Лабораторна робота №11

Звіт

З дисципліни “**Аналіз данних**”

на тему: “ Вступ до Natural Language Processing (NLP)”.

Студента 3 курсу: Групи
МІТ-31 Демиденко
Андрій

Київ - 2024р.

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.

Теоретичне дослідження

1. **Основні етапи NLP:**
    - **Токенізація:** Розбиття тексту на окремі елементи (слова, речення).
    - **Лемматизація та стемінг:**
        - Лемматизація: Зведення слів до їх лем (основної форми), враховуючи граматику.
        - Стемінг: Просте обрізання закінчень без врахування граматики.
    - **Векторизація тексту:**
        - **Bag of Words (BoW):** Створення векторів частотності слів без врахування порядку.
        - **TF-IDF:** Урахування частоти слова у тексті та його важливості в корпусі.
        - **Word Embeddings:** Представлення слів у вигляді щільних векторів у багатовимірному просторі (Word2Vec, GloVe).
    - **Класифікація тексту:** Визначення класу тексту (наприклад, позитивний/негативний відгук).
    - **Розпізнавання сутностей (NER):** Виявлення іменованих об'єктів у тексті (імена, дати, організації).
2. **Ключові моделі для NLP:**
    - Наївний баєсовий класифікатор
    - Логістична регресія
    - LSTM (рекурентна нейронна мережа для роботи з послідовностями)
    - Transformers (моделі на основі архітектури Attention, наприклад, BERT, GPT)

Порівняльний аналіз методів векторизації тексту

Метод	Переваги	Недоліки	Застосування	Складність реалізації
Bag of Words	Простота, швидкість	Ігнорує порядок слів, розмірність зростає	Простий аналіз тексту	Низька
TF-IDF	Враховує важливість слів у документі	Ігнорує семантику	Аналіз документів, пошукові системи	Середня
Word Embeddings	Враховує семантичну близькість слів	Вимагає багато даних для тренування	Машинний переклад, чат-боти	Висока

Огляд інструментів для NLP

Інструмент	Основні функції	Підтримка мов	Простота використання	Особливості
NLTK	Токенізація, стемінг, лемматизація	Багато	Середня	Широкий функціонал, але не завжди оптимальний
SpaCy	NER, токенізація, лемматизація	Англійська та ін.	Висока	Оптимізований для продуктивного використання
Hugging Face	Трансформери, GPT, BERT	Багато	Середня	Сучасні попередньо навчені моделі
Gensim	Word Embeddings, TF-IDF	Англійська	Висока	Сильна підтримка векторизації тексту

Основні результати:

1. **Токенізація:** Виділення окремих слів або фраз із тексту є базовим етапом NLP, який забезпечує основу для подальших обчислень.
2. **Лемматизація та стемінг:** Лемматизація дозволяє звести слово до його початкової форми, враховуючи контекст, тоді як стемінг – більш простий метод, що відкидає закінчення слів.
3. **Векторизація тексту:**
    a. **Bag of Words (BoW):** Проста та ефективна техніка для задач класифікації тексту.
    b. **TF-IDF:** Враховує частоту появи слів у документі та їх унікальність, що робить його більш точним для аналізу документів.
    c. **Word Embeddings (Word2Vec):** Успішно представляє семантику слів і виявляє схожість між ними.
1. **Класифікація тексту:** Застосування наївного баєсового класифікатора дозволило побудувати просту модель для визначення тональності тексту.
2. **Розпізнавання сутностей (NER):** За допомогою бібліотеки SpaCy успішно визначено іменовані сутності в тексті (імена, локації тощо).

Порівняння методів:

- Порівняльний аналіз методів векторизації тексту показав, що кожен підхід має свої переваги та недоліки, і вибір методу залежить від конкретної задачі.
- Сучасні моделі на основі Word Embeddings (Word2Vec, GloVe) та трансформери є найбільш ефективними для складних задач NLP.

Застосування:

NLP має широкий спектр застосувань, включаючи аналіз тональності, створення чат-ботів, автоматизацію перекладів, пошук інформації та рекомендаційні системи.

Висновок

У ході виконання роботи було проведено дослідження основних етапів обробки природної мови (NLP), що включає токенізацію, лемматизацію, стемінг, векторизацію тексту та класифікацію. Також були розглянуті популярні інструменти та бібліотеки для NLP, такі як NLTK, SpaCy, Gensim та Hugging Face Transformers, із зазначенням їх основних переваг, недоліків і сфер застосуванн