

*Побудова
інфраструктури
обробки української
мови*

@vseloved

- *Lisp-програміст*
- *5+ років працюю в NLP*
- *Викладаю (зараз в
Prjctr.com.ua)*

<https://vseloved.github.io>

Дмитро Чаплинський

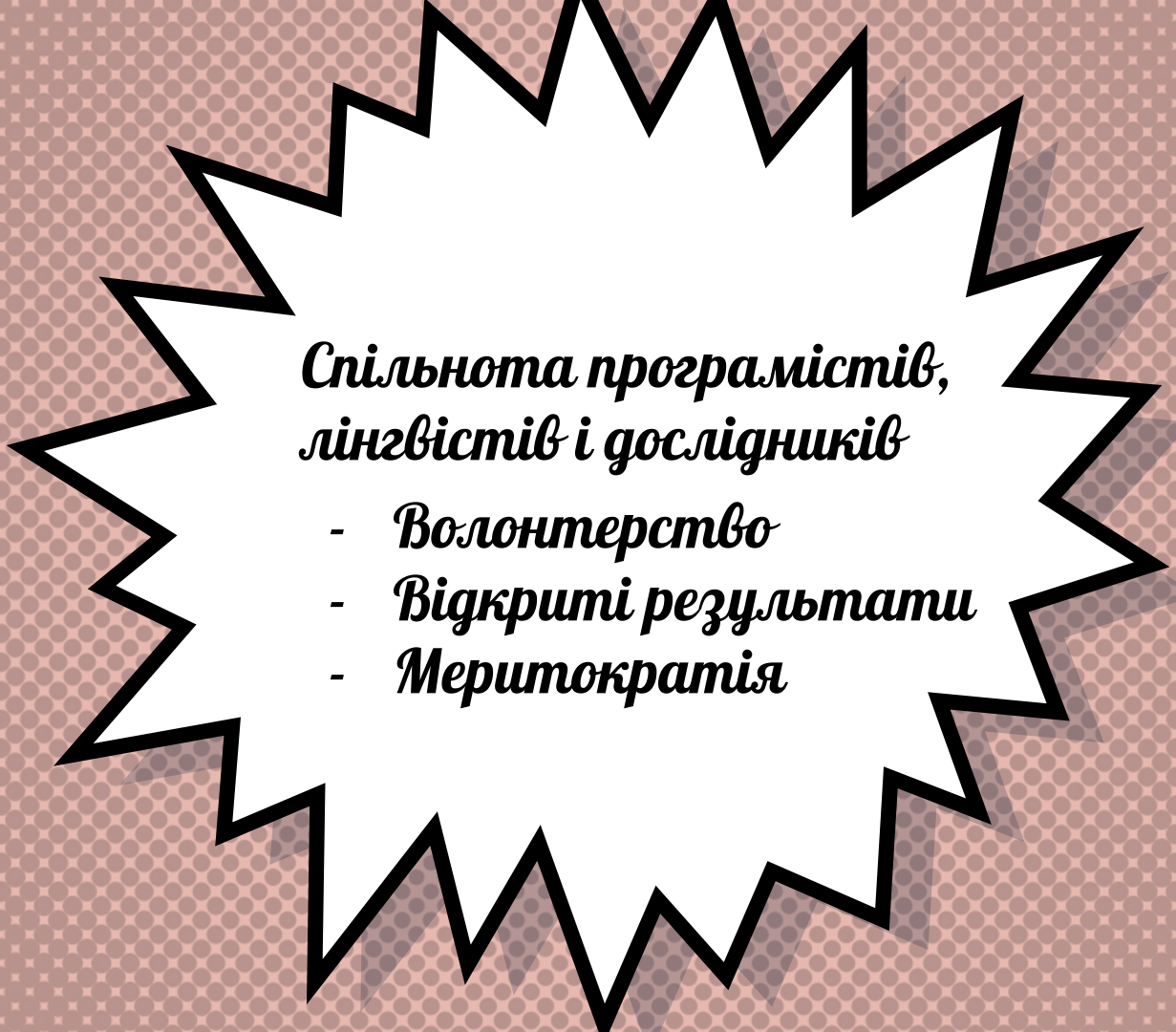
- *Python-програміст*
- *NLP-фанат*
- *Співзасновник*

