ЛАБОРАТОРНА РОБОТА №11

Вступ до Natural Language Processing (NLP)

Виконала: Гульчак дар’я МІТ-31

Теоретичне дослідження етапів та моделей NLP

ОБРОБКА ПРИРОДНОЇ МОВИ (NATURAL LANGUAGE PROCESSING, NLP) Є КЛЮЧОВИМ НАПРЯМКОМ ШТУЧНОГО ІНТЕЛЕКТУ, ЩО ЗАБЕЗПЕЧУЄ РОЗУМІННЯ ТА ГЕНЕРАЦІЮ ЛЮДСЬКОЇ МОВИ КОМП'ЮТЕРАМИ. ПРОЦЕС ОБРОБКИ ТЕКСТУ СКЛАДАЄТЬСЯ З КІЛЬКОХ ПОСЛІДОВНИХ ЕТАПІВ. ПЕРШИМ КРОКОМ ЗАВЖДИ Є ТОКЕНІЗАЦІЯ, ЯКА ПЕРЕДБАЧАЄ РОЗБИТТЯ СУЦІЛЬНОГО ТЕКСТУ НА ОКРЕМІ ЗНАЧУЩІ ОДИНИЦІ — ТОКЕНИ, ЯКИМИ МОЖУТЬ БУТИ СЛОВА, РОЗДІЛОВІ ЗНАКИ АБО ЧАСТИНИ СЛІВ. ПІСЛЯ ЦЬОГО ЧАСТО ЗАСТОСОВУЮТЬСЯ МЕТОДИ НОРМАЛІЗАЦІЇ ТЕКСТУ, ТАКІ ЯК ЛЕМАТИЗАЦІЯ ТА СТЕМІНГ. СЕМІНГ — ЦЕ СПРОЩЕНИЙ ПРОЦЕС ВІДСІКАННЯ ЗАКІНЧЕНЬ СЛІВ ДЛЯ ОТРИМАННЯ ЇХНЬОЇ ОСНОВИ (НАПРИКЛАД, ПЕРЕТВОРЕННЯ "БІГАВ" НА "БІГ"), ТОДІ ЯК ЛЕМАТИЗАЦІЯ Є СКЛАДНІШОЮ ЛІНГВІСТИЧНОЮ ОПЕРАЦІЄЮ, ЩО ПРИВОДИТЬ СЛОВО ДО ЙОГО СЛОВНИКОВОЇ ФОРМИ (ЛЕМИ) З УРАХУВАННЯМ МОРФОЛОГІЇ ТА КОНТЕКСТУ.

НАСТУПНИМ КРИТИЧНО ВАЖЛИВИМ ЕТАПОМ Є ВЕКТОРИЗАЦІЯ ТЕКСТУ, ОСКІЛЬКИ АЛГОРИТМИ МАШИННОГО НАВЧАННЯ НЕ МОЖУТЬ ПРАЦЮВАТИ БЕЗПОСЕРЕДНЬО З ТЕКСТОМ. ВЕКТОРИЗАЦІЯ ПЕРЕТВОРЮЄ ТОКЕНИ НА ЦИФРОВІ ВЕКТОРИ. ДЛЯ ЦЬОГО ВИКОРИСТОВУЮТЬСЯ РІЗНІ ПІДХОДИ: ВІД ПРОСТИХ ЧАСТОТНИХ МЕТОДІВ, ЯК-ОТ BAG OF WORDS ЧИ TF-IDF, ДО СКЛАДНИХ КОНТЕКСТУАЛЬНИХ ВКЛАДЕНЬ (WORD EMBEDDINGS). ОТРИМАНІ ВЕКТОРИ ВИКОРИСТОВУЮТЬСЯ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТУ — ВІДНЕСЕННЯ ДОКУМЕНТА ДО ПЕВНОЇ КАТЕГОРІЇ (СПАМ/НЕ СПАМ, АНАЛІЗ ТОНАЛЬНОСТІ), АБО ДЛЯ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ (NER), ЩО ДОЗВОЛЯЄ АВТОМАТИЧНО ВИДІЛЯТИ З ТЕКСТУ ІМЕНА, ЛОКАЦІЇ, ДАТИ ТА ОРГАНІЗАЦІЇ.

