Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное

учреждение высшего образования

«Воронежский государственный лесотехнический университет

имени Г.Ф. Морозова»

Кафедра Вычислительной техники и информационных систем

(название кафедры)

КУРСОВАЯ РАБОТа

(вид работы)

Модель структурного представления

Текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации

(тема)

09.04.02 Информационные системы и технологии

(код и наименование направления подготовки)

по дисциплине Моделирование информационных систем

|  |  |  |
| --- | --- | --- |
| Студент группы ИС4-191-ОМ  (номер группы)  Руководитель, \_Профессор\_\_\_  (ученая степень, ученое звание) | \_\_\_\_\_\_\_\_\_\_\_  (подпись)  \_\_\_\_\_\_\_\_\_\_  (подпись) | \_М.О. Чайка\_\_\_\_  (инициалы и фамилия)  В.В. Лавлинский  (инициалы и фамилия) |

Воронеж 2019

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 3](#_Toc30413585)

[1 Анализ исходных данных 5](#_Toc30413586)

[1.1 Выделение ключевых для тематики слов в художественных произведениях 9](#_Toc30413587)

[1.2 Выделение ключевых слов в текстах на иностранном языке 12](#_Toc30413588)

[1.2 Выделение ключевых слов в научных статьях 13](#_Toc30413589)

[ЗАКЛЮЧЕНИЕ 15](#_Toc30413590)

# ВВЕДЕНИЕ

Тематика отражает содержание документа и включает в себя множество ключевых слов, находящихся в некоторой зависимости друг от друга. Один из вариантов такой зависимости – весовые коэффициенты, отражающие значимость того или иного слова в конкретной тематике.

Бурное развитие сетевых технологий, в том числе и Интернета, способствуют значительному увеличению доступных информационных ресурсов и объемов передаваемой информации. Зачастую это разнородная, слабо структурированная и избыточная информация, обладающая высокой динамикой обновления. При сегодняшних объемах доступной информации решение задач информационного поиска становится не только приоритетным, но и элементарно необходимым для обеспечения своевременного доступа к интересующей информации.

В данной курсовой работе рассматривается поиск текстовых документов произвольного содержания. Большое влияние на формирование поискового запроса (в поисковых системах Яндекс и Google) оказывает человеческий фактор, а именно: неопытность в работе с поисковыми системами, незнание набора ключевых слов, однозначно определяющих искомую информацию, отсутствие принятой и устоявшейся терминологии в интересующей области. Одним из вариантов решения этой проблемы является поиск документов по образцу, когда человек задает некоторый документ в качестве образца, а система, реализующая данный вариант поиска подбирает документы подобные заданному (подобные по содержанию, тематике).

Подобный подход позволяет существенно увеличить эффективность поиска информации как в интернете, так и в библиотечных каталогах, что делает разработку и совершенствование методов поиска текстов по образцу актуальной задачей.

Цель работы: Усовершенствовать метод тематического анализа неструктурированной текстовой информации для эффективного решения задач поиска документов по образцу.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать эффективность существующих методов и алгоритмов выделения тематики текстов.

2. Усовершенствовать метод автоматического определения тематики текста.

3. Реализовать метод автоматического определения тематики текста и оценить его эффективность

В предложенной курсовой работе рассматриваются методологии поиска произвольной информации в сети интернет или электронных библиотеках. В частности оценка тематической принадлежности текстов к заданному образцу. Под тематикой понимается множество ключевых слов, описывающих, с некоторой степенью адекватности, содержание документа. Тематика – это приближенное представление документа.

# 1 Анализ исходных данных

В обобщенном виде поиск документов по образцу выполняется следующим образом: задается текстовый документ тематически подобные которому тексты требуется найти, задается множество документов среди которых производится поиск, производится оценка схожести текстов, ранжирование в порядке убывания схожести и выборка максимально схожих текстов.



Рисунок 1.1 – Поиск документов по образцу

К методам оценки тематической близости текстов относятся следующие:

Оценка при помощи расстояния Левенштейна

Производится посимвольное сравнение двух текстов, что сужает область применения до заголовков конкретных статей, и вычисляется минимально необходимое число действий (замен, удалений, добавлений символов), необходимых для получения из первого текста второго.

Оценка с использованием коэффициента Танимото.

Данный метод имеет более широкое применение. Он включает в себя сравнение содержимого двух текстов и вычисление коэффициента по формуле: , где a – это количество слов в первом тексте, b – количество слов во втором тексте и c – количество совпадающих символов.

Оба эти метода позволяют обрабатывать большие объемы информации, однако их эффективность снижается прямо пропорционально увеличению объема текста. Таким образом если с их помощью можно с достаточной точностью определить близость нескольких слов или предложений друг к другу, то при анализе больших объемов текстовой информации они смогут определить лишь небольшие отклонения от основного текста, а не установить сходство двух разных текстов. Поэтому данные методы обладают ограниченным потенциалом к использованию в данной работе.

Следующие методы предполагают вначале выделение ключевых слов, определяющих тематику текста и затем сравнение попарно наборов таких слов.

Частотный метод выделения тематики текста.

Для применения этого метода к тексту производится выборка наиболее часто встречающихся в тексте слов, упорядочивание их в порядке возрастания частоты встречаемости и выборка либо определенного количества слов либо слов чье процентное содержание в тексте выше определенного порога.

Графовая модель структурного представления текста произвольного содержания

Суть подхода заключается в моделировании структуры текста информационным потоком и формировании этим потоком ориентированного мультиграфа, вершинами которого являются слова, а ребрами – связи между словами в тексте. Этот мультиграф является информационной структурой текста. Вершины мультиграфа имеющие наибольшее количество связанных с ними ребер представляют собой набор ключевых слов текста. Так же этот метод позволяет в дальнейшем применять метод частотно-контекстной классификации текста, что существенно повышает качество определения темы.

Пример обработки предложения при помощи графовой модели:

Текст: "Лениво дышит полдень мглистый, лениво катится река. И в тверди пламенной и чистой лениво тают облака."

Предложенный фрагмент стихотворения при помощи Графовой модели представления текста будет выглядеть следующим образом:

Множество уникальных информационных элементов

Лениво, чистой

Дышит тают

Полдень облака

мглистый

катится

река

и

в

тверди

пламенной

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

Рисунок 1.2 – Представление текста в виде графа

Для оценивания эффективности работы существующих методов было проведено программное тестирование. Для этого при помощи частотного и графового метода были проанализированы несколько текстов разного содержания. Среди них художественные произведения, исторические статьи, статьи по программированию, статьи и учебники на английском языке.

## 1.1 Выделение ключевых для тематики слов в художественных произведениях

Для проведения анализа использовался полный текст книг Дж. Р. Р. Толкиена, Трилогия Властелин колец, переведенный на русский язык.

