А.Д. ВАРЛАМОВ, Ю.С. ФОМИЧЕВ

Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени А.Г. и Н.Г. Столетовых», г. Муром

Визуализация активности героев для критического анализа литературных текстов

УДК 004.912

В статье разработан алгоритм визуализации активности героев в литературных текстах. Предложено использование этого алгоритма для быстрого поиска в литературных текстах сюжетных событий, сцен, актов, опираясь на участников этих событий – персонажей произведения.

Введение

В настоящее время автоматическая обработка текста – это стремительно развивающаяся область научных исследований, которая направлена на разработку алгоритмов и методов обработки больших объемов неструктурированной информации. Основные трудности, возникающие при решении задач обработки текстов, связаны с необходимостью работы с неструктурированными данными [3,4]. Поэтому еще не удалось достичь универсального подхода к построению алгоритма, так как каждый конкретный алгоритм определяется строем языка.

На практике анализом больших текстов занимаются литературные критики и рецензенты. Их сфера деятельности представляют собой область, находящуюся на грани искусства и науки о литературе. Критики – это люди, которые занимаются оценкой и истолкованием произведений с позиции современности, включая точку зрения насущных проблем духовной и общественной жизни, а также своих личных взглядов и убеждений, утверждают и выявляют творческие принципы различных литературных направлений, оказывают большое влияние на литературное

развитие, а также воздействуют на формирование определенного общественного сознания, опираясь на историю, философию, эстетику и непосредственно саму литературу.

Статьи, написанные русскими и зарубежными критиками, продолжают и сегодня оказывать большое влияние на духовную и нравственную жизнь общества [2]. Они не случайно многие годы входят в состав обязательной программы школьного образования нашей страны. В процессе анализа произведений критикам необходимо не только прочесть большой объем текста, но и многократно возвращаться к его эпизодам для более детального анализа, оценивать своего рода статистику произведения и так далее. Поэтому много времени ими тратится на поиск интересующих их частей произведения.

Современным критикам можно существенно упростить задачу исследования произведений с помощью информационных систем, быстро и эффективно выполняющих поиск по тексту [1, 5], что позволит, в общем, исключить рутинный поиск в тексте при работе рецензента, чтобы сосредоточиться на сути анализа и не отвлекаться по мелочам.

Постановка задачи

Классический шаблонный поиск по ключевым словам или фразам малоэффективен в виду наличия склонений ключевых слов, множества ложных результатов поиска ввиду большого объема произведения, отсутствия целостной картины поиска. Поэтому было принято решение о создании информационной системы автоматизации поиска героев в литературных произведениях с целью их критического анализа. Система должна выполнять обнаружение героев в тексте, вычислять частоту их встречаемости и выводить график по каждому из них с целью визуализации того, в какой части произведения фигурировал тот или иной персонаж, а также агрегировать данныепо нескольким героям для анализа совместной встречаемости персонажей на протяжении всего текста.

Разработка алгоритма

Рассмотрим особенности данных, с которыми предстоит работать. В тексте имена героев пишутся с заглавной буквы.

Помимо имен действующих лиц с большой буквы пишутся первые слова в предложенияхи имена собственные. Необходимо учитывать, что имена героев могут состоять из нескольких идущих подряд слов, начинающихся с заглавной буквы.

Первым этапом при выполнении автоматического анализа произведения предлагается рассчитать, сколько раз в тексте встретилось каждое слово, напечатанное с заглавной буквы. Тем самым можно получить довольно объемный перечень слов. Среди них будут повторяющиеся однокоренные слова с разными окончаниями и имена собственные, которые тоже начинаются с заглавных букв. Большинство этих слов встречаются в тексте лишь несколько раз. Зная это, можно исключить их из общего списка, выполнив пороговую проверку на количество повторов. После этого будет получен предварительный список с героями произведения, но частоты упоминания героев будут не точными. Для этого следует выполнить еще один цикл по всему тексту с поиском именно корней имен каждого героя. Таким образом будут получены корректные данные по количеству встречаемости того или иного персонажа в произведении.

Для визуализации периодичности появления героев необходимо знать, в каком именно части текста они встречались. Для этого литературный текст разбивается на блоки.

Пусть $b = 1, \ldots, B$ – номер блока текста произведения, B – количество блоков текста во всем произведении;

$h = 1, \ldots, H$ – номер литературного героя в текста произведения, H – количество персонажей во всем произведении;

$K_h[b]$ – количество употреблений слова, обозначающего героя hв блоке текста b.

Фактически K_h представляет собой массив, удобный для визуализации в виде графика - диаграммы. Всплески на графике должны указывать на активное упоминание персонажа в данной части произведения. Для анализа степени совместного появления нескольких героев предлагаем использовать среднее гармоническое частот встречаемости отдельных лиц:

$$K_{h1,\ldots,hn}[b] = \frac{n}{\sum_{i=1}^{n} \frac{1}{K_{hi}[b]}} \qquad (1)$$

На основе этих данных был разработан алгоритм визуализации активности героев в литературных текстах, который состоит из следующей последовательности операций:

- Предварительная обработка произведения: разбиение текстана предложения, удаление лишних символов в начале и конце каждого предложения;
- Поиск слов, начинающихся с заглавной буквы, кроме тех, что стоят в начале предложений, и оценка частот их встречаемости в тексте;
- Отсеивание редких слов из результатов поиска;
- Объединение однокоренных слов в списке найденных, и суммирование их частот;
- Пересчет частот слов из списка найденных с учетом первых слов в предложениях;
- Визуализация результатов поиска: построение графика частоты упоминания выбранного героя (или группы героев) во временном срезе.