Канцелярської Сотні

github.com/dchaplinsky

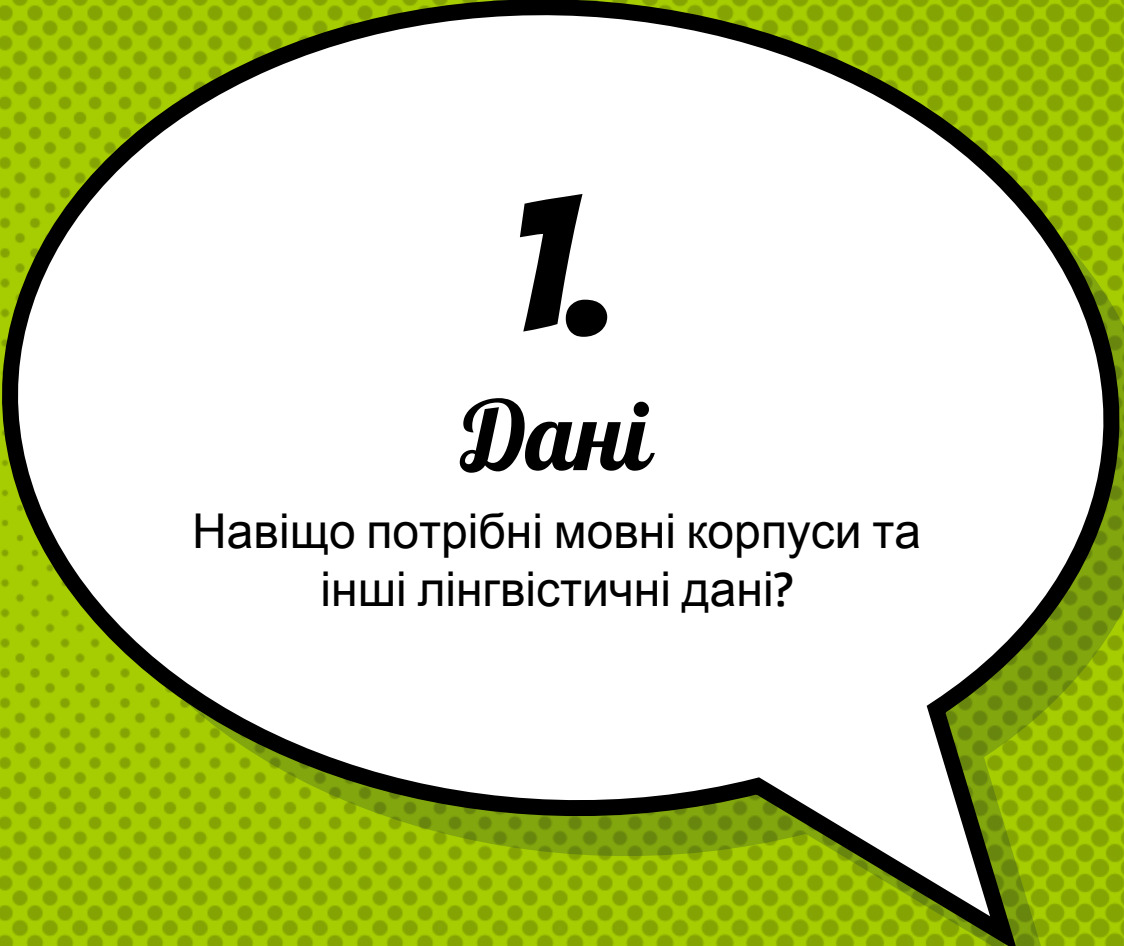


LANG.ORG.UA



*Спільнота програмістів,
лінгвістів і дослідників*

- Волонтерство*
- Відкриті результати*
- Меритократія*



1.

Дані

Навіщо потрібні мовні корпуси та
інші лінгвістичні дані?

Які дані використовуються в NLP

- × Словники
- × Корпуси
- × (Семантичні) мережі та інші бази даних

Навіщо дані потрібні?

- × Пошук прикладів
- × Перевірка гіпотез
- × Тестування алгоритмів і моделей
- × Тренування моделей

DATA IS 10X MORE POWERFUL THAN ALGORITHMS

“The Unreasonable Effectiveness of Data”

<http://youtu.be/yvDCzhbjYWs>



Що є для основних мов?

