****

**KIV/UIR – Semestrální práce**

**Automatická klasifikace dialogových aktů**

**Autor:** David Šavel**Doba řešení:** 20 hodin

# Popis problému

Cílem semestrální práce je napsat program, který se na základě trénovacích dat naučí co nejlépe klasifikovat komiksové věty a hlášky do předem stanovených tříd. Prvním dílčím úkolem je načíst vstupní data a vhodně je číselně reprezentovat za pomocí tří různých algoritmů. Dále je třeba naimplementovat dva klasifikační algoritmy, které se na trénovacích datech naučí klasifikovat testovací data do tříd podle příznaků. Jeden z klasifikační algoritmů musí být Naivní Bayesův klasifikátor. Nakonec je třeba pro každou kombinaci příznakového a klasifikačního algoritmu vypočítat přesnost s jakou data klasifikoval.

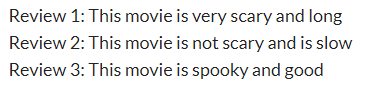
# Analýza problému

## Reprezentace dat

Vstupní data, jak trénovací, tak testovací je třeba vhodně, a hlavně stejným způsobem reprezentovat, aby se například dala počítat vzdálenost dvou vět nebo porovnat jejich zařazení. Nejideálnějším řešením je určitě reprezentovat větu jako pole řetězců, ovšem musí se myslet na to, jak správně z textu tyto řetězce dostat. Oddělit řetězce by v tomto případě nebylo nejmoudřejší, protože by v řetězcích zůstávala interpunkční znaménka a program by vyhodnocoval řetězce jako odlišné, proto je nejlepší interpunkční znaménka odstranit co nejvíce to půjde ještě před spuštěním algoritmu.

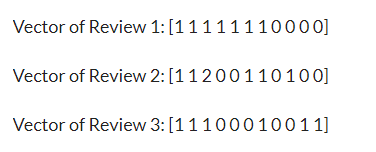
## Algoritmy pro tvorbu příznaků

Vybrat vhodné algoritmy pro tvorbu příznaků z textového dokumentu. Většina nejznámějších algoritmů využívá k tvorbě příznakového vektoru slovník, což je seznam všech různých slov, která se v celém textovém dokumentu vyskytují. Věta je pak reprezentována vůči celému slovníku, tedy vektorem o délce slovníku. Nejznámější algoritmy jsou Bag of Words a TF-IDF. Bag of Words vyjadřuje kolikrát se každé slovo ze slovníku ve větě vyskytuje.

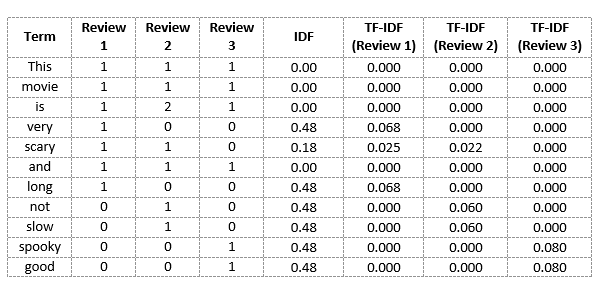


Obsah obrázku stůl

Popis byl vytvořen automaticky



Druhý zmíněný algoritmus je o něco složitější. IDF je pro každé slovo ve slovníku stejné. Jedná se o logaritmus z podílu počtu celkových vět a počtu vět s daným slovem. TF se pak počítá pro každou větu zvlášť a je rovno podílu násobnosti daného slova ve větě a celkovému počtu slov ve větě. U slov, které se ve větě nevyskytují bude celková hodnota nulová stejně jako u Bag of Words.



# Návrh řešení