Дослідження технологій структуризації неструктурованих текстових даних

Виконав: Домілевський О.С, ІПСА, ДА-62
Керівник: к.т.н. Сергеєв-Горчинський О. О.

Об’єкт дослідження

- Неструктуровані дані

Предмет дослідження

- Структурування та аналіз текстових даних
- Платформа Elastic Stack

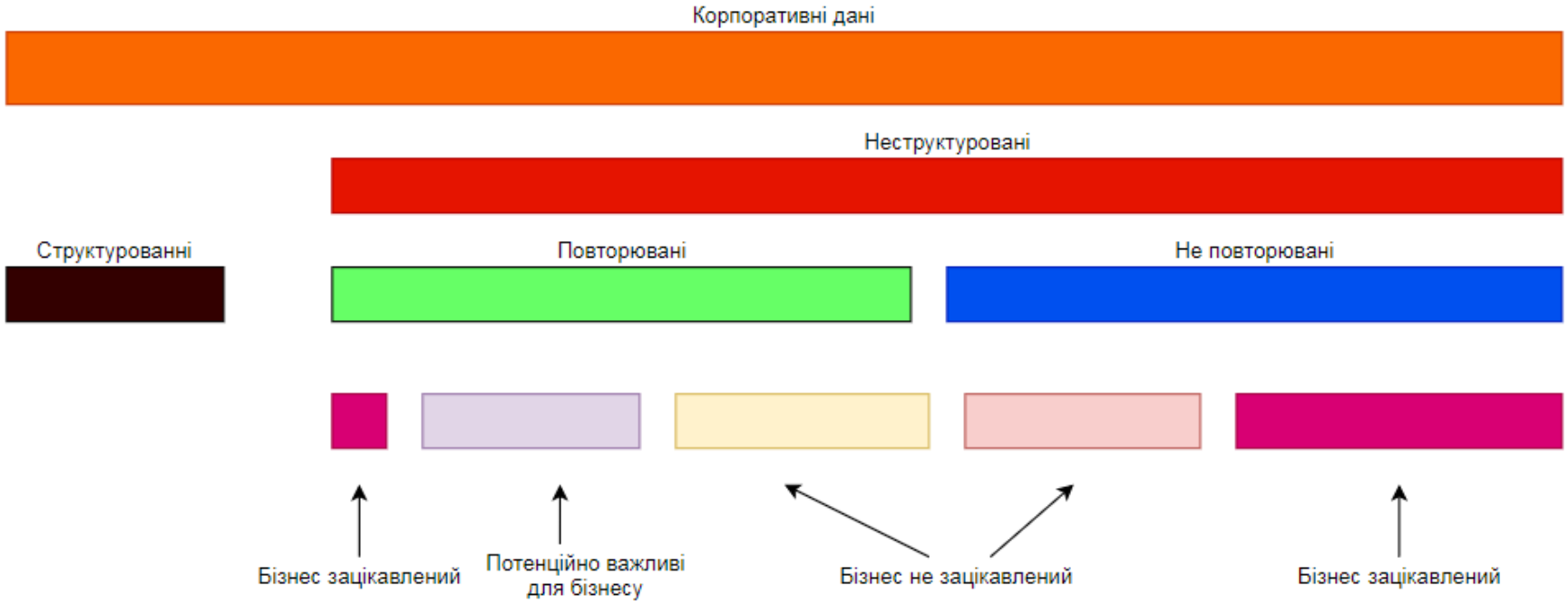
Актуальність

- 80% текстової інформації – неструктурована
- Неможливість вручну обробляти величезні обсяги інформації
- Більшість сучасних IT компаній використовують аналіз текстових даних

Мета

- Розглянути методи структуризації та аналізу текстових даних
- Дослідити компоненти системи Elastic Stack
- Провести аналіз сучасних NLP бібліотек
- Створити додаток для аналізу неструктурованих текстових даних та їх анотування

Дослідження даних в корпорації



Методи аналізу тексту

Основні методи

- Частота слова
- Словосполучення
- Відповідність

Передові методи

- **Класифікація**
  - Аналіз почуттів
  - Тематичний аналіз
  - Виявлення намірів
- **Анотування тексту**
  - Виявлення ключових слів
  - Виявлення особи
  - Розбіжність сенсу слова
- **Кластеризація**

