Оглавление

[Введение 1](#_Toc35198704)

[1. Разработка модели структурного представления и метода тематического анализа текста 2](#_Toc35198705)

[1.1Графовая модель структурного представления текста произвольного содержания 3](#_Toc35198706)

[1.2 Граф 3](#_Toc35198707)

[1.3 Мультиграф 4](#_Toc35198710)

[2. Программа 9](#_Toc35198713)

[2.1 Выделение ключевых слов в текстах на иностранном языке 11](#_Toc35198714)

[ЗАКЛЮЧЕНИЕ 14](#_Toc35198715)

**Практическая реализация модели структурного представления и методы математического анализа**

# 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.

# 1. Разработка модели структурного представления и метода тематического анализа текста

Задача поиска документов по образцу предпологает решение двух основных задач:

1) тематическая классификация текстовой информации;

2) вычисление степени тематической принадлежности текста к заданному классу.

Эти задачи связаны, прежде всего, с анализом текста, а именно, с анализом смыслового содержания текста, его тематической направленности.

* лингвистический анализ;
* статистический анализ.

Первый ориентирован на извлечении смысла текста по его семантической структуре. Второй – по частотному распределению слов в тексте.

В данной работе было принято решение использовать методы статистического анализа в силу их относительной простоты, удобства использования и языковой независимости. Методы лингвистического анализа, хотя и позволяют точнее анализировать текст, выделяя его структурные особенности, но являются более трудоемкими и сложными в использовании.

Однако частотный анализ, используемый в настоящее время при определении тематики документов, не позволяет в полной мере учесть внутреннюю структуру текста, т.к. при таком анализе не учитывается связность и последовательность текста. Хотя именно связность текста (речевого высказывания) считается одним из важнейших условий, необходимых для понимания его смыла и содержания.

## 1.1Графовая модель структурного представления текста произвольного содержания

### 1.2 Граф

Суть предлагаемого подходазаключается 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.

Мультиграф – это граф, который может содержать множество ребер соединяющих одну и туже пару вершин.

Информационный поток – это детерминированный поток событий, принадлежащих некоторому конечному множеству.

Переход к модели структурного представления текста осуществляется следующим образом.

1) Текст рассматривается в виде информационного потока, образованного информационными элементами - словами.

Если последовательно брать слова из текста, начиная с самого первого и кончая последним, то это как раз и будет информационный поток F.

При этом набор всех слов в тексте можно выделить в конечное множество уникальных информационных элементов

Информационный потокF будет представлен в виде последовательного чередования этих элементов(рис. 2.1).

rn-1

r2

r1

Рис. 2.1. Информационный поток

1.3 Мультиграф

Если учесть, что слова в тексте повторяются, то, соответственно, можно допустить, что информационный поток будет многократно проходить через одни и те же информационные элементы, формируя, таким образом, связанную информационную структуру текста.

Фрагмент текста: “Дао, которое может быть выражено словами не есть постоянное Дао. Имя, которое может быть названо, не есть постоянное имя”.

Для вышеприведенного фрагмента текста данная структура будет выглядеть следующим образом (рис. 2.3).

Рис. 2.3. Структура, формируемая информационным потоком

Тезис о том, что информационный поток будет многократно проходить через одни и те же информационные элементы, носит принципиальный характер, при отсутствии циклов графа, образованных информационным потоком, ни о какой формируемой структуре не приходится говорить.

На самом деле структура, представленная на рис.2.3, не отражает в полной мере информационный поток, описывающий фрагмент текста. Необходимо дополнить эту структуру информацией о прохождении потока через каждый из информационных элементов. Для каждого повторного прохождения потока через одну и ту же пару информационных элементов необходимо формировать дополнительные связи – ребра. Тогда структура описывается в виде мультиграфа – графа, который может содержать множество ребер соединяющих одну и ту же пару вершин.

Для удобства отображения такого мультиграфа проиндексируем информационный поток и припишем каждому ребру графа, соединяющего пару вершин, множество индексов, соответствующих прохождению информационного потока через данную пару.

5

17

8

16

7

13

3

12

2

15

11

4

14

6

18

9

10

1

Рис. 2.4. Индексированная структура информационного потока

Весьма показательным является приведенный выше фрагмент текста, особенность его, прежде всего, в том, что для сравнительно небольшого количества слов, входящих в данный фрагмент, существуют циклы. И информационный поток, моделирующий данный фрагмент, формирует явно выраженную структуру. Лингвистический анализ текста, его стилистические особенности и смысловая многозначность выходят за рамки диссертационной работы, но некоторая корреляция, по-видимому, существует между информативностью и структурной насыщенностью - наличием большого числа связей между отдельными элементами. Явно выражен принцип усложнения системы за счет введения дополнительных связей между отдельными элементами.

Введем дополнительные обозначения и определим некоторые важные характеристики информационной структуры.

n(I) = | I | - количество информационных элементов множества I (количество уникальных слов в тексте).

n(F) = | F | - количество информационных элементов набора F (общее количество слов в тексте).

M(I, R) – информационная структура (ориентированный  мультиграф). Является совокупностью I - множества информационных элементов (вершин графа) и R - набора связей между этими элементами (ребер графа).

M(I, R) F.

R - набор связей между парами информационных элементов, может содержать повторяющиеся связи в случае многократного прохождения информационного потока F через одни и те же пары элементов.

Обозначимколичество пар связей как:

d(i) –степеньинформационного элемента

d(i) = n(R(i)),

, .

d(M(I, R))max – максимальная степеньинформационного элемента для информационной структуры M(I, R):

d(M(I, R))max = maxd(i),.

d(M(I, R))min – минимальная степеньинформационного элемента для информационной структуры M(I, R):

d(M(I, R))min = mind(i), .

r+(F, ik, ij, e), r-(F, ik, ij, e) – расстояние между двумя информационными элементами в потокеF, , где:

e –вхождение информационного элемента ikв поток F, его порядковый номер в потоке.

r+ - означает, что расстояние измеряется по ходу информационного потока.

r- - означает, что расстояние измеряется обратно ходу информационного потока.

Расстояние измеряется в количестве информационных элементов находящихся между ikи ij,плюс 1.

Пример:

F = (i2, i5, i1, i4, i5, i2, i3)

r+(F, i2, i3, 1) = 6

r+(F, i2, i3, 2) = 1

r-(F, i1, i2, 1) = 2

r-(F, i1, i3, 1) = 0

r+(F, i1, i3, 1) = 4

0 – означает, что результат не определен

rmin(F, ik, ij) – минимальное расстояние между двумя информационными элементами в потоке, .

Пример:

F = (i2, i6, i1, i4, i5, i7, i2)

rmin(F, i2, i5) = 2

rmin(F, i1, i2) = 2

rmin(F, i1, i8) = 0

Информационный поток относительно некоторого информационного элемента i можно описать как:

F(i, e, [r-, r+]),

гдеe – вхождение информационного элемента ikв поток F, его порядковый номер в потоке; [r-, r+] – окрестность, для которой определяется поток.

