


# ОБРОБКА ПРИРОДНОЇ МОВИ (NLP) У PYTHON. РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ

ЛЕКЦІЯ 5



# РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ

- це підзадача видобування інформації, яка спрямована на пошук і класифікацію іменованих сутностей, згаданих в неструктурованому тексті, по заздалегідь певних категорій, таким як імена людей, організації, місця розташування, медичні коди, вираження часу, кількості, грошові значення, відсотки та інше.
- Для цього ви можете використовувати готову попередньо навчену модель NER за допомогою бібліотеки з відкритим вихідним кодом, таку як Spacy або Stanford CoreNLP.
- Тепер, якщо ви гадаєте, що попередньо навчені моделі NER не дають результату відповідно до ваших очікувань, які ви шукаєте (приклад: тварина, об'єкт), недоступні в попередньо навченій моделі NER, тоді ви можете навчити свою власну кастомну модель.

- 
- Для навчання кастомної моделі NER у вас має бути величезна кількість анотованих даних. Для цього ви повинні використовувати який-небудь інструмент анотації, наприклад:
    - Brat rapid annotation tool
    - GATE
    - WebAnno
  - Наприклад, WebAnno є досить простим, тому що постачається як файл jar, що означає, що вам не потрібно його встановлювати. А також він може використовуватися для складного проекту — кілька користувачів можуть одночасно працювати в одному проекті.

- Щоб запустити це веб-додаток, в консолі використовуйте наступну команду:

```
java -jar webanno-standalone-3.6.7.jar
```

<https://webanno.github.io/webanno/releases/3.6.7/docs/user-guide.html>

- Створіть новий проект, на сторінці налаштувань проектів, у вкладці “Projects” натисніть “Create”. Напишіть яку-небудь назву проекту. (Приклад: «Test\_Annotation») та оберіть “Project type” як annotation.

The screenshot displays the 'Projects Settings' web application. At the top, a red navigation bar contains the application logo, the title 'Projects Settings', a 'Home' link, and user information including 'Help', 'admin', and 'Log out (automatically in 29 min)'. The main content area is titled 'Test\_Annotation' and features a 'Delete' button and a 'Cancel' button. Below the title are several tabs: 'Details', 'Users', 'Documents' (which is active), 'Layers', 'Tagsets', and 'CAS Doctor'. Under the 'Documents' tab, there are sub-tabs for 'Guidelines', 'Constraints', and 'Export'. The 'Documents' section includes a 'Choose Files' button, a text input field containing 'No fi...osen', a 'Format' dropdown menu set to 'Plain text', and an 'Import' button. A list of documents is shown below, with the entry 'vidvidav-ochakivskykh-kotykiv'. At the bottom right of this list is a 'Delete' button. On the left side of the interface, there is a 'Projects' sidebar with a list containing 'Test\_Annotation' and a 'Create' button. Below this is an 'Import Project(s)' section with a 'Browse ...' button, two checkboxes ('Import permissions' is checked, 'Create missing users' is unchecked), and an 'Import' button.

- Після визначення деталей проекту з'являться кілька вкладок, таких як Користувачі, Документи, Шари, Набори тегів та ін. Звідти виберіть вкладку Документи і зробіть наступне: у списку виберіть формат "Plain text", завантажте текстовий файл текстового документа, для якого ми будемо готувати навчальні дані та натисніть "Import". Зразок тексту, який використовувався у вхідному текстовому файлі для підготовки даних навчання, наведено нижче:

Командувач Сил спеціальних операцій США у Європі Девід Тейбор відвідав 73-й морський центр спеціального призначення імені кошового отамана Антіна Головатого.

Про візит американського армійця повідомила пресслужба військової частини в понеділок, 3 травня.

Зазначається, що мета візиту – побачити та оцінити рівень військового співробітництва, підготовки та взаємодії між морськими підрозділами спеціальних операцій України та Сполучених Штатів Америки.

Разом із командувачем ССО ЗСУ генерал-майором Григорієм Галаганом вони оглянули навчально-тренувальну базу центру, обговорили нові можливості для навчання операторів та покращення взаємодії у сфері виконання завдань підрозділами.

“Навчання, які ми проводимо, мають важливе значення для підтримки взаємодії з нашими союзниками, партнерами та друзями. Цей навчальний центр надає нам можливість тренуватися разом і вчитися один у одного в стратегічно важливому регіоні”, – цитує Тейбора пресслужба 73-го морського центру.

- Отже, тепер давайте подивимось, як створити нову сутність. Для цього виконайте наступні кроки:
- 1. Перейдіть на вкладку "Tagsets".
- 2. Виберіть Named Entity зі списку наборів тегів.
- 3. Для створення нового тегу, наприклад "SOF", натисніть «Create» в розділі «Tags».