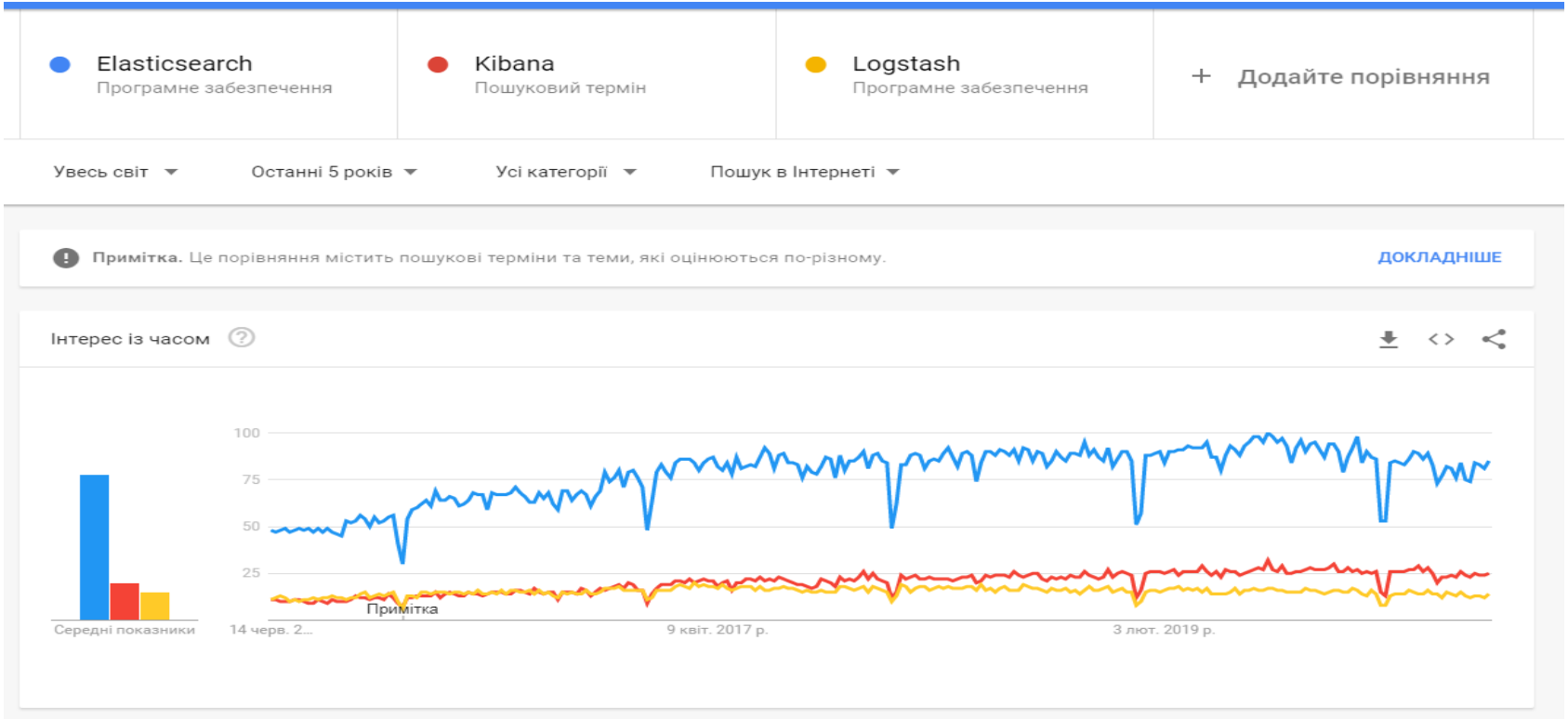
Обробка текстових даних

Слово	Potrer	Lancaster	Snowball	Лематизація	Без стоп слів
Big	big	big	big	Big	Big
Data	data	dat	data	Data	Data
analytics	analyt	analys	analyt	analytic	analytics
can	can	can	can	can	
help	help	help	help	help	help
organizations	organ	org	organ	organization	organizations
better	better	bet	better	better	better
understand	understand	understand	understand	understand	understand
the	the	the	the	the	
information	inform	inform	inform	information	information
contained	contain	contain	contain	contain	contained
within	within	within	within	within	
the	the	the	the	the	
data	data	dat	data	datum	Data