F(i, e, [r-, r+]) = (i - r-, …, i - 2, i - 1, i, i + 1, i + 2, …, i + r+),

где - обозначает индексациюнекоторого информационного элемента в наборе F относительно информационного элемента i;

i + 1 – обозначает информационный элемент, следующий сразу за iв информационном потоке F;

i- 1 – обозначает информационный элемент, предшествующий i, в информационном потоке F.

Такая ситуация возможна, когда заданный информационный элемент находится близко к началу или концу информационного потока, описывающего текст, т.е. элементов i - r-, …, i - 2, i - 1 иi + 1, i + 2, …, i + r+ может просто не существовать, или существовать только часть их, в зависимости от того, насколько близок элемент iк началу или концу текста.

Также такая ситуация возможна в том случае, когда циклы, формируемые потоком, многократно проходящим через i,имеют меньше информационных элементов, чем выбранноеr.

В таких случаях будем принимать равным 0, все элементы, которые не определены в данной окрестности.

Обозначим множество всех информационных потоков относительно информационного элемента i для всех его вхождений в поток F:



Пример:

F = (i2, i5, i1, i4, i5, i2, i3, i8, i5, i1),

d(i5) = 3,

D(F, i5, [2, 2]) = ((0, i2, i5, i1, i4),(i1, i4, i5, i2, i3),(i3, i8, i5, i1, 0)).

На базе представленной модели выполним разработку метода и алгоритмов тематического анализа текста для решения двух основных задач данной работы:

1) тематической классификации текстовой информации;

2) вычисления степени тематической принадлежности текста к заданному классу.

# 2. Программа

Для написания алгоритма был использован высокоуровневый язык python.

С кодом можно ознакомиться по ссылке: <https://github.com/kamome-inc/thematic_text_analysis>