The screenshot displays the 'Projects Settings' interface for a project named 'Test\_Annotation'. The top navigation bar includes 'Projects Settings', 'Home', 'Help', 'admin', and 'Log out (automatically in 26 min)'. The main content area is divided into several sections:

- Projects:** A sidebar on the left shows 'Test\_Annotation' as the selected project.
- Tagsets:** A central list of tagsets includes 'Coreference mentions', 'Coreference relations', 'Dependency flavors', 'Named Entity tags' (highlighted), 'Operation', 'UD Universal Dependencies (v2)', and 'UD Universal POS tags (v2)'. A 'Create' button is at the bottom right of this list.
- Tagset Details:** A form on the right shows the details for the selected tagset: 'Name' (Named Entity tags), 'Language' (uk), and 'Description' (Named Entity annotation). There is a checkbox for 'Annotators may add new tags' and a 'JSON' export option with 'Save', 'Delete', and 'Cancel' buttons.
- Tags:** A list of tags includes 'LOC', 'LOCderiv', 'LOCpart', 'ORG', 'ORGderiv', 'ORGpart', 'OTH', 'OTHderiv', 'OTHpart', and 'PFR'. A 'Create' button is at the bottom right of this list.
- Tag Details:** A form on the right shows the details for the selected tag: 'Name' (SOF) and 'Description' (Special Operations Forces). 'Save' and 'Cancel' buttons are at the bottom right.
- Import Project(s):** A section at the bottom left with a 'Browse ...' button and checkboxes for 'Import permissions' (checked) and 'Create missing users'.
- Import:** A section at the bottom center with a 'Format' dropdown (Choose One), 'Files to import' (Choose Files, No file chosen), and an 'Overwrite' checkbox. A 'Submit' button is at the bottom right.

- Таким же чином ви можете створити свою кастомну сутність.
- Тепер в меню проекту виберіть Анотація. З'явиться нове спливаюче вікно, виберіть документ, який ви хочете анотувати звідти.

The screenshot displays the 'Annotation' web application interface. At the top, there is a red header with the application name 'Annotation', a 'Home' link, and user information: 'Help', 'admin', and 'Log out (automatically in 29 min)'. Below the header is a navigation bar with several tabs: 'Document' (containing 'Open', 'Prev.', 'Next', 'Export', 'Settings'), 'Page' (containing 'First', 'Prev.', 'Go to 1', 'Next', 'Last'), 'Script' (containing 'LTR/RTL'), 'Help' (containing 'Guidelines'), and 'Workflow' (containing 'Reset', 'Finish'). The main content area shows a document titled 'admin: Test\_Annotation/vidvidav-ochakivskykh-kotyktiv' with the text 'Showing 1-5 of 6 sentences [document 1 of 1]'. The document content is displayed in a list of five sentences, each with pink annotations above it. The annotations are labeled with 'SOE', 'LOC', and 'PER'. The first sentence is '1. Командувач Сил спеціальних операцій США у Європі Девід Тейбор відвідав 73-й морський центр спеціального призначення імені кошового отамана Антіна Головатого.' The second sentence is '2. Про візит американського армієця повідомила пресслужба військової частини в понеділок, 3 травня.' The third sentence is '3. Зазначається, що мета візиту — побачити та оцінити рівень військового співробітництва, підготовки та взаємодії між морськими підрозділами спеціальних операцій України та Сполучених Штатів Америки.' The fourth sentence is '4. Разом із командувачем ССО ЗСУ генерал-майором Григорієм Галаганом вони оглянули навчально-тренувальну базу центру, обговорили нові можливості для навчання операторів та покращення взаємодії у сфері виконання завдань підрозділами.' The fifth sentence is '5. "Навчання, які ми проводимо, мають важливе значення для підтримки взаємодії з нашими союзниками, партнерами та друзями.' On the right side of the interface, there is a 'Layer' dropdown menu set to 'Named entity' and an 'Annotation' section that currently displays 'No annotation selected!'.



- Як тільки ви закінчите з анотацією, натисніть «Export», виберіть «WebAnno TSV v3.2» у спливаючому вікні та експорту його.
- Тепер давайте почнемо кодування для створення остаточних форматуваних користувальницьких навчальних даних у форматі Spacy для навчання користувальницької моделі розпізнавання іменованих сутностей (NER) з використанням Spacy.

```
# Підготовка тренувальних даних в форматі Spacy
TRAIN_DATA = []
ent_list = []
from web_anno_tsv import open_web_anno_tsv

tsv = '/content/vidvidav-ochakivskykh-kotyktiv.tsv'

with open_web_anno_tsv(tsv) as f:
    for i, sentence in enumerate(f):
        #print(f"Sentence {i}:", sentence.text)
        ent_list_sen = []
        for j, annotation in enumerate(sentence.annotations):

ent_list_sen.append((annotation.start,annotation.stop,annotation.labe
l))

    ent_list.append(ent_list_sen)
ent_dic = {}
ent_dic['entities'] = ent_list[-1]
# Підготуйте підсумкові дані навчання
TRAIN_DATA.append([sentence.text,ent_dic])
```