Книга 1 – Властелин колец: Братство кольца

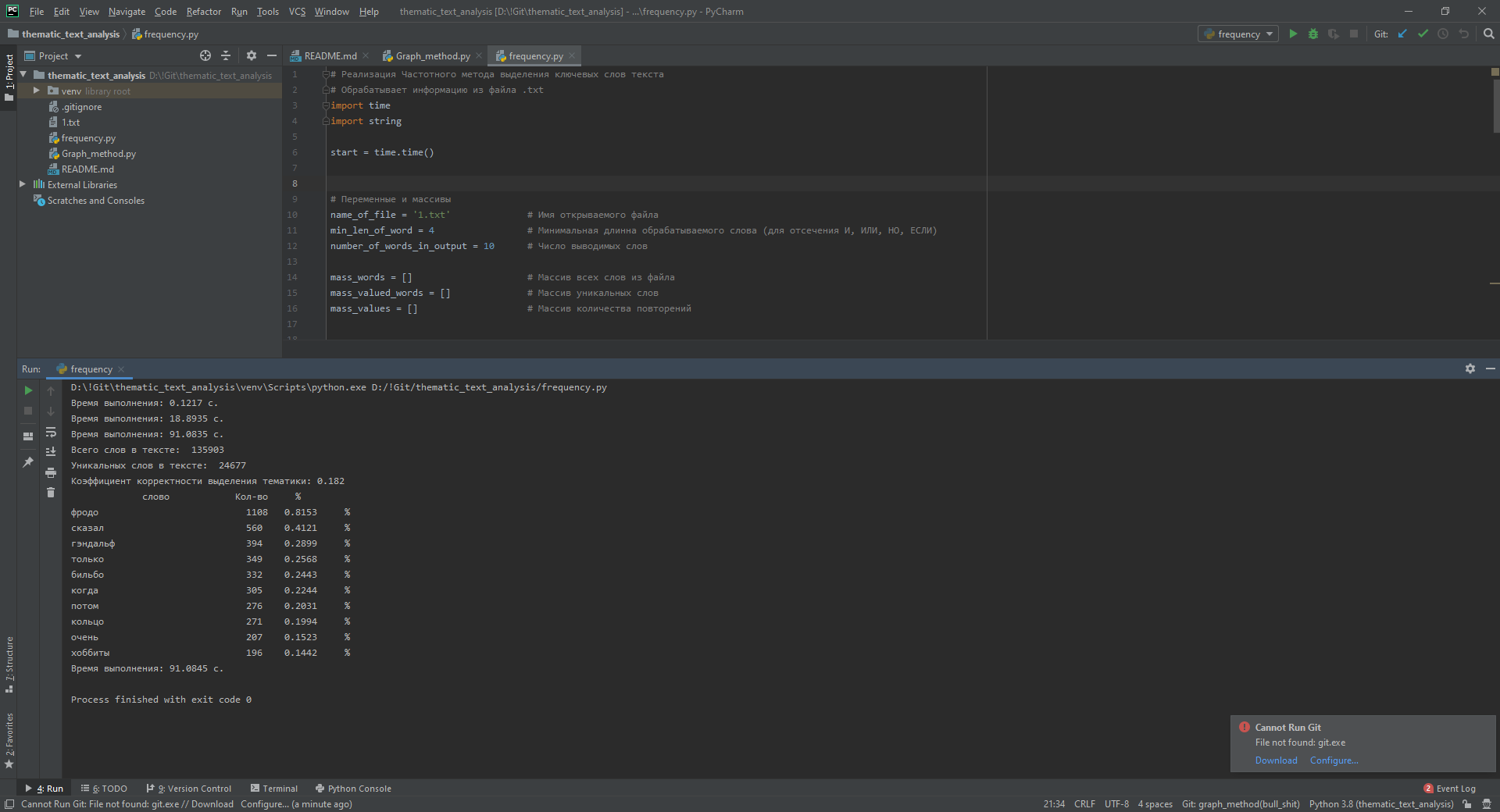


Рисунок 1.3 – Результат анализа частотным методом книги 1

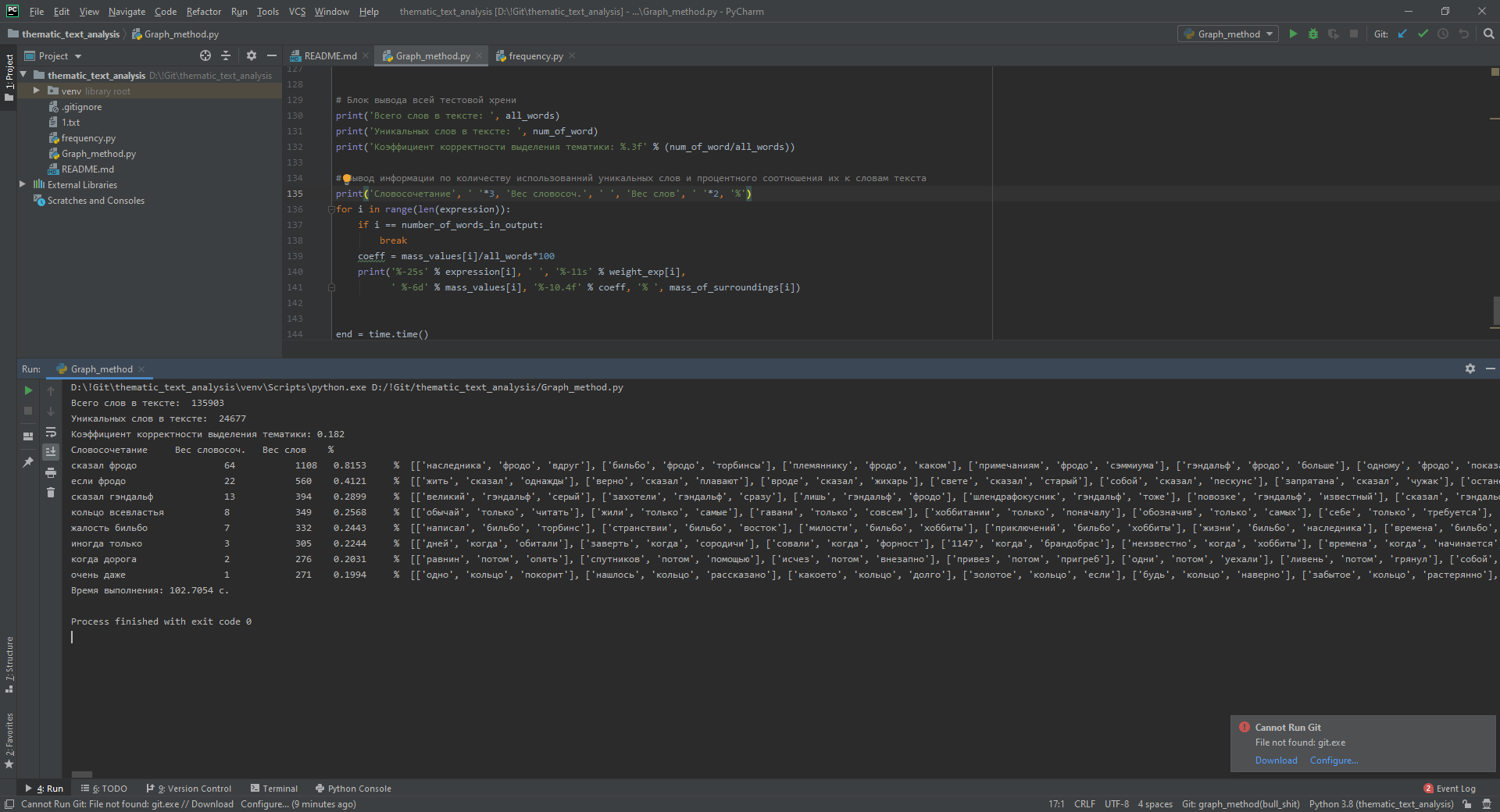


Рисунок 1.4 – Результат анализа графовым методом книги 1

Книга 2 – Властелин колец: Две башни

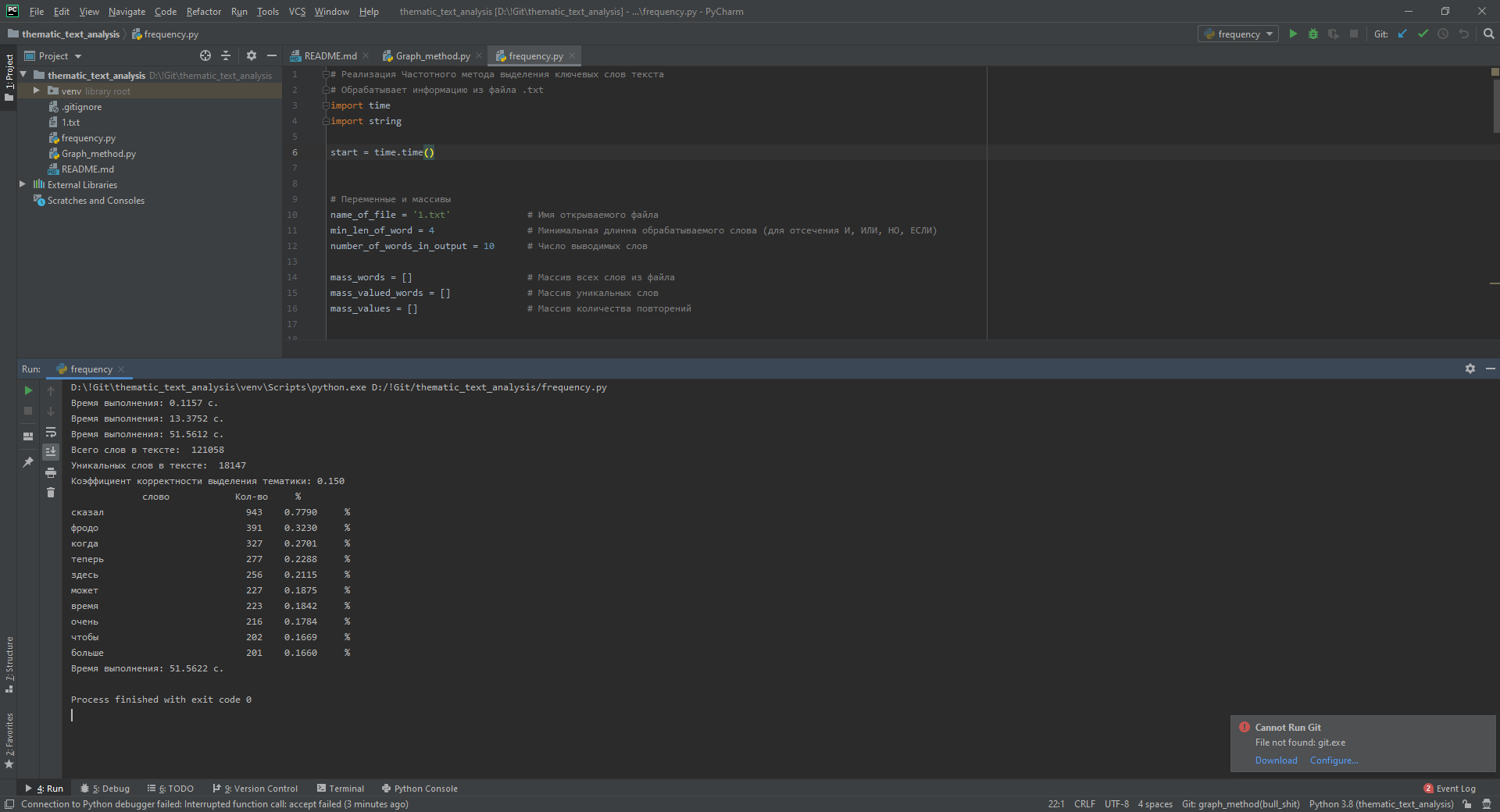


Рисунок 1.5 – Результат анализа частотным методом книги 2

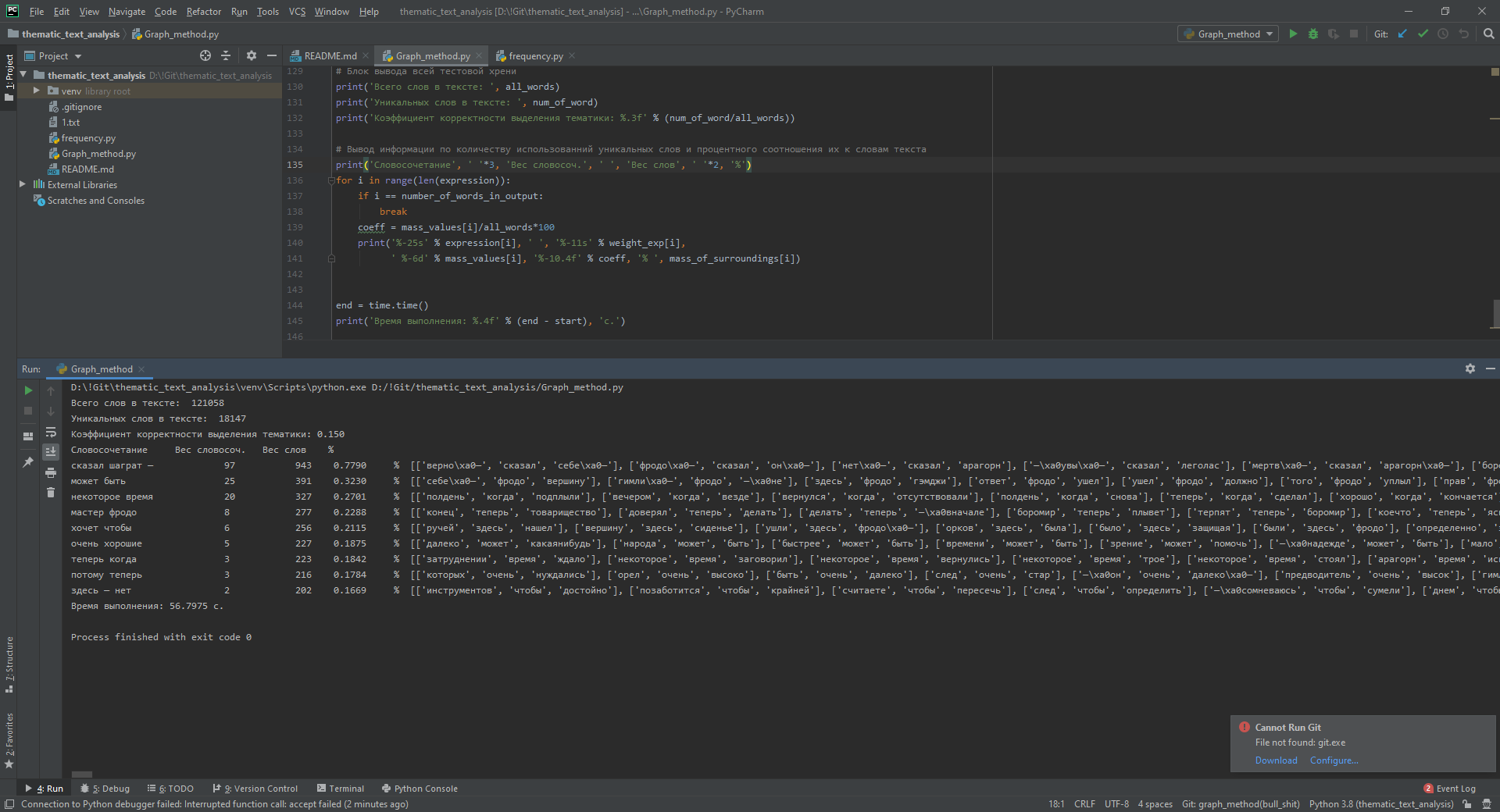


Рисунок 1.6 – Результат анализа графовым методом книги 2

Книга 3 – Властелин колец: Возвращение короля

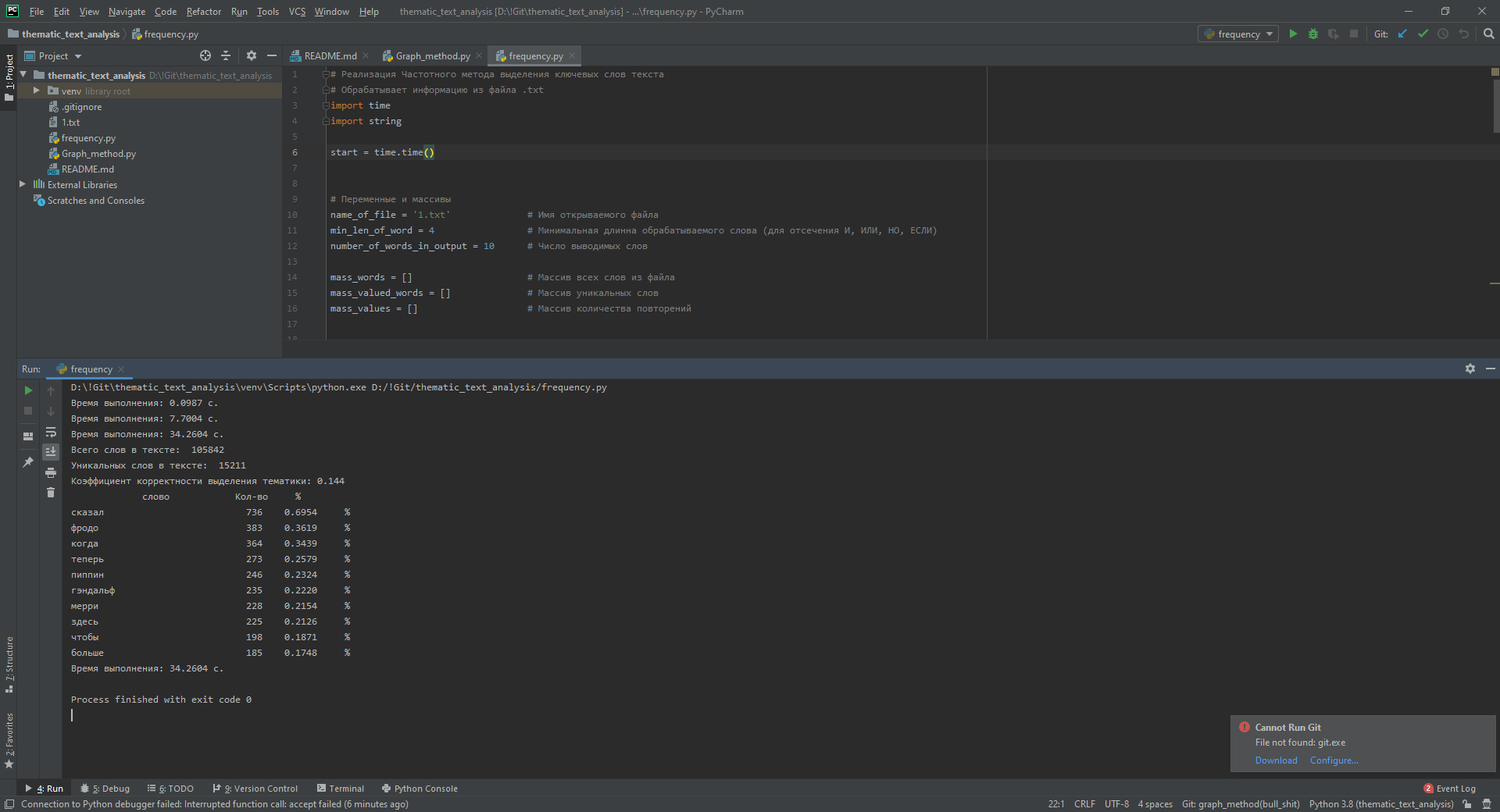


Рисунок 1.7 – Результат анализа частотным методом книги 3

