



NLP: створення парсера для резюме

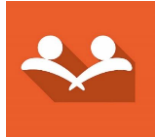
Доповідач: Парнак Д. А.
Керівник: Смиш О. Р.

Міжнародна наукова конференція

Демонстрація перших результатів роботи на міжнародній науковій конференції “Chance for Science”



UNIVERSITÄT
LEIPZIG



Certificate of Participation

Dmytro Parnak

is thanked for his outstanding participation as speaker in the Chance for Science Conference on September 8, 2022

with his presentation

“Ukrainian NLP: creating a CV parser”

September 12, 2022

Date

Carmen Bachmann

Prof. Dr. Carmen Bachmann

Prof. Dr. Carmen Bachmann
Chance for Science
Universität Leipzig
Gönnemannsche Straße 12, D-04103 Leipzig

Актуальність теми



Готові рішення

брак систем для аналізу
текстів резюме,
написаних природною
українською мовою



Багато даних

можливість
безперешкодно
працювати з великими
обсягами даних



Дослідження

проведення нових
досліджень на основі
отриманих даних

Мета дослідження

01

Аналіз

Дослідження
поточного стану
розробки рішень
для обробки текстів,
написаних
природною
українською мовою

02

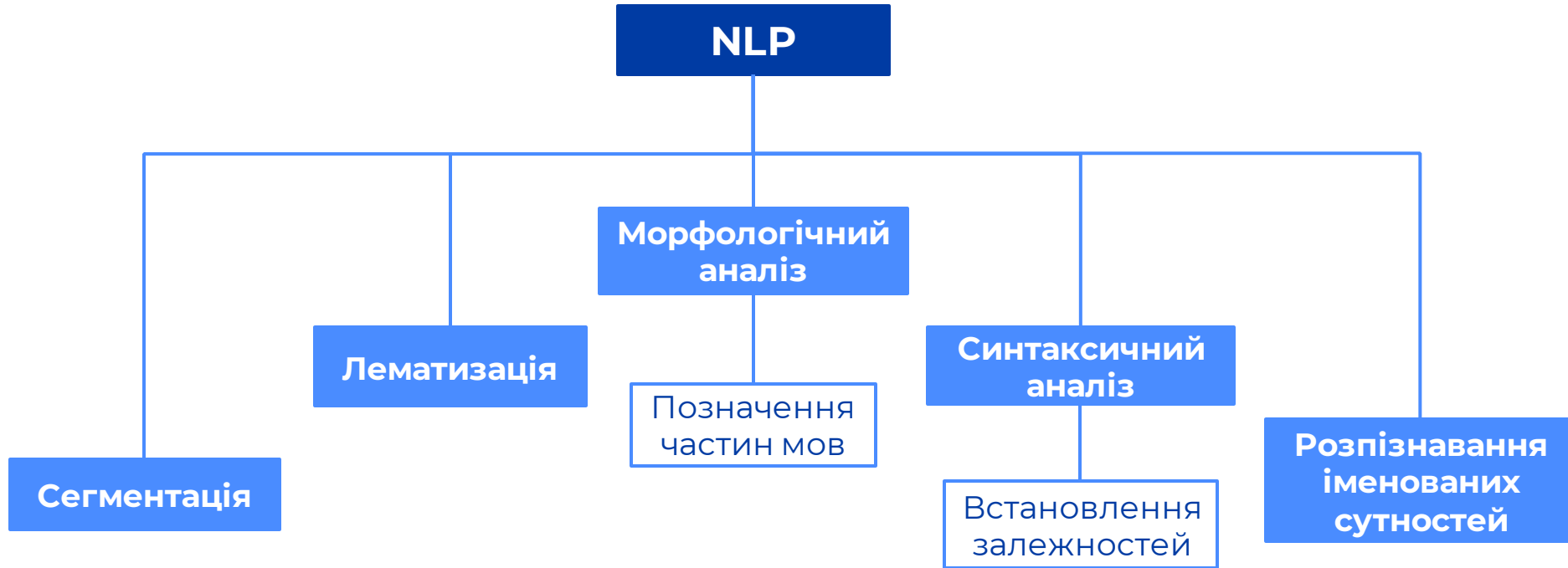
Розробка

Створення парсера
українськомовних
резюме

Завдання роботи

Розробка програми, що здійснюватиме виділення ключової інформації з резюме, написаних українською мовою.