СУЧАСНИЙ NLP БАЗУЄТЬСЯ НА ВИКОРИСТАННІ РІЗНОМАНІТНИХ МОДЕЛЕЙ. СЕРЕД КЛАСИЧНИХ АЛГОРИТМІВ ПОПУЛЯРНИМИ ЗАЛИШАЮТЬСЯ НАЇВНИЙ БАЄСОВИЙ КЛАСИФІКАТОР ТА ЛОГІСТИЧНА РЕГРЕСІЯ, ЯКІ ЕФЕКТИВНІ ДЛЯ ПРОСТИХ ЗАДАЧ КЛАСИФІКАЦІЇ. ГЛИБОКЕ НАВЧАННЯ ПРИВНЕСЛО РЕКУРЕНТНІ НЕЙРОННІ МЕРЕЖІ, ЗОКРЕМА LSTM (LONG SHORT-TERM MEMORY), ЗДАТНІ АНАЛІЗУВАТИ ПОСЛІДОВНОСТІ ТА КОНТЕКСТ. ПРОРИВОМ У ЦІЙ СФЕРІ СТАЛА АРХІТЕКТУРА TRANSFORMERS, ЯКА ЛЕЖИТЬ В ОСНОВІ НАЙСУЧАСНІШИХ МОДЕЛЕЙ, ТАКИХ ЯК BERT ТА GPT. ЦІ МОДЕЛІ ВИКОРИСТОВУЮТЬ МЕХАНІЗМ "УВАГИ" (ATTENTION) ДЛЯ ОБРОБКИ ВСЬОГО ТЕКСТУ ОДНОЧАСНО, ЩО ДОЗВОЛЯЄ ДОСЯГАТИ БЕЗПРЕЦЕДЕНТНОЇ ЯКОСТІ В ГЕНЕРАЦІЇ ТА РОЗУМІННІ МОВИ.

Порівняльний аналіз методів векторизації тексту

У ХОДІ РОБОТИ БУЛО ПОРІВНЯНО ТРИ ОСНОВНІ ПІДХОДИ ДО ПЕРЕТВОРЕННЯ ТЕКСТУ В ЧИСЛА, ЗГІДНО ІЗ ЗАВДАННЯМ. РЕЗУЛЬТАТИ АНАЛІЗУ НАВЕДЕНО В ТАБЛИЦІ НИЖЧЕ.

Метод	**Принцип дії**	**Переваги та Недоліки**	**Складність реалізації**	**Застосування**
Bag of Words (BoW)	Текст представляється як "мішок" слів, де враховується лише їх кількість або наявність, без порядку.	**+** Дуже простий у розумінні. **-** Втрачає контекст і порядок слів; вектори розріджені (багато нулів).	Низька	Проста фільтрація спаму, визначення тематики простих текстів.
TF-IDF	Статистична міра, що оцінює важливість слова в документі відносно всієї колекції документів.	**+** Зменшує вагу часто вживаних загальних слів (стоп-слів). **-** Все ще не враховує семантику та порядок слів.	Низька	Пошукові системи, вилучення ключових слів, інформаційний пошук.
Word Embeddings (Word2Vec, GloVe)	Кожне слово відображається у векторний простір, де семантично схожі слова розташовані поруч.	**+** Вловлює зміст, синоніми та аналогії (король - чоловік + жінка = королева). **-** Потребує великих ресурсів для навчання; складніше інтерпретувати.	Висока	Машинний переклад, чат-боти, глибокий аналіз тональності.

Огляд інструментів та бібліотек для NLP

ДЛЯ ВИКОНАННЯ ЗАВДАНЬ NLP ІСНУЄ НИЗКА ПОТУЖНИХ БІБЛІОТЕК. НИЖЧЕ НАВЕДЕНО ПОРІВНЯННЯ ІНСТРУМЕНТІВ, ЗАЗНАЧЕНИХ У ЗАВДАННІ.