Elastic Stack

Збір	Агрегація обробка	Індексація Зберігання	Аналіз візуалізація

Популярність ELC, Kibana, Logstash



Elastic Search

Аналіз

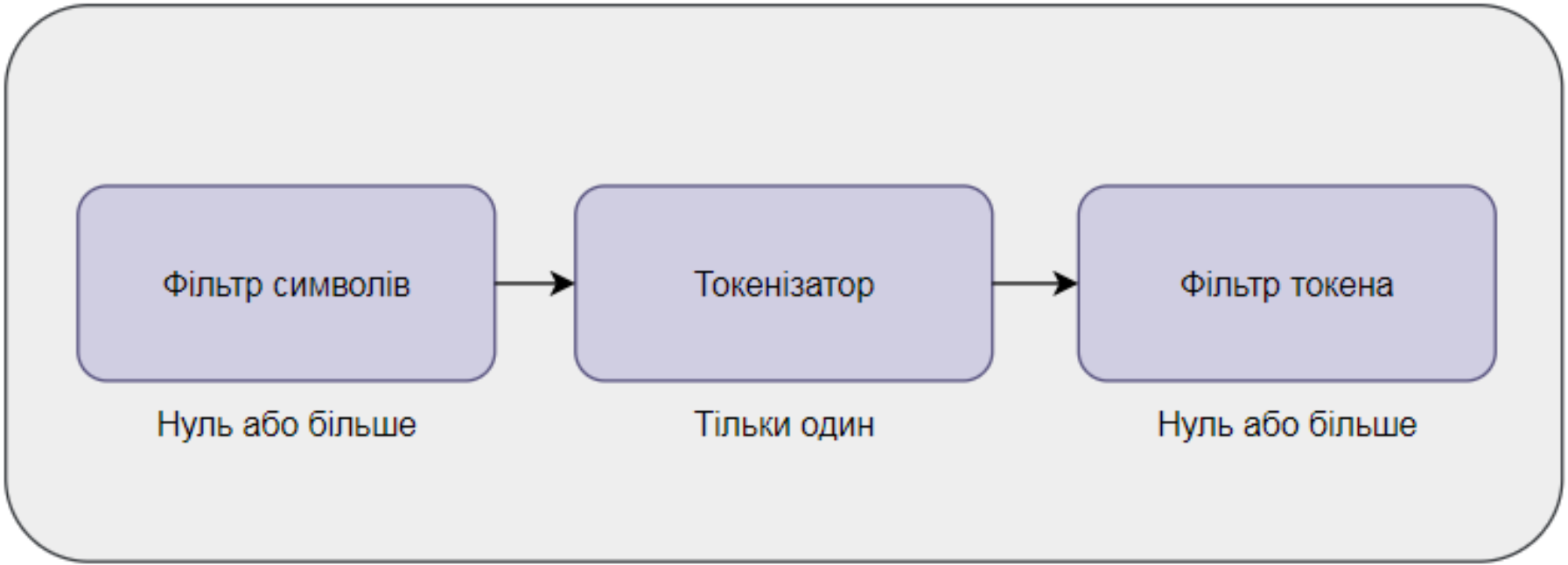
Масштабованість

REST API

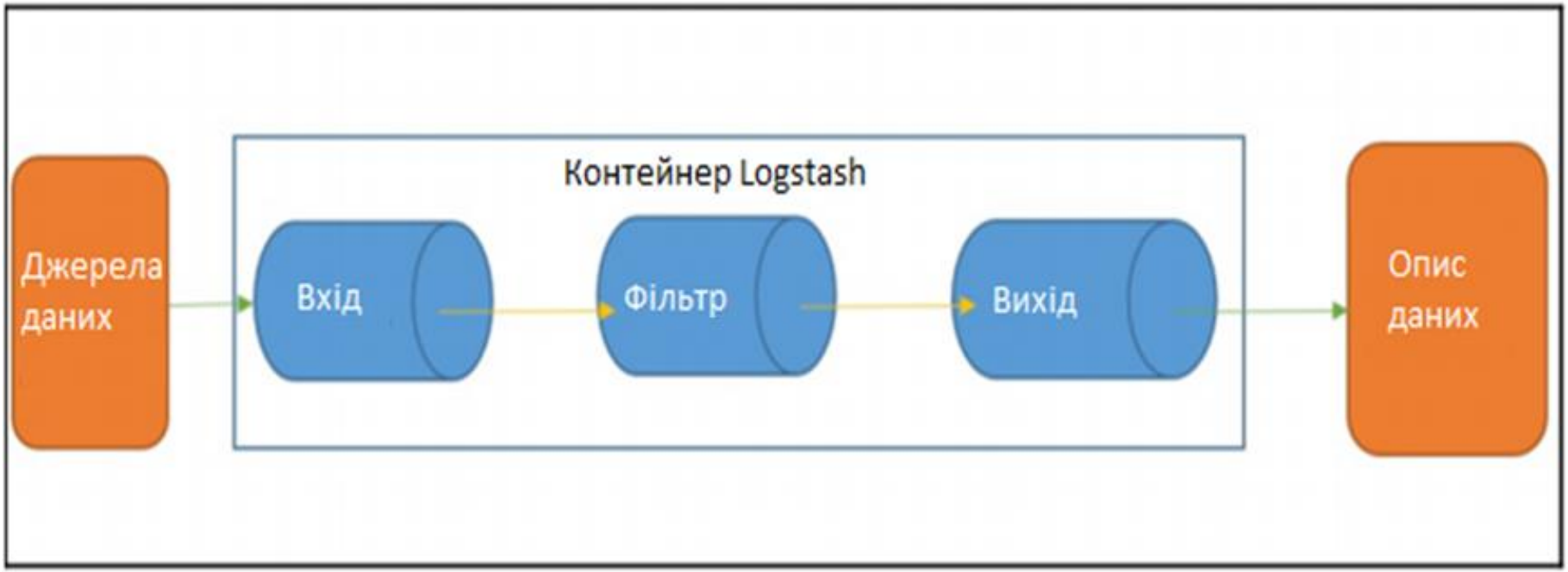
Відмовостійкість

Неструктурованість

Загальний вигляд аналізатора



Logstash


Джерела даних
Контейнер Logstash
Вхід
Фільтр
Вихід
Опис даних

Порівняння можливостей основних бібліотек для NLP

Ім’я	Spark NLP	spaCy	NLTK	CoreNLP
Виявлення речень	+	+	+	+
Токенізація	+	+	+	+
Стематизація	+	+	+	+
Лематизація	+	+	+	+
Позначення часткової мови	+	+	+	+
Іменовані сутності	+	+	+	+
Виявлення залежностей	+	+	+	+
Збіг тексту	+	+	-	+
Збіг дати	+	-	-	+
Поділ на фрагменти	+	+	+	+
Перевірка орфографії	+	-	-	-
Аналіз настроїв	+	-	-	+
Можливість перетренувати моделі	+	+	+	+
Тренувальні моделі	+	+	+	+

Робота програми

Выберите файл: Выберите файл Файл не выбран

Загрузить

Имя файла 1.txt

Тип файла text/plain

Размер в байтах 622

{'text': 'In modern conditions a comprehensive reconstruction of all spheres of public life arises fundamentally new requirements for training future information security professionals in higher education institutions, which differ not only in a high level of professional skills, but also in a harmoniously developed interests, in ability to continuously improve the educational level and react to changes in life. Therefore, a special place in the training of future professionals takes a generation of professionally significant qualities that is the basis for the formation of a coherent worldview of existing reality in youth.\n', 'chain': {'life': {'arise': {'requirement': [1, 13]}}, 'requirement': {'train': {'information': [1, 18]}}, 'institution': {'differ': {'level': [1, 28]}}, 'skill': {'develop': {'interest': [1, 43]}}, 'ability': {'improve': {'level': [1, 49]}}, 'level': {'react': {'change': [1, 54]}}, 'professional': {'take': {'generation': [2, 11]}}, 'worldview': {'exist': {'reality': [2, 30]}}}}

