

Obsah:

0	Úvod	6
1	Znalosti pro udržení konkurenceschopnosti	8
1.1	Základní definice	8
1.2	Typy znalostí	8
1.3	Kvalita informací	10
1.4	Množství informací	11
1.5	Zdroje informací	11
2	Strojové učení v manažerském rozhodování	12
2.1	Information Retrieval (IR)	12
2.2	Data/Text mining (DM/TM)	14
3	Proces zpracování textů	16
3.1	Reprezentace termů	18
3.2	Redukce dimenze prostoru termů	20
3.2.1	Selekce termů	21
3.2.2	Extrakce termů	22
3.3	Kategorizace dokumentů	24
3.3.1	Metoda nejbližšího souseda (K-NN)	24
3.3.2	Měření úhlu mezi vektory (VM)	25
3.3.3	Naivní Bayesův algoritmus (NB)	26
3.4	Shluková analýza	28
3.4.1	Hierarchické shlukování	28
3.4.2	Nehierarchické shlukování	29
3.5	Testování modelů	32
4	Textová kategorizace - experimenty	35
4.1	Česká datová sada - experimenty	35
4.1.1	Výsledky jednotlivých experimentů	37
4.1.2	Vstup 1 – 50 dokumentů, 25 rel a 25 non	38
4.1.3	Vstup 2 – 35 dokumentů, 10 rel a 25 non	39
4.1.4	Vstup 2 – 35 dokumentů, 25 rel a 10 non	39
4.1.5	Celkové hodnocení	40
4.2	Výběr z datasetu 20 Newsgroup	41
5	Shlukování - experimenty	43
5.1	Seminární práce	43
5.2	Výběr z datové sady 20 Newsgroup	43
6	Nástroje pro text mining	45
6.1	Modelování a vyhodnocení v systému WEKA	47
6.2	Modelování a vyhodnocení v systému Rapid Miner	50
6.3	Modelování a vyhodnocení v systémech WordStat a QDA Miner	55
6.4	Modelování a vyhodnocení v systému TextAnalyst	61
6.5	Předzpracování dat v systému ConvertDoc	65
7	Hodnocení vybraných nástrojů	67
7.1	WEKA a Rapid Miner	67
7.2	WordStat a TextAnalyst	67
8	Závěr	69