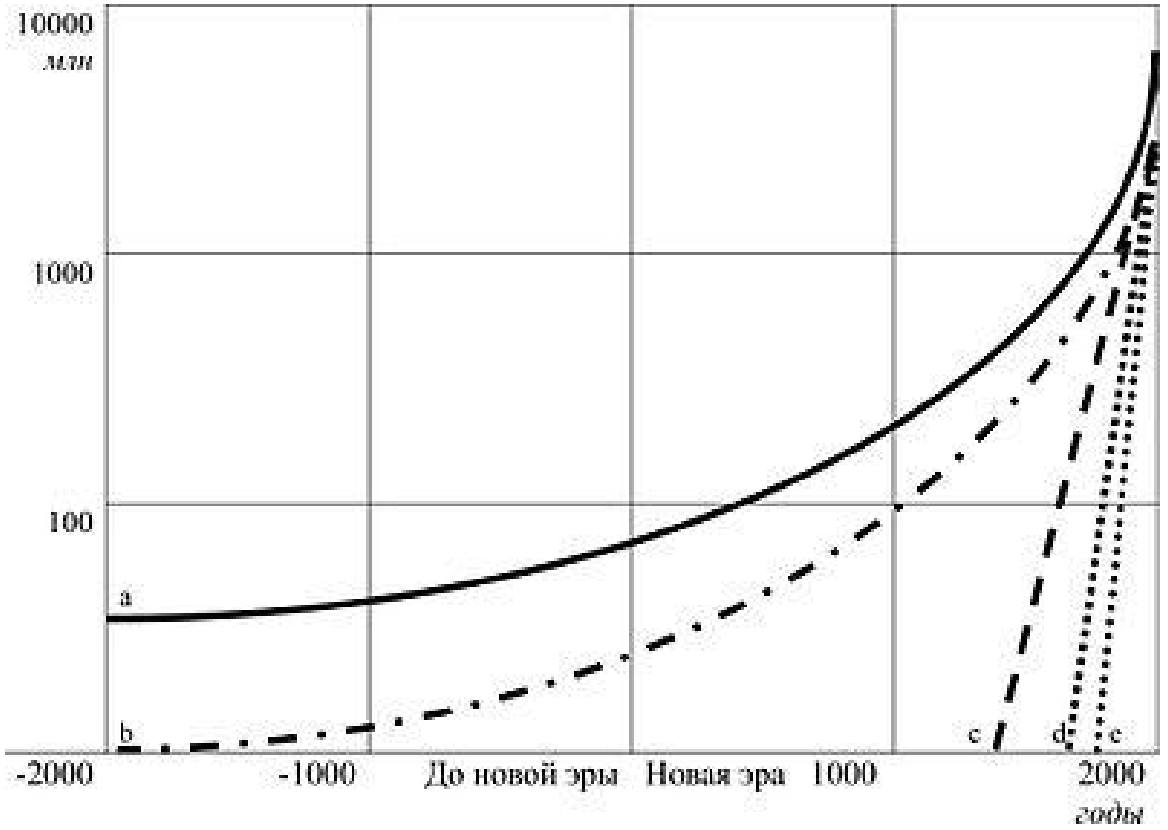
Роботу виконав: ***Федорко Андрій Петрович,***
здобувач освіти Шепетівського НВК №1
ім.Героя України М. Дзявульського
Науковий керівник: ***Мазурець Олександр Вікторович,***
старший викладач кафедри КНІТ
Хмельницького національного університету
Педагогічний керівник: ***Колісецький Вілен Іванович,***
вчитель інформатики Шепетівського НВК №1

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ТЕМАТИЧНОГО СОРТУВАННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Мета - дослідження методів для тематичного сортування текстової інформації та розробка ПЗ для перевірки їх ефективності при автоматизованому сортуванні новин по рубриках.

Актуальність

Об'єм інформації збільшується експоненціально, і виникає потреба в її аналізу і сортуванні. Це заощадить час користувача тому даний напрям досліджень є **актуальним**.



Завдання дослідження

- Аналіз сучасних методів пошуку ключових слів: TF, IDF, DE
- Розробка інформаційної технології
- Побудова математико-алгоритмічних моделей
- Розробка програмного забезпечення
- Визначення рубрики для введеної новини

Обчислення TF

$$TF = \frac{n_t}{\sum_k n_k}$$

TF (term frequency — частота слова) — відношення числа входжень обраного слова до загальної кількості слів документа.

n_t це число входжень слова t в документ

$\sum_k n_k$ — загальна кількість слів в документі

Обчислення IDF

$$IDF = log \frac{|D|}{|(d_i \supset t_i)|}$$

IDF — інверсія частоти, з якою слово зустрічається в документах колекції. Використання IDF зменшує вагу широковживаних слів.

|D| — кількість документів колекції

$|(d_i \supset t_i)|$— кількість документів, в яких зустрічається слово t_i

Обчислення DE

$$\sigma = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle}$$

DE (дисперсійна оцінка) - оцінка дискримінантної сили слів.

DE показує як термін появляється в тексті і враховує його позицію відносно інших слів.

Візуалізація DE (жовтий - поява терміну)

Ключове слово

Стоп-слова

Демонстрація програми

Практичне застосування методики

- Листи до служби підтримки
- Приймальні комісії у ВУЗах
- Новинні агрегатори(RSS-стрічки)
- Пошук спаму(фільтрація e-mail)
- Пошук неблагонадійних сайтів
- Класифікація наукових статей
- Онлайн звернення до лікаря
- Ідентифікація автора твору
- тощо

Висновки

- Досліджено сучасні методи пошуку ключових слів: TF, IDF, DE
- Вперше розроблено інформаційну технологію тематичного сортування текстової інформації
- Опубліковано наукову статтю у фаховому виданні (Вісник ХНУ №5 2019 (277)), включеному в перелік МОН України
- Досліджено практичну ефективність інформаційної технології

97.8%

Середня успішність

Перспектива

- ~~Дослідити пошук ключових слів за допомогою DE~~
- Дослідити використання штучної нейронної мережі
- Повністю автоматизувати роботу програмного додатку.
- Специфікувати і монетизувати програмну систему

Дякую за увагу!