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

Книга 1 – Властелин колец: Братство кольца

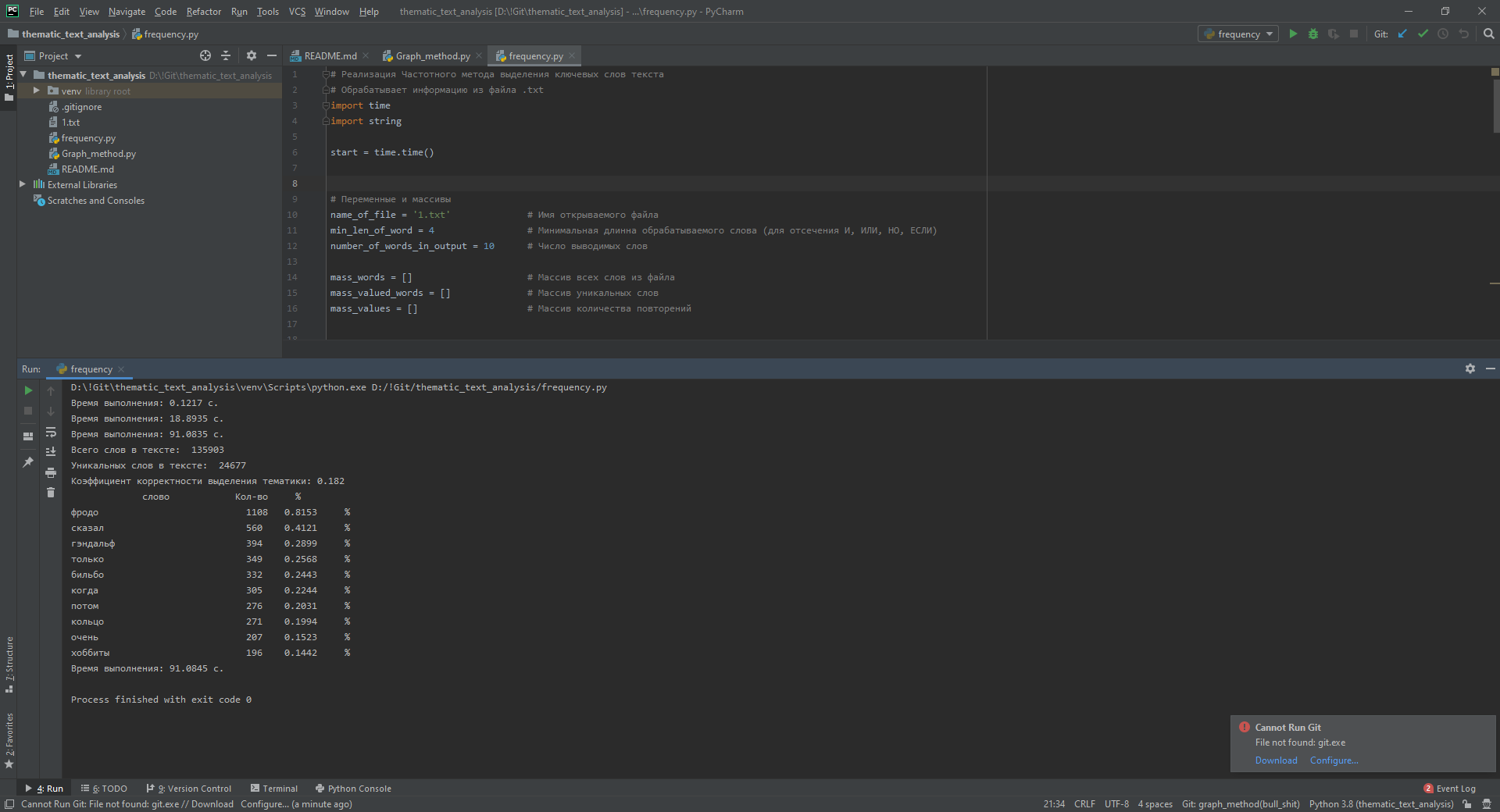


Рисунок 1.3 – Результат анализа частотным методом книги 1

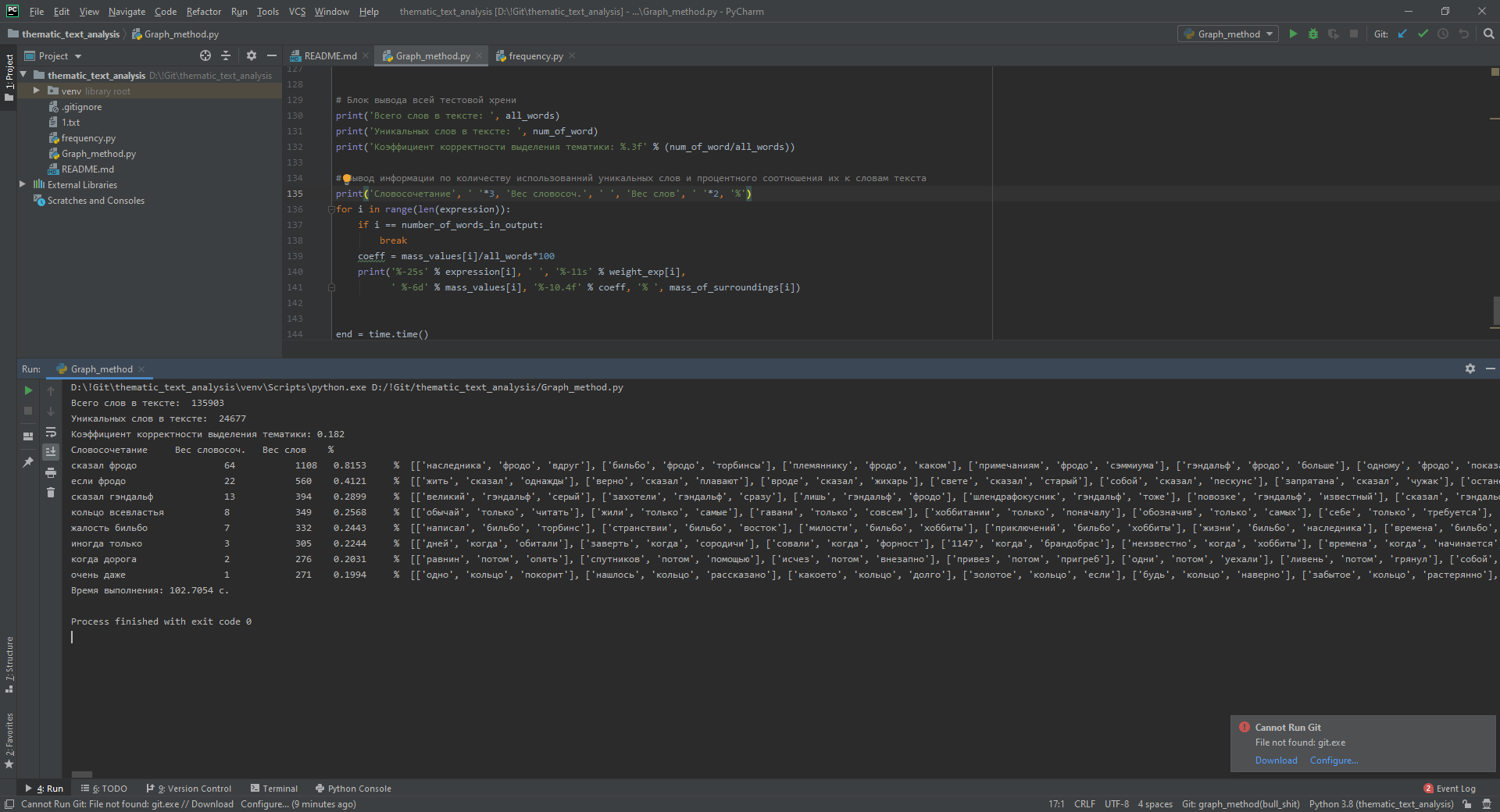


Рисунок 1.4 – Результат анализа графовым методом книги 1