- × Словники, тезауруси, лексикони, газетіри
- × Freebase, Wordnet
- × Brown corpus
- × Treebanks, Propbanks, CCGBanks, AMRBBanks
- × Wikimedia
- × Gigaword
- × Web: Clueweb, Common Crawl
- × Twitter
- × Спеціалізовані корпуси (shared tasks)

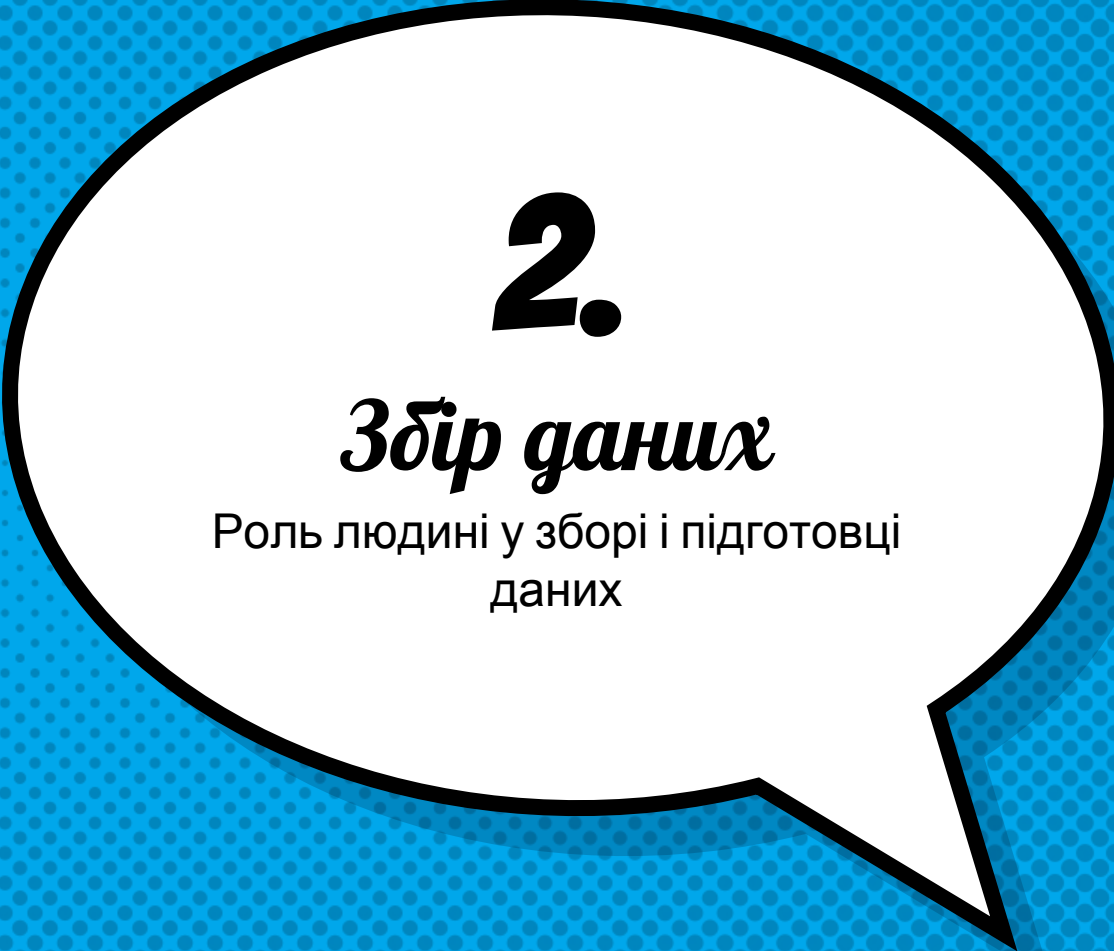
Що є для української?

(This page is intentionally left blank)

Що є для української?

- × ВЕСУМ
- × Wikimedia

**ALL YOUR
DATASETS
ARE BELONG
TO US**



2.

