ВСТУП ДО NATURAL LANGUAGE PROCESSING (NLP)

Виконала : Кубєдінова Поліна

Що таке NLP?

Обробка природної мови (Natural Language Processing, NLP) — це галузь штучного інтелекту, яка займається взаємодією між комп'ютерами та людською мовою. Її мета — навчити машини читати, розуміти та відтворювати людську мову.
NLP дозволяє автоматизувати рутинні задачі, аналізувати величезні масиви текстових даних та створювати інтерфейси для взаємодії "людина–машина".

Основні етапи NLP (Pipeline)

Токенізація (Tokenization): Розбиття тексту на менші частини (речення, слова, символи).

Стемінг та Лематизація (Stemming & Lemmatization): Зведення слів до їхньої основи.

- Стемінг: Відрізає закінчення (грубо). "бігав" > "біг".
- Лематизація: Приводить до словникової форми (аналізує контекст). "кращого" > "хороший".

Векторизація (Vectorization): Перетворення тексту на числовий вигляд (вектори), зрозумілий комп'ютеру.

Розпізнавання сутностей (NER): Виділення з тексту імен, локацій, дат, організацій.
Приклад: "Ілон Маск (PERSON) купив Twitter (ORG)".

Порівняльний аналіз методів векторизації

Характеристика	Bag of Words (BOW)	TF-IDF	Word Embeddings (Word2Vec, GloVe)
Принцип роботи	Підрахунок частоти слів у документі. Порядок слів ігнорується.	Оцінка важливості слова: часто в документі, але рідко в усьому корпусі.	Відображення слів у багатовимірний простір, де схожі слова знаходяться поруч.
Переваги	Проста реалізація, інтуїтивно зрозумілий.	Зменшує вагу загальновживаних слів (стоп-слів), виділяє унікальні терміни.	Вловлює семантичний зміст (король - чоловік + жінка = королева).
Недоліки	Розріджені вектори (багато нулів), втрачає контекст і порядок слів.	Як і BoW, не враховує семантичну схожість слів та їх порядок.	Потребує великих даних для навчання, складніша інтерпретація.
Складність реалізації	Низька.	Низька.	Середня/Висока (або використання готових моделей).
Масштабованість	Погана (розмір вектора росте зі словником).	Погана (залежить від розміру словника).	Висока (фіксований розмір вектора, напр. 300).
Застосування	Проста класифікація спаму.	Пошукові системи, класифікація документів.	Чат-боти, машинний переклад, складний аналіз.

Огляд інструментів та бібліотек для NLP

Інструмент	Основні функції	Підтримка мов	Простота	Особливості
NLTK	Токенізація, стемінг, лінгвістичний аналіз.	Багатомовна (але фокус на англ.)	Середня (академічний стиль)	Чудова для навчання та досліджень, але повільна для продакшну.
SpaCy	NER, POS-tagging, парсинг залежностей, векторизація.	70+ мов (в т.ч. українська).	Висока (зручний API)	Швидка, "Industrial-strength", готова до продакшну.
Hugging Face	Доступ до тисяч попередньо навчених моделей (BERT, GPT, T5).	Майже всі існуючі мови.	Середня (потребує знань DL)	Стандарт індустрії для роботи з трансформерами.
Gensim	Тематичне моделювання, Word2Vec, подібність документів.	Агностична до мови.	Висока	Оптимізована для роботи з великими текстовими корпусами.

Висновки

- NLP — це критично важлива технологія, що поєднує лінгвістику та машинне навчання.
- Вибір методу векторизації (TF-IDF проти Embeddings) залежить від складності задачі: для простих задач достатньо статистики, для розуміння змісту потрібні нейромережі.
- Сучасні бібліотеки (SpaCy, Hugging Face) значно знижують поріг входження в технологію.
- Майбутнє за великими мовними моделями (LLM), які здатні розуміти контекст на рівні людини.