Книга 2 – Властелин колец: Две башни

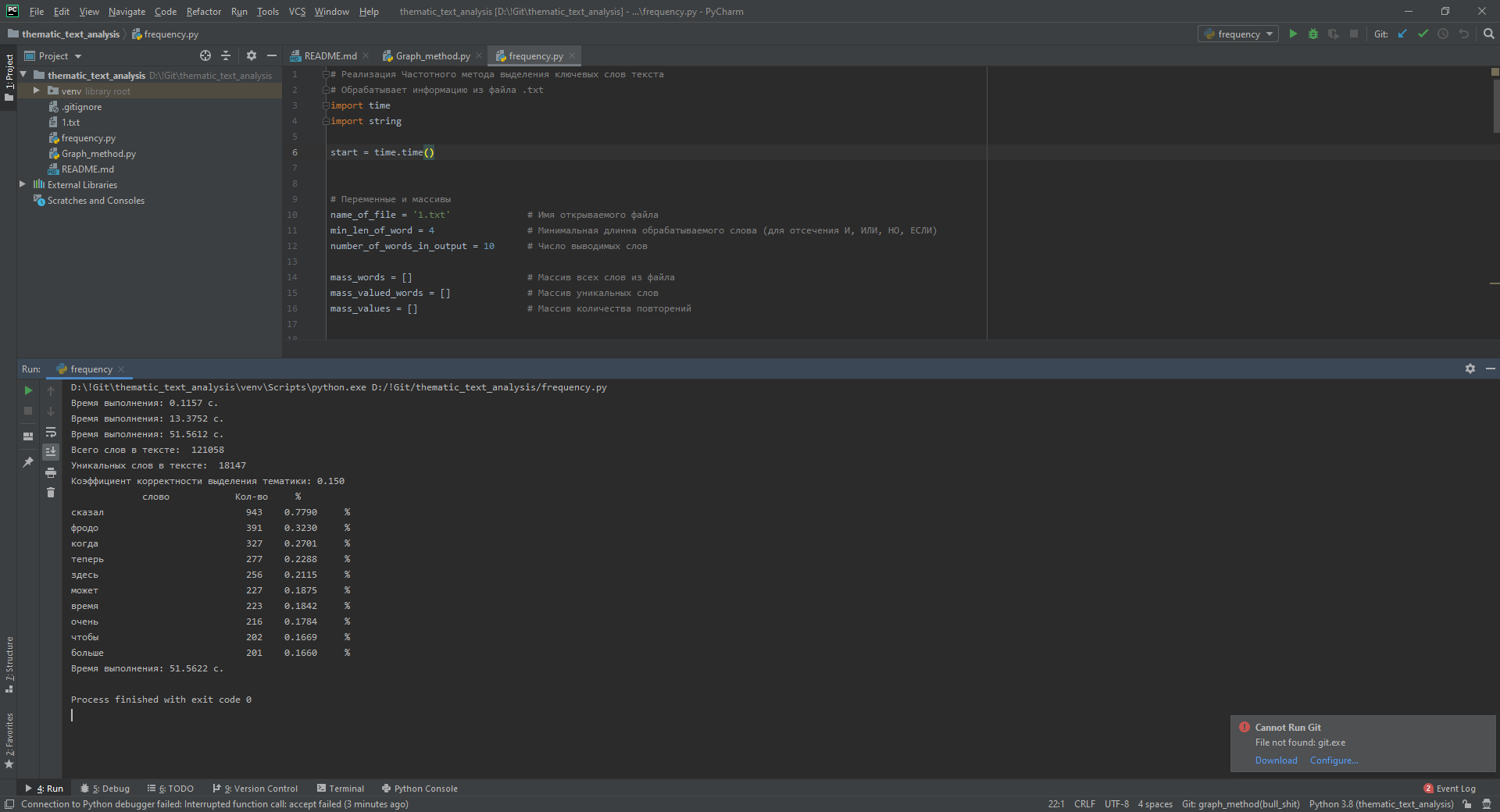


Рисунок 1.5 – Результат анализа частотным методом книги 2

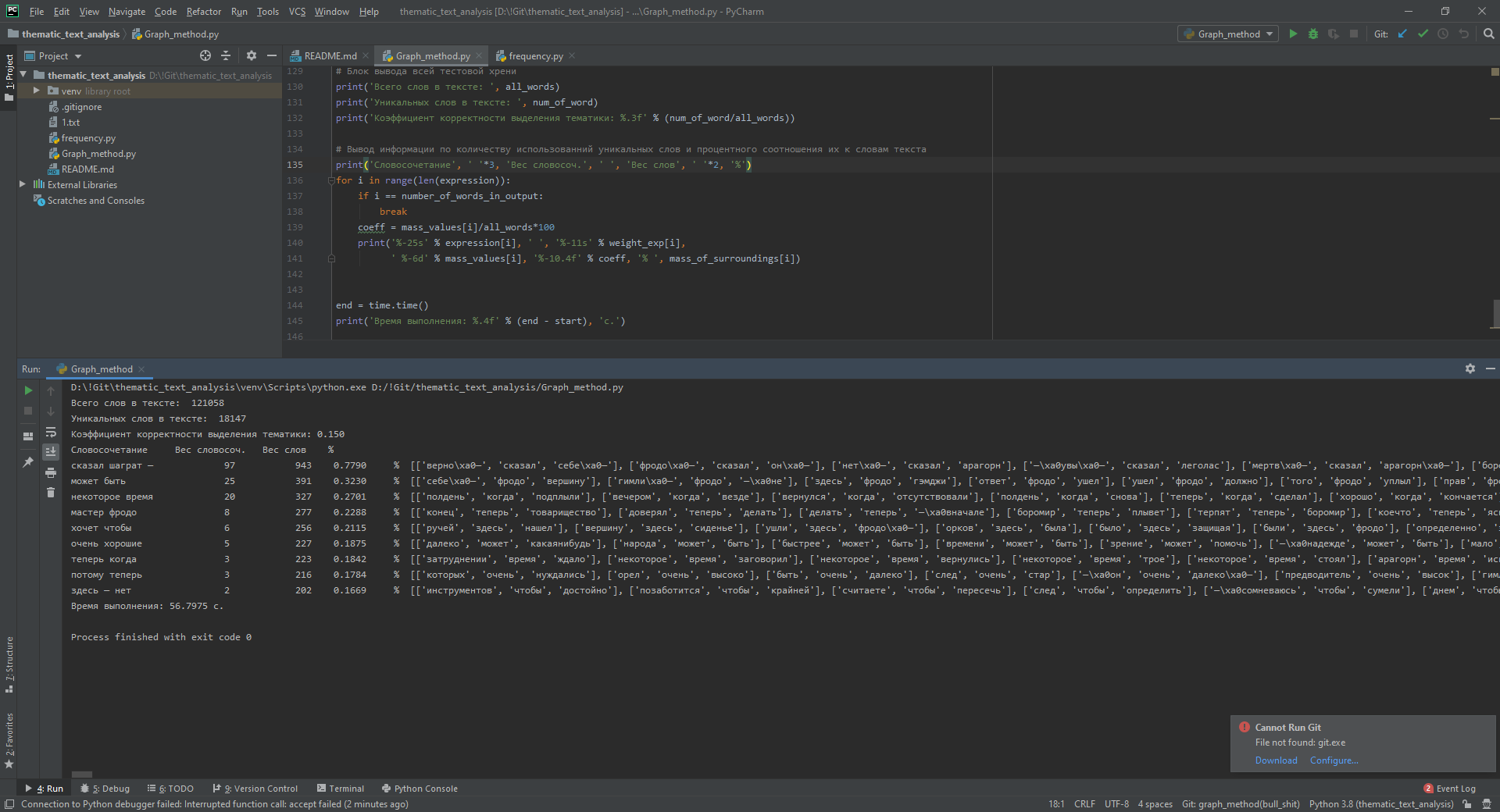


Рисунок 1.6 – Результат анализа графовым методом книги 2