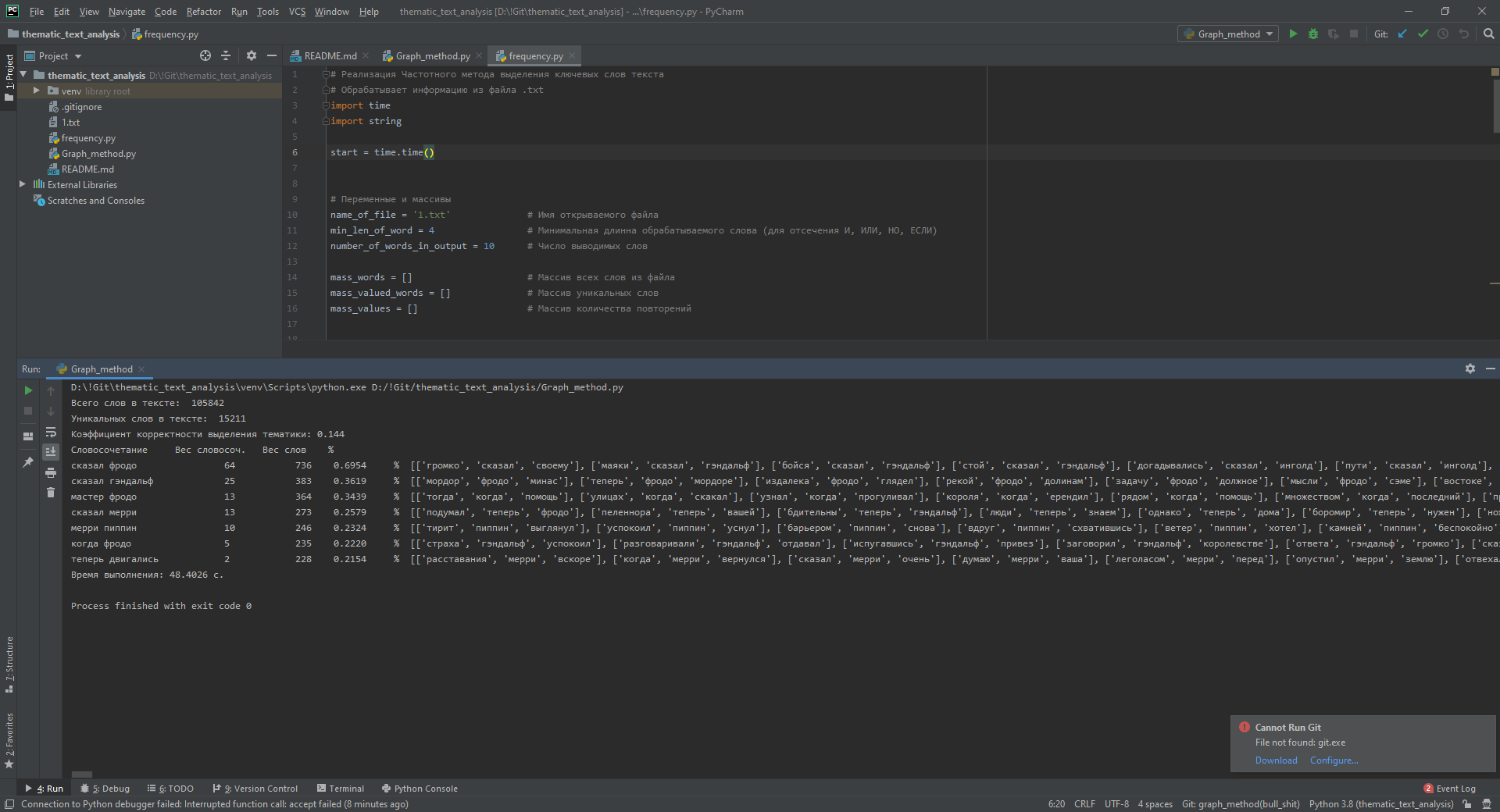


Рисунок 1.8 – Результат анализа графовым методом книги 3

На основе приведенных выше результатов можно сделать следующий вывод: Графовый и частотный метод позволяют достаточно быстро обрабатывать большие объемы текстовой информации, с достаточной точностью выделяют ключевые слова. Однако, как видно на примере анализа второй книги, художественные тексты не слишком корректно обрабатывать, используя только эти два метода. Это связано с тем, что автор, зачастую, старается избегать многократного повторения одинаковых слов, используя синонимы. Из за этого в итоговую выборку ключевых слов или словосочетаний попадают такие выражения: “здесь”, “может”, “чтобы”, “очень хорошие”, “теперь когда” и тд. что снижает качество определения тематики текста.

## 1.2 Выделение ключевых слов в текстах на иностранном языке

Одним из преимуществ данных методов является возможность обработки с такой же эффективностью текстов на иностранных языках, например на рисунках 1.9, 1.10 происходит определение тематики текста на английском языке. В качестве примера использовалась статья Genetic Programming, John R. Koza и часть учебника Julian F.Miller Cartesian Genetic Programming.