Обробка природної мови



Аналіз інструментів

Корпуси

ГРАК

- Генеральний регіонально анований корпус української мови

Золотий морфосинтаксовий стандарт

- Громадська організація “Інститут української”

NLP моделі

UDPipe

- На основі “золотого стандарту”

Stanza

- Універсальні Залежності (Universal Dependencies, UD)

Ukrainian-QA

- Аналог Stanford Question Answering Dataset

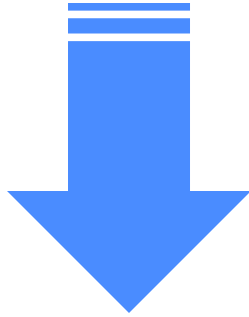


01

Укладання бази даних резюме

Укладання бази даних резюме

Scrapy



Work.ua

Robota.ua

Jooble

Jobs.ua



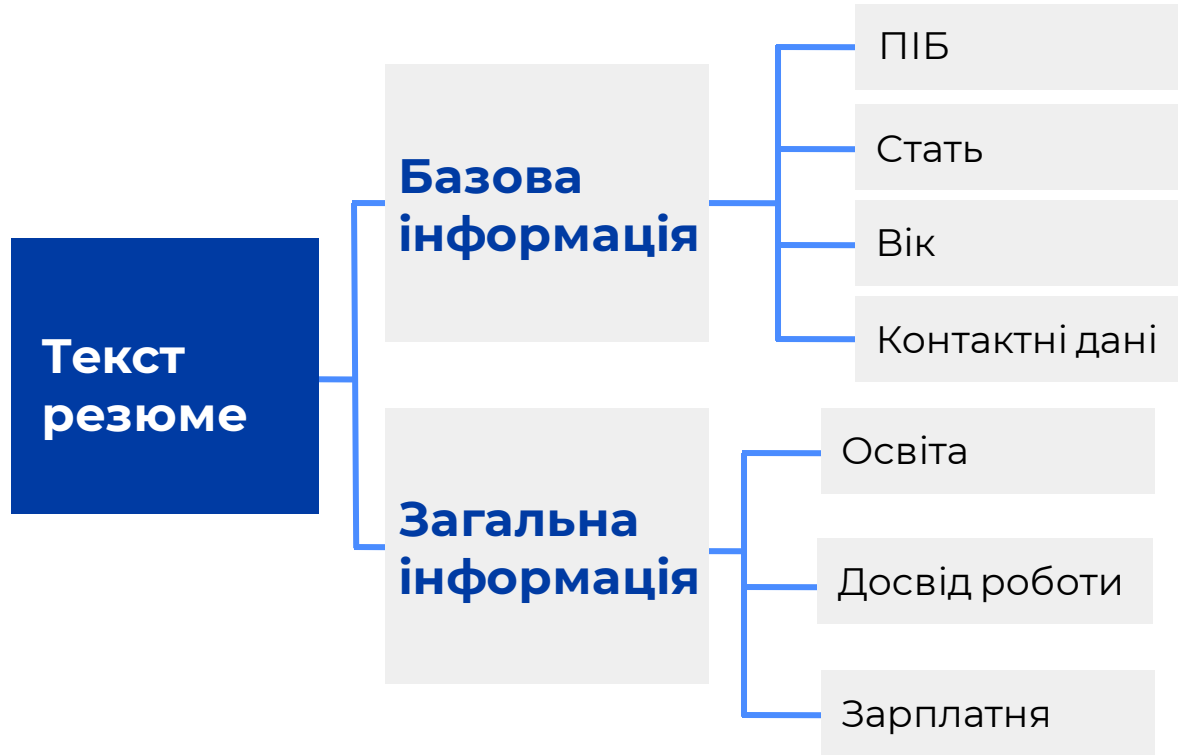
5,226