Книга 3 – Властелин колец: Возвращение короля

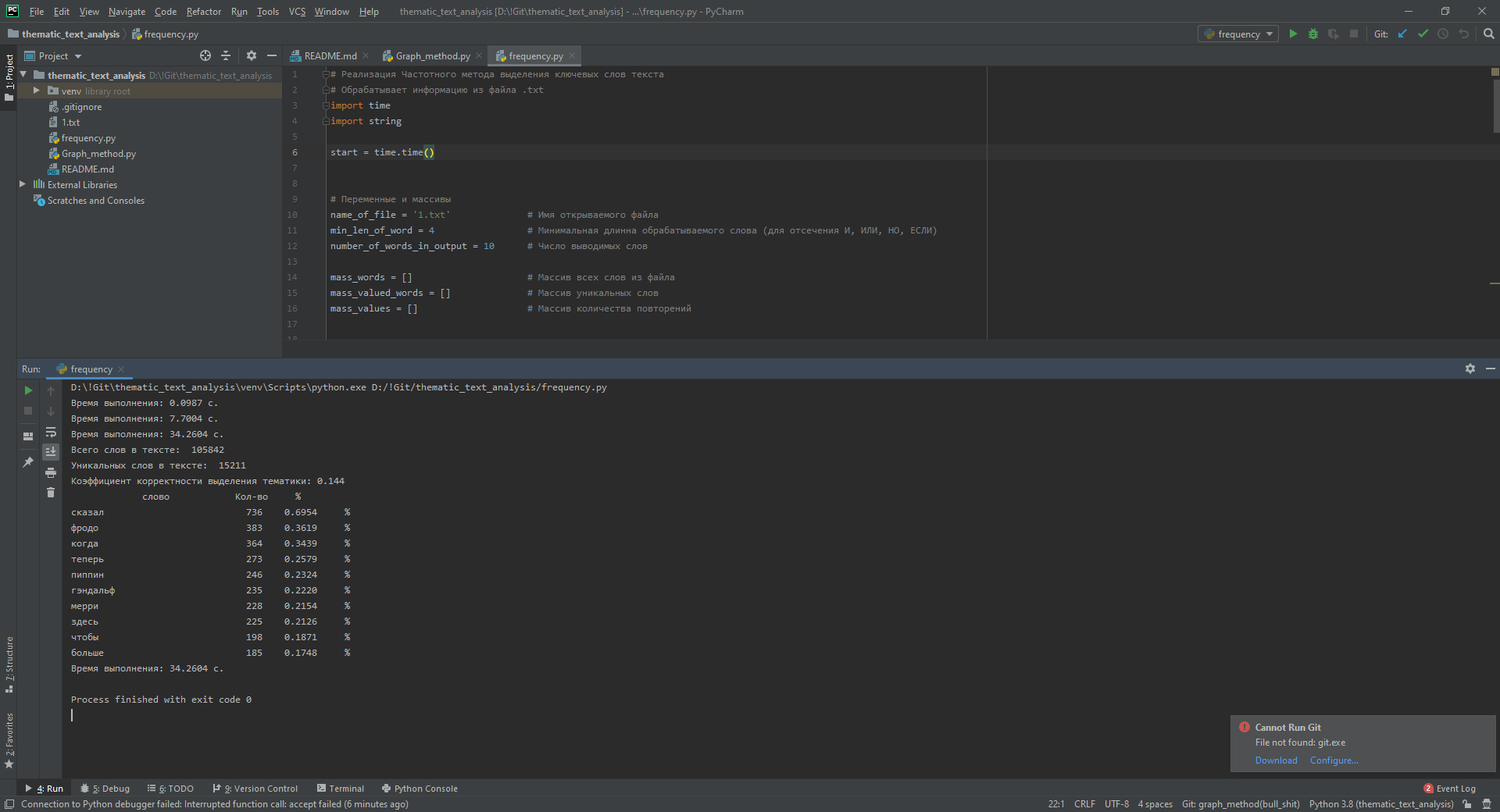


Рисунок 1.7 – Результат анализа частотным методом книги 3

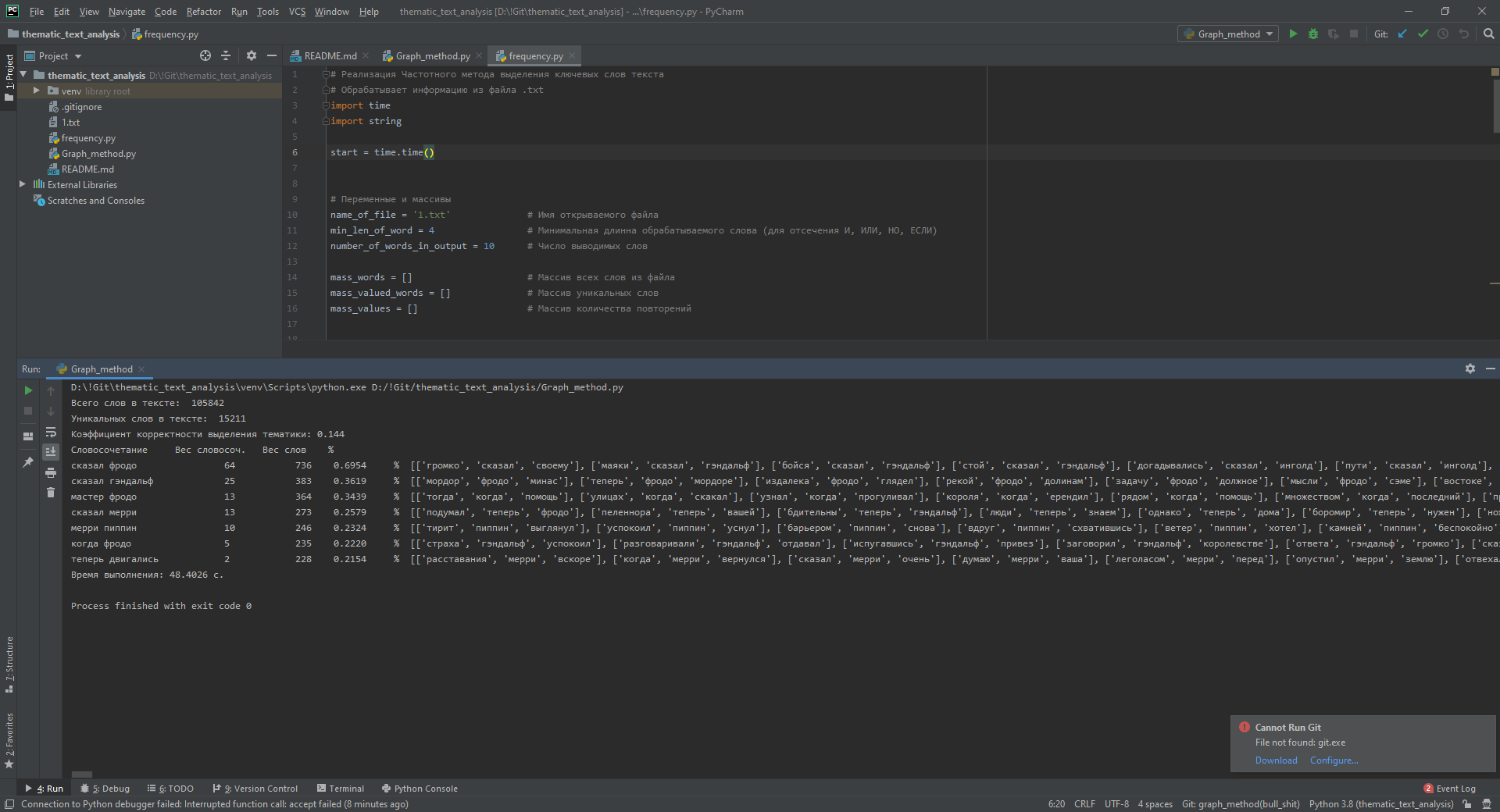


Рисунок 1.8 – Результат анализа графовым методом книги 3

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.

