MUNI SCI

Klasifikace textu s využitím neuronových sítí

Stanislav Zámečník, Vojtěch Šindlář

Ústav matematiky a statistiky, Masarykova univerzita, Brno

16. prosince 2021

Úvod

Cíl: Vytvořit textový klasifikátor, který na základě textu zadaného uživatelem identifikuje téma.

- Omezení klasifikace pouze na tři kategorie novinových článků
- Využití neuronových sítí
- Python

Popis dat

- Vytvoření trénovací množiny na základě "zpravodajského" portálu Daily Mail[1]
- Kategorie *sport*, *travel* a *science*
- Pro stahování článků z webu (tzv. *web scraping*) jsme použili balíček Beautiful Soup
- Čištění dat

[1] `https://www.dailymail.co.uk/home/index.html`

Model

- Použití rekurentní neuronové sítě
- Jedna vstupní vrstva, dvě LTSM vrstvy, jedna hustá vrstva
- Balíčky TensorFlow a Keras pro vytvoření neuronové sítě v Pythonu
- Využití prostředí Google Colab

Ukázka výsledků

- Vytvoření aplikace s použitím balíčku Streamlit
- Využití Git GUI klienta – Fork
- Potenciální umístění aplikace na web (platforma Heroku)
- Odkaz na uložiště na GitHubu:
  `https://github.com/stazam/Datamining2-project`

Shrnutí

- Pro vývoj modelu doporučujeme Google Colab (TPU, GPU)
- Pokud začínáte s gitem nebojte se využít Git GUI klienta
- Pro tvorbu "jednoduchých" machine learningových aplikací zkuste Streamlit
- Pro umístění aplikace na web zkuste Heroku