Министерство науки и высшего образования Российской Федерации
ФГАОУ ВО «Севастопольский государственный университет»
Институт информационных технологий

Кафедра «Информационная безопасность»

РАСЧЁТНО-ГРАФИЧЕСКАЯ РАБОТА

по теме

«Исследование методов векторизации текста
и извлечения признаков»

по дисциплине

«Защита программ и данных»

Выполнил: студент гр. НН/о-ГГ-Г-о
Иванов И. И.
Защитил с оценкой: _____
Принял: доцент Петров П. П.

Севастополь
2022

СОДЕРЖАНИЕ

1. ОЗАГЛАВИТЬ

1.1. Правовое поле использованных библиотек

1.1.1. Библиотеки получения корпуса

– *string* – стандартная библиотека *Python*. Распространяется по лицензии *PSF*;

– *re* – стандартная библиотека *Python*. Распространяется по лицензии *PSF*;

– *SpaCy* – библиотека обработки естественного языка. Распространяется по лицензии *MIT*.

1.1.2. Библиотеки векторизации

– *sklearn* – библиотека машинного обучения. Распространяется по лицензии *BSD-3*;

– *gensim* – библиотека обработки естественного языка и информационного поиска. Распространяется по лицензии *LGPL-2.1*.

1.2. Глоссарий

Текст – это некоторая последовательность предложений, имеющая логическую последовательность и сообщающая какую-либо информацию.

Корпус текстов – это подобранная и обработанная по определенным правилам совокупность текстов, используемая для исследования языка.

Токен – это текстовая единица (слово, словосочетание и т. д.).

1.3. Обозначение входных и выходных данных

pass

1.4. Математические модели методов векторизации

1.4.1. One-hot encoding

pass

1.4.2. TD-IDF

Общая формула показателя IDF выглядит следующим образом:

$$IDF(t, D) = \log \frac{|D| + 1}{DF(t, D) + 1}$$

где

– D – количество документов в корпусе,

– $DF(t, D)$ – количество документов, в которых встречается слово.

Так, если слово встречается во всех документах, то $IDF = 0$. В итоге,

$$TFIDF = IDF \cdot TF$$

1.4.3. CountVectorizer

pass

1.4.4. word2vec

pass

2. ОЗАГЛАВИТЬ

2.1. Методы предварительной обработки и фильтрации

2.1.1. Токенизация

Токенизация представляет из себя процесс разбиения больших участков текста на абзацы, предложения и слова. Данная операция не требует сторонних библиотек и может быть реализована с помощью стандартных модулей языка *Python*.

2.1.2. Лемматизация

2.1.3. Удаление шумовых слов

Под шумовыми словами подразумевают слова, не несущие смысловой нагрузки (междометия, союзы и т. д.). Операция может быть выполнена средствами языка программирования.

3. ОЗАГЛАВИТЬ

3.1.

ЗАКЛЮЧЕНИЕ