Збір даних

Роль людини у зборі і підготовці
даних

DATA SCRAPING

Збирання даних з
відкритих та напіввідкритих
джерел



Джерела даних

- × Періодичні видання
- × Бібліотеки
- × Реєстри та онлайн БД
- × Пошукові системи

SPIDERS ARE CUTE!

```
class PravdaComUaSpider(scrapy.Spider):
    name = "pravda_com_ua"
    allowed_domains = ["pravda.com.ua"]

    def parse(self, response):
        for href in response.css(".article__title a::attr('href')"):
            url = response.urljoin(href.extract())
            yield scrapy.Request(url, callback=self.parse_article)

    def parse_article(self, response):
        if response.url.startswith("http://www.pravda.com.ua/news/"):
            l = ItemLoader(CorpusItem(), response)
            l.add_css("img", ".post_news__photo_img::attr(src)")
            l.add_css("title", ".post_news__title::text")
            l.add_css("tags", ".post__tags__item a::text")
            l.add_css("text", ".post_news__text")
            l.add_css("date_of_publish", '.post_news__date::text')
            l.add_css("source", '.post__source::text')
            l.add_value("url", response.url)
            yield l.load_item()
```



Декілька корисних порад

- × Будьте креативними
- × Поважайте законодавство та авторські права
- × Не перенавантажуйте сайти, з котрих збираєте інформацію
- × Орендуйте сервер
- × Спілкуйтесь з іншими!

DATA ANNOTATION!

- × Шукайте джерела вже проанотованих даних
- × Аотуйте самі
- × Шукайте готові інструменти (Brat, Vulyk)
- × Допомога залу: Mechanical turk, fiverr, Толока



І ще трохи корисних порад

- × Різні задачі — різний рівень анотаторів
- × Будьте готові до витрат
- × Діліться результатами
- × Є можливість — залучайте волонтерів



3.

Вулик

Автоматизація краудсорсингу

THE VULYK

- × Був розроблений у підвалах Канцелярської сотні
- × Є фреймворком для краудсорсінгу
- × Можна додати будь який тип завдання, розробивши невеличкий плагін
- × Написан на Python/Flask/Mongo, open source





Vulyk: crowdsourcing platform

Інструкція

Завдання

Топ-10 волонтерів

Logout

Опрацював всього: 1 Позиція у рейтингу: 3

- 1 Одіозний український тенор Володимир Гришко вкотре опинився у центрі скандалу
- 2 Тенор поскаржився на те , що львівська громада змусила його відмовитись від концерту у Львівській філармонії , оскільки напередодні розгону Євромайдану Гришко виступив на Антимайдані
- 3 А у випадку , якби Гришко таки не скасував концерт , львів'яни пообіцяли влаштувати акцію протесту біля філармонії
- 4 Згодом не відбувся концерт народного артиста у Трускавці
- 5 Обурення пана Гришка виглядає щонайменше дивно , адже після тодішнього скасування концерту тенор , чи то з переляку , чи з ввічливості , запевнив кореспондента « ВЗ » , що позицію львів'ян підтримує

Vulyk NER tagging (based on BRAT)



Vulyk: crowdsourcing platform

[Інструкція](#)

[Завдання](#)

[Топ-10 волонтерів](#)

[Logout](#)

