

Виділення числівників у тексті

Мета лабораторної роботи:

*Розробити програму, яка автоматично
знаходить числівники в тексті, визначає
їхні характеристики (тип, будову,
відмінок) і надає зручний інтерфейс для
взаємодії з користувачем.*

Використані бібліотеки

1. SpaCy

SpaCy – це бібліотека для обробки природної мови (NLP). Вона використовується для аналізу тексту, виявлення частин мови, таких як числівники, та визначення граматичних характеристик слів. SpaCy забезпечує потужні інструменти для розуміння структури тексту та його морфологічного аналізу.

2. Gradio

Gradio – це бібліотека для створення простих і зручних веб-інтерфейсів. Вона дозволяє легко інтегрувати програму з інтерфейсом користувача, що робить можливим взаємодію з програмою через веб-браузер без необхідності встановлення додаткового ПЗ. Завдяки Gradio користувач може вводити текст і бачити результати роботи програми в реальному часі.

Опис коду

```
!pip install spacy  
!pip install gradio  
!python -m spacy download uk_core_news_sm
```

1. Встановлення бібліотек

!pip install spacy

встановлюємо бібліотеку **spacy**, яка потрібна для обробки природної мови.

!pip install gradio

встановлюємо бібліотеку **Gradio**, яка використовується для створення веб-інтерфейсу.

!python -m spacy download uk_core_news_sm

завантажуємо українську мовну модель **uk_core_news_sm** для бібліотеки **spacy** для роботи з текстами українською мовою.

Опис коду

Основна частина коду складається з п'яти функцій:

1. `extract_numerals` - знаходить числівники в тексті та повертає їх список.
2. `determine_numeral_type` - визначає тип числівника.
3. `determine_numeral_structure` - визначає будову числівника.
4. `extract_numeral_features` - показує характеристики числівників (тип, будова, відмінок).
5. `create_interface` - створює інтерфейс для взаємодії з користувачем через Gradio.

Створення словників

```
common_numerals = ["нуль", "один",
"два", "три", "четири", "п'ять",
"шість", "сім", "вісім", "дев'ять",
"десять",
"одинадцять", "дванадцять",
"тринадцять", "чотирнадцять",
"п'ятнадцять",
"шістнадцять",
"сімнадцять", "вісімнадцять",
"дев'ятнадцять", "двадцять",
"тридцять",
"сорок",
"п'ятдесят", "шістдесят", "сімдесят",
"вісімдесят", "дев'яносто", "сто",
"двісті", "триста",
"чотириста", "п'ятсот", "шістсот",
"сімсот", "вісімсот",
"дев'ятсот",
"тисяча", "мільйон", "мільярд"]
```

Словник зі словами, які не завжди коректно опрацьовуються

```
num_type_translation = {"Card": "кількісний", "Ord": "порядковий"}  
  
pos_translation = {"NUM": "числівник", "NOUN": "іменник", "ADJ": "прикметник", "PROPN": "власна назва", "ADV": "прислівник"}  
  
case_translation = {  
    "Nom": "називний", "Gen": "родовий", "Dat": "давальний", "Acc": "зناхідний",  
    "Ins": "орудний", "Loc": "місцевий", "Voc": "кличний"  
}
```

Словники, які використовуються для перекладу лінгвістичних характеристик з англійських абревіатур

Функція extract_numerals

Функція для виділення числівників у тексті

```
def extract_numerals(text):
    if not text.strip():
        raise ValueError("Помилка:
введіть непорожній текст для
аналізу.")

    doc = nlp(text)

    numerals = []
    current_numeral = []

    for token in doc:
        if token.pos_ == "NUM" or
token.morph.get("NumType") in
(["Card"], ["Ord"]) or
token.dep_ == "nummod":
```

1. Перевірка на порожній текст;

2. Аналіз тексту за допомогою spaCy;

Створюється об'єкт doc, який дозволяє обробляти текст та аналізувати його на рівні токенів.

3. Ініціалізація списків;

- numerals – список для збереження всіх числівників, знайдених у тексті.
- current_numeral – список для збереження поточного числівника, який може складатися з одного або кількох слів.

