	0,1702	0,2362
65	0,1273	0,1225	0,1954	0,0948	0,0935	0,1681	0,1478	0,1456	0,2003	0,3918	0,3343	0,4723	0,1430	0,1452	0,2084
66	0,1093	0,1048	0,1702	0,0844	0,0829	0,1508	0,1226	0,1217	0,1741	0,3469	0,2933	0,4255	0,1229	0,1243	0,1841
67	0,0953	0,0905	0,1509	0,0753	0,0741	0,1368	0,1054	0,1040	0,1530	0,3102	0,2587	0,3861	0,1053	0,1073	0,1651
68	0,0830	0,0785	0,1346	0,0674	0,0661	0,1250	0,0907	0,0892	0,1348	0,2775	0,2299	0,3512	0,0906	0,0937	0,1521
69	0,0726	0,0681	0,1210	0,0609	0,0597	0,1152	0,0775	0,0767	0,1200	0,2527	0,2067	0,3198	0,0850	0,0883	0,1447
70	0,0598	0,0563	0,1097	0,0530	0,0526	0,1067	0,0602	0,0615	0,1075	0,2261	0,1831	0,2940	0,0808	0,0843	0,1402
71	0,4967	0,5063	0,2366	0,6084	0,6225	0,2683	0,6219	0,5999	0,3211	0,8824	0,8484	0,6506	0,7218	0,7094	0,4368
72	0,5830	0,5942	0,3001	0,5863	0,6005	0,2787	0,6808	0,6574	0,3799	0,9115	0,8843	0,7175	0,7740	0,7597	0,5025

Tabela 5.23. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkowaniu klasy negatywnej do 80%, część 3.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
37	0,9284	0,9116	0,8836	0,4979	0,5156	0,4892	0,8756	0,8423	0,8468	0,5688	0,6024	0,5947	0,9166	0,8951	0,9060
38	0,9297	0,9130	0,8841	0,5027	0,5197	0,4913	0,8777	0,8447	0,8477	0,5740	0,6066	0,5972	0,9181	0,8968	0,9065
39	0,9308	0,9143	0,8847	0,5072	0,5236	0,4933	0,8796	0,8469	0,8485	0,5789	0,6107	0,5995	0,9195	0,8984	0,9070
40	0,9319	0,9156	0,8852	0,5116	0,5274	0,4952	0,8815	0,8490	0,8493	0,5836	0,6145	0,6017	0,9208	0,9000	0,9075
41	0,9329	0,9167	0,8856	0,5157	0,5309	0,4969	0,8832	0,8510	0,8500	0,5881	0,6182	0,6038	0,9221	0,9014	0,9079
42	0,9338	0,9178	0,8860	0,5197	0,5344	0,4985	0,8849	0,8529	0,8506	0,5924	0,6217	0,6057	0,9233	0,9028	0,9083
43	0,9347	0,9189	0,8863	0,5235	0,5377	0,5000	0,8865	0,8547	0,8512	0,5965	0,6251	0,6075	0,9244	0,9041	0,9086
44	0,9356	0,9199	0,8867	0,5272	0,5408	0,5015	0,8881	0,8564	0,8518	0,6004	0,6283	0,6092	0,9255	0,9054	0,9090
45	0,9364	0,9208	0,8870	0,5307	0,5438	0,5029	0,8895	0,8581	0,8523	0,6041	0,6314	0,6108	0,9265	0,9066	0,9093
46	0,9371	0,9217	0,8873	0,5341	0,5467	0,5041	0,8909	0,8597	0,8528	0,6078	0,6344	0,6123	0,9275	0,9077	0,9096
47	0,9378	0,9226	0,8875	0,5374	0,5495	0,5053	0,8923	0,8612	0,8533	0,6112	0,6373	0,6137	0,9284	0,9088	0,9099
48	0,9385	0,9234	0,8877	0,5405	0,5522	0,5065	0,8935	0,8626	0,8537	0,6146	0,6401	0,6151	0,9293	0,9099	0,9101
49	0,9392	0,9241	0,8880	0,5436	0,5548	0,5075	0,8948	0,8640	0,8541	0,6178	0,6427	0,6164	0,9302	0,9109	0,9104
50	0,9398	0,9249	0,8881	0,5465	0,5573	0,5086	0,8959	0,8654	0,8545	0,6209	0,6453	0,6176	0,9310	0,9119	0,9106
51	0,9227	0,9018	0,8836	0,5207	0,5225	0,4977	0,8742	0,8375	0,8483	0,5647	0,5757	0,5953	0,8958	0,8681	0,9014
52	0,9112	0,8877	0,8800	0,5013	0,5036	0,4903	0,8626	0,8238	0,8445	0,5380	0,5495	0,5839	0,8814	0,8514	0,8976
53	0,8979	0,8724	0,8742	0,4790	0,4821	0,4812	0,8484	0,8071	0,8397	0,5117	0,5243	0,5696	0,8670	0,8349	0,8920
54	0,8846	0,8569	0,8673	0,4540	0,4576	0,4700	0,8310	0,7872	0,8333	0,4861	0,4997	0,5543	0,8531	0,8190	0,8861
55	0,8684	0,8379	0,8585	0,4260	0,4297	0,4562	0,8096	0,7631	0,8245	0,4577	0,4715	0,5363	0,8371	0,8002	0,8786
56	0,8470	0,8136	0,8466	0,3957	0,3994	0,4396	0,7850	0,7351	0,8127	0,4242	0,4387	0,5139	0,8165	0,7767	0,8680
57	0,8182	0,7805	0,8299	0,3655	0,3693	0,4191	0,7598	0,7069	0,7968	0,3855	0,4010	0,4867	0,7899	0,7471	0,8539
58	0,7818	0,7397	0,8073	0,3330	0,3373	0,3955	0,7312	0,6752	0,7773	0,3425	0,3585	0,4541	0,7580	0,7114	0,8350
59	0,7373	0,6916	0,7773	0,2982	0,3027	0,3689	0,6972	0,6375	0,7545	0,2972	0,3138	0,4176	0,7197	0,6702	0,8106
60	0,6843	0,6340	0,7394	0,2615	0,2660	0,3396	0,6580	0,5953	0,7269	0,2521	0,2677	0,3777	0,6759	0,6231	0,7803
61	0,6239	0,5695	0,6944	0,2235	0,2274	0,3079	0,6129	0,5465	0,6937	0,2102	0,2235	0,3364	0,6280	0,5715	0,7443
62	0,5584	0,5024	0,6430	0,1863	0,1901	0,2745	0,5639	0,4949	0,6559	0,1752	0,1882	0,2957	0,5804	0,5217	0,7036
63	0,5002	0,4419	0,5924	0,1522	0,1549	0,2402	0,5122	0,4421	0,6113	0,1505	0,1622	0,2608	0,5402	0,4799	0,6645
64	0,4508	0,3937	0,5438	0,1276	0,1298	0,2086	0,4669	0,3979	0,5650	0,1312	0,1425	0,2315	0,5052	0,4444	0,6277
65	0,4090	0,3523	0,4989	0,1098	0,1120	0,1818	0,4276	0,3600	0,5221	0,1146	0,1252	0,2056	0,4734	0,4136	0,5916
66	0,3730	0,3171	0,4588	0,0957	0,0976	0,1601	0,3938	0,3287	0,4832	0,1005	0,1103	0,1844	0,4455	0,3874	0,5588
67	0,3460	0,2906	0,4268	0,0837	0,0858	0,1424	0,3648	0,3020	0,4496	0,0888	0,0986	0,1678	0,4252	0,3682	0,5313
68	0,3242	0,2707	0,4017	0,0733	0,0751	0,1268	0,3388	0,2785	0,4182	0,0793	0,0895	0,1558	0,4094	0,3531	0,5100
69	0,3156	0,2632	0,3864	0,0638	0,0659	0,1138	0,3182	0,2604	0,3901	0,0757	0,0860	0,1487	0,4035	0,3476	0,4959
70	0,3091	0,2583	0,3759	0,0524	0,0561	0,1019	0,2972	0,2412	0,3659	0,0729	0,0835	0,1443	0,3995	0,3428	0,4866
71	0,9215	0,9005	0,7425	0,7796	0,7644	0,4549	0,9649	0,9501	0,8153	0,8456	0,8411	0,5793	0,9783	0,9695	0,8813
72	0,9411	0,9242	0,7961	0,7471	0,7288	0,4528	0,9572	0,9394	0,8139	0,8207	0,8146	0,5778	0,9733	0,9627	0,8803