Опрацював всього: 1 Позиція у рейтингу: 3

Так

, удень

вона

яскраво-червона

, а

ввечері

— біла

білий

adj

f

v_naz

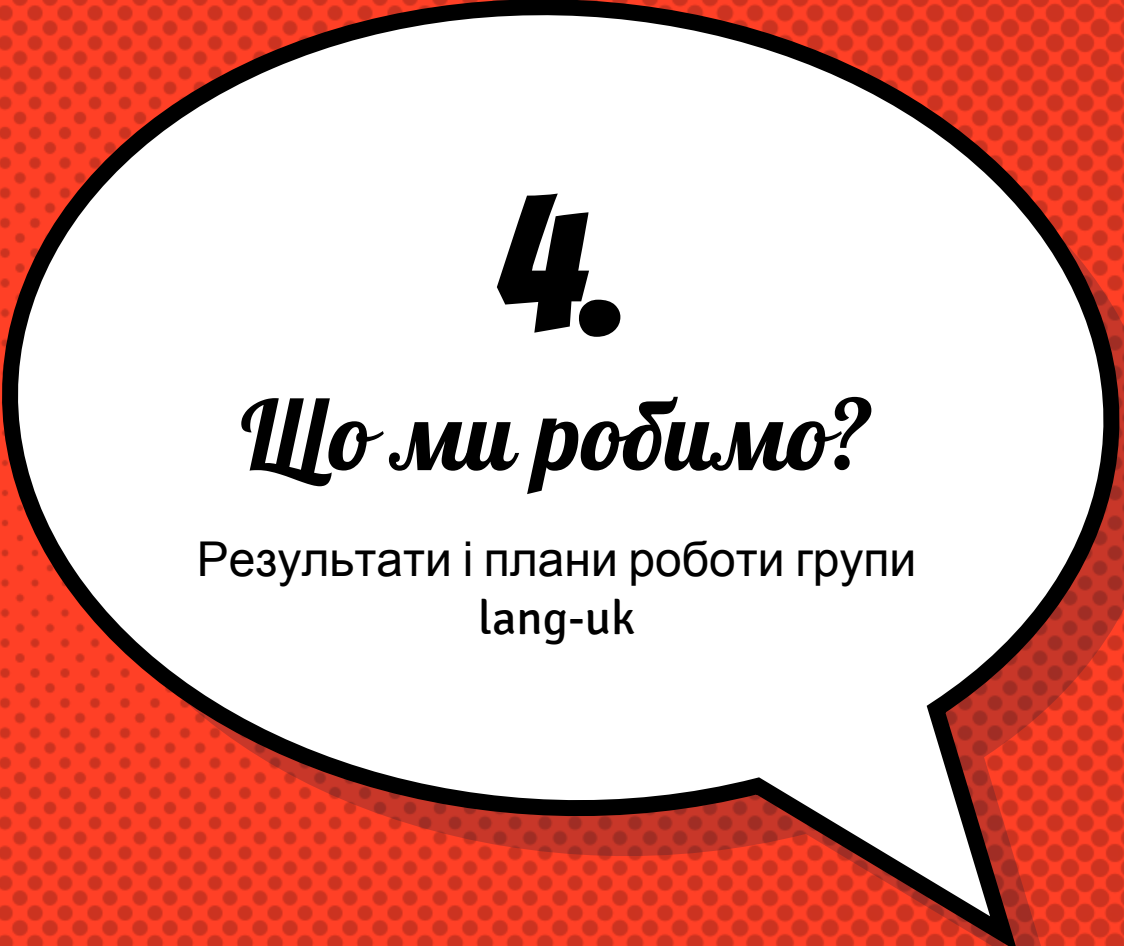
compb

Коректний варіант відсутній

Пропустити завдання

Надіслати форму

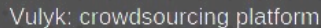
vulyk POS tagging



4.

Що ми робимо?

Результати і плани роботи групи
lang-uk



Оцененов всього: 25 Позиція у рейтингу: 1

NER-корпус (<https://github.com/lang-uk/ner-uk>)

На травневій сесії депутати Сумської міської ради ухвалили рішення про утворення комісії з питань у міських

клінічних лікарнях №1, №5 та №4 трьох Центрів первинної медико-санітарної допомоги

- ✗ 229 текстів з корпусу БрУК, ~200k tokenів, 6751 NER-сутність
- ✗ ~ 1 місяць роботи 3-х анотаторів і одного редактора³, однак за внесення
- ✗ ~ пару десятків тисяч гривень

Нині в Сумах налічується 39 сімейних лікарів при потребі 181

× Натренована модель MITIE -

[http://lang.org.ua/static/downloads/ner_models/uk_model.dat.](http://lang.org.ua/static/downloads/ner_models/uk_model.dat)

Це можуть бути і заробітна плата, яка в пілотних областях України, що впроваджують форми сімейної медицини, сягає 5-7 тис. грн. Створення

Центрів первинної медико-санітарної допомоги – це лише перший етап реформування галузі охорони здоров'я на теренах міста

ПЕРС
Ольга Братушка запевнила, що реформа не передбачає жодних скорочень лікарів та медперсоналу, а жителі обстежуватимуться і отримуватимуть медичну допомогу там, де й раніше

Крім того, планується відкрити 8 амбулаторій у віддалених районах міста, зокрема на Баранівці, Басах, Роменському мікрорайоні тощо