4. Проходження через кожен токен у тексті;

- Токен вважається числівником, якщо:
- його частина мови є "NUM".
- в морфологічних атрибутих він має ознаку "NumType" із значенням "Card" або "Ord".
- токен є у списку числівників.

5. Додавання числівника до списку:

6. Створення списку числівників як текст:

Третій модуль

sounds - характеризує звуки та підраховує кількість голосних і приголосних у тексті

```
sounds = {
    "а": {
        "тип": "голосна",
        "ряд": "заднього ряду",
        "підняття": "низького
підняття",
        "відкритість": "відкрита",
        "лабіалізація": "нелабіалізована"
    },
    "б": {
        "тип": "приголосна",
        "з акустичного погляду": "шумна, дзвінка",
        "за місцем творення": "губно-
губна",
        "за способом творення": "
```

Словник характеристик звуків

```
def sound_characteristics(sound):
    characteristics =
sounds.get(sound)
    if characteristics:
        return f"Характеристики звуку
[{sound}]:\n" +
"\n".join([f"{key.capitalize()}: {value}" for key, value in
characteristics.items()])
    else:
        return f"Графема [{sound}] не
позначає звуків"

def analyze_text(text):
    analyze_result = []
    for element in text.lower():
        if element in sounds:
```

Код для третього модулю

Функція `create_interface`

Створення інтерфейсу

```
with gr.Blocks() as demo:  
    gr.Markdown("""<h1  
style="text-align: center;">Виділення  
числівників у тексті</h1>""")  
    text_input =  
        gr.Textbox(label="Введіть текст  
(включаючи числівники для аналізу)",  
        lines=4, placeholder="Наприклад: У на-  
        налічується більше чотирьохсот корів,  
        триста свиней і близько ста двадцяти  
        овець.")  
    extract_button =  
        gr.Button("Виділити числівники")  
    result_output =  
        gr.Textbox(label="Знайдені  
числівники", interactive=False)  
    characteristics_button =
```

1. Створюється блок `gr.Blocks` з заголовком та текстовим полем для введення тексту.
2. Створюються кнопки:
 - "Виділити числівники" – для запуску функції виділення числівників.
 - "Вивести характеристики числівників" – для виведення характеристик знайдених числівників.
3. Логіка кнопок:
 - `handle_extract` – виділяє числівники і робить кнопку для виведення характеристик видимою.
 - `handle_characteristics` – виводить характеристики числівників.

Як користуватися програмою

1. Введення тексту

Введіть текст безпосередньо в текстове поле.

2. Натисніть кнопку "Виділити числівники".

У полі "Знайдені числівники" буде відображені всі числівники, знайдені у введеному тексті.

3. Натисніть кнопку "Вивести характеристики числівників"

У полі "Характеристики числівників" будуть показані деталі числівників, включаючи їх тип, будову, та відмінок.

Демонстрація роботи на тестових прикладах

Виділення числівників у тексті

Введіть текст (включаючи числівники для аналізу)

"У нашій країні зараз проводиться масштабне будівництво нової дороги, яка буде проходити через кілька регіонів і з'єднає понад двісті населених пунктів. Загальна довжина дороги складе близько тисячі двохсот кілометрів, а для її будівництва знадобиться понад мільйон тонн асфальту і близько семисот тисяч тонн бетону. Крім того, залучено понад три тисячі працівників, включаючи будівельників, інженерів і водіїв."

Визначники

Знайдені числівники

кілька, двісті, двохсот, мільйон, семисот тисяч, три тисячі

Вивести характеристики числівників

Характеристики числівників

кілька:

кілька -> Тип: кількісний, Будова: простий, Відмінок: знахідний

двісті:

двісті -> Тип: кількісний, Будова: простий, Відмінок: знахідний

двохсот:

двохсот -> Тип: кількісний, Будова: простий, Відмінок: родовий

мільйон:

мільйон -> Тип: кількісний, Будова: простий, Відмінок: знахідний

семисот тисяч:

семисот -> Тип: кількісний, Будова: складений, Відмінок: родовий

тисяч -> Тип: кількісний, Будова: складений, Відмінок: родовий

три тисячі:

три -> Тип: кількісний, Будова: складений, Відмінок: знахідний

тисячі -> Тип: кількісний, Будова: складений, Відмінок: знахідний