

Spis treści

WSTĘP	13
WYSZUKIWANIE INFORMACJI W SIECI ROZLEGŁEJ: KATALOGI STRON WWW A WYSZUKIWARKI INTERNETOWE.....	16
CEL I STRUKTURA PRACY.....	21
ZWIĄZKI NLP Z INFORMACJĄ NAUKOWĄ	25
1.1. WPROWADZENIE TEORETYCZNE DO PRZETWARZANIA JĘZYKA NATURALNEGO.....	25
1.2.1. Ustalenia terminologiczne związane z nazwą badań nad tekstami języka naturalnego.....	26
1.2. WYBRANE KIERUNKI BADAWCZE.....	28
1.2.1. Wyszukiwanie informacji w dokumentach.....	28
1.2.2. Grupowanie dokumentów (klasteryzacja).....	33
1.2.2.1. Grupowanie oparte o wzorce.....	36
1.2.2.2. Grupowanie bezwzorcowe.....	36
USTALENIA TERMINOLOGICZNE ORAZ WYBRANE METODY KOMPUTEROWEGO PRZETWARZANIA JĘZYKA NATURALNEGO	39
2.1. TERMINY PRZYJĘTE W KSIĄŻCE.....	41
2.2. ANALIZA KWANTYTATYWNA TEKSTÓW.....	45
2.2.1. Jednostki badania kwantytatywnego tekstów.....	49
2.2.2. Cechy statystyczne jednostek leksykalnych.....	54
2.2.3. Zależności leksykalne.....	57

2.3.	WYBRANE METODY REPREZENTACJI TREŚCI DOKUMENTÓW.....	62
2.3.1.	Zbiór słów (<i>bag-of-words</i>).....	65
2.3.2.	Lista frekwencyjna.....	66
2.3.3.	Reprezentacja wektorowa.....	67
2.4.	WYBRANE SPOSOBY OKREŚLANIA WAGI SŁÓW.....	71
2.5.	OPTIMALIZACJA LINGWISTYCZNA TREŚCI DOKUMENTU.....	73
2.5.1.	Przygotowanie dokumentów do indeksowania treści.....	73
2.5.2.	Usunięcie wyrazów mało znaczących.....	74
2.5.3.	Wyznaczanie rdzenia wyrazu.....	76
2.5.3.1.	Metody wskazywania wspólnego rdzenia.....	78
2.5.4.	Wskazywanie lematu słowoformy.....	80

ZASADY POSTĘPOWANIA BADAWCZEGO I OPIS PRZYGOTOWANEGO

SYSTEMU.....	83	
3.1.	PRZEDMIOT, CEL I METODOLOGIA BADAŃ.....	84
3.1.1.	Przedmiot badań.....	84
3.1.2.	Cele i hipotezy badawcze.....	85
3.2.	ZASTOSOWANE METODY, TECHNOLOGIE I NARZĘDZIA BADAWCZE....	88
3.2.1.	Zastosowane technologie.....	88
3.2.2.	Język Python.....	90
3.3.	ORGANIZACJA I PRZEBIEG BADAŃ.....	92
3.3.1.	Przygotowanie dokumentów do analizy.....	92
3.3.2.	Klasyfikacja zawartości pliku.....	96
3.3.3.	Usunięcie wyrazów nierelevantnych.....	100
3.3.4.	Ustalenie podstawowej postaci wyrazów.....	101
3.3.5.	Zliczenie wystąpień danego słowa.....	105
3.3.6.	Analiza słów wyróżnionych.....	105
3.3.7.	Metody ustalania wagi słów.....	106
3.3.8.	Porównanie zestawów słów kluczowych ustalanych tradycyjnie i automatycznie.....	108
3.4.	PREZENTACJA MATERIAŁU BADAWCZEGO.....	109
3.4.1.	Korpus tekstów.....	109
3.4.2.	Teksty z zakresu informacji naukowej i bibliologii.....	110
3.4.2.1.	Artykuły z czasopism.....	114
3.4.2.2.	Artykuły z materiałów konferencyjnych.....	120
3.4.3.	Subkorpus ekonomia i zarządzanie.....	122
3.4.4.	Słowa kluczowe.....	123
3.4.4.1.	Słowa kluczowe wybierane przez autorów.....	123