Tabela 5.24. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkowaniu klasy negatywnej do 80%, część 4.

nie, w niektórych przypadkach wyrost sięgał całego rzędu, a w przypadku funkcji *Expected Frequency* okazał się niespodziewanie ogromny – kilkudziesięciokrotny.

Zmianie uległa zestaw funkcji osiągających najlepsze wyniki i jest następujący: *Frequency*, *Expected Frequency*, *Mutual Expectation*, *W Specific Correlation*, *Specific Frequency Biased Mutual Dependency*, *T-Score*, *Loglikelihood*, *W Order*, *W Term Frequency Order* oraz funkcje parametryczne – *Specific Exponential Correlation* i *W Specific Exponential Correlation*. Zauważyć można, że zestaw funkcji jest bardzo podobno do tego z poprzednich badań, ale w stosunku do badania poprzedniego grono najlepszych zostało rozszerzone o miary *Mutual Expectation* i *Specific Frequency Biased Mutual Dependency*.

Wyniki badań jakości wyników dla miar asocjacyjnych po podpróbkowaniu klasy negatywnej do 95%

Cztery tabele ??, ??, ?? oraz ?? prezentują jakość wyników osiągniętych przez 72 funkcje w 30 różnych badaniach – 30 zestawów danych pozyskanych z korpusu *KIPI*. Indeksy miar i typów badań pozostały takie same jak w poprzednich badaniach.

Różnica pomiędzy tym, a poprzednim badaniem jest taka, że w tym przypadku krotki z klasy negatywnej stanowią około 95% wszystkich kandydatów, a nie tylko 80%. Motywacją do tego badania były stosunkowo dobre wyniki przy podpróbkowaniu na poziomie 80% i chęć sprawdzenia jak zmieniają się wyniki przy kilkukrotnym ograniczeniu liczby jednostek wielowyrazowych wśród wszystkich kandydatów.

Zgodnie z oczekiwaniami jakość wyników spadła, co było do przewidzenia biorąc pod uwagę czterokrotne zmniejszenie liczby instancji klasy pozytywnej – wyrażeń wielowyrzowych, i zwiększenie liczby instancji z klasy negatywnej w tym problemie binarnym. Natomiast zestaw funkcji osiągających najlepsze wyniki nie uległ zmianie w stosunku do badania poprzedniego – podpróbkowania klasy negatywnej na poziomie 80%.

Najlepszy z osiągniętych wyników w tym badaniu uplasował się na poziomie około 93% najlepszego z rezultatów osiągniętych badaniu poprzednim. Jednak zaznaczyć trzeba, że w ogólności jakość wyników uległa dużym zmianom, w niektórych przypadkach nawet około dwu- i półkrotnie obniżając wynik przykładowo z 0,8174 do 0,3148 czyli o około 61%.