Загрузка ещё одного файла

Выберите файл: Выберите файл Файл не выбран

Загрузить

Приклад зберігання в базі Elastic Search

Підсумки виконаної роботи

1. Проаналізовано методи аналізу тексту
2. Проаналізовано систему Elastic Stack
3. Порівняно бібліотеки NLP
4. На основі бібліотеки spacy створено веб-додаток, який аналізує текстову інформацію та зберігає в Elastic Search

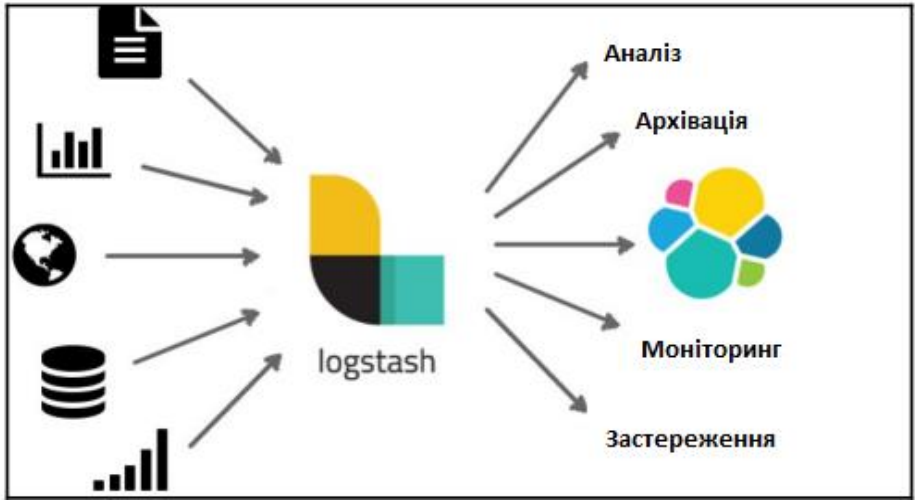
Майбутні напрямки роботи та досліджень

- Проаналізувати варіанти оптимізації ресурсів, необхідних для роботи стеку
- Реалізувати Веб-сервіс у вигляді плагіну
- Забезпечити функцію створення шаблонів для аннотування текстових даних
- Забезпечити функцію автоматичного завантаження тексту з cloud-платформ