## 

## 2.1 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.

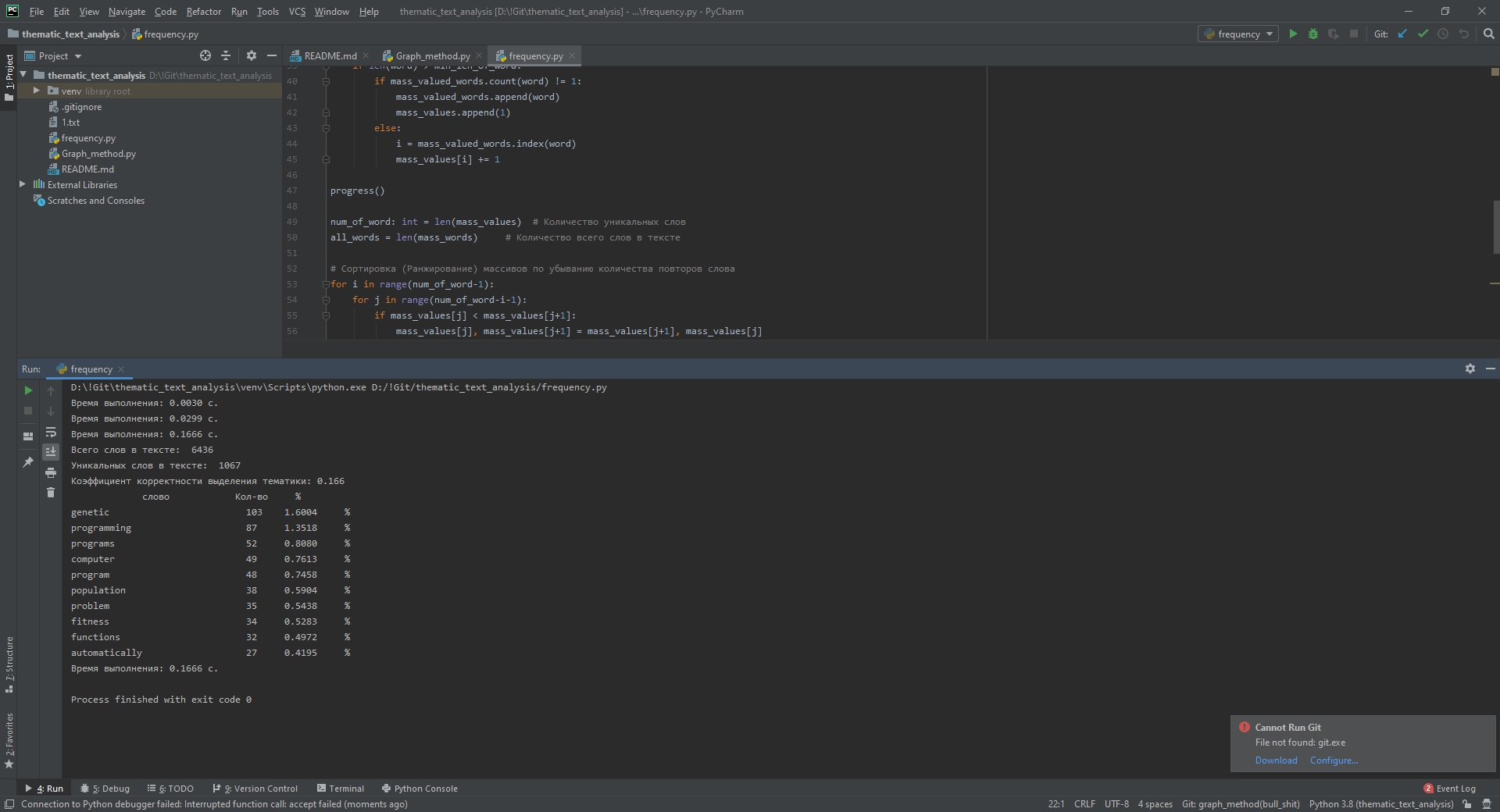


Рисунок 1.9 – Результат анализа статьи Genetic Programming

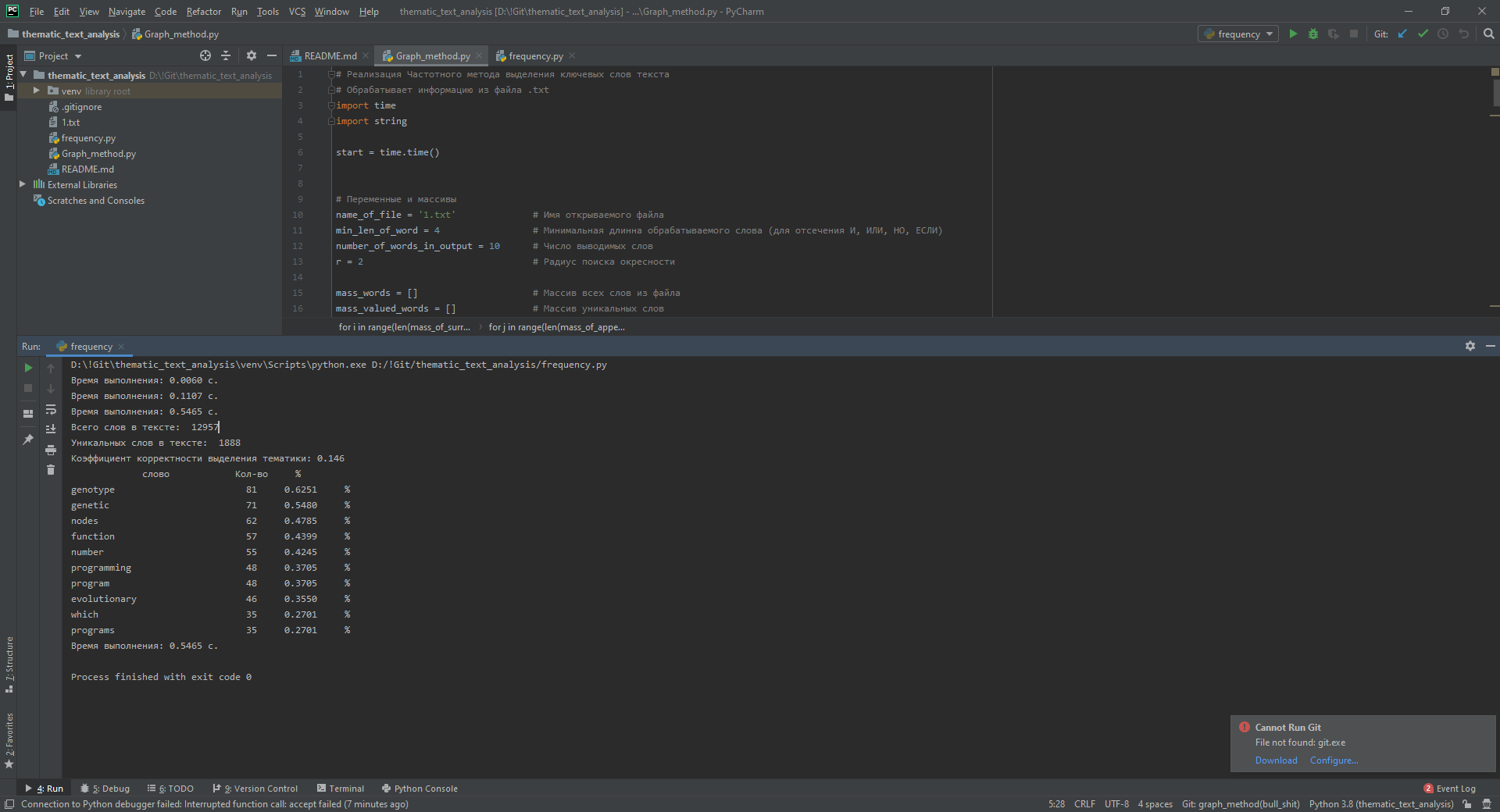


Рисунок 1.10 – Результат анализа учебника Cartesian Genetic Programming

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

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.