Wnioskiem z tego i poprzedniego badania może być to, że nawet proste funkcje takie jak zwykła częstość czy wartość oczekiwana częstości mogą osiągać bardzo dobre wyniki jeśli zbiór danych jest klasowo zbalansowany. Dodatkowo w tym problemie zbalansowanie klas wydaje się mieć znaczący wpływ na jakość generowanych rozwiązań biorąc pod uwagę wyniki niniejszego badania, a także badania poprzedniego i pierwszego.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,3540	0,3708	0,1683	0,3038	0,3154	0,1581	0,4210	0,4046	0,1995	0,7649	0,7165	0,5163	0,5299	0,5152	0,2929
2	0,3148	0,3084	0,0645	0,3698	0,3672	0,0824	0,2539	0,2134	0,0614	0,4959	0,4206	0,2062	0,3796	0,3399	0,1151
3	0,0214	0,0219	0,0700	0,0180	0,0192	0,0637	0,0202	0,0237	0,0585	0,0829	0,0686	0,1694	0,0310	0,0375	0,0898
4	0,0949	0,1151	0,1732	0,0687	0,0809	0,1473	0,0823	0,1228	0,1444	0,3365	0,3044	0,3999	0,1250	0,1831	0,2123
5	0,0949	0,1151	0,1733	0,0686	0,0810	0,1474	0,0823	0,1229	0,1443	0,3365	0,3042	0,3998	0,1251	0,1831	0,2123
6	0,0919	0,1225	0,1851	0,0641	0,0843	0,1587	0,0775	0,1262	0,1470	0,3372	0,3076	0,4099	0,1086	0,1769	0,2047
7	0,0458	0,0526	0,1260	0,0327	0,0379	0,1089	0,0428	0,0578	0,1008	0,1771	0,1517	0,2880	0,0526	0,0724	0,1273
8	0,0750	0,0854	0,1435	0,0500	0,0577	0,1216	0,0701	0,0921	0,1151	0,2668	0,2342	0,3265	0,0837	0,1145	0,1524
9	0,0292	0,0273	0,0771	0,0240	0,0232	0,0678	0,0271	0,0292	0,0657	0,1017	0,0833	0,1875	0,0415	0,0448	0,1004
10	0,0292	0,0273	0,0770	0,0240	0,0232	0,0678	0,0271	0,0292	0,0657	0,1017	0,0833	0,1876	0,0415	0,0448	0,1004
11	0,2723	0,2800	0,2139	0,2002	0,2110	0,1937	0,2988	0,3171	0,2105	0,6614	0,6191	0,5451	0,3769	0,4053	0,2940
12	0,0362	0,0401	0,1110	0,0285	0,0324	0,0975	0,0320	0,0417	0,0875	0,1339	0,1148	0,2499	0,0489	0,0660	0,1311
13	0,3070	0,3176	0,2074	0,2276	0,2342	0,1918	0,3660	0,3621	0,2202	0,7149	0,6700	0,5649	0,4439	0,4422	0,3059
14	0,0919	0,1225	0,1852	0,0641	0,0843	0,1588	0,0774	0,1262	0,1471	0,3371	0,3076	0,4099	0,1086	0,1768	0,2047
15	0,2349	0,2697	0,2286	0,1566	0,1897	0,2047	0,2323	0,2838	0,2060	0,6158	0,5776	0,5453	0,2729	0,3525	0,2758
16	0,3179	0,3298	0,1810	0,2382	0,2426	0,1667	0,3911	0,3780	0,2050	0,7315	0,6828	0,5299	0,4765	0,4645	0,2957
17	0,0906	0,1204	0,1838	0,0627	0,0819	0,1566	0,0766	0,1251	0,1462	0,3350	0,3054	0,4082	0,1075	0,1746	0,2037
18	0,0912	0,1213	0,1844	0,0634	0,0832	0,1580	0,0774	0,1259	0,1468	0,3362	0,3069	0,4094	0,1084	0,1766	0,2043
19	0,0903	0,1201	0,1836	0,0625	0,0816	0,1563	0,0767	0,1250	0,1462	0,3349	0,3053	0,4082	0,1075	0,1745	0,2036
20	0,3073	0,3215	0,2171	0,2356	0,2497	0,2053	0,3677	0,3680	0,2243	0,7243	0,6810	0,5749	0,4533	0,4589	0,3101
21	0,2457	0,2777	0,2291	0,1656	0,1976	0,2061	0,2460	0,2927	0,2090	0,6282	0,5897	0,5509	0,2898	0,3630	0,2800
22	0,2552	0,2846	0,2293	0,1741	0,2047	0,2071	0,2585	0,3005	0,2115	0,6388	0,6001	0,5555	0,3057	0,3725	0,2837
23	0,2634	0,2907	0,2292	0,1820	0,2112	0,2078	0,2696	0,3074	0,2136	0,6479	0,6090	0,5594	0,3202	0,3809	0,2870
24	0,2706	0,2960	0,2288	0,1894	0,2170	0,2081	0,2794	0,3136	0,2154	0,6558	0,6167	0,5626	0,3334	0,3885	0,2899
25	0,2768	0,3007	0,2283	0,1960	0,2222	0,2082	0,2882	0,3190	0,2169	0,6627	0,6234	0,5651	0,3453	0,3953	0,2924
26	0,2823	0,3049	0,2276	0,2020	0,2270	0,2081	0,2961	0,3239	0,2181	0,6689	0,6294	0,5671	0,3559	0,4014	0,2946
27	0,2871	0,3087	0,2267	0,2074	0,2313	0,2078	0,3031	0,3283	0,2191	0,6743	0,6346	0,5687	0,3655	0,4070	0,2965
28	0,2913	0,3121	0,2258	0,2123	0,2353	0,2074	0,3094	0,3323	0,2200	0,6792	0,6393	0,5699	0,3742	0,4121	0,2982
29	0,2951	0,3151	0,2249	0,2167	0,2389	0,2068	0,3151	0,3360	0,2207	0,6836	0,6436	0,5709	0,3819	0,4167	0,2997
30	0,2985	0,3179	0,2239	0,2208	0,2422	0,2062	0,3202	0,3393	0,2212	0,6876	0,6474	0,5715	0,3889	0,4210	0,3010
31	0,3015	0,3204	0,2228	0,2245	0,2453	0,2056	0,3249	0,3423	0,2217	0,6912	0,6509	0,5720	0,3953	0,4250	0,3021
32	0,3043	0,3228	0,2218	0,2279	0,2481	0,2048	0,3292	0,3451	0,2220	0,6945	0,6541	0,5723	0,4012	0,4286	0,3031
33	0,3068	0,3249	0,2208	0,2310	0,2507	0,2041	0,3332	0,3477	0,2223	0,6976	0,6571	0,5724	0,4065	0,4320	0,3039
34	0,3091	0,3269	0,2198	0,2339	0,2532	0,2033	0,3368	0,3501	0,2225	0,7004	0,6598	0,5724	0,4114	0,4351	0,3047
35	0,3112	0,3287	0,2187	0,2366	0,2555	0,2026	0,3401	0,3523	0,2226	0,7031	0,6623	0,5723	0,4159	0,4381	0,3053
36	0,3131	0,3304	0,2177	0,2391	0,2576	0,2018	0,3432	0,3544	0,2227	0,7055	0,6646	0,5721	0,4202	0,4409	0,3059