3.4.4.2. Słowa kluczowe wskazane przez indeksatorów.....	126
3.4.4.3. Słowa kluczowe generowane automatycznie.....	129

ANALIZA ORAZ INTERPRETACJA MATERIAŁU BADAWCZEGO

I WYNIKÓW BADAŃ.....	133
4.1. ANALIZA GŁÓWNEGO KORPUSU TEKSTÓW.....	133
4.1.1. Czasopisma.....	133
4.1.2. Materiały konferencyjne.....	139
4.1.3. Analiza całego korpusu.....	142
4.2. ANALIZA KORPUSU POMOCNICZEGO.....	149
4.3. SŁOWA KLUCZOWE UZYSKANE W WYNIKU INDEKSOWANIA TRADYCYJNEGO I AUTOMATYCZNEGO.....	152
4.3.1. Waga słów wyróżnionych w tekście.....	159
4.3.2. Słowa kluczowe wskazywane automatycznie.....	160
4.4. OCENA ZASTOSOWANYCH METOD USTALANIA WAGI SŁOWA.....	163
PODSUMOWANIE.....	165
POSTULATY TECHNOLOGICZNE.....	170
Standardy metainformacji.....	170
Formaty zapisu dokumentów.....	170
PROPOZYCJE DALSZYCH BADAŃ.....	172
BIBLIOGRAFIA.....	173
SPIS TABEL.....	185
SPIS ILUSTRACJI.....	189
SPIS WYKRESÓW.....	191
INDEKS RZECZOWY.....	193

Table of contents

INTRODUCTION.....	13
INFORMATION SEARCHING IN GLOBAL NETWORK: WEB-SITES CATALOGUES AND SEARCH ENGINES.....	16
AIM AND STRUCTURE OF BOOK.....	21
NLP AND INFORMATION SCIENCE.....	25
1.1. INTRODUCTION TO THEORY OF NATURAL LANGUAGE PROCESSING...25	
1.2.1. Terminology connected with the name of natural language texts research.....	26
1.2. CHOSEN RESEARCH DIRECTIONS.....	28
1.2.1. Information retrieval in documents.....	28
1.2.2. Documents clustering.....	33
1.2.2.1. Clustering based on standards.....	36
1.2.2.2. Clustering non-based on standards.....	36
USED TERMINOLOGY AND AUTOMATIC NLP TOOLS.....	39
2.1. TERMS USED IN BOOK.....	41
2.2. QUANTITATIVE TEXTS ANALYSIS.....	45
2.2.1. Entities of quantitative texts analysis.....	49
2.2.2. Statistical properties of lexical units.....	54
2.2.3. Lexical relations.....	57
2.3. CHOSEN METHODS OF TEXT REPRESENTATION.....	62
2.3.1. Bag-of-words.....	65