02

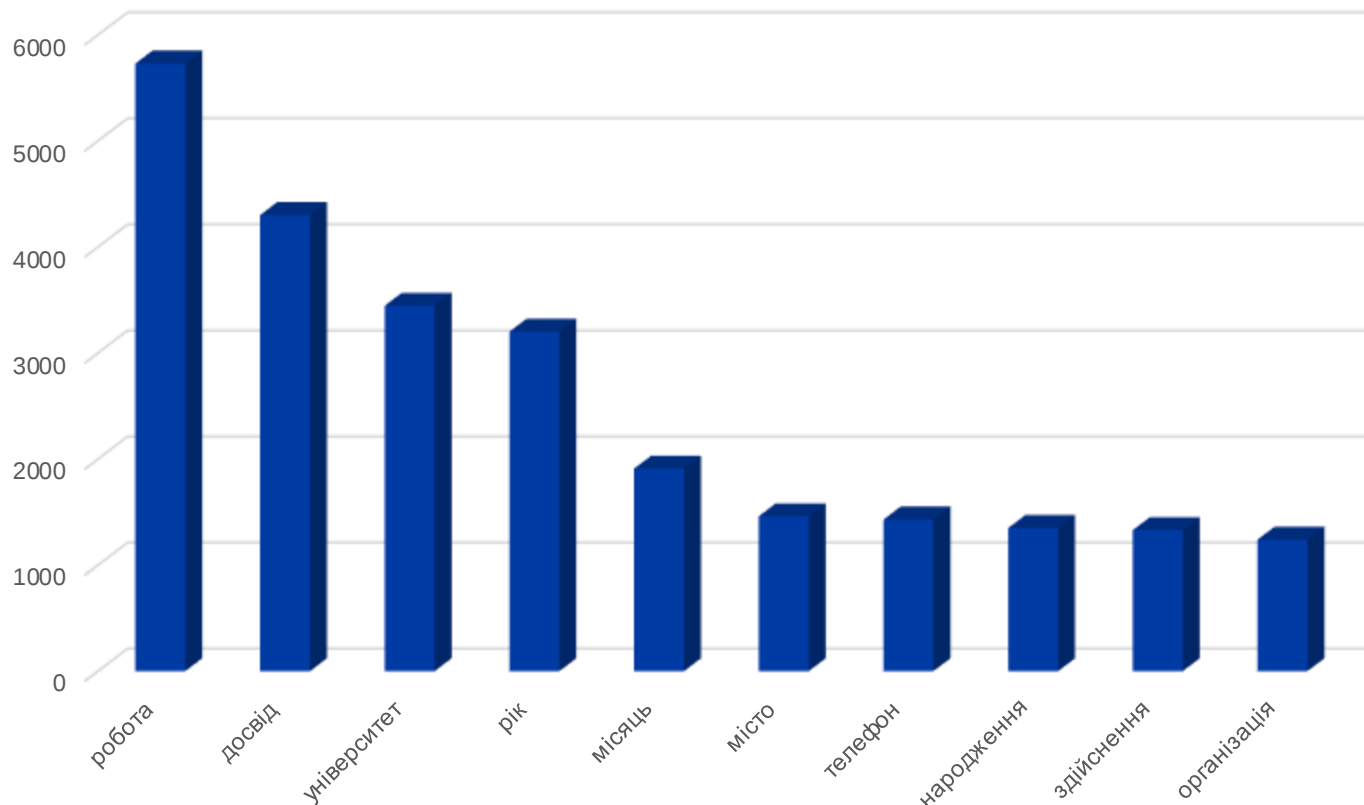
Аналіз резюме



Аналіз типових резюме



Найбільш повторювані слова у резюме





03

Розробка парсера

CoNLL-U -> JSON

```
{
  "sent_id": "2",
  "text": "Досвід роботи: Контент-менеджер з 03.2020 по 08.2021 (1 рік  
5 місяців)",
  "tokens": [
    {
      "id": "1",
      "form": "Досвід",
      "lemma": "досвід",
      "upos": "NOUN",
      "xpos": "Ncmsnn",
      "feats": "Animacy=I nan | Case=Nom | Gender=Masc | Number=Sing",
      "head": "0",
      "deprel": "root",
      "deps": " _",
      "misc": " _"
    }
  ],
}
```