## TRAIN\_DATA

```
[['Командувач Сил спеціальних операцій США у Європі Девід Тейбор відвідав 73-й  
{ 'entities': [(0, 10, 'SOF'),  
              (11, 35, 'SOF'),  
              (36, 39, 'LOC'),  
              (42, 48, 'LOC'),  
              (49, 61, 'PER'),  
              (71, 157, 'SOF')] }],  
 ['Про візит американського армієця повідомила пресслужба військової частини в  
{ 'entities': [] }],  
 ['Зазначається, що мета візиту – побачити та оцінити рівень військового співро  
{ 'entities': [(58, 85, 'SOF'),  
              (125, 158, 'SOF'),  
              (159, 166, 'LOC'),  
              (170, 195, 'LOC')] }],  
 ['Разом із командувачем ССО ЗСУ генерал-майором Григорієм Галаганом вони оглян  
{ 'entities': [(9, 21, 'SOF'),  
              (22, 29, 'SOF'),  
              (30, 45, 'SOF'),  
              (46, 65, 'PER'),  
              (80, 113, 'SOF')] }],  
 ['"Навчання, які ми проводимо, мають важливе значення для підтримки взаємодії  
{ 'entities': [] }],  
 ['Цей навчальний центр надає нам можливість тренуватися разом і вчитися один у  
{ 'entities': [] }]
```

- Тепер давайте спробуємо навчити нову свіжу модель NER, використовуючи підготовлені призначені для користувача дані NER. Визначте змінні, необхідні для обробки навчальної моделі.

```
model = None
model_dir=Path("model_ner")
n_iter=100

if model is not None:
    nlp = spacy.load(model)
    print("Loaded model '%s'" % model)
else:
    nlp = spacy.blank('uk')
    print("Created blank 'uk' model")

#Потім завантажте порожню модель для процесу, що виконує дію NER, і
#налаштуйте конвеєр тільки з NER за допомогою функції create_pipe.

if 'ner' not in nlp.pipe_names:
    ner = nlp.create_pipe('ner')
    nlp.add_pipe(ner, last=True)
else:
    ner = nlp.get_pipe('ner')

for _, annotations in TRAIN_DATA:
    for ent in annotations.get('entities'):
        ner.add_label(ent[2])

other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 'ner']
with nlp.disable_pipes(*other_pipes): # only train NER
    optimizer = nlp.begin_training()
    for itn in range(n_iter):
        random.shuffle(TRAIN_DATA)
        losses = {}
        for text, annotations in tqdm(TRAIN_DATA):
            nlp.update(
                [text],
                [annotations],
                drop=0.5,
                sgd=optimizer,
                losses=losses)
        print(losses)
```

Щоб протестувати навчену модель,

```
for text, _ in train_data:
    doc = nlp(text)
    print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
    print("Tokens", [(t.text, t.ent_type_, t.ent_iob) for t in doc])
```

```
Entities [{"Командувач", 'SOF'}, {"Сил спеціальних операцій", 'SOF'}, {"США", 'LOC'}, {"Європі", 'LOC'}, {"Девід Тейбор", 'PER'},
Tokens [{"Командувач", 'SOF', 3}, {"Сил", 'SOF', 3}, {"спеціальних", 'SOF', 1}, {"операцій", 'SOF', 1}, {"США", 'LOC', 3}, {"у",
Entities [{"командувачем", 'SOF'}, {"ССО ЗСУ", 'SOF'}, {"генерал-майором", 'SOF'}, {"Григорієм Галаганом", 'PER'}, {"навчально-тре
Tokens [{"Разом", '', 2}, {"із", '', 2}, {"командувачем", 'SOF', 3}, {"ССО", 'SOF', 3}, {"ЗСУ", 'SOF', 1}, {"генерал", 'SOF', 3},
Entities []
Tokens [{"", '', 2}, {"Навчання", '', 2}, {"", "", 2}, {"які", '', 2}, {"ми", '', 2}, {"проводимо", '', 2}, {"", "", 2}, {"мають
Entities []
Tokens [{"Цей", '', 2}, {"навчальний", '', 2}, {"центр", '', 2}, {"надає", '', 2}, {"нам", '', 2}, {"можливість", '', 2}, {"тренув
Entities [{"військового співробітництва", 'SOF'}, {"підрозділами спеціальних операцій", 'SOF'}, {"України", 'LOC'}, {"Сполучених Ш
Tokens [{"Зазначається", '', 2}, {"", "", 2}, {"що", '', 2}, {"мета", '', 2}, {"візиту", '', 2}, {"-", "", 2}, {"побачити", '', 2
Entities []
Tokens [{"Про", '', 2}, {"візит", '', 2}, {"американського", '', 2}, {"армія", '', 2}, {"повідомила", '', 2}, {"пресслужба", ''
```

Нарешті, збережіть модель на свій шлях, який зберігається в змінній `model_dir`.

```
if model_dir is not None:
    model_dir = Path(model_dir)
    if not model_dir.exists():
        model_dir.mkdir()
    nlp.to_disk(model_dir)
    print("Saved model to", model_dir)

model = spacy.load(model_dir)
```