## KIV/UIR - Semestrální práce pro ak. rok 2016/17

# Klasifikace dokumentů

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

- použijte korpus dokumentů v českém jazyce, který je k dispozici na http://home.zcu.cz/~pkral/sw/ (uvažujte pouze první třídu dokumentu podle názvu, tedy např. dokument 05857_zdr_ptr_eur.txt náleží do třídy „zdr“ - zdravotnictví.)
- 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: trénovací_množina, testovací_množina, parametrizační_algoritmus, klasifikační_algoritmus, název_modelu
    program natrénuje klasifikátor na dané trénovací množině, použije zadaný parametrizační/klasifikační algoritmus, zároveň vyhodnotí úspěšnost klasifikace a natrénovaný model uloží do souboru pro pozdější použití (např. s GUI).
  - spuštění s jedním parametrem = 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). Otestujte všechny konfigurace klasifikátorů (tedy celkem 6 výsledků).

**Bonusové úkoly**:

- vyzkoušejte již nějakou hotovou implementaci klasifikátoru (Weka, apod.) a výsledky srovnejte s Vaší implementací
- vyzkoušejte klasifikaci bez učitele (např. k-means) a výsledky porovnejte s výsledky klasifikace s učitelem
- vyzkoušejte klasifikaci anglických dokumentů, korpus Reuters je k dispozici na adrese http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html