



Semestrální práce KIV/UIR

Klasifikace dokumentů

Pavel Třeštík
A17B0380P

21. května 2020

Obsah

1	Zadání	1
2	Analýza úlohy	3
2.1	Algoritmy pro tvorbu příznaků	3
2.1.1	Bag of Words	3
2.1.2	Bigram	4
2.1.3	TF-IDF	5
2.2	Klasifikační algoritmy	5
2.2.1	Naive Bayes	5
2.2.2	k-nearest neighbors	6
3	Implementace	7
4	Uživatelská dokumentace	7
5	Závěr	7

1 Zadání

Ve zvoleném programovacím jazyce navrhnete a implementujete program, který umožní klasifikovat textové dokumenty do tříd podle jejich obsahu, např. počasí, sport, politika, apod. Při řešení budou splněny následující podmínky:

- Použijte data z českého historického periodika Posel od Čerchova“, která jsou k dispozici na <https://drive.google.com/drive/folders/1mQbBNS43gWFRMHDYdSkQug47cuhPTsHJ?usp=sharing>. V původní podobě jsou data k dispozici na <http://www.portafontium.eu/periodical/posel-od-cerchova-1872?language=cs>.
- Pro vyhodnocení přesnosti implementovaných algoritmů bude NUTNÉ vybrané dokumenty ručně označkovat. Každý student ručně anotuje 10 stran zadaného textu – termín 31.3.2020. Za dodržení termínu obdrží student bonus 10b.
- Přiřazení konkrétních textů jednotlivým studentům spolu s návodem na anotaci a příklady je uloženo spolu s daty na výše uvedené adrese, konkrétně:
 - 0 - vzorová složka (takhle by měl výsledek vypadat)
 - 1, 2, .. , 15, 101, 102, .. - data k anotaci
 - přiřazení souboru studentum.xlsx - určení, jaké soubory má jaký student anotovat. Až budete mít anotaci hotovou, doplňte sem informaci.
 - Anotační příručka - návod, jak články anotovat.
 - Klasifikace dokumentů - kategorie.xlsx - seznam kategorií k anotaci s příklady.
 - sem prace20.pdf - Zadání semestrální práce
- implementujte alespoň tři různé algoritmy (z přednášek i vlastní) pro tvorbu příznaků reprezentující textový dokument.
- implementujte alespoň dva různé klasifikační algoritmy (klasifikace s učitelem):
 - Naivní Bayesův klasifikátor

– klasifikátor dle vlastní volby

- funkčnost programu bude následující: – spuštění s parametry:
název klasifikátoru, soubor se seznamem klasifikačních tříd, trénovací množina, testovací množina, parametrizační algoritmus, klasifikační algoritmus, název modelu
program natrénuje klasifikátor na dané trénovací množině, použije zadaný parametrizační a klasifikační algoritmus, zároveň vyhodnotí úspěšnost klasifikace a natrénovaný model uloží do souboru pro pozdější použití (např. s GUI). – spuštění s jedním parametrem:
název klasifikátoru, název modelu
program se spustí s jednoduchým GUI a uloženým klasifikačním modelem. Program umožní klasifikovat dokumenty napsané v GUI pomocí klávesnice (resp. překopírované ze schránky).
- ohodnoťte kvalitu klasifikátoru na dodaných datech, použijte metriku přesnost (accuracy), kde jako správnou klasifikaci uvažujte takovou, kde se klasifikovaná třída nachází mezi anotovanými. Otestujte všechny konfigurace klasifikátorů (tedy celkem 6 výsledků).

Poznámky:

- pro vlastní implementaci není potřeba čekat na dokončení anotace. Pro průběžné testování můžete použít korpus současné češtiny, který je k dispozici na <http://ctdc.kiv.zcu.cz/> (uvažujte pouze první třídu dokumentu podle názvu, tedy např. dokument 05857 zdr ptr eur.txt náleží do třídy zdr – zdravotnictví). ”
- další informace, např. dokumentace nebo forma odevzdávání jsou k dispozici na CW pod záložkou Samostatná práce.

2 Analýza úlohy

2.1 Algoritmy pro tvorbu příznaků

Existuje řada metod získávání informací (modelů) z rozdílných typů dat. Pro tuto práci nás ale zajímají pouze jazykové pravděpodobnostní modely. Ty se řadí do několika tříd: unigram, n-gram, exponenciální, neuronové sítě a ostatní (takové, které nepřípadají do ani jedné z přechozích skupin). Ve skutečnosti ale některé třídy jsou založeny na jiné třídě. Na příklad unigram je ve skutečnosti n-gram pro jednotlivá slova. Mezi nejvýznačnější reprezentanty patří Bag of Words, Bigram, Trigram a Word2vec.

2.1.1 Bag of Words

Prvním z algoritmů, které jsem vybral je Bag of Words. Jedná se o unigram, tudíž každý text/ dokument je reprezentován jednotlivými slovy. Slova jsou ukládána do slovníku a jednotlivé klasifikační třídy/ dokumenty jsou reprezentovány vektory o délce slovníku. Hodnoty vektorů jsou reprezentací četností výskytu slov v textu/ dokumentu pro každou klasifikační třídu (nebo jednotlivé dokumenty). Toto ovšem vytváří problém, kdy při velké množině slov ve slovníku, se jich pouze malá část vyskytuje ve vektoru reprezentující klasifikační třídu/ dokument.