Tabela 5.25. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkiowaniu klasy negatywnej do 95%, część 1.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0,8385	0,8006	0,6404	0,3636	0,3498	0,2063	0,7750	0,7224	0,5873	0,4598	0,4527	0,3037	0,8458	0,8037	0,7114
2	0,6381	0,5671	0,3218	0,3201	0,2765	0,0740	0,6306	0,5486	0,3162	0,4627	0,4248	0,1453	0,7453	0,6867	0,4573
3	0,1350	0,1116	0,2656	0,0170	0,0207	0,0549	0,1066	0,0848	0,2082	0,0256	0,0331	0,0881	0,1780	0,1458	0,3417
4	0,4865	0,4370	0,5526	0,0609	0,0892	0,1293	0,3531	0,3068	0,4378	0,0918	0,1376	0,1973	0,5199	0,4602	0,6122
5	0,4866	0,4370	0,5526	0,0609	0,0892	0,1293	0,3532	0,3069	0,4378	0,0918	0,1376	0,1972	0,5199	0,4602	0,6121
6	0,4690	0,4241	0,5483	0,0576	0,0940	0,1369	0,3529	0,3098	0,4522	0,0807	0,1370	0,1967	0,5049	0,4478	0,6088
7	0,2355	0,2014	0,3743	0,0323	0,0443	0,0957	0,1991	0,1645	0,3337	0,0414	0,0601	0,1259	0,2871	0,2427	0,4507
8	0,3436	0,3007	0,4339	0,0501	0,0673	0,1065	0,2812	0,2381	0,3708	0,0616	0,0897	0,1468	0,3892	0,3351	0,5076
9	0,1637	0,1342	0,2918	0,0219	0,0240	0,0584	0,1279	0,1008	0,2263	0,0332	0,0382	0,0935	0,2112	0,1727	0,3697
10	0,1637	0,1342	0,2918	0,0219	0,0240	0,0584	0,1279	0,1008	0,2263	0,0332	0,0382	0,0935	0,2112	0,1727	0,3697
11	0,7539	0,7150	0,6717	0,2200	0,2422	0,2085	0,6209	0,5687	0,5829	0,2786	0,3182	0,2943	0,7283	0,6785	0,7165
12	0,2185	0,1862	0,3754	0,0261	0,0351	0,0811	0,1628	0,1349	0,2949	0,0393	0,0562	0,1264	0,2732	0,2304	0,4545
13	0,7769	0,7356	0,6807	0,2825	0,2822	0,2270	0,6663	0,6115	0,6098	0,3332	0,3408	0,3128	0,7263	0,6736	0,7249
14	0,4689	0,4241	0,5483	0,0576	0,0940	0,1369	0,3530	0,3099	0,4522	0,0807	0,1370	0,1967	0,5049	0,4478	0,6088
15	0,7089	0,6699	0,6659	0,1603	0,2120	0,2037	0,5720	0,5227	0,5810	0,1911	0,2721	0,2751	0,6825	0,6312	0,7088
16	0,7912	0,7480	0,6527	0,3070	0,2976	0,2115	0,6965	0,6389	0,5870	0,3595	0,3565	0,3011	0,7442	0,6901	0,7053
17	0,4648	0,4194	0,5461	0,0570	0,0928	0,1357	0,3497	0,3066	0,4500	0,0795	0,1342	0,1948	0,4973	0,4396	0,6050
18	0,4684	0,4239	0,5477	0,0577	0,0944	0,1372	0,3533	0,3106	0,4523	0,0815	0,1396	0,1984	0,5108	0,4553	0,6119
19	0,4646	0,4192	0,5459	0,0570	0,0927	0,1357	0,3496	0,3064	0,4499	0,0794	0,1341	0,1947	0,4970	0,4393	0,6049
20	0,7960	0,7596	0,6898	0,3017	0,3083	0,2377	0,6985	0,6495	0,6250	0,3851	0,4067	0,3334	0,7909	0,7494	0,7437
21	0,7190	0,6805	0,6707	0,1720	0,2202	0,2076	0,5830	0,5337	0,5869	0,2037	0,2815	0,2803	0,6910	0,6403	0,7133
22	0,7278	0,6896	0,6747	0,1829	0,2275	0,2111	0,5929	0,5435	0,5920	0,2159	0,2901	0,2849	0,6987	0,6484	0,7172
23	0,7353	0,6975	0,6780	0,1928	0,2341	0,2141	0,6016	0,5523	0,5962	0,2278	0,2979	0,2890	0,7055	0,6557	0,7206
24	0,7419	0,7044	0,6806	0,2019	0,2400	0,2166	0,6094	0,5601	0,5999	0,2390	0,3049	0,2927	0,7118	0,6622	0,7234
25	0,7477	0,7103	0,6827	0,2101	0,2453	0,2189	0,6164	0,5671	0,6030	0,2498	0,3114	0,2960	0,7174	0,6682	0,7258
26	0,7529	0,7157	0,6845	0,2176	0,2502	0,2208	0,6228	0,5735	0,6056	0,2597	0,3173	0,2989	0,7226	0,6737	0,7279
27	0,7575	0,7204	0,6859	0,2243	0,2546	0,2224	0,6287	0,5793	0,6077	0,2687	0,3227	0,3015	0,7273	0,6787	0,7296
28	0,7617	0,7247	0,6870	0,2304	0,2586	0,2237	0,6341	0,5847	0,6096	0,2769	0,3277	0,3037	0,7317	0,6833	0,7311
29	0,7654	0,7285	0,6877	0,2360	0,2623	0,2249	0,6391	0,5897	0,6111	0,2844	0,3323	0,3058	0,7358	0,6876	0,7323
30	0,7688	0,7320	0,6884	0,2411	0,2658	0,2259	0,6438	0,5943	0,6124	0,2913	0,3366	0,3076	0,7396	0,6916	0,7333
31	0,7720	0,7352	0,6888	0,2458	0,2689	0,2267	0,6481	0,5986	0,6134	0,2976	0,3406	0,3091	0,7432	0,6954	0,7342
32	0,7748	0,7381	0,6890	0,2501	0,2719	0,2274	0,6522	0,6027	0,6143	0,3034	0,3443	0,3105	0,7465	0,6989	0,7349
33	0,7775	0,7409	0,6892	0,2542	0,2747	0,2280	0,6560	0,6065	0,6150	0,3088	0,3478	0,3117	0,7497	0,7022	0,7355
34	0,7799	0,7434	0,6892	0,2579	0,2772	0,2284	0,6596	0,6101	0,6156	0,3138	0,3511	0,3128	0,7527	0,7054	0,7360
35	0,7822	0,7457	0,6892	0,2614	0,2797	0,2288	0,6630	0,6135	0,6160	0,3184	0,3542	0,3138	0,7555	0,7083	0,7364
36	0,7843	0,7479	0,6891	0,2647	0,2819	0,2291	0,6663	0,6167	0,6164	0,3228	0,3571	0,3146	0,7582	0,7111	0,7368

