УДК 004.55

ПРОЕКТУВАННЯ ПРОГРАМНОЇ СИСТЕМИ КЛАСТЕРИЗАЦІЇ КОРОТКИХ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ ЧАСТОТИ ВХОДЖЕННЯ ТЕРМІНІВ

Мілейко І.І.
Керівник: професор каф. СПЗ Кунгурцев О.Б.
Державний університет «Одеська політехніка», УКРАЇНА

АНОТАЦІЯ. Робота присвячена розробці програмної, системи, що допоможе користувачу розбити масив документів на кластери, а також вивченню аналогів. В основі ідеї лежить поняття близькості документів, що розраховується на основі частоти входження термінів (іменників) в документ.

Вступ. Кожна людина працює з великою кількістю документів. Якщо раніше це були паперові документи, то у час розвитку інформаційних технологій, це найчастіше електроні документи. Щоб полегшити роботу з ними, має сенс виконати логічне розподілення їх по темах, або за схожістю. Також це спростить розробку словників [1]. Саме для цього потрібно розробити програмну систему, що має змогу розділяти документи. Тому **метою роботи** стало зменшення часу, необхідного для розподілення документів за схожістю, а також прискорення виділення з документів іменників та їх підрахунку.

Основна частина роботи. Для визначення запланованих функцій програми, був проведений конкурентний аналіз програм-аналогів зі складання глосарію (табл. 1), що є близькою задачею для однієї з функцій розроблюваної програми – виділення іменників. Жоден з проаналізованих конкурентів не пропонує групування документів по кластерах. Крім цього, всі програмні системи перед аналізом не виключають зі списку слів слова, що не є іменниками. Таким чином до користувача попадає багато зайвого. Також, крім системи, що використовується для тлумачення слів, жоден аналог не приводить слова до початкової форми. Тим самим підрахунок частоти повторень може бути помилковим.

Таблиця 1.1 – Конкурентний аналіз

	Виділення термінів	Підрахунок термінів	Розбиття документів на кластери	Фільтрація іменників	Приведення слів в початкову форму
Моя система	+	+	+	+	+
Word Tabulator	+	+	-	-	-
MonoConc	-	+	-	-	-
Concordancer	-	+	-	-	-
Lingvo	-	-	-	-	+

Для того, щоб визначити перелік сервісів, що повинна виконувати система та її реакцію на різні вхідні дані та поведінки, була складена діаграма варіантів використання (Рис.1).

Можна виділити 2 основних етапи роботи системи:

- Додавання документів, який передбачає вибір документів для кластеризації та обробка їх програмою для подальшої роботи. В даному випадку при обробці кожного документа потрібно:

1) Виділити з документа іменники в початковій формі.
2) Підрахувати загальну кількість іменників та кількість кожного іменника окремо.
3) Зберегти цю інформацію для кожного документа окремо.

- Кластеризація [2]. В основі алгоритму кластеризації лежить поняття близькості документів. Близькість двох документів залежить від кількості загальних іменників. Умовно процес кластеризації можна поділити на 2 етапи:

1) Первісна кластеризація. На основі близькості документів, складаємо такі кластери, щоб для кожної пари документів, що входить у кластер, близькість була не менша ніж деяка заздалегідь задана мінімальна близькість.

2) Закріплення результату. Для кожного кластеру знаходимо ядро – документ, у якого максимальна близькість з іншими документами того ж кластеру. Потім якщо є документи, у яких

близькість з ядром іншого кластера вище, ніж з документом первісного кластера, то відносимо ці документи до більш близького кластеру.

