Klasifikace telefonických hovorů mezi volajícím a operátorem podle jejich témat

Jaromír Novotný[1]

1 Úvod

Cílem experimentu je otestovat klasifikační metody – s učením s učitelem i s učením bez učitele – použité ke kategorizaci textových přepisů telefonických hovorů podle jejich témat. Byly vybrány dvě základní metody: Lineární Support Vector Machine (SVM) a K-means. Výsledky obou metod jsou v tomto experimentu porovnány a ohodnoceny.

2 Použitá data a jejich příprava

Datová množina používaná v experimentu byla vytvořena přepisy telefonických hovorů získaných centrem jazykové poradny (CJP) Ústavu českého jazyka Akademie Věd České republiky. Tato množina je vytvořena z unikátních jazykových dat v Českém jazyce a to tak, že jsou nahrávány telefonické hovory mezi volajícím a operátorem CJP. Hovory se týkají dotazů ohledně gramatiky v českém jazyce.

Příprava dat začíná zmenšením veškerých velkých znaků na malé a všechny číselné znaky jsou nahrazeny univerzálním symbolem. Dále je provedena lemmatizace (MorphoDiTa Straková et al. (2014) volně dostupný balíček pro Python a proces odstranění stop-slov (vybrání nejlepších slov s nejvyšší hodnotou mutual information – MI). Ve chvíli, kdy máme takto připravená data, lze provést jejich reprezentaci a to za pomoci TF-IDF vah (počítané stejným způsobem jako v Novotný et al. (2017)) a doc2vec vah (popsáno v Lau et al. (2016)). Posléze je též provedena redukce dimenzí za pomoci metody Latent Semantic Analysis (LSA).

3 Klasifikační metody a možnosti jejich ohodnocení

Jako zástupce metody typu učení s učitelem byla vybrána základní Lineární SVM metoda a jako zástupce metod typu učení bez učitele byla vybrána základní metoda K-means.

Pro provedené experimenty byla vybrána nejednodušší míra a to přesnost (Accuracy).

4 Experiment

Všechna data v Tabulce 1 a 2 jsou tvořena přepisy rozhovorů mezi operátorem a volajícím: *PCT (přepisy telefonických hovorů) mono* – pouze mono nahrávky obsahující 607 částí hovorů rozdělených do 20 kategorií; *PCT mono malé* – pouze mono nahrávky obsahující 504 částí hovorů rozdělených do 8 kategorií; *PCT stereo* – pouze stereo nahrávky obsahující 3128 částí hovorů rozdělených do 20 kategorií; *PCT stereo malé* – pouze stereo nahrávky obsahující 2866 částí hovorů rozdělených do 10 kategorií; *PCT vše* – jak mono tak stereo nahrávky obsahující 3713 částí hovorů rozdělených do 20 kategorií; *PCT vše malé* – jak mono tak stereo

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

nahrávky obsahující 3343 částí hovorů rozdělených do 10 kategorií.

	Přesnost metod [%] Lineární SVM metoda s reprezentacemi				
	TF-IDF	*TF-IDF (LSA)*	*doc2vec*	*doc2vec (LSA)*	*TF-IDF + doc2vec*
PCT mono	76.58	75.20	69.87	66.45	76.84
PCT mono malé	82.94	81.19	73.25	69.21	82.78
PCT stereo	76.56	70.33	69.28	66.93	73.44
PCT stereo malé	79.61	73.39	72.15	70.40	77.16
PCT vše	77.92	71.14	70.86	68.36	75.19
PCT vše malé	78.89	71.56	71.51	68.60	74.90

Tabulka 1: Výsledky s použitím Linearního SVM

	Přesnost metod [%] K-means metoda s reprezentacemi				
	TF-IDF (TF-IDF)	*TF-IDF (LSA)*	*doc2vec*	*doc2vec (LSA)*	*TF-IDF + doc2vec*
PCT vše	31.29	32.12	31.51	28.65	32.53
PCT vše malé	40.34	38.79	38.68	38.54	42.08

Tabulka 2: Výslekdy s použitím K-means

5 Závěr

Z experimentu je patrné, že pro klasifikaci přepsaných telefonických hovorů podle jejich témat byl nalezen vhodný postup a metoda typu učení s učitelem viz. Tabulka 1. Co se týče metody učení bez učitele (viz. Tabulka 2) nebylo dosaženo tak kvalitních výsledků v porovnání s metodou typu učení s učitelem což bylo předvídáno. Budoucím cílem bude vylepšení postupu přípravy a vylepšení metod typu učení bez učitele aby dosahovali alespoň podobné kvality výsledků jako metody typu učení s učitelem.

Poděkování

Příspěvek byl podpořen grantovým projektem SVK1-2018-024

Literatura

Straková, J.; Straka, M. & Hajič, J. (2014) Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pp. 13-18

Novotný, J. & Ircing, P. (2017) Unsupervised Document Classification and Topic Detection. *International Conference on Speech and Computer*. pp. 748-756

Lau, J. H. & Baldwin, T. (2016) An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv*.