Юбертекст

- × Our take on Gigaword
- × WIP
- × Джерела: Wikipedia, періодика, худліт, НАНУ, Рада, судові рішення, (you name it)
- × На даний момент зібрано: 67М речень, в яких $\sim \frac{2}{3}$ В слів
- × База для побудови статистичних моделей (нграми, вектори)

Тональний словник

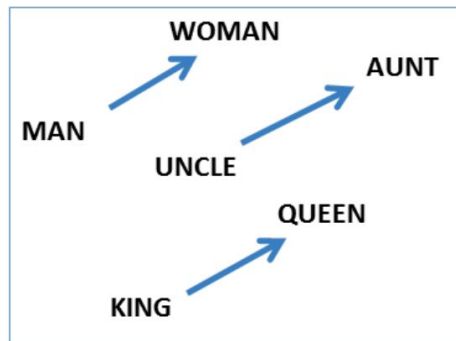
(<https://github.com/lang-uk/tone-dict-uk>)

- × ~3,5 тисячі слів (~20 тисяч словоформ)
- × No linguists were hurt during the creation of this data



Вектори слів

- × 300d word2vec, lexvec, GloVe побудовані на Юберкорпусі
- × WIP: moar vectors



Допомога групі БрУК

Вони створюють:

- × ВЕСУМ
- × БрУК
- × Український LanguageTool

Ми допомагаємо:

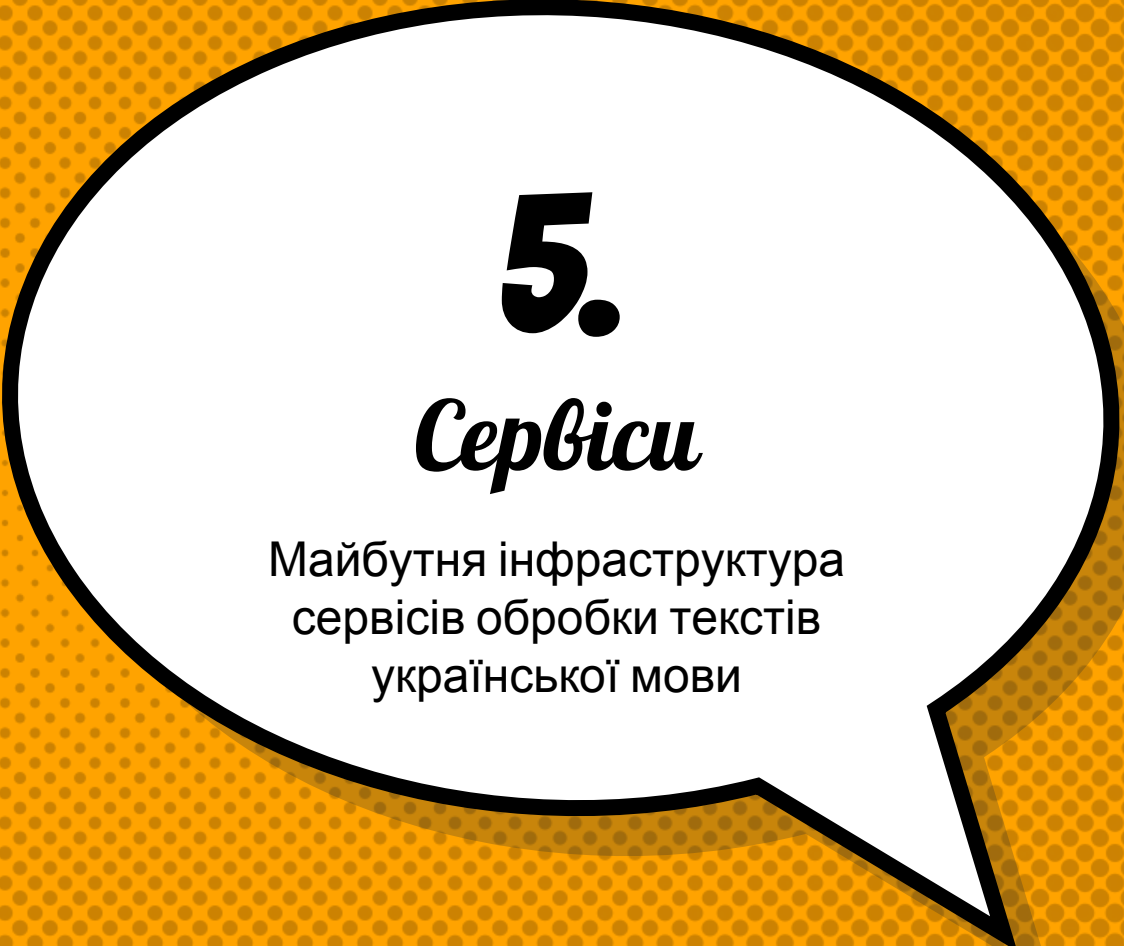
- × Збирати і обробляти тексти
- × Організувати краудсорсинг