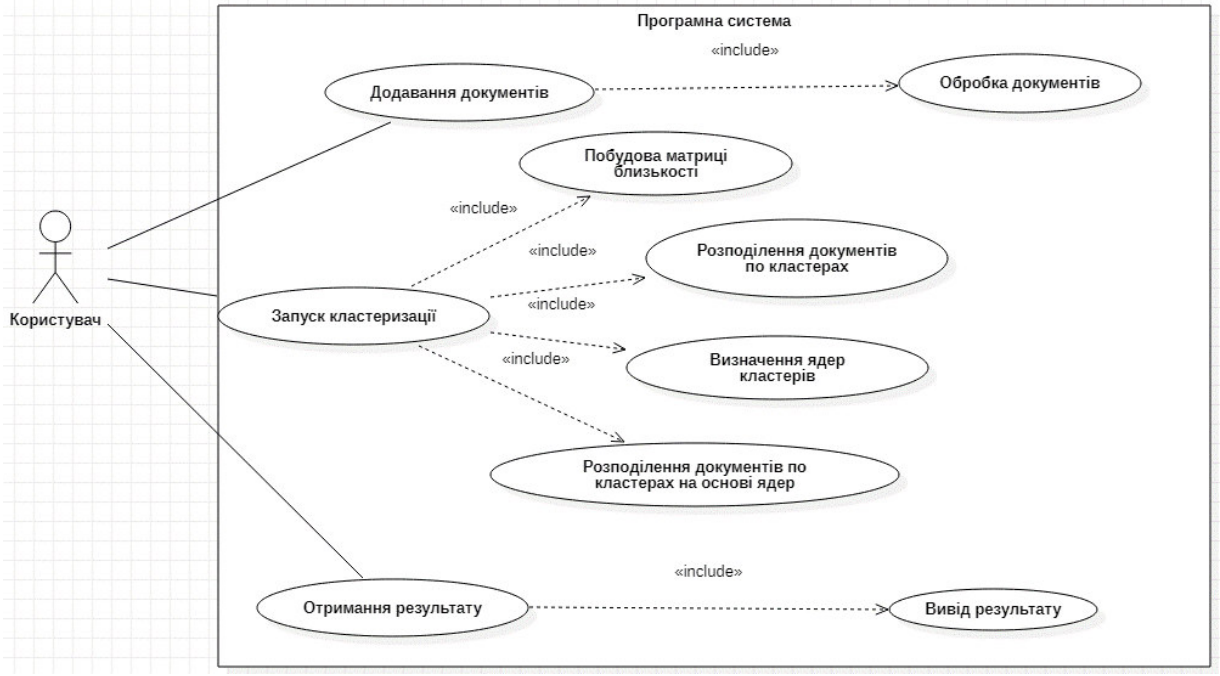

Рисунок 1 – Діаграма варіантів використання

Щоб побачити приклад роботи системи, наведена діаграма послідовностей, що демонструє роботу системи під час виконання прецеденту «Додавання документів» (рис. 2). На ній видно, якими повідомленнями обмінюються між собою користувач (або UI), система та обробник, який відповідає за обробку вхідних файлів.

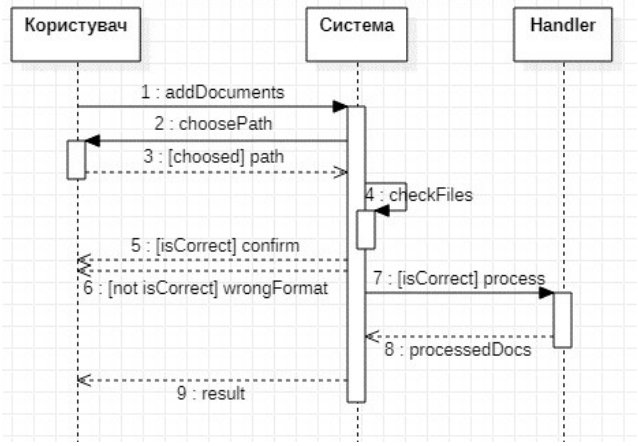

Рисунок 2 – Діаграма послідовності «Додавання документів»

Висновки. У роботі виконані основні етапи проектування програмного забезпечення для кластеризації документів з наведенням основних *UML*-діаграм, що демонструють його архітектуру. Це програмне забезпечення повинно значно скоротити час, що користувач витрачає на роботу з документами.

СПИСОК ЛІТЕРАТУРИ

1. Кунгурцев, А. Б. Метод построения словарей предметных областей для извлечения фактов из текстов на естественном языке / А. Б. Кунгурцев, С. Н. Бородавкин, А. П. Голуб // Вост.-Европ. журн. передовых технологий. - 2010. - № 1/4 (43). – С. 32-36

2. Обзор алгоритмов кластеризации данных [Електронний ресурс]. – Режим доступу: URL: https://habr.com/ru/post/101338/.