FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

Klasifikace textových dokumentů bez učitele

Jaromír Novotný[1]

1 Úvod

Cílem je příprava vstupních textových dat a následná klasifikace dokumentů za použití metod učení bez učitele. Jedna důležitá část předzpracování dat je převedení vstupních textů do tvaru vektorové reprezentace. Existuje mnoho klasifikačních metod, z nichž jsou vybrány dvě – "klasický" K-means shlukovací algoritmusaLatent Dirichlet Allocation (LDA) přizpůsobená ke klasifikaci dokumentů.

Nakonec vybrané klasifikační metody jsou porovnány s již používanými metodami a i s metodami využívající informaci učitele. Porovnání je provedeno na $20NewsGroup$ anglickém data setu složeného z elektronické korespondence. Pro zajímavost jsou metody ozkoušeny na CNO českém data setu složeného z online novinových článků.

2 Příprava data setů, jejich zpracování a ohodnocení

Pro účely klasifikace byl vybrán často používaný anglický data set $20NewsGroup$. Menší data sety ($Binary20NG$, $20NG$, $10NG$, $Binary_{0,1,2}$, $5Multi_{0,1,2}$, $10Multi_{0,1,2}$) podle Slonim et al. (2002) získány z původního $20NewsGroup$. Další data set CNO a byl získán z databáze novinových článků $JMZW$ Západočeské Univerzity (pouze malá část z celkové databáze). Před tím, než se vytvoří reprezentační vektory je důležité řádně data sety (texty) předzpracovat. Nejdříve všechny znaky velkých písmen jsou převedena na znaky malých písmen a veškeré znaky čísel jsou převedena na jednotný symbol. Dále je provedena lemmatizace, na což byl použit algoritmus MorphoDiTa (2017) zpřístupněný v pythonu přes balíček ufal. Pro odstranění stop slov je využita inverzní frekvence dokumentů (idf), po té zbylá slova mohou být použita jako reprezentanty v reprezentačních vektorech. Další možnost je vybrání n nejlepších slov jako reprezentanty podle Vzájemné Informace (Mutual Information – MI) podle Siolas et al. (2000). Předzpracovaná vstupní data lze převést do gensim gensim (2017) *slovníku* a vytvoření *korpusu* (použitím doc2bow funkce gensim (2017) slovníku přes všechny reprezentační vektory lemmat). Tento *slovníku* a *korpusu* následně využívá LDA metoda. Metoda $K-means$ využívá matici $tf-idf$ vah, jenž lze vytvořit z předzpracovaných vstupních dat (k vytvoření je využit python balíček sklearn). Dále je provedeno snížení dimenze získané matice vah za pomoci metody Latentní sémantické analýzy LSA (využit modul TruncatedSVD z python balíčku sklearn). Výstupní matice z metody LSA je následně používána jako vstup metody $K-means$. Z mnoha možností ohodnocení konečných výsledků použitých metod byly vybrány následující: Accuracy, přesnost (P) a úplnost (R). Míry P a R jsou počítány jako mikro-průměry podle Slonim et al. (2002).

[1] student navazujícího doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika a řídící technika, specializace Umělá inteligence, e-mail: fallout7@kky.zcu.cz

3 Výsledky

Tabulka 1: Výsledky testovaných metod a metod v Siolas et al. (2000).

Metody	Accuracy míra [%]
KNN+P	80.13
SVM+P	88.52
LDA	56.46
K-Means+LSA	75.47

Tabulka 2: Výsledky metod na CNO data setu.

	Přesnost P metod [%]		
	LDA	K-means+LSA (dim. 2000)	K-means+LSA (dim. 31)
CNO	14.67	42.59	41.72

$20NewsGroup$ sub-sets	Přesnost P metod [%]				
	sIB	sK-means	LDA	K-means+LSA	K-means+LSA n repr.)
Průměr z $20NG$ & $10NG$	68.50	65.20	22.84	41.69	43.43
Průměr z *malých*	83.30	47.60	55.02	74.74	77.28

Tabulka 3: Porovnání testovacích metod s metodami v Slonim et al. (2002).

Průměr z *malých* v Tabulce 3 obsahuje pod-sety $Binary_{0,1,2}$, $5Multi_{0,1,2}$, $10Multi_{0,1,2}$.

4 Závěr

Byl předveden postup jak připravit vstupní textová data pro použití zpracování textu (hlavně v případě shlukování dokumentů se stejným nebo podobným tématem). Metoda LDA, jak je vidět v Tabulkách 1, 2 a 3, nedosahuje vhodných výsledků (v aktuální podobě). V Tabulce 1 se porovnává náš přístup učení bez učitele s metodami učení s učitelem a se sémantickými kernely. V Tabulce 2 lze vidět výsledky našich metod na CNO data setu. Výsledné hodnoty v Tabulce 3 zobrazují porovnání s metodami Slonim et al. (2002). Je vidět, že náš postup zlepšuje klasické $K-means$, ovšem zaostává za sIB (sequential Information Bottleneck). V budoucnu bude dobré se zaměřit na lepší přípravu vstupních dat, vyzkoušet jiné metody a popřípadě využít kombinaci více metod pro vylepšení jejich jednotlivých nedostatků.

Poděkování

Příspěvek byl podpořen grantovým projektem SVK1-2017-021

Literatura

Slonim, N. F., N. T. (2002) *Unsupervised Document Classification using Sequential Information Maximization*, Jerusalem, Israel, SIGIR'02

G. Siolas, F. d'Alché-Buc (2000)*Support Vector Machines based on a Semantic Kernel for Text Categorization*, Universitk Pierre et Marie Curie.

Python balíček gensim dostupný na: `https://radimrehurek.com/gensim/`.

MorphoDiTa, Institute of Formal and Applied Linguistics, Charles University, Czech Republic Faculty of Mathematics and Physics, `http://ufal.mff.cuni.cz/morphodita`