Лабораторна робота №3. Класифікація текстів

Мета: виконати обчислювальні експерименти з різними моделями класифікації та векторизації

Теоретичні відомості

Класифікація текстів принципово не відрізняється від будь-якої задачі класифікації. Основні етапи полягають у наступному.

- Попередня обробка (перетворення, лематизація, токенізація).
- Векторизація.
- Виділення ознак (якщо потрібно).
- Класифікація.

Методи класифікації. Глибинні нейронні мережі, такі як згорткові нейронні мережі, видаються найкращим варіантом для класифікації текстів, оскільки таки типи мереж можуть додатково виділяти ознаки, які потім використовуються для розпізнавання.

Отже, шари згортки Conv1D (https://keras.io/api/layers/convolution_layers/convolution1d/) або Conv2D (https://keras.io/api/layers/convolution_layers/convolution2d/) визначають суттєві ознаки, а подальші Dense (https://keras.io/api/layers/core_layers/dense/) шари виконують безпосередньо класифікацію.

Метрики якості класифікації.

Матриця розбіжностей (Confusion matrix). Ефективність прогнозування розробленої моделі за кожним класом можна розглянути за допомогою матриці розбіжностей. Індекси за строками позначають фактичні класи, а за стовпцями – спрогнозовані. Таким чином, матриця розбіжностей ідеального класифікатора є діагональною.

Програмна реалізація:
http://scikitlearn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

F1 метрика. Відносна числова метрика. За сенсом близька до поняття відносної точності.

https://uk.wikipedia.org/wiki/F-%D0%BC%D1%96%D1%80%D0%B0

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Завдання для самостійної роботи

1. Використовуючи приклади та результати лабораторних робіт 1,2 підготуйте дані для класифікації. Виконайте поділ на train\validation\test набори даних Spooky Author Identification або First GOP Debate Twitter Sentiment.

2. Використайте декілька методів векторизації для підготовки даних на попередньому етапі (Count Vectorizer, TF-IDF та один з Word2vec, Bert, Glove, FastText, ELMo).

3. Виконати класифікацію за допомогою згорткової мережі. Знайти оптимальні гіперпараметри мережі (кількість шарів, розмір матриці згортки, кількість нейронів тощо) із застосування підходу Grid Search або Random Search.

4. За допомогою обчислень Confusion matrix та F1 порівняти результати класифікації для різних методів векторизації.

Запитання для самоконтролю

1. В чому особливість методів глибинного навчання?

2. Які параметри системи класифікації текстів є визначальними?

3. Яка з метрик якості класифікації є більш інформативною?