УДК 519.715
Сидор В. –ст. гр. СНнм-61
Тернопільський національний технічний університет імені Івана Пулюя

ОБҐРУНТУВАННЯ ВИБОРУ ІНСТРУМЕНТІВ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТУ

Науковий керівник: к.т.н., доцент Фриз М. Є.

Sydor V.
Ternopil Ivan Puluj National Technical University

SUBMISSION OF SELECTION OF TOOLS FOR CLASSIFICATION OF TEXT

Supervisor: associate professor Fryz M. Ye.

Ключові слова: Інструменти, Класифікація
Keywords: Tools, Classification

У доповіді буде розглянуто бібліотеки та інструменти для роботи із текстовою класифікацією. Машинне навчання останніми роками розвивається дуже швидкими темпами. В 2018-2019 роках виділяються лідерами за популярністю над іншими мовами програмування для розробки машинного навчання як R, Python та програма RapidMiner. Але якщо взяти запити за популярністю від Google Trends то інтерес до Python набагато більший ніж до R. Коли мова йде про проект з використанням машинного навчання, то і Python, і R мають свої особливі переваги. Завдяки широкій доступності пакетів більшість загальних завдань, пов'язаних з однією з цих мов, виконуються в обох випадках.

Тим не менш, Python працює краще в маніпуляціях даними та повторюваними завданнями - і це, безумовно, правильний вибір, якщо планується побудувати цифровий продукт на основі машинного навчання. Але, якщо проект перебуває на ранніх стадіях, і потрібно розробити інструмент для спеціального аналізу та вивчення набору даних, хорошим вибором є R - якщо немає хороших знань в Python.

В Python велика кількість бібліотек для роботи на всіх етапах розробки проекту із використання текстової класифікації. Розглянуто найбільш популярні бібліотеки які використовуються на даний час. Для класифікації використано комбінацію інструментів як NLTK та Scikit-learn. NLTK використовується для створення і оброблення текстової інформації та створення базових класифікаторів. Scikit-learn – для розширення різноманітності класифікаторів, що супроводжується дружнім API для NLTK. Крім створення інформації, потрібно ще також візуалізувати отримані результати. В простих випадках використовують бібліотеку Matplotlib, але використаємо сучасне та інтерактивне представлення інформації із онлайн доступом - хорошим вирішенням даної проблеми буде бібліотека Plotly. Даний набір інструментів дозволяє створити надійний класифікатор та красиво візуалізувати отриману інформацію.

Література

1. Christopher D. Manning. Introduction to Information Retrieval / D. Manning Christopher - Cambridge University Press, 2013. – 581 с.

2. Muller A. Introduction to Machine Learning with Python: A Guide for Data Scientists / Andreas Muller., 2016. – 418 с.