Попередня обробка текстових даних

Мета роботи

Ознайомитись з основними підходами для попередньої обробки текстових даних.

Рекомендована література

1. Natural Language Processing with Transformers. Revised Edition. Lewis Tunstall, Leandro von Werra.
2. Practical Natural Language Processing with Python. 1st Ed. Mathangi Sri
3. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems, Sowmya Vajjal
4. Practical Natural Language Processing: A Pragmatic Approach to Processing and Analyzing Language Dat, Sowmya Vajjala, Bodhisattwa Majumder
5. Natural Language Processing Projects. 1st Ed. Akshay Kulkarni, Adarsha Shivananda
6. Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning, Delip Rao.

Хід роботи

Дані

В роботі використовується публічний набор даних з https://www.kaggle.com/datasets, https://paperswithcode.com, або https://huggingface.co.

Підготовчий етап

Провести аналіз вибраного набору даних, визначити вхідні та вихідні параметри, візуалізувати залежності входів на виходу, спробувати виявити основні залежності, детектувати аномалії, неповні зразки тощо у даних.

Реалізація моделі

1. Використати алгоритми/моделі tf-idf, BoW, Word2Vec, Doc2Vec для feature extraction.
2. Використати щонайменьше 2 алгоритми з пункту 1 для вирішення задачі summarization.
3. Маючи самарі з пункту 2, побудувати/використати алгоритм для Named Entity Recognition.

Аналіз результатів

1. Проаналізувати результати ембедингів за допомогою cross-one-out та пошуком сусідів.
2. Використати метрики оцінки summarization та обгрунтувати їх використання (чому ці доречні, а інші - ні).
3. Використати метрики оцінки Named Entity Recognition та обгрунтувати їх використання (чому ці доречні, а інші - ні).

Студенти, що розраховують на високий бал мають:
1. Відкоментувати код (що кожна строка робить, окрім import / library)
2. Якісно візуалізувати результати та проміжні етапи (за необхідністю)
3. Обрати порівняно складний датасет, який потребує додаткового очищення
4. Обмежити використання готових рішень/моделей чи датасетів.

Контрольні питання

1. Які основні метрики Named Entity Recognition та summarization існують?
2. У чому полягає різниця між doc2vec та word2vec?
3. Чи завжди Word2Vec є більш доречним підходом у порівнянні до BoW чи tf-idf, чому?