Tabela 5.26. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbce-
niu klasy negatywnej do 95%, część 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
37	0,3148	0,3320	0,2168	0,2414	0,2596	0,2010	0,3460	0,3563	0,2227	0,7078	0,6668	0,5719	0,4241	0,4434	0,3064
38	0,3165	0,3335	0,2158	0,2436	0,2615	0,2002	0,3487	0,3581	0,2227	0,7099	0,6688	0,5716	0,4277	0,4459	0,3068
39	0,3180	0,3349	0,2149	0,2456	0,2633	0,1994	0,3512	0,3598	0,2227	0,7119	0,6707	0,5712	0,4311	0,4481	0,3071
40	0,3194	0,3362	0,2140	0,2476	0,2649	0,1987	0,3535	0,3614	0,2226	0,7138	0,6725	0,5708	0,4343	0,4503	0,3074
41	0,3207	0,3375	0,2131	0,2494	0,2665	0,1979	0,3557	0,3629	0,2225	0,7156	0,6742	0,5703	0,4373	0,4523	0,3077
42	0,3220	0,3386	0,2122	0,2511	0,2680	0,1972	0,3578	0,3643	0,2224	0,7173	0,6758	0,5699	0,4401	0,4543	0,3079
43	0,3231	0,3397	0,2114	0,2527	0,2694	0,1965	0,3598	0,3656	0,2223	0,7189	0,6772	0,5694	0,4428	0,4561	0,3081
44	0,3242	0,3407	0,2106	0,2542	0,2707	0,1958	0,3616	0,3669	0,2221	0,7204	0,6787	0,5689	0,4453	0,4578	0,3082
45	0,3253	0,3417	0,2098	0,2556	0,2719	0,1951	0,3634	0,3681	0,2220	0,7218	0,6800	0,5683	0,4477	0,4595	0,3083
46	0,3262	0,3426	0,2090	0,2570	0,2731	0,1944	0,3651	0,3692	0,2218	0,7232	0,6813	0,5678	0,4499	0,4610	0,3084
47	0,3271	0,3435	0,2083	0,2583	0,2743	0,1938	0,3666	0,3703	0,2217	0,7244	0,6824	0,5673	0,4521	0,4625	0,3085
48	0,3280	0,3443	0,2076	0,2595	0,2753	0,1932	0,3682	0,3713	0,2215	0,7257	0,6836	0,5667	0,4541	0,4639	0,3085
49	0,3288	0,3451	0,2069	0,2607	0,2764	0,1926	0,3696	0,3723	0,2213	0,7268	0,6846	0,5662	0,4561	0,4653	0,3086
50	0,3296	0,3459	0,2062	0,2618	0,2773	0,1920	0,3709	0,3732	0,2211	0,7279	0,6857	0,5657	0,4579	0,4666	0,3086
51	0,2950	0,3040	0,2108	0,2121	0,2176	0,1945	0,3574	0,3543	0,2208	0,7044	0,6592	0,5669	0,4292	0,4284	0,3047
52	0,2787	0,2856	0,2138	0,1946	0,1988	0,1967	0,3468	0,3448	0,2209	0,6911	0,6461	0,5681	0,4110	0,4107	0,3019
53	0,2572	0,2612	0,2151	0,1752	0,1777	0,1967	0,3340	0,3331	0,2204	0,6742	0,6291	0,5679	0,3884	0,3883	0,2968
54	0,2364	0,2401	0,2130	0,1584	0,1609	0,1921	0,3191	0,3192	0,2186	0,6540	0,6082	0,5652	0,3643	0,3664	0,2888
55	0,2258	0,2293	0,2071	0,1483	0,1507	0,1838	0,3007	0,3019	0,2152	0,6275	0,5818	0,5585	0,3391	0,3420	0,2781
56	0,2180	0,2205	0,2044	0,1390	0,1408	0,1782	0,2782	0,2802	0,2092	0,5926	0,5470	0,5461	0,3094	0,3124	0,2643
57	0,2075	0,2088	0,2011	0,1271	0,1281	0,1709	0,2519	0,2540	0,1996	0,5492	0,5030	0,5248	0,2745	0,2780	0,2465
58	0,1918	0,1922	0,1930	0,1113	0,1116	0,1594	0,2246	0,2270	0,1863	0,5036	0,4570	0,4933	0,2349	0,2391	0,2237
59	0,1705	0,1702	0,1793	0,0927	0,0918	0,1437	0,1953	0,1976	0,1696	0,4521	0,4060	0,4532	0,1933	0,1982	0,1973
60	0,1446	0,1442	0,1609	0,0725	0,0717	0,1247	0,1648	0,1671	0,1516	0,3969	0,3499	0,4093	0,1509	0,1560	0,1689
61	0,1164	0,1152	0,1390	0,0542	0,0533	0,1043	0,1339	0,1358	0,1317	0,3366	0,2914	0,3577	0,1124	0,1169	0,1405
62	0,0872	0,0860	0,1153	0,0421	0,0418	0,0852	0,1029	0,1047	0,1109	0,2741	0,2308	0,3040	0,0807	0,0840	0,1150
63	0,0612	0,0584	0,0925	0,0346	0,0343	0,0694	0,0746	0,0757	0,0908	0,2115	0,1723	0,2515	0,0606	0,0625	0,0945
64	0,0441	0,0422	0,0738	0,0290	0,0290	0,0579	0,0532	0,0534	0,0742	0,1626	0,1308	0,2079	0,0483	0,0501	0,0791
65	0,0350	0,0340	0,0600	0,0248	0,0248	0,0501	0,0402	0,0404	0,0613	0,1328	0,1057	0,1732	0,0394	0,0414	0,0667
66	0,0289	0,0281	0,0506	0,0216	0,0217	0,0445	0,0321	0,0327	0,0513	0,1110	0,0881	0,1470	0,0329	0,0348	0,0579
67	0,0244	0,0237	0,0438	0,0190	0,0191	0,0397	0,0266	0,0272	0,0435	0,0939	0,0744	0,1265	0,0282	0,0299	0,0517
68	0,0210	0,0202	0,0385	0,0170	0,0170	0,0363	0,0223	0,0231	0,0381	0,0808	0,0641	0,1114	0,0245	0,0264	0,0473
69	0,0183	0,0174	0,0347	0,0154	0,0155	0,0335	0,0191	0,0198	0,0338	0,0713	0,0563	0,0997	0,0230	0,0249	0,0444
70	0,0155	0,0149	0,0318	0,0139	0,0141	0,0314	0,0157	0,0166	0,0307	0,0629	0,0499	0,0912	0,0220	0,0239	0,0425
71	0,1766	0,1841	0,0659	0,2489	0,2583	0,0793	0,2672	0,2528	0,0999	0,6136	0,5520	0,3038	0,3646	0,3483	0,1489
72	0,2291	0,2402	0,0883	0,2321	0,2408	0,0825	0,3127	0,2952	0,1234	0,6794	0,6198	0,3665	0,4199	0,4003	0,1829

