Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №2

По дисциплине: «Естественно-языковой интерфейс ИС»

Тема: «Разработка системы автоматического реферирования документов»

Выполнил:

Студент 4 курса

Группы ИИ-21

Романко Н. А.

Проверила:

Булей Е. В.

Брест 2024

Цель: освоить на практике основные принципы автоматического реферирования документов.

Ход работы:

№	Язык текста	Стратегия поиска	Язык
4	Русский, немецкий	Sentence Extraction + ML	Научные статьи по computer science, Сочинения по литературе

- на входе – на входе – текстовые документы одинакового размера (например, 10 страниц формата А4), содержащие тексты из предметных областей на естественных языках согласно варианту подлежащие процедуре автоматического реферирования;
- на выходе – активная ссылка на исходный документ и построенный, в соответствии с вариантом реферат документа, состоящий из 2-х разделов: 1 - классического реферата и реферата в виде списка ключевых слов (по методу Sentence extraction) ; 2 – реферата, построенного с применением машинного обучения (ML).
- наличие средств сохранения в файл и распечатки полученной на выходе информации;
- интерфейс системы должен быть предельно простым и доступным для пользователей любого уровня, содержать понятный набор инструментов и средств, а также help-средства.

Код программы:

Реферирование при помощи TF-IDF:

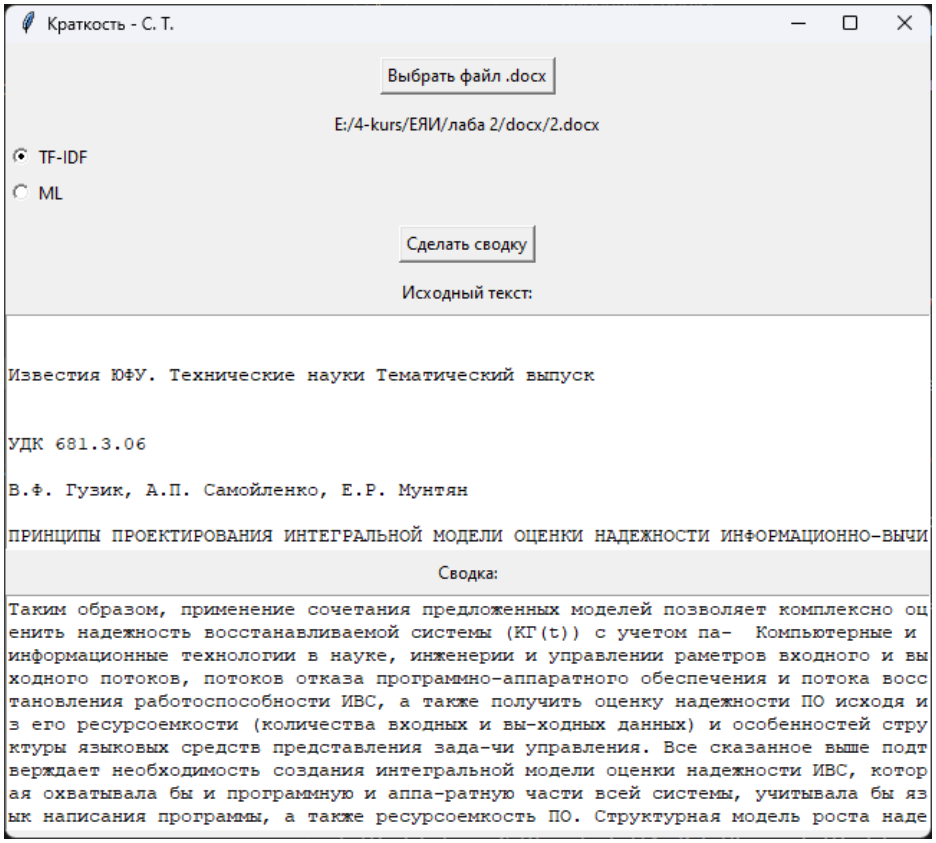
```
def extract_sentences_tfidf(text):
    sentences = sent_tokenize(text)
    stop_words = set(stopwords.words('russian'))
    cleaned_sentences = [' '.join([word for word in word_tokenize(sentence.lower()) if word.isalpha()
and word not in stop_words]) for sentence in sentences]

    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(cleaned_sentences)

    return sentences, tfidf_matrix.sum(axis=1)
```

