SPAM filtr

Zadání úlohy

- https://www.kaggle.com/uciml/sms-spam-collection-dataset

- cíl: vytvořit model predikující, zda je zpráva SPAM či není

Převod textu na vektor

- bag of words
- TF-IDF
- word embeddings

Bag of words

- vezmou se všechna různá slova z celého datasetu a z nich se vytvoří slovník
  - může to být i např. 10 tisíc různých slov
- pro každé slovo se vytvoří nová feature
  - takže z jednoho textového pole nám vznikne např. 10 tisíc příznaků
- hodnotou této nové feature pro danou instanci je, kolikrát se dané slovo v této instanci vyskytuje
- protože drtivá většina hodnot je prázdných, používá se typicky sparse matrix
- třída CountVectorizer, udělá vše potřebné za nás
  - řada parametrů k vyzkoušení: min_df, max_df, ngram_range, stop_words, ...

TF-IDF

- některá slova se mohou vyskytovat takřka ve všech instancích, takže jejich významnost pro konkrétní instanci nebude patrně příliš vysoká
- TF = term frequency
- IDF = inverse document frequency
- výsledkem je číslo 0 až 1, vyšší hodnoty mají slova, která mají velký význam
- k dispozici jako TfidfTransfomer, nebo jako TfIdfVectorizer (nahrazuje CountVectorizer)

Úkoly

- vyberte vhodné metriky úspěchu s ohledem na úlohu a dataset

- vyzkoušejte různé parametry převodu textu na vektor (n_gramy, stop words, ...])

- vyzkoušejte SVM a grid search pro nalezení vhodných hyperparametrů
- vyzkoušejte Random Forest
- vyzkoušejte Logistickou regresi
- vyzkoušejte ensemble metodu Voting Classifiers
- vyzkoušejte ensemble metodu Gradient Boosting (příp. XGBoost)

- vyzkoušejte přidat další atributy
  - např. počet slov ve zprávě