# Лабораторна 11

Заяць Діана МІТ - 31

# Вступ до Natural Language Processing (NLP)

## Мета

Познайомитися з основними поняттями, методами та підходами у сфері обробки природної мови (NLP). Провести порівняльний аналіз популярних алгоритмів та інструментів, а також підготувати презентацію на цю тему.

---

# Опис завдання

## Теоретичне дослідження

1. **Основні етапи NLP**:
    - **Токенізація**: Розділення тексту на слова або інші значущі одиниці.
    - **Лемматизація та стемінг**: Зведення слів до їх базових форм.
    - **Векторизація тексту**:
        - Bag of Words (BOW)
        - TF-IDF
        - Word Embeddings (Word2Vec, GloVe)
    - **Класифікація тексту**: Визначення категорій тексту.
    - **Розпізнавання сутностей (NER)**: Виділення іменованих об’єктів у тексті.
2. **Ключові моделі для NLP**:
    - Наївний баєсовий класифікатор
    - Логістична регресія
    - LSTM
    - Transformers
    - GPT

---

## Порівняльний аналіз методів векторизації тексту

| Метод | Переваги | Недоліки | Складність реалізації | Застосування | Складність обробки великих данних |
|---|---|---|---|---|---|
| Bag of Words (BOW) | Простота реалізації | Ігнорує семантику | Низька | Класифікація, аналіз тональності | Висока |
| TF-IDF | Враховує частоту слів у контексті документа | Ігнорує порядок слів | Середня | Пошукові системи, аналіз тексту | Середня |
| Word Embeddings | Відображає семантичні зв’язки між словами | Необхідність великих обсягів даних | Висока | Рекомендаційні системи, чат-боти | Низька |

---

## Огляд інструментів для NLP

| Інструмент | Основні функції | Підтримка мов | Простота використання | Особливості |
|---|---|---|---|---|
| NLTK | Токенізація, стемінг, NER | Багато | Середня | Добре підходить для навчання |
| SpaCy | NER, векторизація, залежнісні дерева | Багато | Висока | Швидкий та ефективний |
| Hugging Face Transformers | Transformers, GPT, BERT | Багато | Висока | Готові моделі для різних задач |
| Gensim | Word2Vec, LDA, тематичне моделювання | Багато | Середня | Зосереджений на Word Embeddings |

---

## Приклади можливих застосувань NLP

- **Аналіз тональності**: Виявлення емоцій у відгуках користувачів.
- **Чат-боти**: Автоматизоване спілкування з користувачами.
- **Рекомендаційні системи**: Персоналізовані пропозиції.

---

## Основні результати порівняння

1. **Методи векторизації тексту**:
    - BOW та TF-IDF підходять для простих задач класифікації.
    - Word Embeddings краще для задач, що потребують семантичного аналізу.
2. **Інструменти NLP**:
    - NLTK добре підходить для початківців.
    - SpaCy та Hugging Face ефективні для комерційних проектів.

## Висновки

- Для простих задач використовуються BOW або TF-IDF.
- Для складних задач із великими обсягами даних доцільно застосовувати Word Embeddings.
- Hugging Face Transformers є найбільш універсальним інструментом для роботи з сучасними моделями.

## Приклади застосувань

- Наївний баєсовий класифікатор: Аналіз тональності відгуків.
- Transformers: Автоматизація перекладу текстів.
- Gensim: Тематичне моделювання для великих текстових корпусів.