

Попередня обробка даних

Мета роботи

Ознайомитись з основними підходами для попередньої обробки текстових даних та зображень.

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 Data, 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.
7. <https://medium.com/coinmonks/review-srcnn-super-resolution-3cb3a4f67a7c>
8. Deep Learning Cookbook: Practical recipes to get started quickly 1st Edition, Douwe Osinga
9. Fundamentals of Deep Learning. Designing Next-Generation Machine Intelligence Algorithms. 2nd Edition, Joe Papa

Хід роботи

Дані

В роботі використовується публічний набір 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 щонайменше 2 алгоритми з пункту 1 для вирішення задачі summarization.
3. Використати SRCNN для покращення якості зображення.
4. Використати Segmentation алгоритм на 2х різних датасетах (оригінальному датасеті для SRCNN та покращеному).

Аналіз результатів

1. Проаналізувати результати ембедингів за допомогою cross-one-out та пошуком сусідів.
2. Використати метрики оцінки summarization та обґрунтувати їх використання (чому ці доречні, а інші - ні).
3. Оцінити якість роботи SRCNN.
4. Порівняти якість роботи Segmentation алгоритмів на двох датасетах.

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 для оцінки якості 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?
4. Які особливості роботи з Object Detection або Segmentation алгоритмами Ви можете назвати з точки зору попередньої обробки даних, наприклад, чи важливою є якість зображення, колір тощо?