Лабораторна робота 11: Вступ до Natural Language Processing (NLP)

Мета

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

Звіт

1. Основні етапи NLP

1.1 Токенізація
Процес розбиття тексту на менші одиниці, такі як слова чи речення.

- **Приклад**: *"Це тест." → ["Це", "тест", "."]*
- Використання: Аналіз тексту, побудова частотних словників.

1.2 Лемматизація та стемінг

- **Лемматизація**: Приведення слів до базової форми (леми), враховуючи контекст.
  Приклад: *"біг", "бігла" → "бігти"*.
- **Стемінг**: Видалення закінчень для отримання "основи" слова.
  Приклад: *"running", "runner" → "run"*.
- **Порівняння**: Лемматизація точніша, але повільніша, ніж стемінг.

1.3 Векторизація тексту
Перетворення тексту на числові представлення:

- **Bag of Words (BOW)**: Проста модель частотного представлення слів.
- **TF-IDF**: Враховує важливість слів у документі та в корпусі.
- **Word Embeddings**: Високорозмірні вектори, що враховують семантичну схожість слів (Word2Vec, GloVe).

1.4 Класифікація тексту
Розподіл тексту на категорії (наприклад, спам/не спам). Моделі: логістична регресія, SVM, LSTM, Transformers.

1.5 Розпізнавання сутностей (NER)
Завдання автоматичного виділення імен, локацій, дат тощо з тексту.
Приклад: *"Іван працює в Києві" → ["Іван" - PERSON, "Київ" - LOCATION]*.

---

2. Порівняльний аналіз методів векторизації тексту

Метод	Переваги	Недоліки	Складність	Застосування
Bag of Words (BOW)	Простота, швидкість.	Не враховує контекст.	Низька	Класифікація тексту, базовий аналіз.
TF-IDF	Враховує важливість слів.	Не враховує порядок слів і контекст.	Середня	Аналіз тональності, інформаційний пошук.
Word Embeddings	Семантична схожість, контекст.	Потребує багато даних для тренування.	Висока	Чат-боти, рекомендаційні системи.

---

3. Огляд інструментів для NLP

Інструмент	Основні функції	Підтримка мов	Простота	Особливості
NLTK	Токенізація, стемінг, NER.	Багато мов.	Середня	Освітній інструмент, багато прикладів.
SpaCy	Токенізація, лематизація, NER, векторизація.	Англійська, інші.	Висока	Оптимізована для швидкості, готові моделі.
Hugging Face	Підтримка моделей Transformers, GPT.	Багато мов.	Висока	Легке використання сучасних NLP-моделей.
Gensim	Word2Vec, LDA, векторизація тексту.	Багато мов.	Середня	Добре підходить для великих текстових корпусів.

---

4. Застосування NLP у різних галузях

1. **Аналіз тональності**: Визначення емоцій тексту (позитивні, негативні).
2. **Чат-боти**: Взаємодія з користувачами через текстові повідомлення.
3. **Пошукові системи**: Ранжування результатів пошуку.
4. **Рекомендаційні системи**: Аналіз тексту для персоналізації рекомендацій.
5. **Розпізнавання медичних даних**: Автоматизація аналізу клінічних записів.

Висновки

1. **Методи векторизації тексту**:
    - Для невеликих завдань підходять BOW та TF-IDF.
    - Для складних задач з контекстом (наприклад, NER) краще використовувати Word Embeddings.
2. **Інструменти**:
    - Для освітніх цілей та експериментів підійде **NLTK**.
    - Для великих та швидких проєктів краще використовувати **SpaCy** або **Hugging Face Transformers**.
3. **Застосування**:
    - Кожен метод чи інструмент має свою область застосування, яка залежить від завдання та обсягу даних.