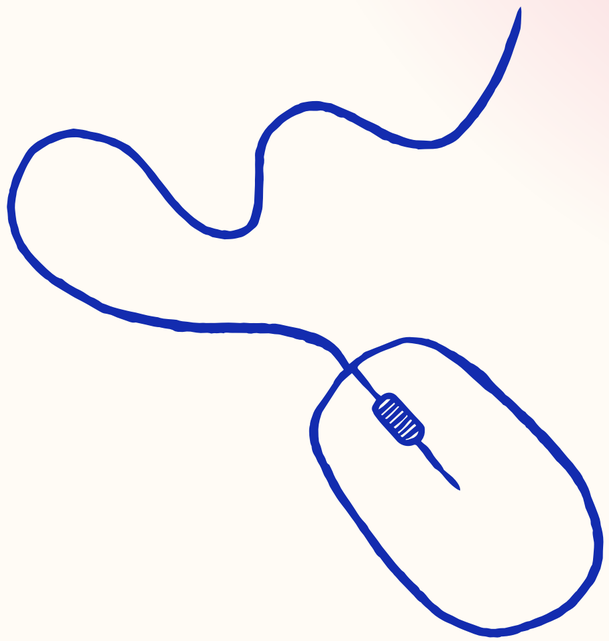
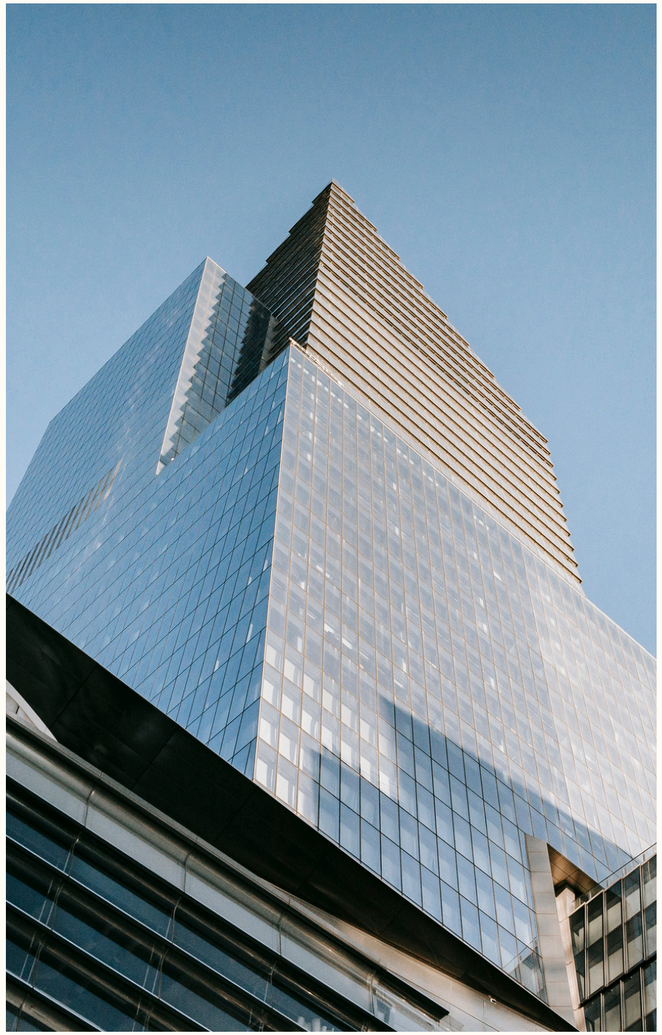
Вступ до Natural Language Processing (NLP)



інтелектуальний аналіз даних

Панченко
Владислава
МІТ-31

01 - Основні етапи NLP



Токенізація
це розбиття тексту на окремі компоненти (слова, речення, символи)

Методи токенізації:

- Простий поділ за пробілами.
- Використання регулярних виразів.
- Готові бібліотеки (NLTK, SpaCy).

Застосування: підготовка тексту для подальшої обробки

Лемматизація та стемінг

Стемінг – це обрізання слів до кореня (наприклад, "running" → "run").

Лемматизація – це перетворення слова до початкової форми з урахуванням граматики.

Критерій	Стемінг	Лемматизація
Точність	Низька	Висока
Швидкість	Висока	Низька
Залежність від контексту	Немає	Присутня

Векторизація тексту

Методи:

- Bag of Words (BOW): представлення тексту у вигляді матриці частот слів.
- TF-IDF: оцінка важливості слів у текстах колекції.
- Word Embeddings (Word2Vec, GloVe): числове представлення слів у багатовимірному просторі.

Класифікація тексту

Мета: призначення тексту до однієї чи кількох категорій.
Приклади задач: аналіз тональності, визначення теми.

Розпізнавання сутностей (NER)

Мета: виділення іменованих сутностей (імена, дати, локації).
Застосування: чат-боти, автоматизація обробки тексту.

02 – Порівняльний аналіз методів векторизації тексту

Метод	Переваги	Недоліки	Складність реалізації	Обробка великих даних	Застосування
Bag of Words	Простота, ефективність для малих наборів даних	Ігнорує порядок слів, створює великі розріджені матриці	Низька	Середня	Аналіз тональності, класифікація текстів
TF-IDF	Оцінює важливість слів у документах, знижує вагу поширених слів	Складність обчислень для великих корпусів	Середня	Середня	Інформаційний пошук, класифікація
Word Embeddings	Урахування контексту, компактність, висока якість представлення	Потребує великих даних для навчання, складність моделювання	Висока	Низька	Чат-боти, NER, рекомендаційні системи

03 – Огляд інструментів для NLP

Інструмент	Основні функції	Підтримка мов	Простота використання	Особливості
NLTK	Токенізація, лемматизація, стемінг, класифікація	Багато мов	Середня	Підходить для навчання, але має повільну обробку
SpaCy	NER, лемматизація, векторизація, підтримка мов	Основні мови	Висока	Швидка обробка, моделі для глибокого навчання
Hugging Face Transformers	Робота з моделями Transformers (BERT, GPT тощо)	100+ мов	Середня	Найкраще підходить для роботи з великими мовними моделями
Gensim	Word Embeddings (Word2Vec, LDA)	Багато мов	Висока	Ефективність для великих корпусів, тематичне моделювання

04 – Приклади застосувань NLP у різних галузях

Задача	Приклад застосування
Аналіз тональності	Відстеження відгуків у соціальних мережах
Чат-боти	Автоматизація обслуговування клієнтів
Рекомендаційні системи	Персоналізація товарів чи контенту
Розпізнавання сутностей	Автоматичний аналіз юридичних документів

05 – Висновки

Основні результати

Методи векторизації тексту:

- Word Embeddings демонструють найкращі результати у врахуванні контексту.
- TF-IDF залишається популярним для інформаційного пошуку завдяки простоті.

Інструменти NLP:

- SpaCy добре підходить для швидкої обробки тексту.
- Hugging Face Transformers є лідером у застосуванні великих мовних моделей.

Приклади застосувань

TF-IDF підходить для пошукових систем, тоді як Word Embeddings оптимальні для чат-ботів і рекомендаційних систем.

Висновки

Прості задачі (наприклад, класифікація) ефективно виконуються за допомогою NLTK чи SpaCy.
Складні задачі (наприклад, генерація тексту) вимагають використання Hugging Face Transformers.