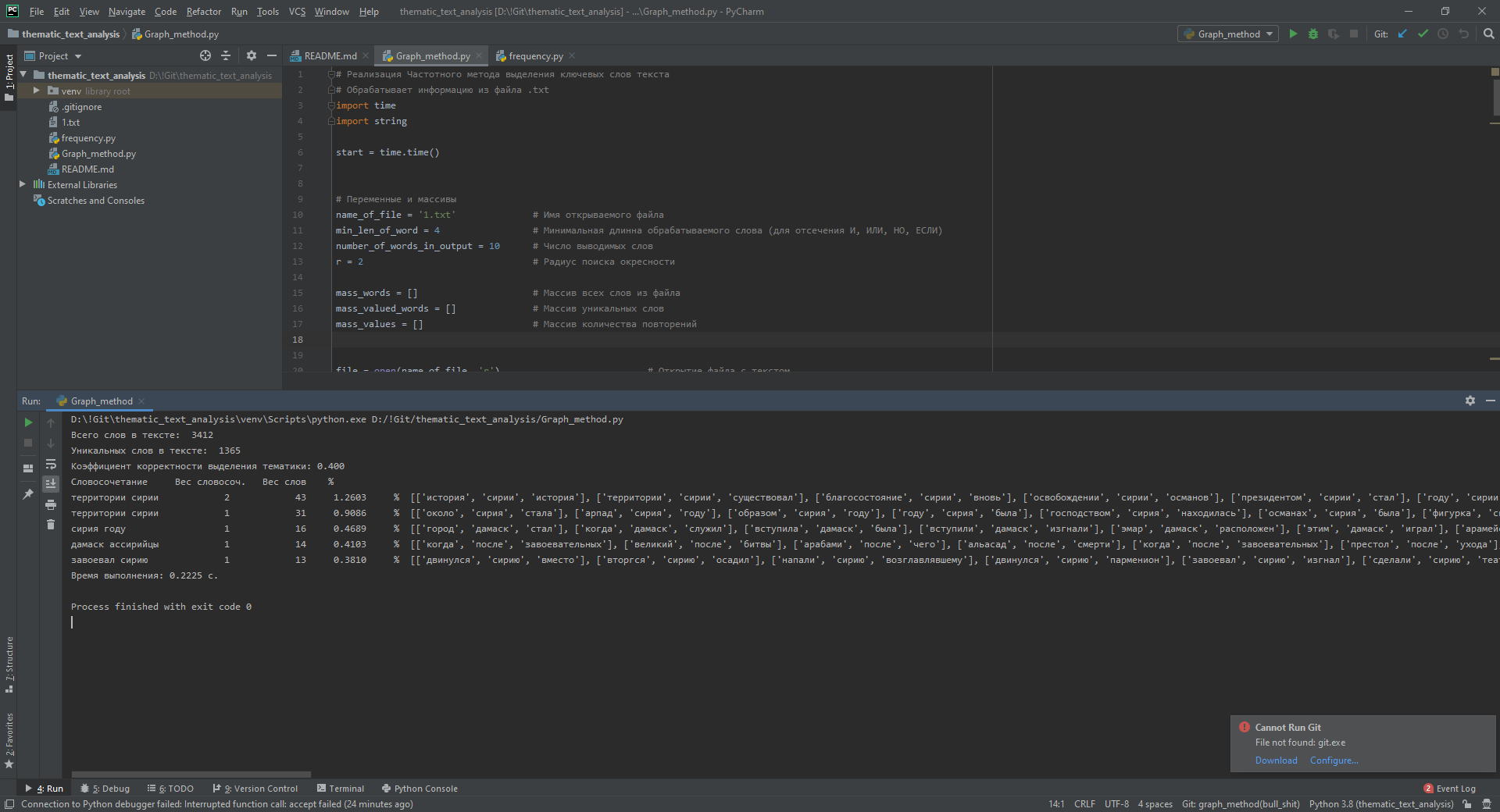


Рисунок 1.11 – Результат анализа графовым методом статьи википедии

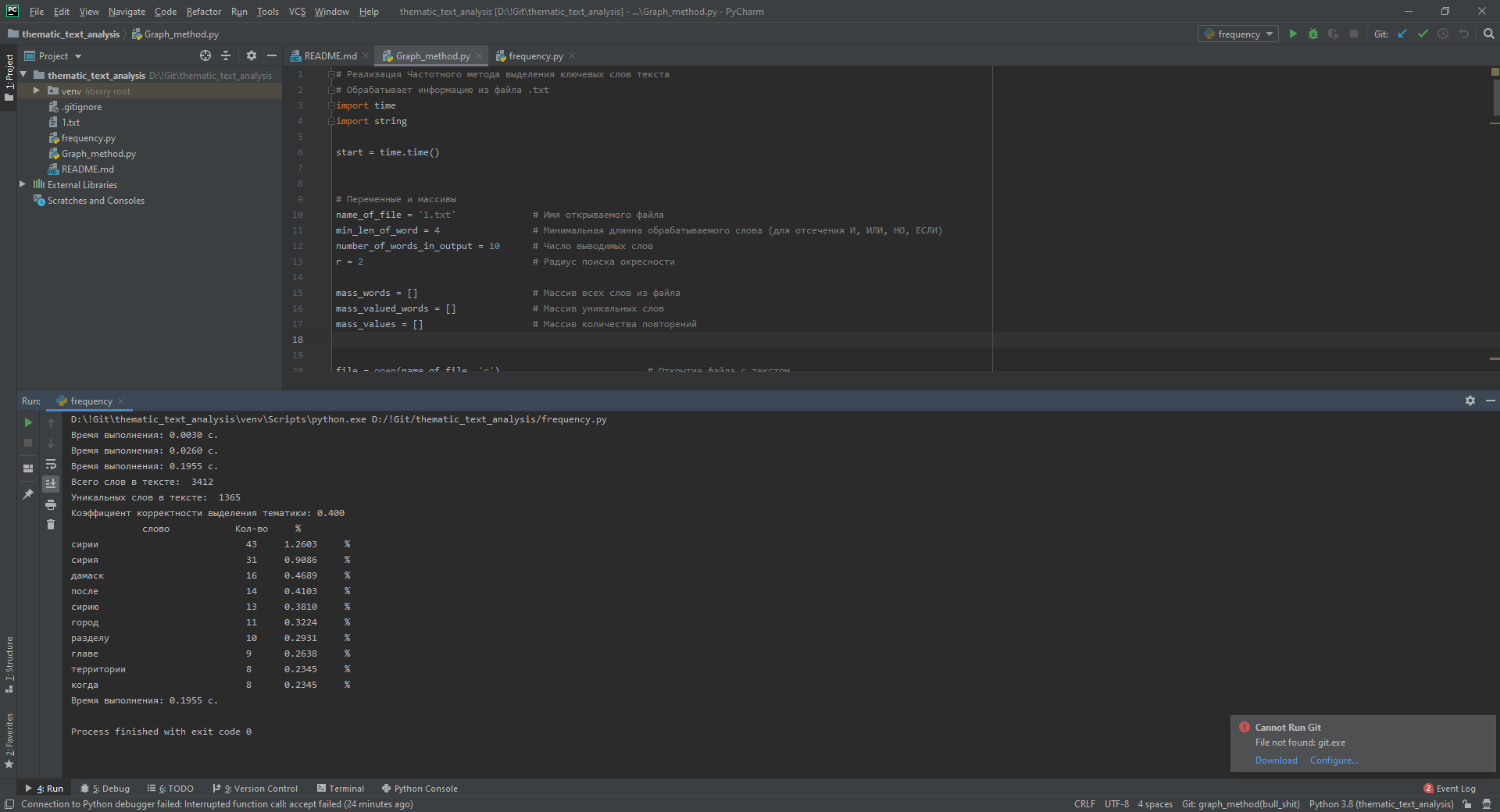


Рисунок 1.12 – Результат анализа частотным методом статьи википедии

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

