

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 nerelewantnych.....	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 DAJSZYCH BADAŃ.....	172
BIBLIOGRAFIA.....	173
SPIS TABEL.....	185
SPIS ILUSTRACJI.....	189
SPIS WYKRESÓW.....	191
INDEKS RZECZOWY.....	193