Tabela 5.27. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbce-
niu klasy negatywnej do 95%, część 3.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
37	0,7863	0,7499	0,6889	0,2677	0,2841	0,2293	0,6693	0,6197	0,6167	0,3269	0,3599	0,3154	0,7607	0,7138	0,7370
38	0,7882	0,7518	0,6886	0,2706	0,2861	0,2295	0,6723	0,6226	0,6169	0,3307	0,3625	0,3160	0,7631	0,7163	0,7372
39	0,7900	0,7535	0,6883	0,2733	0,2880	0,2296	0,6750	0,6254	0,6170	0,3344	0,3650	0,3166	0,7654	0,7187	0,7373
40	0,7916	0,7552	0,6880	0,2758	0,2898	0,2297	0,6777	0,6280	0,6171	0,3378	0,3673	0,3171	0,7676	0,7210	0,7374
41	0,7932	0,7568	0,6876	0,2782	0,2915	0,2298	0,6802	0,6305	0,6171	0,3410	0,3695	0,3175	0,7697	0,7232	0,7374
42	0,7946	0,7583	0,6873	0,2805	0,2931	0,2298	0,6827	0,6329	0,6171	0,3441	0,3717	0,3179	0,7716	0,7253	0,7375
43	0,7960	0,7597	0,6869	0,2827	0,2947	0,2297	0,6850	0,6352	0,6171	0,3469	0,3737	0,3183	0,7735	0,7273	0,7374
44	0,7974	0,7610	0,6864	0,2848	0,2962	0,2297	0,6872	0,6374	0,6170	0,3497	0,3756	0,3186	0,7754	0,7292	0,7374
45	0,7986	0,7623	0,6860	0,2868	0,2976	0,2296	0,6893	0,6394	0,6170	0,3524	0,3775	0,3188	0,7771	0,7310	0,7373
46	0,7998	0,7635	0,6856	0,2887	0,2989	0,2295	0,6913	0,6414	0,6168	0,3549	0,3793	0,3190	0,7788	0,7328	0,7372
47	0,8009	0,7646	0,6851	0,2904	0,3002	0,2294	0,6933	0,6434	0,6167	0,3573	0,3810	0,3192	0,7804	0,7345	0,7371
48	0,8020	0,7657	0,6847	0,2922	0,3014	0,2293	0,6951	0,6452	0,6166	0,3596	0,3826	0,3193	0,7819	0,7361	0,7370
49	0,8031	0,7667	0,6842	0,2938	0,3025	0,2292	0,6969	0,6470	0,6164	0,3618	0,3841	0,3195	0,7833	0,7376	0,7369
50	0,8041	0,7677	0,6838	0,2954	0,3036	0,2291	0,6986	0,6487	0,6162	0,3640	0,3856	0,3196	0,7848	0,7391	0,7367
51	0,7624	0,7202	0,6812	0,2721	0,2726	0,2270	0,6497	0,5948	0,6090	0,3168	0,3254	0,3101	0,7053	0,6522	0,7222
52	0,7433	0,6994	0,6795	0,2601	0,2614	0,2262	0,6305	0,5751	0,6069	0,2989	0,3081	0,3061	0,6824	0,6284	0,7179
53	0,7166	0,6711	0,6745	0,2464	0,2483	0,2245	0,6075	0,5516	0,6032	0,2779	0,2879	0,3001	0,6541	0,5996	0,7115
54	0,6891	0,6437	0,6638	0,2312	0,2337	0,2215	0,5817	0,5254	0,5971	0,2577	0,2686	0,2900	0,6278	0,5728	0,6991
55	0,6602	0,6132	0,6481	0,2140	0,2173	0,2165	0,5525	0,4969	0,5878	0,2359	0,2473	0,2767	0,5997	0,5435	0,6819
56	0,6236	0,5744	0,6250	0,1945	0,1980	0,2086	0,5173	0,4619	0,5728	0,2101	0,2220	0,2603	0,5648	0,5074	0,6600
57	0,5778	0,5262	0,5951	0,1733	0,1766	0,1967	0,4784	0,4231	0,5496	0,1813	0,1927	0,2388	0,5231	0,4638	0,6290
58	0,5208	0,4684	0,5545	0,1513	0,1547	0,1805	0,4397	0,3833	0,5171	0,1493	0,1601	0,2134	0,4712	0,4123	0,5892
59	0,4572	0,4047	0,5037	0,1281	0,1308	0,1620	0,3959	0,3390	0,4787	0,1171	0,1278	0,1853	0,4150	0,3576	0,5414
60	0,3859	0,3348	0,4447	0,1039	0,1061	0,1424	0,3486	0,2907	0,4355	0,0873	0,0966	0,1559	0,3554	0,2988	0,4846
61	0,3124	0,2643	0,3813	0,0798	0,0816	0,1214	0,2971	0,2415	0,3866	0,0652	0,0716	0,1272	0,2993	0,2465	0,4249
62	0,2449	0,2016	0,3198	0,0583	0,0598	0,1003	0,2455	0,1931	0,3353	0,0514	0,0567	0,1039	0,2543	0,2084	0,3699
63	0,1984	0,1617	0,2677	0,0436	0,0447	0,0810	0,2042	0,1590	0,2850	0,0425	0,0472	0,0866	0,2210	0,1799	0,3226
64	0,1667	0,1354	0,2282	0,0350	0,0362	0,0662	0,1740	0,1343	0,2427	0,0358	0,0402	0,0740	0,1956	0,1578	0,2859
65	0,1412	0,1139	0,1959	0,0292	0,0305	0,0554	0,1507	0,1158	0,2097	0,0304	0,0346	0,0642	0,1742	0,1401	0,2545
66	0,1222	0,0985	0,1715	0,0248	0,0260	0,0474	0,1314	0,1008	0,1846	0,0263	0,0301	0,0568	0,1584	0,1271	0,2312
67	0,1093	0,0876	0,1546	0,0213	0,0226	0,0411	0,1168	0,0892	0,1633	0,0233	0,0271	0,0515	0,1478	0,1182	0,2139
68	0,1005	0,0805	0,1426	0,0183	0,0196	0,0362	0,1046	0,0798	0,1462	0,0210	0,0248	0,0477	0,1410	0,1126	0,2014
69	0,0966	0,0775	0,1349	0,0160	0,0175	0,0327	0,0960	0,0733	0,1333	0,0203	0,0239	0,0452	0,1381	0,1103	0,1931
70	0,0946	0,0755	0,1296	0,0139	0,0155	0,0299	0,0889	0,0681	0,1240	0,0198	0,0234	0,0436	0,1364	0,1088	0,1876
71	0,7156	0,6623	0,4065	0,4408	0,4221	0,1694	0,8427	0,7991	0,5197	0,5471	0,5358	0,2432	0,8995	0,8653	0,6434
72	0,7730	0,7249	0,4789	0,3927	0,3739	0,1679	0,8163	0,7679	0,5195	0,4991	0,4867	0,2414	0,8805	0,8413	0,6430

Tabela 5.28. Wyniki badań miar dwuelementowych dla korpusu *KIPI* poddanemu podpróbkowaniu klasy negatywnej do 95%, część 4.

