титульний аркуш ТЧ
індивідуальне завдання на КР(МР)
календарний план
зміст
анотація
вступ

7) основна частина (її розділи): 30-50 ст.;

– аналіз існуючих методів (алгоритмів) вирішення поставленої задачі;

– обґрунтування вибору рішення;

– вибір принципу дії системи чи обґрунтування методик;

– розробка структурної і (або) функціональної схеми;

– розробка принципової схеми;

– експериментальні дослідження;

– метрологічні характеристики;

– алгоритмічне та програмне забезпечення;

висновки
література
глосарій
додатки

Анотація

Мета даної роботи – розробити метод і відповідне програмне забезпечення для вирішення задачі побудови термінології в колекції текстів наукової тематики.

Дане дослідження є складовою частиною циклу робіт проведених на кафедрі з тематики, присвяченої побудові пошукової системи і репозіторію наукових праць.

З алгоритмічної точки зору результатом буде такий алгорітм, що на вході отримує колекцію документів у форматах pdf | doc, а на виході віддає файл що містить RDF граф з термінологією що зустрічається в даній колекції.

Як складовий модуль системи пошуку і каталогізації наукових праць, на модульному рівні дистрибутив програми буде надано у вигляді jar-архіву з відповідним API.

Для кінцевого користувача і для тестування ефективності побудови термінології експертами буде розроблено веб-інтерфейс до компонентів програми, де в користувача будуть можливості завантажити в систему архівом колекцію документів, запустити алгоритм побудови термінології (можливо з введенням параметрів алгориту), передивитись і скачати файл-результат у вигляді RDF, а також продивитись термінологію, її входження в документи колекції і пов'язані з результатом роботи алгоритму метадані терміну.

Функціональна схема роботи алгоритму побудови термінології буде включати в себе наступні етапи і технології:

- початкова індексація документів колекції за допомогою рішень з відкритим кодом (Apache Hadoop, Lucene, Cloudera SaaS);
- застосування алгоритмів розбору NLP (лінгвістичних аналізаторів) на реченнях і абзацах текстів, зберігання отриманих метаданих у спеціалізованих індексах;
- експериментальні дослідження щодо формулювання множини ефективних запитів до використаних пошукових систем і сховищ даних, що будуватимуть зв’язні висловлювання щодо термінів, опис і використання лінгвістичних евристик;
- розробка і застосування ітеративного алгоритму побудови термінології, що зможе покращувати результати при додаванні нових документів до колекції, а також після кожної ітерації;
- використання як сховища даних RDF графової або RDF баз даних, використання в алгоритмі запитів мовами Cypher, SPARQL;
- розробка RESTful API і веб-інтерфейсу користувача до досліджуваної системи;
- підбір тестових колекцій наукових документів і написання тестових пакетів для контролю ефективності і точності алгоритму;
- експериментальні дослідження алгоритму з варіацією параметрів;
- встановлення залежностей між параметрами, публікація найбільш вдалих налаштувань як окремих методів API;

- розгортання готової системи на сервері НаУКМА для публічного доступу;