2.3.2.	Frequency list.....	66
2.3.3.	Vector model.....	67
2.4.	WORDS WEIGHTING RULES.....	71
2.5.	LINGUISTIC OPTIMISATION OF DOCUMENT TEXT.....	73
2.5.1.	Preparing documents to indexing.....	73
2.5.2.	Removing low frequency words.....	74
2.5.3.	Stemming.....	76
2.5.3.1.	Methods of stemming.....	78
2.5.4.	Lamas.....	80
RESEARCH METHODOLOGY AND RESEARCH SYSTEM DESCRIPTION.....		83
3.1.	SUBJECT, GOAL AND RESEARCH METHODOLOGY.....	84
3.1.1.	Subject of research.....	84
3.1.2.	Research goals and hypothesis.....	85
3.2.	USED METHODS, TECHNOLOGIES AND RESEARCH TOOLS.....	88
3.2.1.	Used technologies.....	88
3.2.2.	Python programming language.....	90
3.3.	PREPARING AND PROGRESS OF RESEARCH.....	92
3.3.1.	Pre-preparing of documents.....	92
3.3.2.	Classification of file contents.....	96
3.3.3.	Removing of irrelevant words.....	100
3.3.4.	Common form of words.....	101
3.3.5.	Word appearance counting.....	105
3.3.6.	Analysis of distinguished words.....	105
3.3.7.	Word weighting methods.....	106
3.3.8.	Comparison of automatic and traditional indicated keywords sets.....	108
3.4.	RESEARCH MATERIAL DESCRIPTION.....	109
3.4.1.	Texts corpora.....	109
3.4.2.	Information sciences texts corpora.....	110
3.4.2.1.	Articles from magazines.....	114
3.4.2.2.	Articles from conference papers.....	120
3.4.3.	Economy and management texts sub-corpora.....	122
3.4.4.	Keywords.....	123
3.4.4.1.	Keywords indicated by authors.....	123
3.4.4.2.	Keywords indicated by indexators.....	126
3.4.4.3.	Keywords indicated automatically.....	129

ANALYSIS AND INTERPRETATION OF RESEARCH MATERIAL AND RESULTS...	133
4.1. MAIN TEXTS CORPORA ANALYSIS.....	133
4.1.1. Magazines.....	133
4.1.2. Conference papers.....	139
4.1.3. Full corpora analysis.....	142
4.2. SUB-CORPORA ANALYSIS.....	149
4.3. KEYWORDS INDICATED ON THE BASIS OF TRADITIONAL AND AUTOMATIC INDEXING.....	152
4.3.1. Weight of distinguished words.....	159
4.3.2. Automatic indicated keywords.....	160
4.4. EVALUATION OF USED METHODS OF WORD WEIGHTING.....	163
CONCLUSIONS.....	165
TECHNOLOGICAL POSTULATES.....	170
Metainformation standards.....	170
Documents formats.....	170
FURTHER RESEARCH SUGGESTIONS.....	172
LITERATURE.....	173
TABLES INDEX.....	185
ILLUSTRATIONS INDEX.....	189
CHARTS INDEX.....	191
SUBJECT INDEX.....	193



Piotr Malak jest asystentem w Instytucie Informacji Naukowej i Bibliologii UMK w Toruniu oraz członkiem Polskiego Towarzystwa Informatycznego. Jego zainteresowania badawcze dotyczą zarządzania informacją, wyszukiwania informacji w dokumentach, inżynierii lingwistycznej oraz zarządzania czasem i zadaniami. Entuzjasta i praktyk lifehacking'u oraz efektywnego zarządzania czasem.

Prowadził wykłady gościnne na Uniwersytecie w Ankarze, Hogeschool van Amsterdam w Amsterdamie oraz Uniwersytecie Wileńskim.

Książka prezentuje wyniki badań porównawczych nad skutecznością metod automatycznych i kognitywnych w tworzeniu charakterystyk wyszukiwawczych za pomocą słów kluczowych. Na potrzeby badań wykorzystany został autorski system analizy kwantytatywnej tekstów języka polskiego posługujący się metodami statystycznymi do ustalenia i oceny frekwencji wyrażen językowych w korpusie tekstów.

We wprowadzeniu teoretycznym czytelnik płynnie przechodzi od teorii badań nad przetwarzaniem języka naturalnego, poprzez problematykę nazewnictwa, omówienie jednostek badania kwantytatywnego tekstów, cech statystycznych jednostek leksykalnych i wybrane sposoby reprezentacji treści dokumentów do metod optymalizacji tekstu na potrzeby automatycznego przetwarzania.

Książka, z założenia obejmująca zaawansowane zagadnienia wyszukiwania informacji, kierowana jest do badaczy i pracowników sektora informacyjnego oraz użytkowników informacji cyfrowej. Świetnie sprawdzi się także jako pomoc podczas zajęć poświęconych wyszukiwaniu informacji oraz w codziennej praktyce dla osób związanych z informacją cyfrową, bibliotekarzy i badaczy procesów przetwarzania informacji.