Розробка парсера

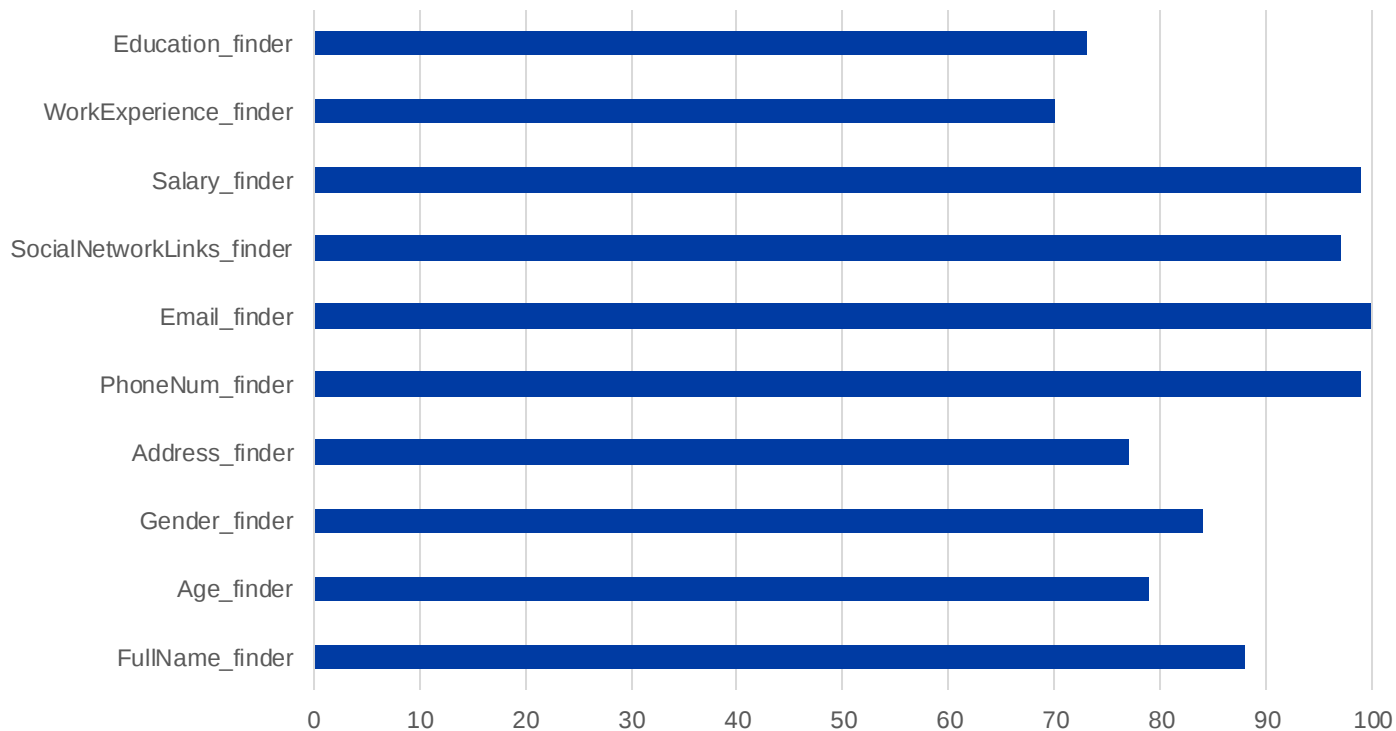
	UDPipe	Stanza	Ukrainian-QA	Регулярні вирази	Інші методи
ПІБ	●	●			
Стать	●	●			
Вік	●				●
Номер телефону, email, соцмережі				●	
Адреса	●				●
Зарплатня				●	
Освіта	●				●
Досвід роботи	●		●		