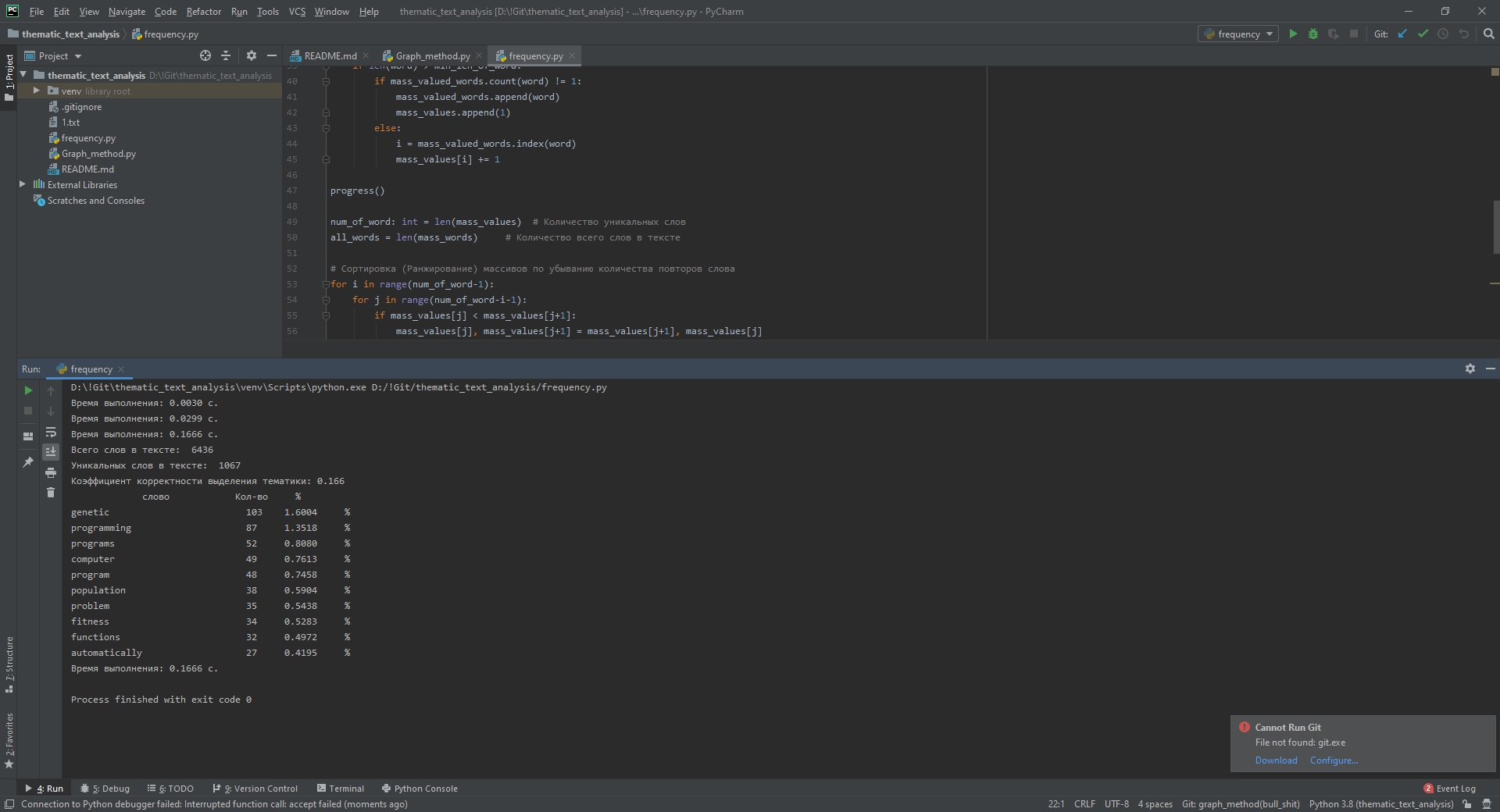


Рисунок 1.9 – Результат анализа статьи Genetic Programming

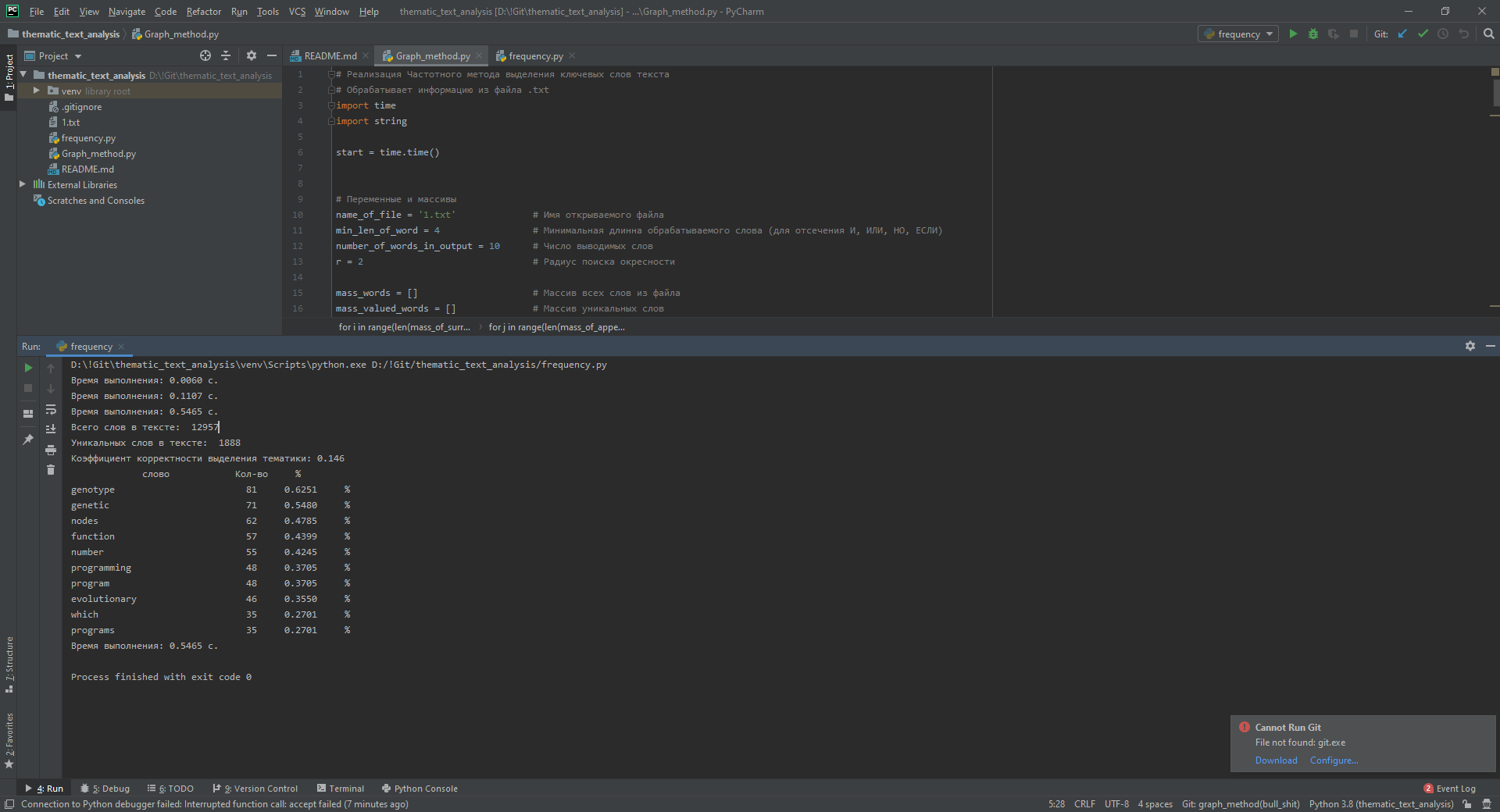


Рисунок 1.10 – Результат анализа учебника Cartesian Genetic Programming

Таким образом не обязательно знать язык на котором написаны тексты для того что бы приблизительно определить их тематическую близость.

## 1.2 Выделение ключевых слов в научных статьях

Рассмотрим работу методов на примере исторических статей по истории древней Сирии:

Были обработаны два разных текста по данной тематике. На рисунках 1.11 и 1.12 представлены результаты обработки статьи из википедии.