Wyniki badań jakości rozwiązań generowanych przez sieci neuronowe

Wyniki badań dla sieci neuronowych po podpróbkowaniu klasy negatywnej

5.5. MIARY WIELOELEMENTOWE, KORPUS KIPI

5.5.1. Przygotowanie i zbadanie podkorpusów KIPI

5.5.2. Szczegółowy opis przebiegu tej części badań

5.5.3. Wyniki

5.6. METODY MASZYNOWEGO UCZENIA, KORPUS KIPI

5.6.1. Opis sieci neuronowa

Wykorzystana w badaniach sieć neuronowa została opisana we wcześniejszej części tej pracy przy okazji omawiania pakietu narzędzi *MWeXtractor*.

5.6.2. Przebieg badań

5.6.3. Wyniki dla kolokacji dwuelementowych

5.6.4. Wyniki dla kolokacji wieloelementowych

5.7. PODEJŚCIE MIAR MIESZANYCH

Podejście miar mieszanych polegało na wykorzystaniu kombinacji liniowe rankingów wygenerowanych przez pewien zestaw poszczególnych funkcji asocjacyjnych.

5.7.1. Optymalizacja wag dla rankingów w modelu kombinacji liniowej

Narzędziem wykorzystanym w optymalizacji był opisany we wcześniejszej części tej pracy algorytm ewolucyjny. Jako zbiór uczący wybrano korpus *KIPI*, a testowym był cały zbiór *KGR7*. Oba wykorzystane zbiory danych były przygotowane do tego zadania w inny sposób niż opisano w części pracy traktującej o wykorzystanych zbiorach danych. Nie były one tagowane od nowa, a jedynie zastosowana była konwersja tagsetów z *KIPI* do *NKJP*. Zauważyć jednak należy, że zarówno zbiór testowy jak i uczący były przygotowane w taki sam sposób.

Proces optymalizacji wag był wykonany dwuetapowo. Pierwszym krokiem było dostrajanie parametrów algorytmu genetycznego, a drugim optymalizacja wag dla rankingów. Parametry, które autor pracy optymalizował to: szansa krzyżowania, rozmiar turnieju – dotyczy operatora selekcji, szansa na mutację genotypu osobnika. Zbadany zostały wartości parametrów przedstawione w tabeli ??.

parametr	wartości
rozmiar turnieju	[2:5], skok co 1
szansa krzyżowania	[0.4:0.8], skok co 0.05
szansa mutacji	[0.03:0.11], skok co 0.01

Tabela 5.29. Sprawdzone wartości parametrów algorytmu genetycznego

Optymalizacja parametrów algorytmu została wykonana na korpusie uczącym przy progu odcięcia na poziomie dziesięciotysięcznej pozycji rankingu. Każda wartość parametru była dostrajana przez 50 iteracji dla rozmiaru populacji równego 25. Konkretny parametr był dostrajany przy utrzymaniu stałych wartości innych parametrów przez cały proces jego optymalizacji. Ręczna ocena wyników polegała na sprawdzeniu, dla których parametrów wzrost jakości rozwiązania jest najszybszy, ale przy jednoczesnym stosunkowo wolnym zbieganiu się najlepszego, średniego i najgorszego z rozwiązań. Ostatecznie wybrany został zestaw parametrów zaprezentowany poniżej:

$$\text{rozmiarturnieju} = 5, \text{szansakrzyżowania} = 0.75, \text{szansamutacji} = 0.05$$

Po ustaleniu parametrów dla algorytmu ewolucyjnego wykonany został drugi krok – optymalizacja wag dla rankingów. Optymalizacja wag rankingów polegała na wybraniu zestawu miar, które dawały dobre wyniki w ekstrakcji miar dwuelementowych, a następnie wykorzystaniu ich w procesie tworzenia rankingów do kombinacji liniowej. Dobrane zostały następujące miar:

1. *Loglikelihood*,
2. *Mutual Expectation*,
3. *Specific Frequency Biased Mutual Dependency*,
4. *Jaccard*,
5. *W Specific Correlation*,
6. *W Specific Exponential Correlation* z parameterami 1.35, 1.375 oraz 1.4 stosowanymi zamiennie.

Algorytm genetyczny dostrajał wagi przeszukując przestrzeń rozwiązań, wybierając jakiś zestaw wag i sprawdzając rozwiązanie. Zastosowany agregator to *maksymalna suma*, a funkcje normalizujące to *Borda score* oraz *Zipf's Borda score*. Użyta miara oceny to *Average precision on hit* przy długości rankingu równej 10000. Liczba osobników w populacji to 25, a liczba iteracji była równa przynajmniej 100.

Podjętych zostało wiele prób optymalizacji wag dla dwóch zakresów możliwych wag dla każdego z rankingów – od 0 do 1 oraz od -1 do 1, dwóch różnych sposobów punktowania rankingów oraz dla zmiennej wartości parametru funkcji *W Specific Exponential Correlation*.

5.7.2. Wyniki i obserwacje po zastosowaniu optymalizacji

Badania wykazały bardzo słaby wynik dla zbioru testowego, w uczącym wynik był poprawiony o około 8%, ale w przypadku testowego uplasował się na pozycji około 20-25% gorszym niż najlepsze funkcje.

