Лабораторна робота №3
Грищенко Юрій, ПЗС-2
NER (Named-entity recognition)

Працюватимемо з наступним текстом:

У романі «Місто» Валер'ян Підмогильний описав селянську українську молодь, яка на початку 1920-х років тисячами потягнулась у міста, щоб завоювати і зробити своїм українське місто, влити в нього свіжу селянську кров, зліквідувати антагонізм між українським містом і селом.
Твір не був подібний до традиційної народницької прози XIX ст., бо автор орієнтувався на європейський роман 19-початку 20 століття, засвоївши традицію романістики Оноре де Бальзака, Гі де Мопассана, Анатоля Франса, Джека Лондона, а також вітчизняну — Агатангела Кримського, Володимира Винниченка.
Після 1991 року неодноразово перевидавався різними українськими видавництвами. 2012 року перевиданий в Києві видавництвом «Український письменник» у збірці «Третя революція».[3] 2015-го перевиданий у видавництві Знання. У 2017 році роман перевиданий видавництвом «Основи».

1. Stanza
Вставновимо Stanza:

```
python -m venv .env
source .env/bin/activate
pip install stanza
```

Необхідні моделі та пакети коду для роботи з українським текстом встановлюються в runtime програми. Приклад простого коду:

```
from pathlib import Path
import stanza

text = Path("lab3.txt").read_text()

stanza.download("uk")
nlp = stanza.Pipeline("uk")
doc = nlp(text)

for entity in doc.entities:
    print(entity.text, entity.type)
```

Цей код автоматично будує для нас pipeline, тобто бібліотека сама завантажує необхідні пакети і відповідно створює «ланцюг» процесорів. При запуску бачимо використані пакети:

```
==========================
| Processor | Package     |
```

```
---------------------------
| tokenize  | iu          |
| mwt       | iu          |
| pos       | iu_charlm   |
| lemma     | iu_nocharlm |
| depparse  | iu_charlm   |
| ner       | languk      |
===========================
```

Отримуємо:

```
Місто ORG
Валер'ян Підмогильний PERS
Оноре де Бальзака PERS
Гі де Мопассана PERS
Анатоля Франса PERS
Джека Лондона PERS
Агатангела Кримського PERS
Володимира Винниченка PERS
Києві LOC
Український письменник» ORG
Третя революція MISC
Основи ORG
```

Бачимо пропущене видавництво «Знання» (єдине видавництво, яке в тексті зазначене без лапок), та зайвий символ у назві «Український письменник»» (помилка нейронної мережі?)

Цікаво, що правильно визначено власну назву «Місто», хоча це загальновживане слово. Але неправильно визначено тип сутності (це назва твору, а не ORGанізація).

Інші власні назви визначено правильно. (Збірка «Третя революція» визначена як «інше» (MISC), що по суті не є помилкою)

2. spaCy

```
import spacy
from pathlib import Path

text = Path("lab3.txt").read_text()

nlp = spacy.load("uk_core_news_sm")
doc = nlp(text)

for ent in doc.ents:
    print(ent.text, ent.label_)
```

Отримуємо:

```
Місто ORG
Валер'ян Підмогильний PER
Оноре де Бальзака PER
Гі де Мопассана PER
Анатоля Франса PER
```

```
Джека Лондона PER
Агатангела Кримського PER
Володимира Винниченка PER
Києві LOC
Український письменник ORG
Основи ORG
```

Ті самі помилки, але є відмінності:

- «Український письменник» цього разу не містить зайвих символів
- Збірник «Третя революція» та видавництво «Знання» не визначені

Цікаво, що обидві моделі визначили «Місто» як організацію, хоча в тексті явно вказано, що це роман. Бачимо, що ці інструменти є досить ефективними, але не ідеальними.