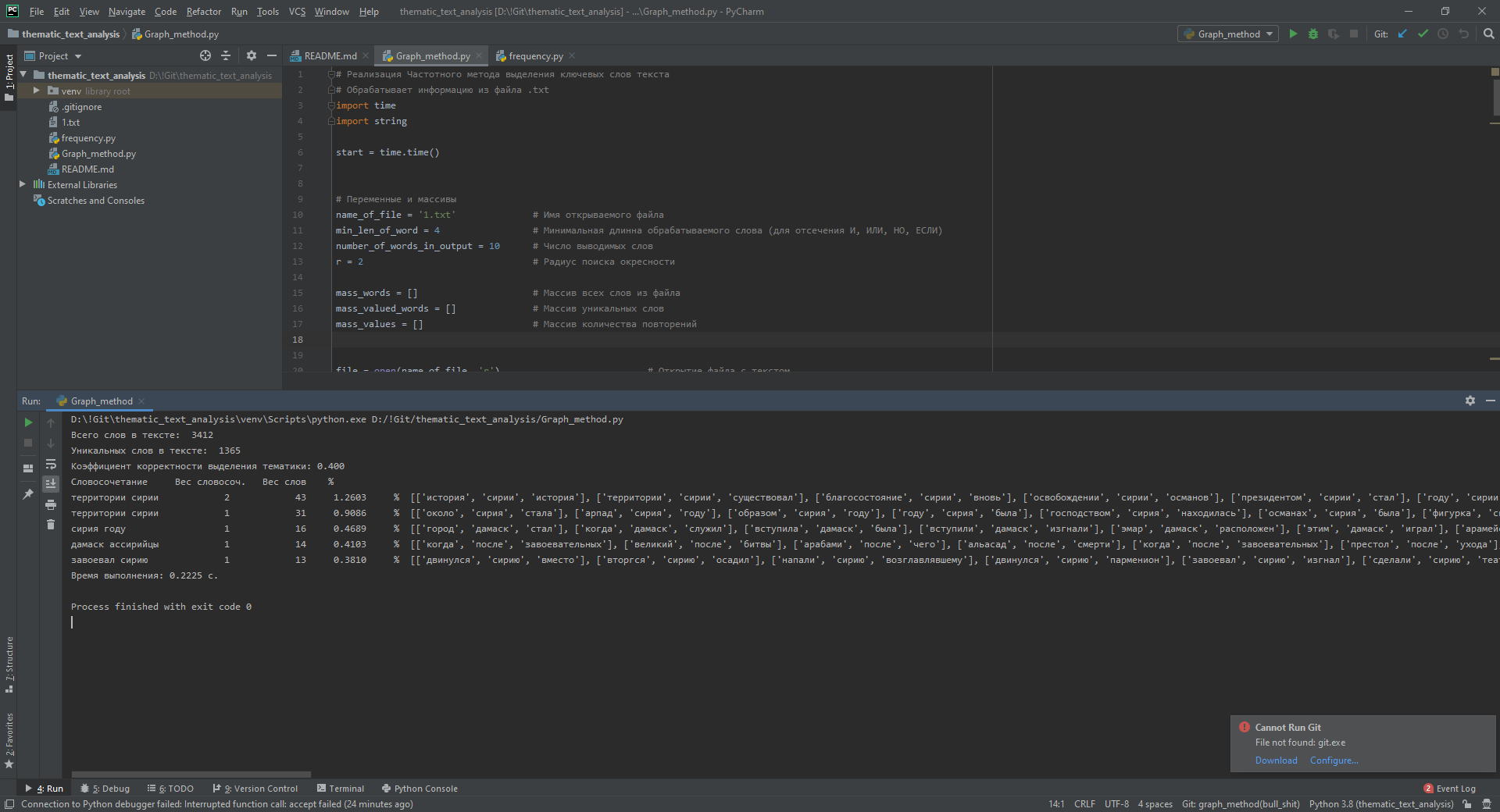


Рисунок 1.11 – Результат анализа графовым методом статьи википедии

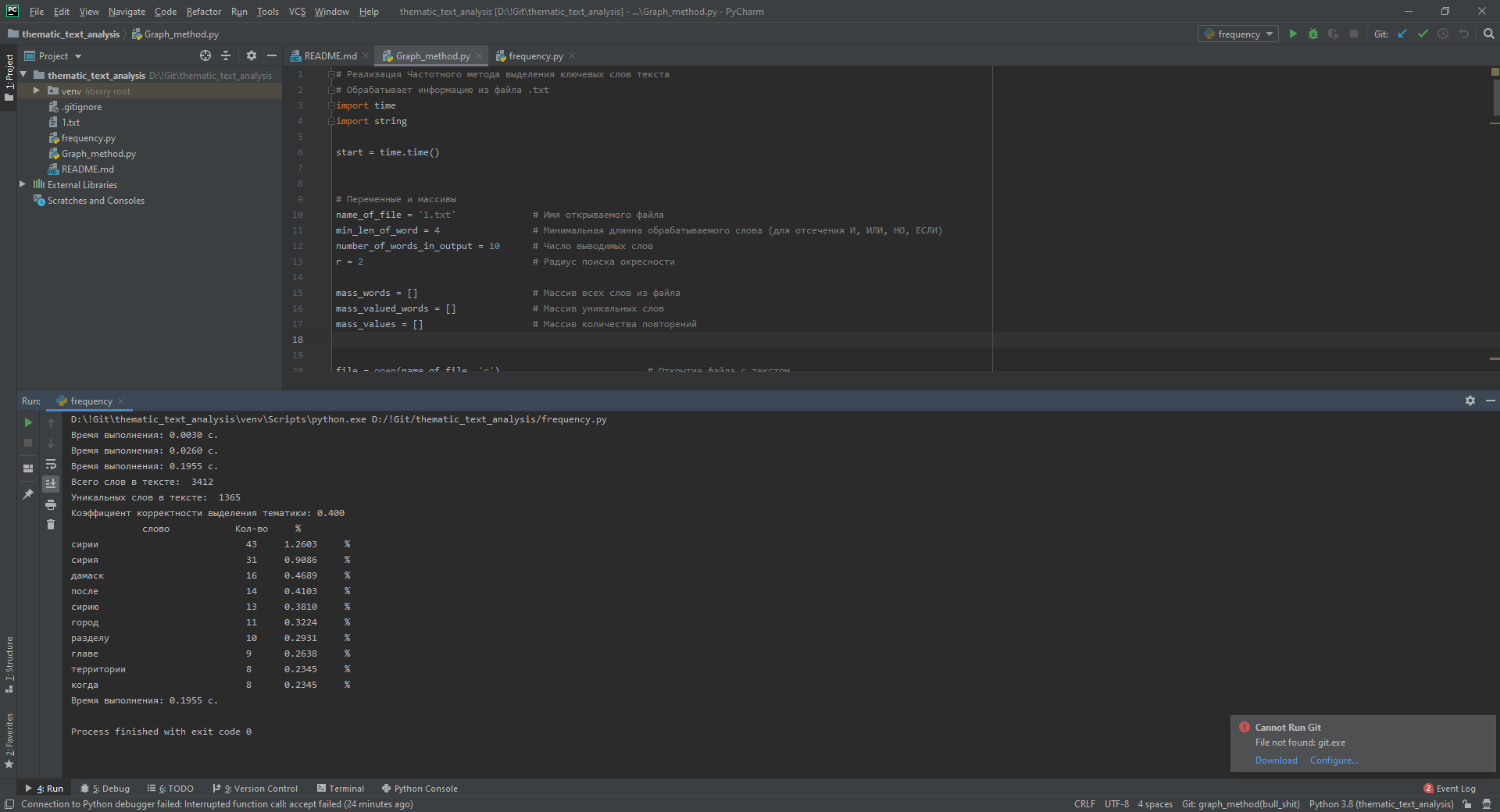


Рисунок 1.12 – Результат анализа частотным методом статьи википедии