Реферирование при помощи модели mBART:

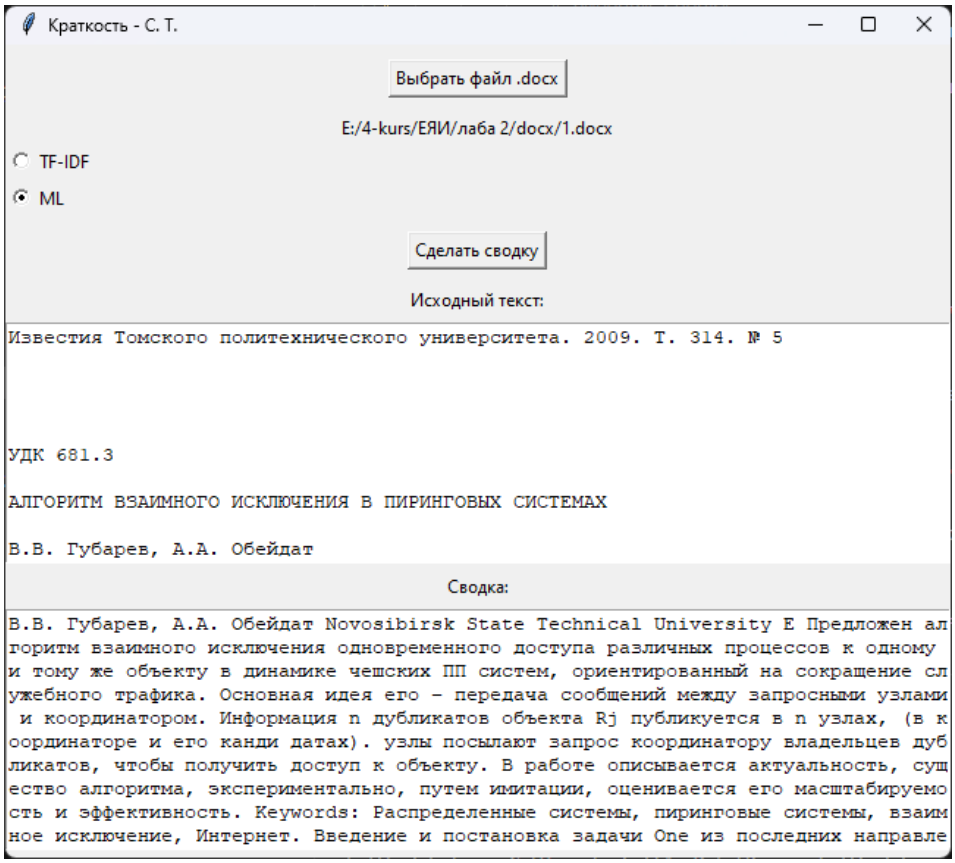
```
def generate_ml_summary(text):
    model_name = "facebook/mbart-large-50-many-to-many-mmt"
    model = MBartForConditionalGeneration.from_pretrained(model_name)

    tokenizer = MBart50TokenizerFast.from_pretrained(model_name)
    tokenizer.src_lang = "ru_RU"
    inputs = tokenizer.encode(text, return_tensors="pt", max_length=1024, truncation=True)
    summary_ids = model.generate(inputs, max_length=250, min_length=40, length_penalty=2.0,
num_beams=4, early_stopping=False, forced_bos_token_id=tokenizer.lang_code_to_id["ru_RU"])
    summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
    return summary
```

Результат:

Реферирование при помощи TF-IDF:



Реферирование при помощи модели mBART:



Вывод: в ходе выполнения лабораторной работы освоил принципы автоматического реферирования документов на естественном языке при помощи метода Sentence Extraction и при помощи ML-методов.