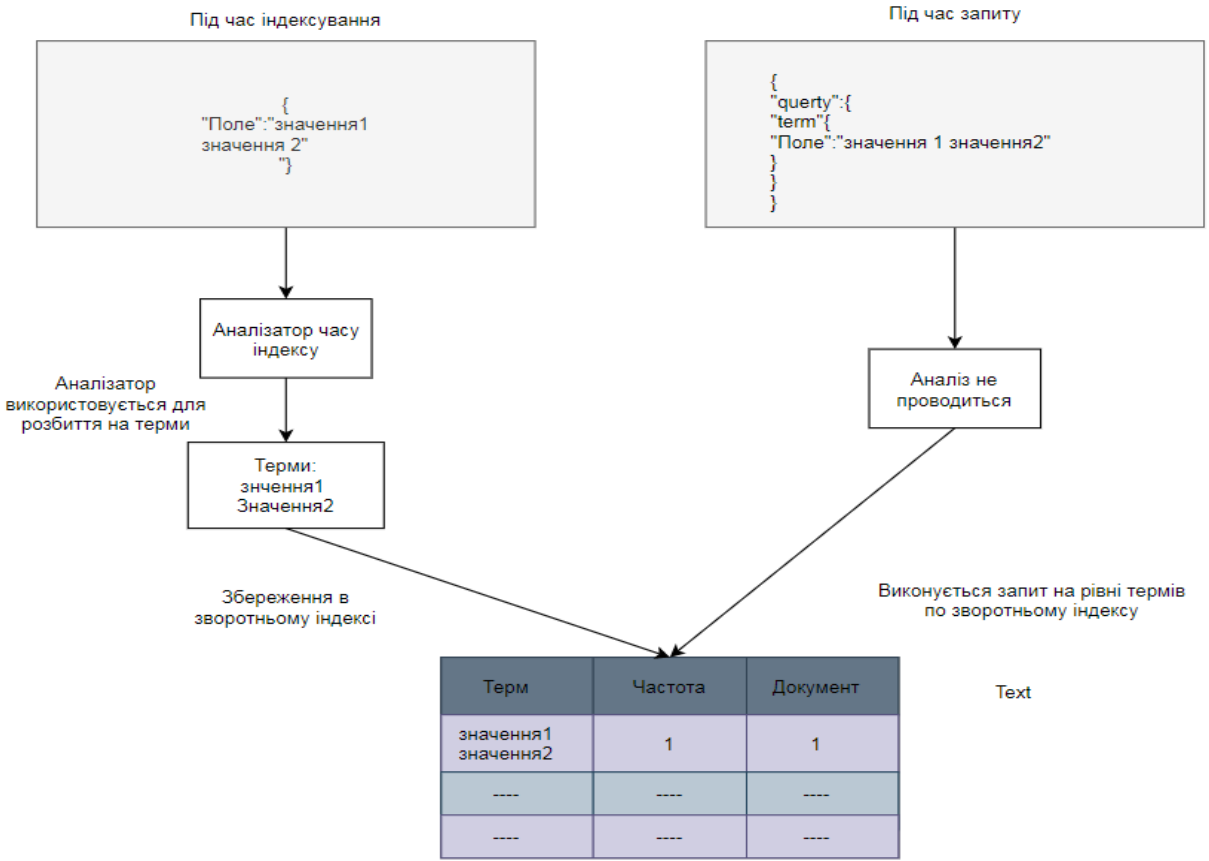
Дякую за увагу!

