TF-IDF ДЛЯ ОЧИЩЕННЯ ТЕКСТОВИХ ДАНИХ У КЛАСИФІКАЦІЙНИХ ЗАДАЧАХ З ВИКОРИСТАННЯМ NLP

Шидловський Тимофій Максимович
студент групи ІАСм1-23-1.4д,
Київського столичного університету імені Бориса Грінченка,
науковий керівник – к.т.н. Яскевич В.О.

Одна з проблем NLP (Natural Language Processing) – це очищення текстових даних перед навчанням та прогнозуванням. Стандартні методи очищення включають приведення тексту до нижнього регістру, видалення пунктуації та стоп-слів, токенізацію, а також стемінг або лематизацію. Ці підходи спрямовані на зменшення шуму та приведення даних до єдиного формату.

Проте, в задачах класифікації тексту можна додатково покращити очищення за допомогою алгоритму TF-IDF (Term Frequency-Inverse Document Frequency). Зазвичай TF-IDF використовують для оцінки важливості слів в документах, відносно всього тексту [1]. Іншими словами, він допомагає знайти та оцінити, які слова краще всього характеризують документ. Але, в задачах класифікації можна застосовувати TF-IDF для визначення значущості слів для кожного з класів.

Такий підхід дозволяє присвоїти різні ваги словам під час навчання моделі, підкреслюючи ті, що будуть найбільш релевантними до певного класу та зменшуючи вплив більш поширених слів. Також, можна перед навчанням прибирати слова, які не будуть мати достатньої ваги в жодному з класів. Таким чином, модель може зосередитися на найбільш значущих ознаках, що потенційно підвищить точність класифікації.

Однак, цей метод може і погіршити результати моделі, якщо вона використовує шари (layers), які враховують контекст попередніх слів, такі як LSTM (Long Short-Term Memory). Видаляючи слова за допомогою TF-IDF, ми ризикуємо втратити важливу контекстуальну інформацію, оскільки певні слова (які були поширені у корпусі) можуть суттєво змінювати контекст/сенс речення. Наприклад, слово «не» може мати малу вагу при використанні TF-IDF, але це слово ~~*не*~~ може повністю змінити сенс речення.

Тому при застосуванні TF-IDF для очищення даних необхідно балансувати між зменшенням шуму та збереженням контексту. Але при правильному використанні можна не тільки покращити результати нейромережі, а й ще зменшити її розмір та збільшити її швидкість.

ДЖЕРЕЛА

1. TF-IDF in NLP (Term Frequency Inverse Document Frequency) | by Abhishek Jain | Medium, 2024, URL: https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d