5.8. ZBIORCZE WYNIKI BADAŃ ORAZ WNIOSKI

SPIS RYSUNKÓW

3.1.	Słowo <i>strong</i> w odniesieniu do słowa <i>opposition</i>	17
3.2.	Słowo <i>strong</i> w odniesieniu do słowa <i>support</i>	17
3.3.	Słowo <i>strong</i> w odniesieniu do słowa <i>for</i>	18
4.1.	Przykładowy schemat procesu badania jakości rankerów	36
4.2.	Przykładowy schemat procesu wydobywania wyrażień wielowyrazowych	39

SPIS TABEL

3.1.	Przykład wykorzystania okna przesuwne	15
3.2.	Wyniki ekstrakcji kolokacji dla metody wariancji i odległości słów	19
3.3.	Wyniki badań dwóch generalizacji funkcji <i>Mutual Information</i>	27
4.1.	Parametry programu <i>Miner</i>	38
4.2.	Parametry programu <i>FeatureMaker</i>	40
4.3.	Parametry programu <i>Digger</i>	41
4.4.	Parametry dodatkowe programu <i>Digger</i>	43
4.5.	Format składowania krotek w pliku tekstowym	44
4.6.	Przykładowe indeksy dla 4-hipermacierzy krotek	46
4.7.	Tablica wielodzielna dla krotki trójelementowej	48
4.8.	Tablica wielodzielna dla krotki trójelementowej	50
5.1.	Podkorporusy i statystyki korpusu <i>KGR7</i>	70
5.2.	72
5.3.	Statystyki podzbioru danych <i>KIPI 2W</i>	73
5.4.	Statystyki podzbioru danych <i>KIPI 2R</i>	74
5.5.	Statystyki podzbioru danych <i>KIPI 2W1H</i>	75
5.6.	Statystyki podzbioru danych <i>KIPI 2R1H</i>	77
5.7.	Zestaw przeprowadzonych badań dla funkcji dwuelementowych na korpusie <i>KIPI</i> . .	80
5.8.	Zestaw zbadanych funkcji dwuelementowych na korpusie <i>KIPI</i>	81
5.9.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> , część 1	82
5.10.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> , część 2	83
5.11.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> , część 3	84
5.12.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> , część 4	85
5.13.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 10 części, i poddanemu dyspersji miarą TF-IDF, część 1	88
5.14.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 10 części, i poddanemu dyspersji miarą TF-IDF, część 2	89
5.15.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 10 części, i poddanemu dyspersji miarą TF-IDF, część 3	90
5.16.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 10 części, i poddanemu dyspersji miarą TF-IDF, część 4	91
5.17.	Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 20 części, i poddanemu dyspersji miarą TF-IDF, część 1	93

5.18. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 20 części, i poddanemu dyspersji miarą TF-IDF, część 2	94
5.19. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 20 części, i poddanemu dyspersji miarą TF-IDF, część 3	95
5.20. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> podzielonego na 20 części, i poddanemu dyspersji miarą TF-IDF, część 4	96
5.21. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 80%, część 1	98
5.22. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 80%, część 2	99
5.23. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 80%, część 3	100
5.24. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 80%, część 4	101
5.25. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 95%, część 1	103
5.26. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 95%, część 2	104
5.27. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 95%, część 3	105
5.28. Wyniki badań miar dwuelementowych dla korpusu <i>KIPI</i> poddanemu podpróbkowaniu klasy negatywnej do 95%, część 4	106
5.29. Sprawdzone wartości parametrów algorytmu genetycznego	108

BIBLIOGRAFIA

- [1] Christopher D. Manning i Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, Massachusetts Institute of Technology, wydanie drugie, 1999.
- [2] Stefan Evert, *The Statistics of Word Cooccurrences Word Pairs and Collocations*, Uniwersytet Struttgart, 2004.
- [3] Pavel Pecina i Pavel Schlesinger, *Combining Association Measures for Collocation Extraction*, Charles University, Prague, 2006.
- [4] Aleksander Buczyński, *Pozyskiwanie z Internetu tekstów do badań lingwistycznych*, Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski, 2004.
- [5] Pavel Pecina, *Lexical association measures and collocation extraction*, 2009.
- [6] Pavel Pecina, *Reference Data for Czech Collocation Extraction*, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- [7] Firth, J. R. *A synopsis of linguistic theory 1930–55. In Studies in linguistic analysis*, The Philological Society, Oxford. 1957.
- [8] Adam Radziszewski, Adam Wardyński, Tomasz Śniatowski, *WCCL: A Morpho-syntactic Feature Toolkit*, Institute of Informatics, Wrocław University of Technology, 2011.
- [9] Frank Smadja, *Retrieving Collocations from Text: Xtract*, Columbia University, 1993.
- [10] Anca Dinu, Liviu P. Dinu, Ionut T. Sorodoc, *Aggregation methods for efficient collocation detection*, University of Bucharest, 2014.
- [11] <http://plwordnet.pwr.wroc.pl/wordnet/>.
- [12] Lek-Heng Lim, *Tensors and hypermatrices w Handbook of Linear Algebra* pod redakcją L. Hogben (Ed.), wydanie drugie, CRC Press, Boca Raton, FL, 2013.
- [13] Aristomenis Thanopoulos, Nikos Fakotakis, George Kokkinakis, *Comparative Evaluation of Collocation Extraction Metrics*, Wire Communications Laboratory Electrical i Computer Engineering Dept., University of Patras, 265 00 Rion, Patras, Greece.
- [14] Tim Van de Cruys, *Two Multivariate Generalizations of Pointwise Mutual Information*, University of Cambridge, United Kingdom.
- [15] Joaquim Ferreira da Silva, Gabriel Pereira Lopes *A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora*, Universidade Nova de Lisboa.
- [16] Mariusz Paradowski, *Opracowanie formalnej analizy zależności pomiędzy współczynnikami służącymi do wykrywania wyrażeń wielowyrazowych oraz na tej podstawie opracowanie współczynnika uogólniającego*, 2014.
- [17] Sasa Petrovic, Jan Snajder, Bojana Dalbelo Basic, *Extending lexical association measures for collocation extraction*, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia, 2009.

- [18] Michael W. Berry, *Survey of Text Mining - Clustering, Classification, and Retrieval*, TODO 2003.
- [19] *Mutual Expectation: A Measure for Multiword Lexial Unit Extraction*
- [20] Łukasz Kłyk, *Metody sztucznej inteligencji w zwiększaniu skuteczności klasyfikatora*, Praca dyplomowa - magisterska, Politechnika Wrocławska, Wrocław 2013.
- [21] Strona internetowa autorów korpusu KIPi, www.korpus.pl, dostępna dnia 2014-11-29.
- [22] Adam Przepiórkowski *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*, IPI PAN, Warszawa 2004.
- [23] Stefan Th. Fries, *Dispersions and adjusted frequencies in corpora*, University of California, Santa Barbara.