# Identifikace rodného jazyka pisatelů na základě anglicky psaných esejů

Robert Brada[1]

## 1 Úvod

Identifikací rodného jazyka rozumíme úlohu, ve které chceme automaticky určit rodný jazyk autora čistě na základě textu, který je autorem napsán v jiném než rodném jazyce (v této práci angličtina). Cílem této práce je navrhnout systém umělé inteligence, který bude v této úloze dosahovat co nejlepšího skóre.

Uplatnění výsledků lze najít například v oblasti výuky cizího jazyka. Je zřejmé, že lidé se stejným rodným jazykem budou náchylní k tomu, aby dělali podobné chyby při studiu cizího jazyka. Naučí-li se systém takové vlastnosti rozpoznávat, může být jazykovým studentům poskytnuta cílená zpětná vazba ohledně jejich chyb, což značně zefektivní výuku.

Tato úloha se řeší metodou strojového učení, což je oblast umělé inteligence zabývající se technikami, které umožňují počítačovému systému učit se, aniž by musel být naprogramován pro konkrétní úlohu.

## 2 Postup řešení

K naučení systému je třeba mít dostatek dat. My budeme pracovat s datovým korpusem, který obsahuje 12 100 anglicky napsaných esejů. Tyto eseje psali autoři s 11 různými rodnými jazyky, kterými jsou: arabština, čínština, francouzština, němčina, indština, italština, japonština, korejština, španělština, telugština, turečtina.

Samotné texty ovšem musíme předzpracovat. Vstupem do algoritmů strojového učení musí být vektor fixní délky (tzv. příznakový vektor). Texty tedy musíme na takové vektory převést. Správná volba příznakového vektoru je klíčová při návrhu modelu a má zásadní vliv na dosažené výsledky. Příznaky mohou tvořit například slova, dvojice slov, slovní druhy a podobně. Jak převést texty na vektory pomocí tzv. *Bag Of Words* modelu lze vidět na následujících příkladu:

***Příklad:***

K dispozici je datová sada obsahující 2 texty, které chceme převést na vektory:

*Text 1:* Nevím, zda je to možné.
*Text 2:* Je možné, že je to pravda.

| | je | možné | nevím | pravda | to | zda | že |
|---|---|---|---|---|---|---|---|
| Text 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Text 2 | 2 | 1 | 0 | 1 | 1 | 0 | 1 |

**Tabulka 1:** Příklad vytvoření příznakového vektoru

[1] student bakalářského studijního programu Aplikované vědy a informatika, obor Kybernetika a řídicí technika, e-mail: bradar@students.zcu.cz

Kromě zmíněného *Bag Of Words* modelu lze využít i *doc2vec* modelu, který je složitější a využívá neuronových sítí. Tento model je blíže popsán v originální práci.

Úkolem je tedy převést na příznakové vektory všechny texty z datového korpusu. Dalším krokem je texty rozdělit na trénovací a testovací množinu. Texty z trénovací množiny můžeme použít k natrénování klasifikátoru *(klasifikátor = algoritmus, který určuje, do které z kategorií text patří)*. Texty z testovací množiny využijeme k určení přesnosti klasifikace.

## 3 Výsledky

Klasifikátorů, které lze použít, je celá řada a jejich výběr výrazně ovlivňuje přesnost klasifikace. Testuje se tedy jak použitý model tvorby příznakových vektorů, tak použitý klasifikátor. V tabulce jsou uvedeny nejlepší výsledky pro jednotlivé modely a také přesnost výchozího modelu, ze kterého se vycházelo. Modely a klasifikátory mají další parametry, jejichž detailnější popis lze najít v originální práci.

| Model | Klasifikátor | Přesnost klasifikace |
|---|---|---|
| Bag Of Words (výchozí) | SVC | 71,0 % |
| Bag Of Words | SVC | 84,3 % |
| doc2vec | SVM | 69,6 % |

**Tabulka 2:** Výsledky klasifikace

## 4 Závěr

Cílem práce bylo najít systém, který bude mít co nejlepší přesnost klasifikace. Z dosažených výsledů je zřejmé, že takovým systémem je ten využívající *Bag Of Words* model, který měl o 13,3% lepší přesnost klasifikace než výchozí model. Tento model využíval jako příznaky trigramy slov a 11-gramy znaků. Výchozí model využíval jako příznaky jednotlivá slova.

Přesnost klasifikace ovšem není jediným meřítkem, podle kterého se řídit, pokud chceme systém využít v praxi. Může docházet například k přetrénování, dále potřebujeme nějak interpretovat důležitost příznaků, podle kterých se klasifikátor rozhoduje a podobně. Tyto aspekty jsou rozebrány v originální práci. Závěrem je, že klasifikátor s nejlepší přesností je výrazně přetrénován a i systém využívající *doc2vec* model poskytuje i přes horší přesnost klasifikace velmi užitečné informace.

## Literatura


Ircing, P. *Vektorová sémantika a vyhledávání informací.* Slajdy z přednášek předmětu Vyhledávání informací (IR) na ZČU.

Le, Q., a Mikolov, T. (2014) *Distributed Representations of Sentences and Documents.* Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing.

Simeone, O. (2017) *A Brief Introduction to Machine Learning for Engineers.* King's College London, Department of Informatics.

Tetreault, J., Blanchard, D., a Cahill, A. (2013) *A Report on the First Native Language Identification Shared Task.* Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp 48-57.