Лабораторна робота №11: Вступ до Natural Language Processing (NLP)

ВИКОНАВ: ЩЕРБАНЬ ДМИТРО

ТЕОРЕТИЧНІ ОСНОВИ ТА КЛЮЧОВІ МОДЕЛІ NLP

ПРОЦЕС ОБРОБКИ ПРИРОДНОЇ МОВИ ВКЛЮЧАЄ КІЛЬКА КРИТИЧНИХ ЕТАПІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ, ТАКИХ ЯК ТОКЕНІЗАЦІЯ, ЛЕММАТИЗАЦІЯ ТА СТЕМІНГ, ЯКІ ПЕРЕТВОРЮЮТЬ "СИРИЙ" ТЕКСТ У ФОРМАТ, ПРИДАТНИЙ ДЛЯ АНАЛІЗУ. ДЛЯ ВИРІШЕННЯ ЗАВДАНЬ КЛАСИФІКАЦІЇ ТА РОЗПІЗНАВАННЯ СУТНОСТЕЙ (NER) ВИКОРИСТОВУЮТЬСЯ РІЗНОМАНІТНІ МОДЕЛІ: ВІД КЛАСИЧНИХ АЛГОРИТМІВ НА КШТАЛТ НАЇВНОГО БАЄСОВОГО КЛАСИФІКАТОРА ТА ЛОГІСТИЧНОЇ РЕГРЕСІЇ ДО СУЧАСНИХ АРХІТЕКТУР ГЛИБОКОГО НАВЧАННЯ, ТАКИХ ЯК LSTM, TRANSFORMERS ТА GPT.

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ВЕКТОРИЗАЦІЇ ТЕКСТУ

ВЕКТОРИЗАЦІЯ ТЕКСТУ Є КЛЮЧОВИМ ЕТАПОМ ПЕРЕТВОРЕННЯ СЛІВ У ЧИСЛОВІ ВЕКТОРИ, І ДЛЯ ЦЬОГО ВИКОРИСТОВУЮТЬСЯ ТАКІ ПІДХОДИ, ЯК BAG OF WORDS (BOW), TF-IDF ТА WORD EMBEDDINGS (WORD2VEC, GLOVE). У ХОДІ РОБОТИ БУЛО СТВОРЕНО ПОРІВНЯЛЬНУ ТАБЛИЦЮ, ДЕ ЗАЗНАЧЕНО ПЕРЕВАГИ, НЕДОЛІКИ ТА СКЛАДНІСТЬ РЕАЛІЗАЦІЇ КОЖНОГО МЕТОДУ. АНАЛІЗ ПОКАЗАВ, ЩО ПРОСТІ МЕТОДИ НА ЗРАЗОК BOW ЛЕГШІ В РЕАЛІЗАЦІЇ, АЛЕ WORD EMBEDDINGS КРАЩЕ ПЕРЕДАЮТЬ СЕМАНТИЧНИЙ ЗМІСТ СЛІВ.

ОГЛЯД ІНСТРУМЕНТІВ ТА БІБЛІОТЕК ДЛЯ NLP

ДЛЯ ВИКОНАННЯ ЗАДАЧ NLP ІСНУЄ ШИРОКИЙ СПЕКТР ПРОГРАМНИХ ІНСТРУМЕНТІВ, СЕРЕД ЯКИХ НАЙБІЛЬШ ПОПУЛЯРНИМИ Є NLTK, SPACY, HUGGING FACE TRANSFORMERS ТА GENSIM. КОЖЕН ІЗ ЦИХ ІНСТРУМЕНТІВ МАЄ СВОЇ ОСОБЛИВОСТІ: НАПРИКЛАД, NLTK ДОБРЕ ПІДХОДИТЬ ДЛЯ НАВЧАННЯ ТА ДОСЛІДЖЕНЬ, SPACY ЗАБЕЗПЕЧУЄ ВИСОКУ ШВИДКІСТЬ ОБРОБКИ, А HUGGING FACE НАДАЄ ДОСТУП ДО ПЕРЕДОВИХ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ. ВИБІР КОНКРЕТНОЇ БІБЛІОТЕКИ ЗАЛЕЖИТЬ ВІД ВИМОГ ДО ПІДТРИМКИ МОВ ТА ПРОСТОТИ ВИКОРИСТАННЯ.

ВИСНОВКИ ТА СФЕРИ ЗАСТОСУВАННЯ

ТЕХНОЛОГІЇ NLP АКТИВНО ЗАСТОСОВУЮТЬСЯ В РІЗНИХ ГАЛУЗЯХ ДЛЯ ТАКИХ ЗАДАЧ, ЯК АНАЛІЗ ТОНАЛЬНОСТІ ВІДГУКІВ, РОЗРОБКА ЧАТ-БОТІВ ТА СТВОРЕННЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ. ЗА РЕЗУЛЬТАТАМИ ПОРІВНЯННЯ МОЖНА ЗРОБИТИ ВИСНОВОК, ЩО ДЛЯ КОЖНОЇ КОНКРЕТНОЇ ЗАДАЧІ ІСНУЮТЬ СВОЇ ОПТИМАЛЬНІ ІНСТРУМЕНТИ ТА МЕТОДИ ВЕКТОРИЗАЦІЇ, ЯКІ ЗАБЕЗПЕЧУЮТЬ НАЙКРАЩИЙ БАЛАНС МІЖ ПРОДУКТИВНІСТЮ ТА ТОЧНІСТЮ.