При обработке текста из энциклопедии Кольера частотным методом был получен следующий результат (Рисунок 1.13)

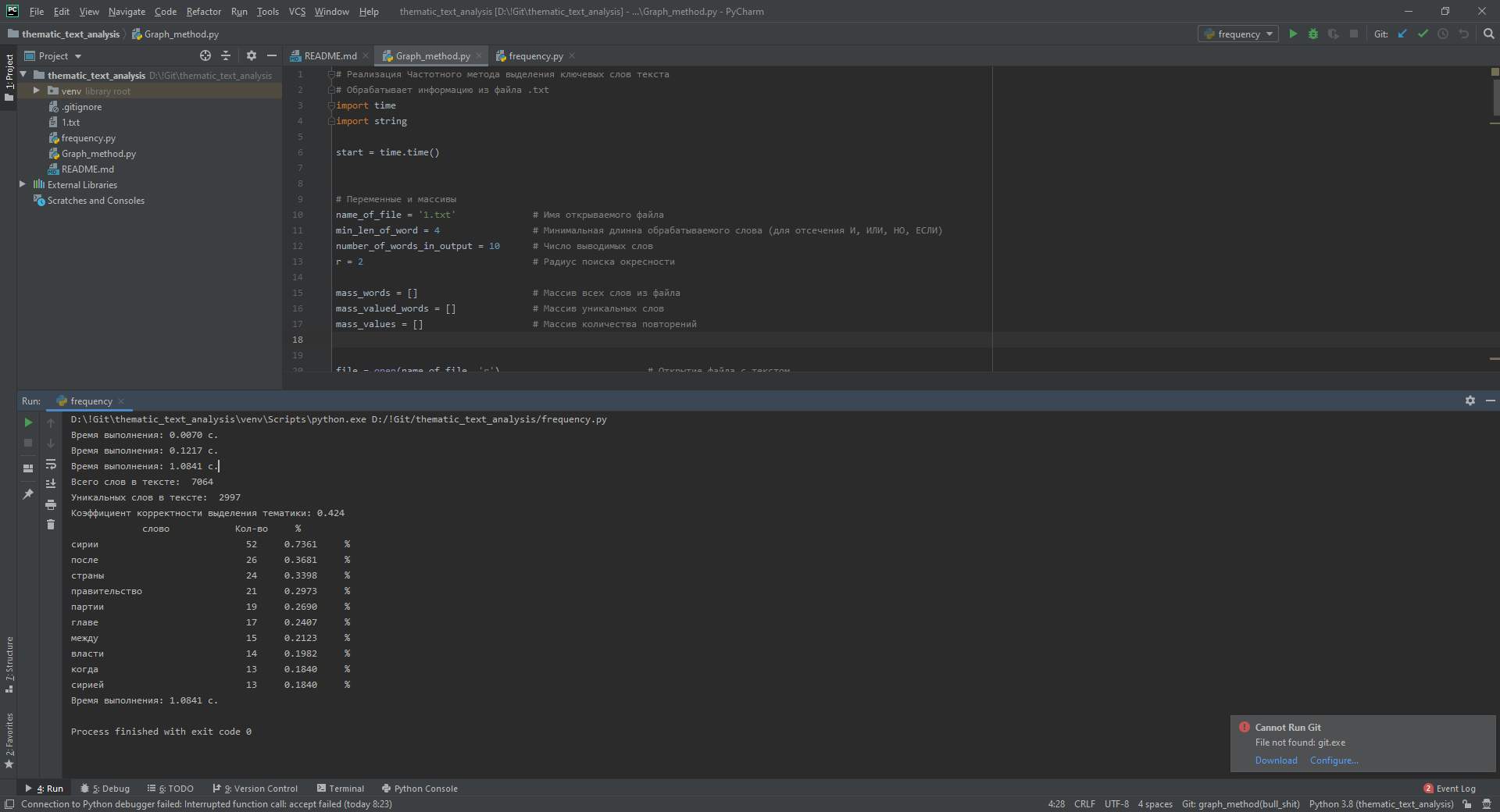


Рисунок 1.13 – Результат анализа частотным методом текста из энциклопедии Кольера

