Добрий день. Я Федорко Андрій учень Шепетівського НВК №1. Тема наукової роботи інформаційна технологія тематичного сортування текстової інформації. Науковий керівник Мазурець Олександр старший викладач кафедри КНІТ Хмельницького національного університету. Педагогічний керівник Колісецький Вілен.

**[Слайд 2 Мета]**

Метою роботи є розробка інформаційної технології для тематичного сортування текстової інформації та розробка ПЗ для перевірки його ефективності при автоматизованому сортуванні новин по рубриках.

Об’єктом дослідження є ключові слова як еквівалент семантичного вмісту цифрових текстів.

Предметом дослідження – процеси визначення ключових слів у цифровому текстовому контенті новин.

Автоматизація сортування текстової інформації є ефективним інструментом, що заощаджує час користувача та підвищує якість роботи новинних агрегаторів, тому даний напрям досліджень є актуальним.

**[Слайд 3 Задачі дослідження]**

Для досягнення поставленої мети визначені наступні задачі дослідження:

* Аналіз сучасних методів пошуку ключових слів:  TF, IDF, TF-IDF
* Розробка інформаційної технології та побудова математико-алгоритмічних моделей для визначення приналежності введеної новини до актуальних рубрик новин
* Розробка програмного забезпечення
* Його тестування та оптимізація

**[Слайд 4 Обчислення TF]**

Очевидно що ключові слова зустрічаються частіше інших слів. Формула TF дозволяє вирахувати частоту слова. TF - відношення числа входжень обраного слова до загальної кількості слів документа.

**[Слайд 5 Обчислення IDF]**

Нажаль в тексті дуже часто зустрічаються сполучники, хоча вони не є ключовими словами. Їх відмінною рисою є те що вони зустрічаються у всіх текстах, навідмінну від ключових слів які зустрічаються тільки в текстах їхньої теми. Оцінка IDF дає більшу вагу специфічним словам. Отже, IDF – інверсія частоти, з якою слово зустрічається в документах колекції, тому використання IDF зменшує вашу широковживаних слів.

*Логарифм слугує в якості нормалізатора.*

**[Слайд 6 Обчислення TF\*IDF]**

Якщо ми перемножимо ці два терміна, то отримаємо показник який показує на скільки слово є ключовим.

У кінці ми порівнюємо ключові слова із текстом який ми ввели. І дізнаємося до якої категорії він відноситься.

**[Слайд 7 Інформаційна технологія]**

Вхідними даними є множина новин для формування ключових слів нової рубрики, або тестова новина для аналізу.

На першому етапі визначаємо ключові слова з використанням оцінок TFIDF.

Обмежуємо їх кількість, наприклад, до 30.

На другому етапі реалізується визначення приналежності даної новини до рубрик новин відповідно до алгоритму.

На основі розроблених інформаційної технології тематичного сортування текстової інформації та підходу до визначення множин ключових слів для рубрик новин було створено два програмних продукти, систему визначення множин ключових слів для рубрик новин та систему тематичного сортування новин.

**[Слайд 7 Демонстрація програми 5хв]**

Перейдемо до демонстрації програми.

Програма складається з двох модулів.

Перший модуль створює список ключових слів для кожної категорії.

**[Модуль 2]**

Другий модуль визначає до якої категорії належить введений текст.

Давайте спробуємо ввести якийсь текст з категорії Економіка.

***Перехід на сайт новин і копіювання інформації з блоку Економіка.***

Натискаємо Аналізувати і програма в зручному форматі показує до якої категорії ймовірніше всього відноситься текст. Два центральних стовпчики показують к-сть співпадінь і відсоток. Найточніше працює третій стовпчик, тому що ...

**[Модуль 1]**

Тепер перейдемо до першого модуля і розберемо його детальніше.

Для того щоб подивитись TF-IDF категорії, потрібно вибрати її і натиснути на відповідну кнопку.

Якщо текстове поле пусте, то програма покаже вже пораховані дані.

Для того щоб створити нову категорії потрібно просто ввести її і натиснути TF, IDF, TF-IDF

Програма інтуїтивно зрозуміла, і дає підказки, які допомагають розібратися з нею.

Якщо натиснути Save програма збереже дані у файл у мультимовний формат JSON.

**[Висновки]**

Одержані результати показали, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології тематичного сортування текстової інформації, успішно виконала сортування новин за рубриками, й середня успішність сортування за рубриками склала 94,4%.

**[Плани на майбутнє]**

* Дослідити пошук ключових слів за допомогою DE
* Використання штучної нейронної мережі для вирішення поставленої задачі
* Повністю автоматизувати роботу програмного додатку