Příklad Bag of Words

Zdroj (Wikipedia): [url](#)

Texty

(1) John likes to watch movies. Mary likes movies too.

(2) Mary also likes to watch football games.

Rozdělení textů

BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};

BoW2 = {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

Slovník textů

BoW3 = {"John":1,"likes":3,"to":2,"watch":2,"movies":2,"Mary":2,"too":1,"also":1,"

Vektory reprezentující texty

(1) [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

(2) [0, 1, 1, 1, 0, 1, 0, 1, 1, 1]

2.1.2 Bigram

Druhým algoritmem pro tvorbu příznaků bude **Bigram**. Ačkoliv **Bag of Words** je ve skutečnosti n-gram pro $n = 1$ a **Bigram** je n-gram pro $n = 2$ je v těchto algoritmech značný rozdíl. **Bigram** páruje slova a počítá pravděpodobnost páru jako podmíněnou pravděpodobnost obou slov páru.

Párování slov je ukázáno na následujícím příkladu Ukázka 1 Pravděpodobnost výrazu je spočítána podle Vztahu 2.

Věta

Tato věta je příklad.

Bigramy

bigrams = {"S, Tato", "Tato, věta", "věta, je", "je, příklad", "příklad, /S"}

S = začátek věty /S = konec věty

Obrázek 1: Příklad párování slov

$$P(W_n|W_{n-1}) = \frac{P(W_{n-1}, W_n)}{P(W_{n-1})}$$

Obrázek 2: Pravděpodobnost výrazu bigramu
Zdroj: <https://en.wikipedia.org/wiki/Bigram>

2.1.3 TF-IDF

Posledním vybraným algoritmem je TF-IDF. Na rozdíl od předchozích dvou vybraných algoritmů nevytváří příznaky z textu, ale upravuje již existující příznaky. Tudiž je využíván spolu s **Bag of Words**. Algoritmus ohodnotí jednotlivá slova váhou místo jejich četností. Čím častěji se slovo v textech/dokumentech vyskytuje, tím nižší váhu má po TF-IDF. Tento algoritmus svým způsobem "vytváří" takzvané **stop words**. To jsou slova, které v jazyce pro klasifikaci nemají žádný význam (na příklad pro Český jazyk slova jako: a, pro, nebo...). Přidat url na příklad a jak počítat??

2.2 Klasifikační algoritmy

Výběr klasifikačního algoritmu značně závisí na účelu našeho projektu. Z mnoha klasifikátorů jako je **Naive Bayes**, **Support Vector Machine (SVM)**, **Lineární regrese**, **Neural Networks** a dalších, potřebujeme vybrat podle potřeb. Nejdůležitějšími vlastnostmi klasifikátoru při výběru jsou jeho přesnost (accuracy), složitost a rychlost.

2.2.1 Naive Bayes

Je jeden z nejrychlejších a nejjednodušších klasifikátorů. Jeho nedostatkem je, že potřebuje poměrně přesná trénovací data, aby jeho klasifikace byly co možná nejpřesnější.

Klasifikátor vybere nejpravděpodobnější třídu podle následujícího vztahu (viz Obrázek 3).

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^n p(x_i | C_k).$$

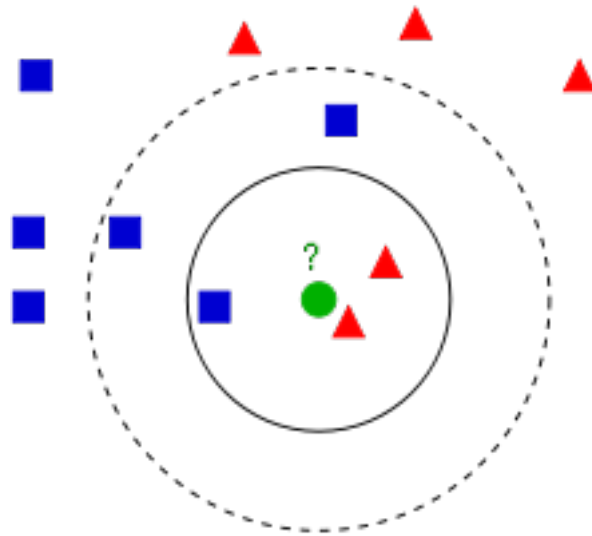
Obrázek 3: Rovnice vybrání třídy

Zdroj: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

2.2.2 k-nearest neighbors

Převede klasifikovaný text/ dokument na vektor a spočítá vzdálenost ke každému trénovacímu textu/ dokumentu. Poté přiřadí klasifikovanému textu třídu, která má nejvíce zástupců v určité blízkosti. Tento algoritmus je velmi jednoduchý a poměrně rychlý. Nevýhodou je přesnost, která pokud trénovací soubory nemají jednoznačně určenou třídu, není velmi vysoká.

Příklad klasifikace zelené tečky na Obrázku 4. Zelená tečka bude klasifikována jako červený trojúhelník, protože v okolí jsou 2 červené trojúhelníky a pouze 1 modrý čtverec.



Obrázek 4: Příklad klasifikace knn

Zdroj:

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

- 3 Implementace**
- 4 Uživatelská dokumentace**
- 5 Závěr**