Інструмент	Основні функції та особливості	Підтримка мов	Простота використання	Типове застосування
NLTK	Найстаріша бібліотека для навчання та наукових досліджень. Величезний набір корпусів та алгоритмів.	Багатомовна, але часто потребує ручного налаштування моделей.	Середня (академічний синтаксис).	Навчання, лінгвістичний аналіз, прототипування.
SpaCy	Орієнтована на "industrial-strength" NLP. Дуже швидка, має вбудовані конвеєри обробки.	Відмінна підтримка багатьох мов (включно з українською).	Висока (зручний API).	Продакшн-системи, вилучення інформації, швидкий парсинг.
Hugging Face Transformers	Доступ до тисяч попередньо навчених моделей (BERT, GPT, T5). Де-факто стандарт для сучасного NLP.	Підтримує понад 100 мов завдяки мультимовним моделям.	Середня (потребує розуміння Deep Learning).	Складні задачі: генерація тексту, переклад, відповіді на запитання.
Gensim	Спеціалізується на тематичному моделюванні та семантичній подібності (Word2Vec, Doc2Vec).	Мовонезалежна (працює з математичними векторами).	Висока для специфічних задач.	Рекомендаційні системи, пошук схожих документів, кластеризація.

Висновки

ЗА РЕЗУЛЬТАТАМИ ПРОВЕДЕНОГО ДОСЛІДЖЕННЯ МОЖНА ЗРОБИТИ ВИСНОВОК, ЩО ВИБІР МЕТОДУ ТА ІНСТРУМЕНТУ В NLP ПРЯМО ЗАЛЕЖИТЬ ВІД ПОСТАВЛЕНОЇ ЗАДАЧІ. ДЛЯ ПРОСТИХ ЗАВДАНЬ КЛАСИФІКАЦІЇ, ДЕ НЕ ВАЖЛИВИЙ ГЛИБОКИЙ КОНТЕКСТ, КЛАСИЧНІ МЕТОДИ ВЕКТОРИЗАЦІЇ (TF-IDF) У ПОЄДНАННІ З БІБЛІОТЕКАМИ NLTK АБО SCIKIT-LEARN Є ДОСТАТНІМИ ТА ШВИДКИМИ. ВОНИ НЕ ВИМАГАЮТЬ ЗНАЧНИХ ОБЧИСЛЮВАЛЬНИХ ПОТУЖНОСТЕЙ.

ОДНАК ДЛЯ ПОБУДОВИ СУЧАСНИХ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ, ТАКИХ ЯК ЧАТ-БОТИ, СИСТЕМИ МАШИННОГО ПЕРЕКЛАДУ ЧИ ГЛИБОКОГО АНАЛІЗУ ТЕКСТІВ, НЕОБХІДНЕ ВИКОРИСТАННЯ ЩІЛЬНИХ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ (EMBEDDINGS) ТА ТРАНСФОРМЕРНИХ МОДЕЛЕЙ. У ЦЬОМУ ВИПАДКУ НАЙКРАЩИМ ВИБОРОМ Є БІБЛІОТЕКА HUGGING FACE TRANSFORMERS. ДЛЯ РОЗРОБКИ ШВИДКИХ ТА НАДІЙНИХ ІНЖЕНЕРНИХ РІШЕНЬ У ПРОДАКШНІ ОПТИМАЛЬНИМ ВИБОРОМ Є SPACY ЗАВДЯКИ ЇЇ ШВИДКОДІЇ ТА ЗРУЧНОСТІ. GENSIM ЗАЛИШАЄТЬСЯ НЕЗАМІННИМ ІНСТРУМЕНТОМ ДЛЯ СПЕЦИФІЧНИХ ЗАДАЧ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ ТА РОБОТИ З ВЕКТОРНИМИ ПРОСТОРАМИ СЛІВ