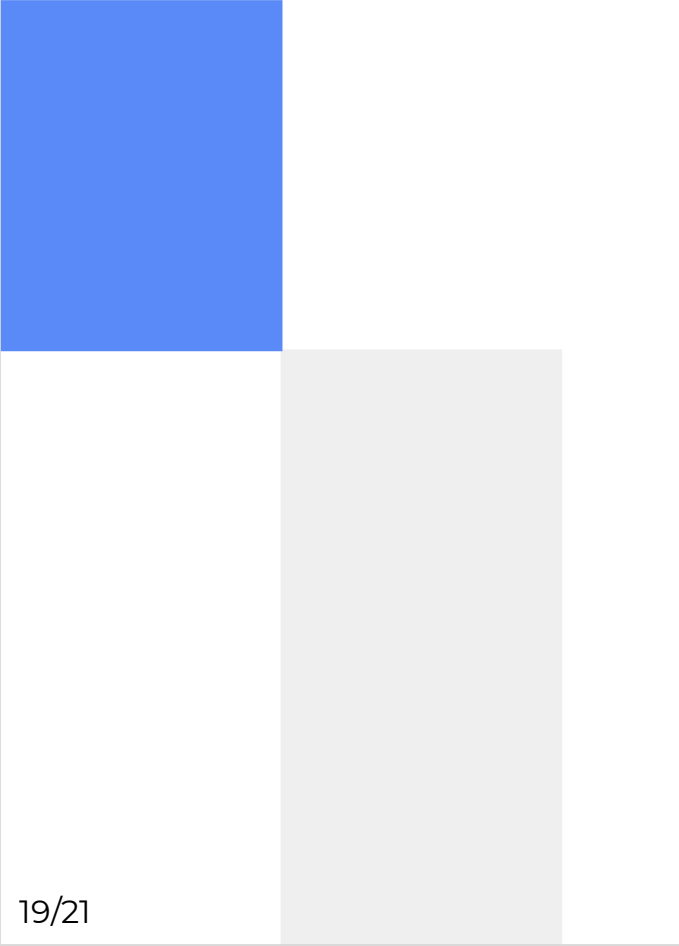
04

Аналіз точності



Пройдено тестів

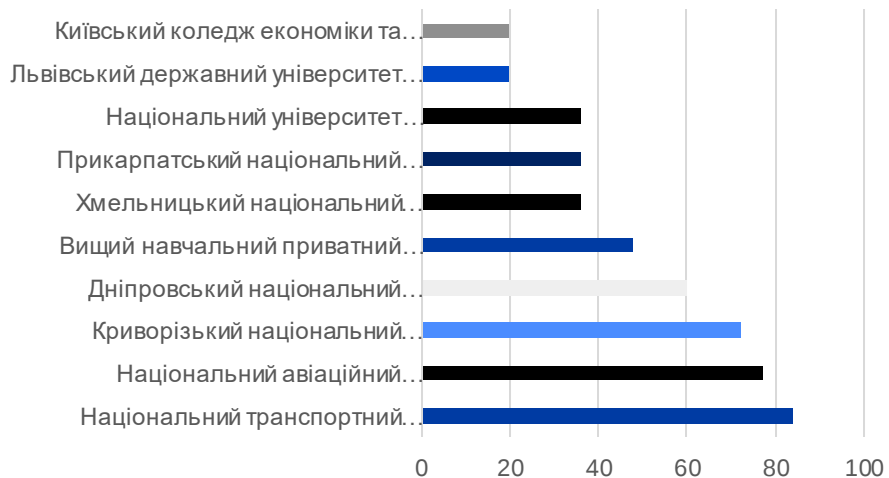




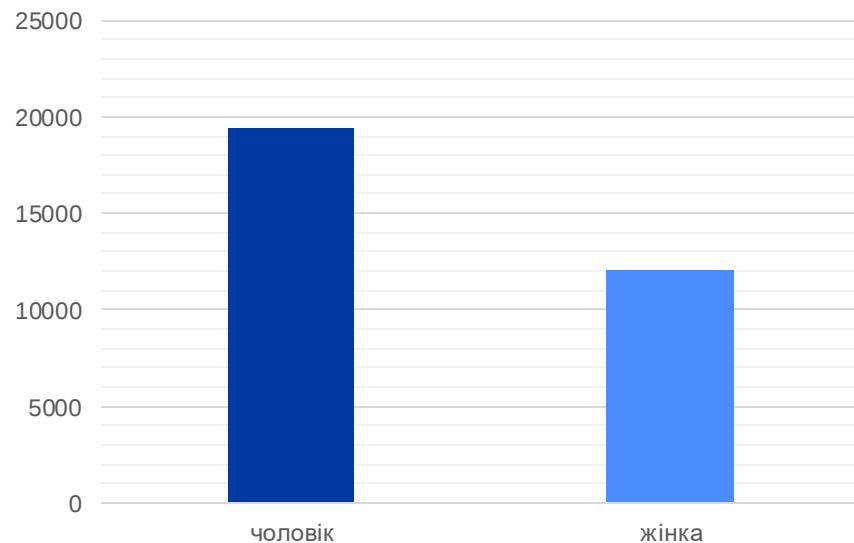
05

Додаткові дослідження

10 найбільш повторюваних навчальних закладів у резюме



Бажана заробітна плата



На основі 1000 резюме

Висновки

- Здійснено аналіз інструментів, які застосовуються для роботи з обробкою природної української мови
- Проаналізовано структуру резюме, виявлено закономірності
- Створено систему для автоматизованого парсингу резюме
- Встановлено точність системи та перспективи для проведення майбутніх статистичних досліджень

**Дякую
за
увагу!**