Технології, що використовуються

Типова взаємодія користувача з Kibana та Logstash



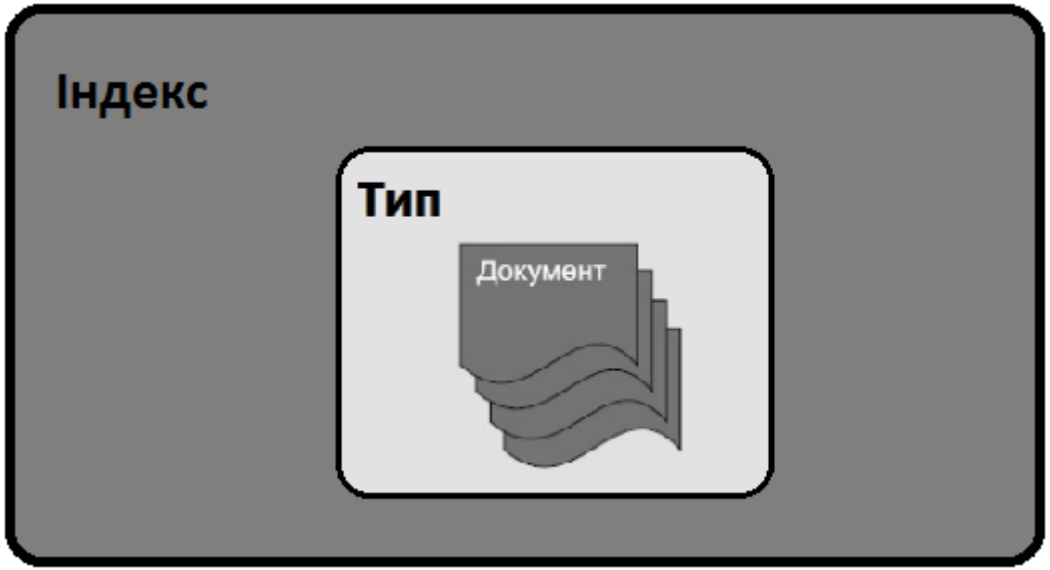
Виконання Term запиту



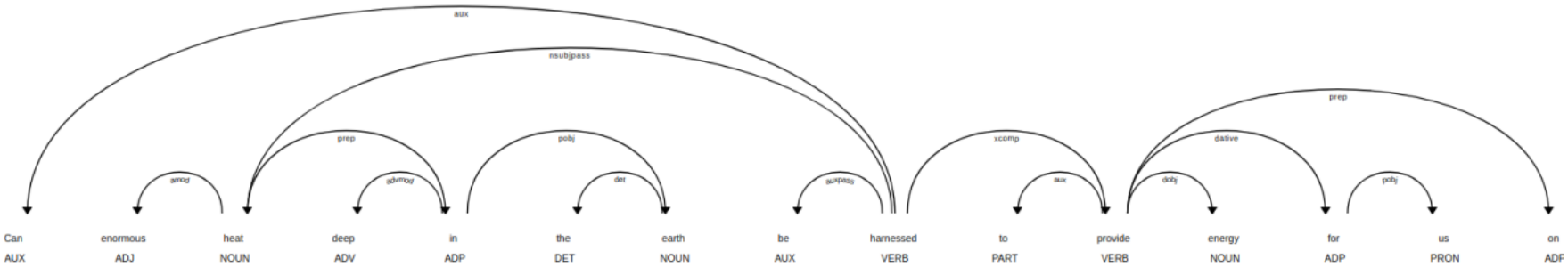
Типи агрегацій, що підтримує ElasticSearch