Корпус зв'язків

- × WIP
- × Типи зв'язків: родинні, організаційні, місцеві, часові і т.і. (загалом, декілька десятків)
- × Дістати з Wikipedia

Open-source & open-serve

- × Токенізація для Python,.Net,Lisp
- × Плагіни для Вулика
- × Ідентифікація мови тексту
- × Сервіс NER
- × ... та інший WIP



5.

Сервіси

Майбутня інфраструктура
сервісів обробки текстів
української мови

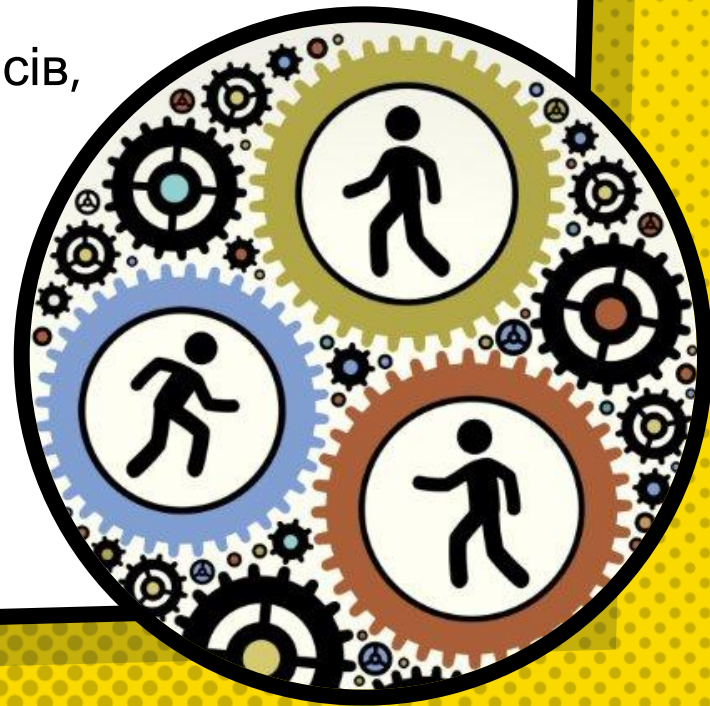


...Or C? Or C++? Groovy? GO?

(MICRO)SERVICES

Ми створюємо екосистему мікросервісів,
що дозволить:

- × Спробувати різні NLP сервіси у будь-яких ваших проектах
- × Не треба нічого встановлювати
- × Не треба нічого сплачувати!



Ми починаємо з

- × Токенізації
- × Лематизації
- × Визначення мови
- × NER для трьох мов
- × Word embeddings векторів
- × Та пошуку подібних слів



Колл?



JOIN US!

<http://lang.org.ua>

<https://github.com/lang-uk>



Лінгвісти

- × Постановка задачі
- × Документація для анотації
- × Редакторська робота
- × Власне анотація
- × Лінгвістичні вишукування



Дослідники

- × Пошук нових способів видобутку даних
- × Побудова моделей
- × Експерименти з даними



Програмісти

- × Автоматизація
- × Оптимізація
- × Scraping
- × Розробка мікросервісів



Студенти

- × Курсові та дипломні на будь-який смак
- × Ми готові менторити



KUDOS!

Волонтери

- × Анатолій
Стегній
- × Сергій
Шеховцов
- × Олесь Петрів
- × Микита
Павлюченко

Постачальники

- × Олександр
Маріковський
- × В'ячеслав
Тихонов
- × Олександр
Швець

Благодійники

- × Зеник
Матчишин
- × Макс Іщенко
- × Сергій Волошин
- × Ілля Сіменко
- × Олександр
Панчик

???