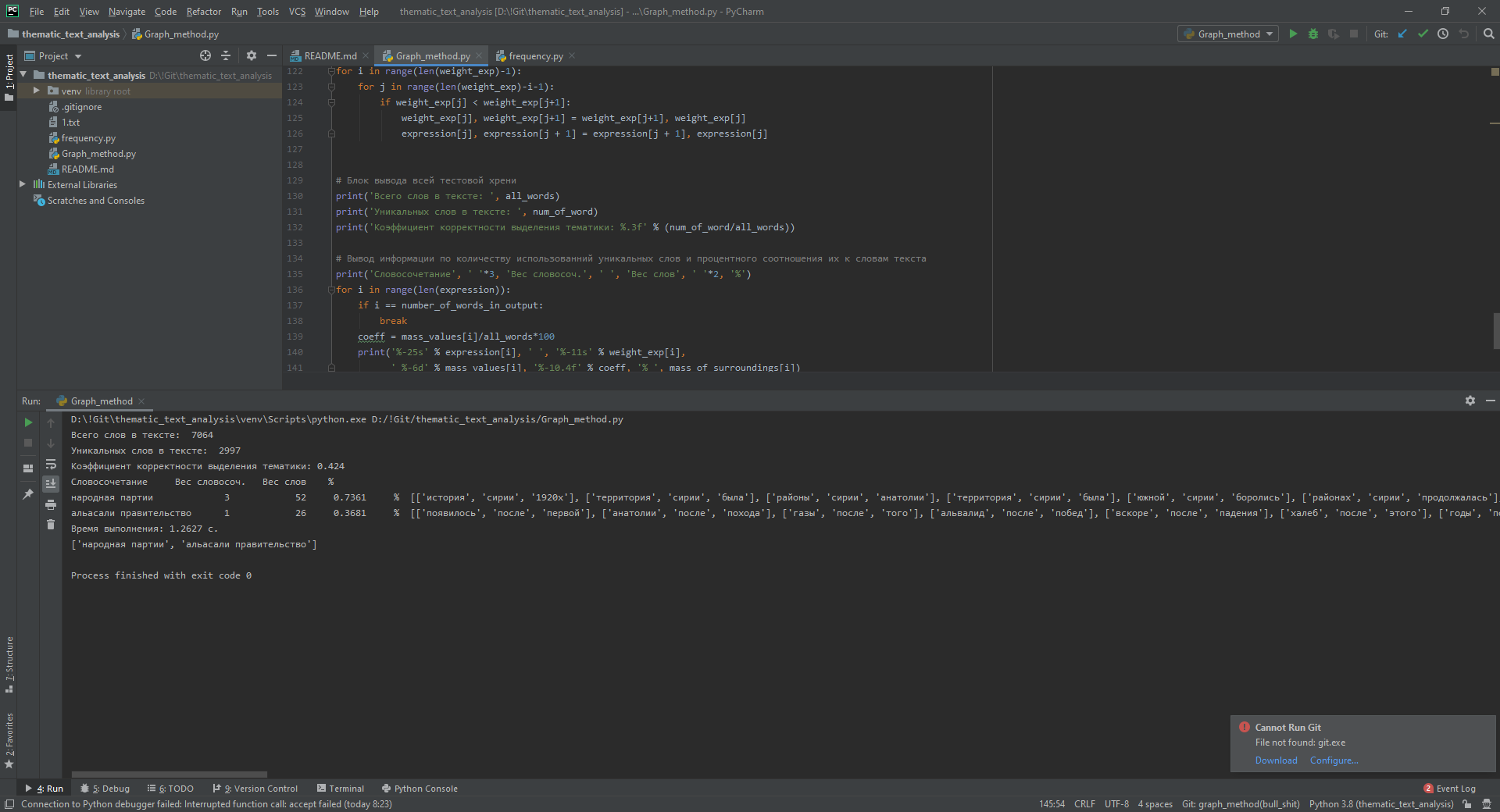


Рисунок 1.14 – Результат анализа графовым методом текста из энциклопедии Кольера

На основе полученных результатов можно сказать что данные методы позволяют присоединить оба текста к тематической группе «Сирия», но не более того. Так же на рисунках 1.12 и 1.13 виден один из недостатков этого подхода - данные методы рассматривают разные формы одного слова (например Сирия, Сирии, Сирией) как разные слова и учитывают каждое из них отдельно. В случае если бы эти слова рассматривались как одинаковые в результате могли появиться дополнительные слова, позволяющие уточнить тематику текста.

# ЗАКЛЮЧЕНИЕ

На основе проведенного исследования можно сделать вывод о том что частотный метод выделения тематики текста и, основанный на нем, графовый метод позволяют справляться с поставленной задачей и обладают рядом преимуществ, таких как скорость работы и низкая зависимость от языка обрабатываемого текста. Однако они обладают и рядом качественных недостатков, которые сказываются на точности определения тематики текста.

Таким образом цель магистерской диссертации – Усовершенствование метод тематического анализа неструктурированной текстовой информации для эффективного решения задач поиска документов по образцу, является актуальной как с практической так и с научной точек зрения.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Адарюков В.И. Исследование и разработка машинно-ориентированного метода инфологического моделирования информационно-поисковых систем фактографического типа: Диссертационная работа к.т.н.: 05.13.06 / Ленинградский электротехнический институт им В.И. Ульянова (Ленина). – СПб., 1988. – 256 с.
2. Ахутина Т. В. Порождение речи. Нейро-лингвистический анализ синтаксиса - М.: МГУ, 1989. – 215 с.
3. Белянин В.П. Введение в психолингвистику. – Изд. 2-е, испр. и доп., – М.: ЧеРо, 2000. – 128 с.
4. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.
5. Добрынин В.Ю., Некрестьянов И.С., Задача выбора тематических коллекций, релевантных запросу. // Труды Всероссийской научно-методической конференции “Интернет и современное сообщество”, Санкт-Петербург, декабрь 1998.
6. Дубинский А.Г. Разработка моделей и совершенствование структуры систем информационного поиска в глобальной компьютерной сети: Диссертационная работа к.т.н.: 05.13.06 / Днепропетровский национальный университет. – Днепропетровск, 2002.
7. Дубинский А.Г. Проблема автоматизации поиска информации в глобальной сети // Проблемы автоматизации информационных технологий. – Днепропетровск, 1999. - С. 40-48.
8. Дубинский А.Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. - 2001. - №4. - С. 77-83.
9. Ермаков А.Е. Полнотекстовый поиск: проблемы и их решение // Мир ПК. – 2000. - N5.
10. Ермаков А.Е. Неполный синтаксический анализ текста в информационно-поисковых системах // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2002. В двух томах. Т.2. “Прикладные проблемы”. – Москва, 2002. - С. 180-185.
11. Ермаков А.Е., Плешко В.В. Ассоциативная модель порождения текста в задаче классификации // Информационные технологии. - 2000. - N 12.
12. Иванов В., Некрестьянов И., Пантелеева Н. Расширение представления документов при поиске в Веб // Труды четвертой всероссийской конференция RCDL'2002. В двух томах. Т.2. - Дубна, 2002. - C. 55-68.
13. Когаловский М. Р. Перспективные технологии информационных систем. – М.: ДМК Пресс; М.: Компания АйТи, 2003. – 288 с.
14. Когаловский М.Р. Энциклопедия технологий бах данных. – М.: Финансы и статистика, 2002. – 800 с.