- Сегментні
- Метричні
- Матричні
- Агрегації контейнерів

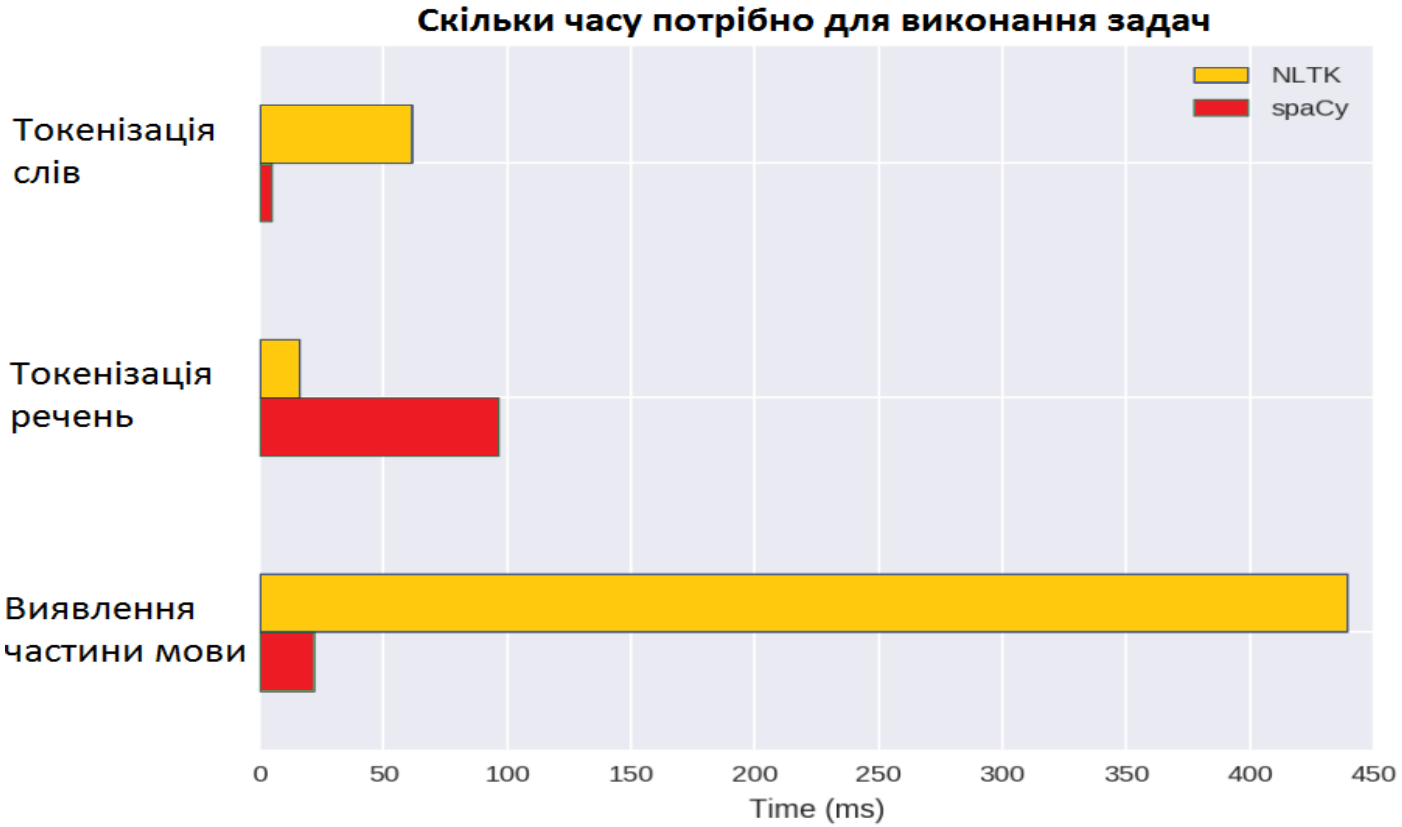
Загальна структура індексу

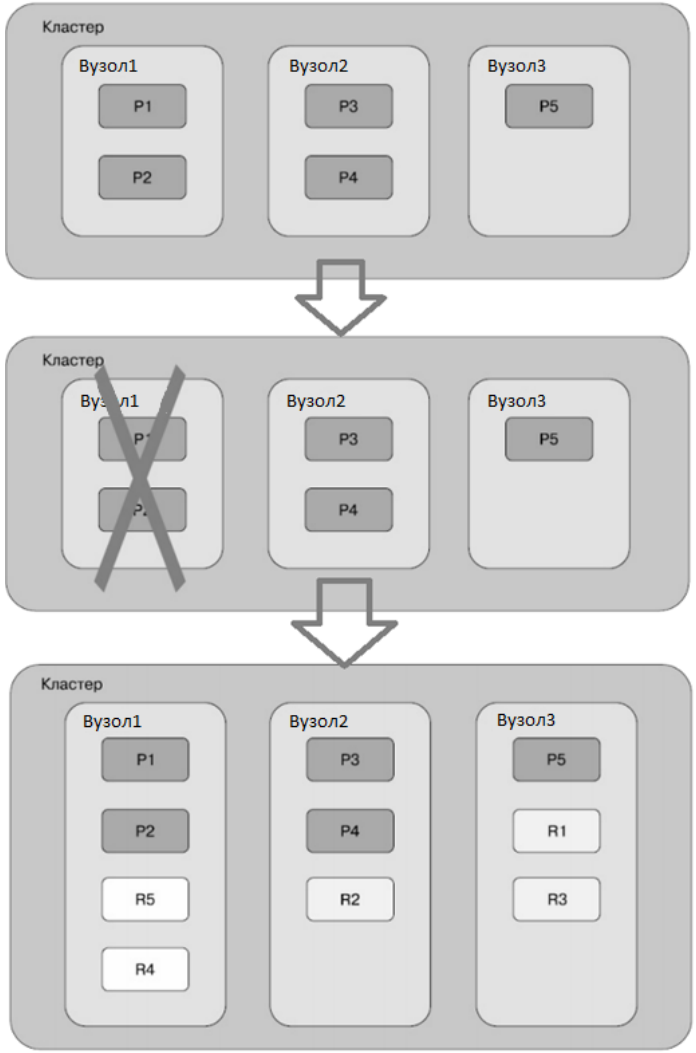


Приклад розбиття на частини мови



Перевірка швидкості бібліотек spacy і NLTK




Кластер
Вузол1
P1
P2
Вузол2
P3
P4
Вузол3
P5
Кластер
Вузол2
P3
P4
Вузол3
P5
Кластер
Вузол1
P1
P2
R5
R4
Вузол2
P3
P4
R2
Вузол3
P5
R1
R3