**[Представлення]**

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

Об’єм інформації збільшується

**[Слайд 2 Мета]**

Метою роботи є розробка інформаційної технології ~~для тематичного сортування текстової інформації~~ та розробка ПЗ для перевірки його ефективності за автоматизованого сортування новин по рубриках.

Об’єктом дослідження є ключові слова як еквівалент семантичного вмісту (*~~значення~~*) 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.

Актуальність сказати про Екзобайт

**[Слайд 3 Практичне застосування]**

Листи до служби підтримки

Приймальні комісії у ВУЗах

Новинні агрегатори(RSS-стрічки)

Пошук спаму(фільтрацію)

Пошук неблагонадійних сайтів

Класифікація наукових статей

Визначити автора твору

Щодо останнього це дуже цікаво. Наприклад Тарас Григорович слово “люде” пише з буквою “е”. І це є його особливістю. В свою чергу Віктор Федорович, слово “професор”, пише з двома буквами “ф” і це також є його особливістю.

**[Слайд 3 Завдання дослідження]**

Для досягнення поставленої мети визначені наступні завдання дослідження:

* Аналіз сучасних методів пошуку ключових слів:  TF, IDF, DE
* 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]**

Що не враховує TF це те, що тексті дуже часто зустрічаються слова які не несуть самостійного смислового навантаження. Отже, IDF дає більшу вагу специфічним словам. ~~Отже, IDF – інверсія частоти, з якою слово зустрічається в документах колекції, тому використання IDF зменшує вагу широковживаних слів.~~

*Логарифм слугує в якості нормалізатора.*

**[Слайд 7 Обчислення DE]**

Ми вирішили додати ДО тому що попередні дві технології не враховують позицію.  
DE показує дискримінантну силу слів, тобто відстань між появами сусідніх слів.

Якщо слово розподілено рівномірно, то відстані для кожних пар будуть рівнію

**[Слайд 8 Visualisation of DE]**

Наскільки слово є розсіяним по тексту. Якщо поява слова нерівномірна, то воно є ключовим.

Отже коли TFIDF знаходить ключові слова за частотою їх появи, і не враховує порядок, то ДО навпаки, враховує позицію і не залежить від частоти.

В результаті ми отримаємо систему яка досить точно визначає ключові слова, які допоможуть визначити приналежність тексту до певної категорії.

**~~[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 було створено два програмних продукти, систему визначення множин ключових слів для рубрик новин та систему тематичного сортування новин.~~

**[Слайд 8 Демонстрація програми 5хв]**

Перейдемо до демонстрації програми.

Програма складається з двох модулів.

Перший модуль створює список ключових слів для кожної категорії.

*Клікнути на ComboBox, щоб показати які є категорії*

*Щоб створити нову категорію потрібно ввести її назву і по черзі натиснути всі кнопки.*

Програма під час користування дає підказки і також наявна інструкція користування нею.

*Якщо категорія введена раніше, то програма покаже вже пораховані данні*

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

**[Модуль 2]**

Другий модуль визначає до якої категорії належить введений текст.

Давайте спробуємо ввести якийсь текст.

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

Категорія нехай буде Економіка.

*Коментую свої дії*

Натискаємо Аналізувати і програма в зручному форматі показує до якої категорії ймовірніше всього відноситься текст. Два центральних стовпчики показують к-сть і відсоток співпадінь. Найточніші дані у третьому стовпчику тому що він показує суму коефіцієнтів TFIDF у слів які співпали.

**[Висновки]**

Досліджено сучасні методи пошуку ключових слів:TF, IDF, TF-IDF

Вперше розроблено інформаційну технологію тематичного сортування текстової інформації

Досліджено практичну ефективність інформаційної технології тематичного сортування текстової інформації

Результати показали, що програмна система на 97,8% успішно виконала сортування новин за рубриками. Тобто поставлені цілі були досягнуті.

Вкінці хочемо повідомити про наші плани на майбутнє.

**[Плани на майбутнє]**

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

Дякую за увагу! Готовий з радістю відповісти на ваші запитання.