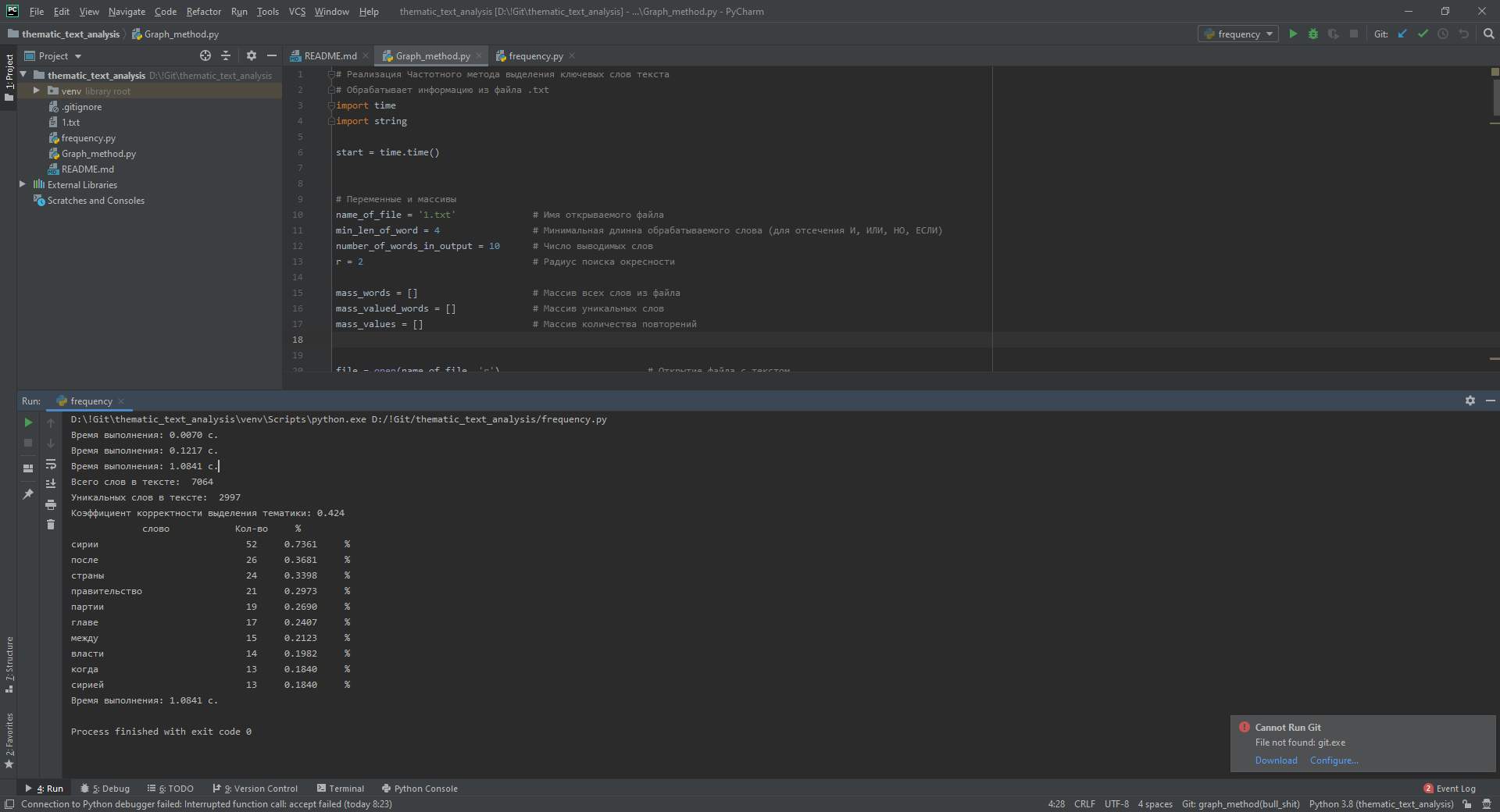


Рисунок 1.13 – Результат анализа частотным методом текста из энциклопедии Кольера

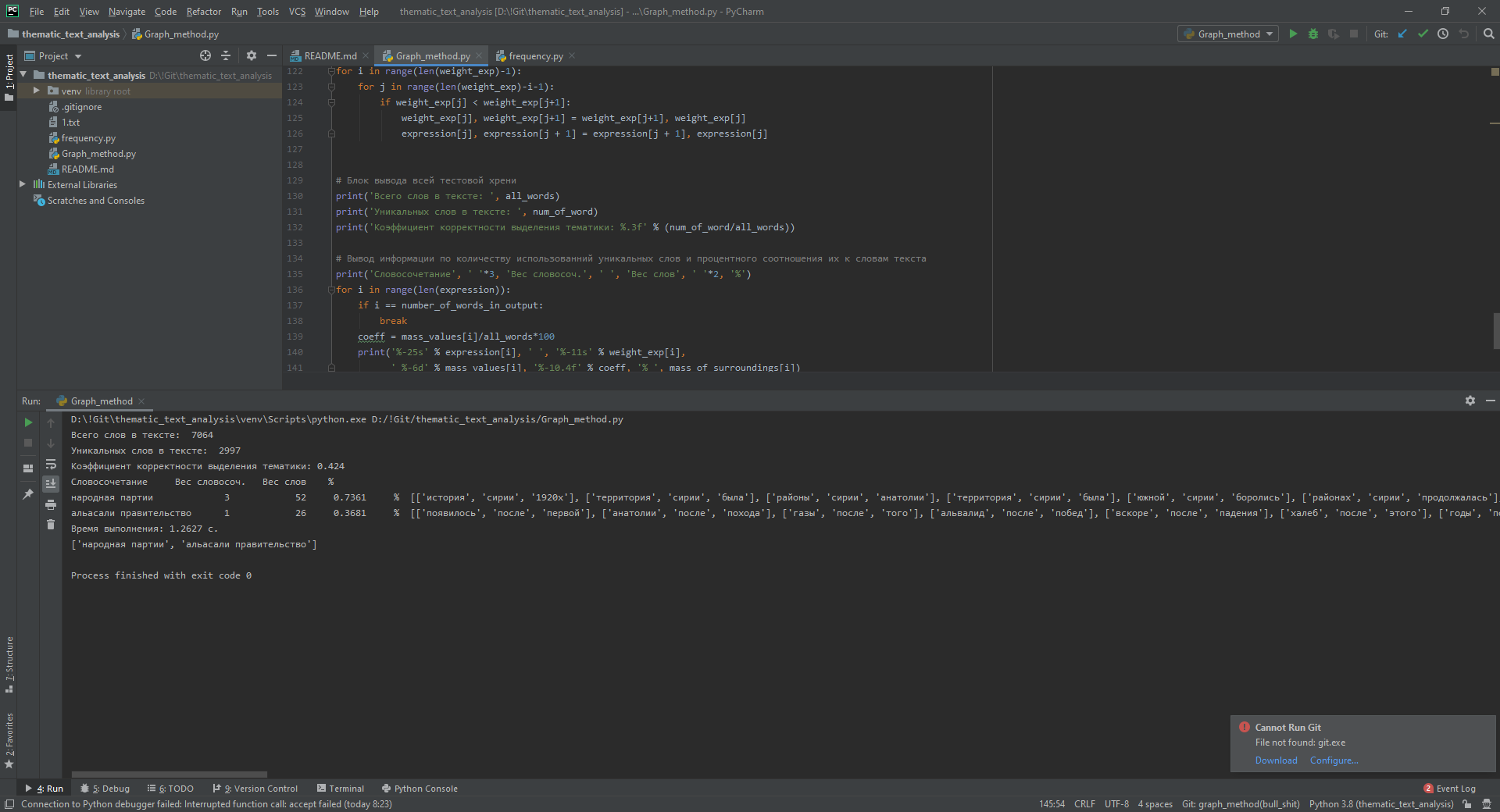


Рисунок 1.14 – Результат анализа графовым методом текста из энциклопедии Кольера

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.