Результаты исследований

Тестирование проводилось по двум литературным произведениям. Первым анализируемым произведением стала известная сказка Алексея Николаевича Толстого «Золотой ключик, или Приключения Буратино». В таблице 1 приведены частоты появления героев сказки.

Таблица 1

Персонаж	**Количество упоминаний**	**Персонаж**	**Количество упоминаний**
Буратино	339	Базилио	19
Карло	111	Тортила	18
Карабас Барабас	96	Жаба	9
Пьеро	86	Арлекин	9
Мальвина	80	Говорящий Сверчок	7
Артемон	67	Сова	6
Дуремар	38	Богомол	4
Лиса Алиса	29	Сизый Нос	3
Джузеппе	22	Сплюшка	3

Отметим, что все ключевые герои сказки присутствуют в списке. В качестве тестового эксперимента отыщем фрагмент произведения, в котором Карабас Барабас обсуждает с Дуремаром тайну, которую рассказала черепаха Тортила.

Частоты упоминания этих трех персонажей показаны на рисунках 1-3.

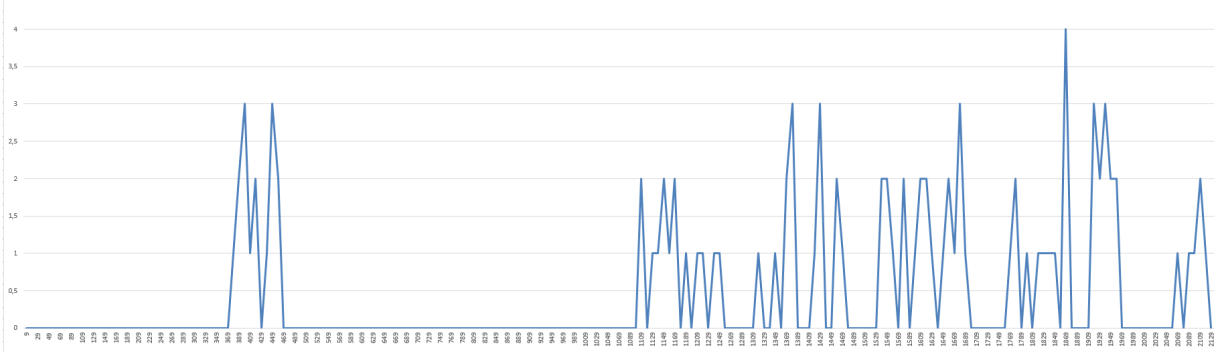

Рис. 1.Частота упоминания героя «Карабас Барабас».

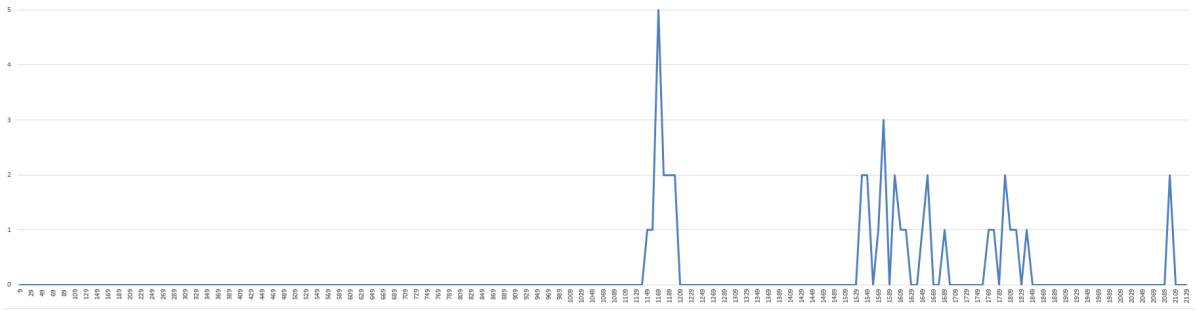

Рис.2. Частота упоминания героя «Дуремар».

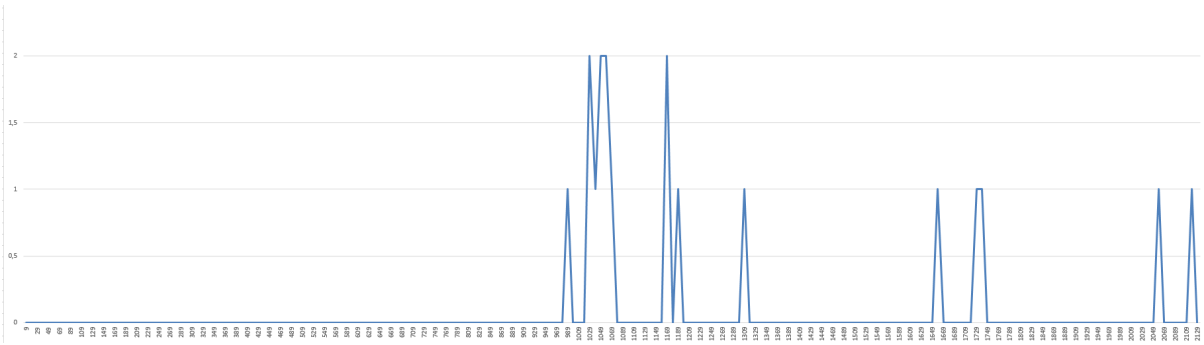

Рис. 3. Частота упоминания героя «Тортила».

Применив формулу (1), получим график пересечений героев в блоках текста, который отображен на рисунке (4).

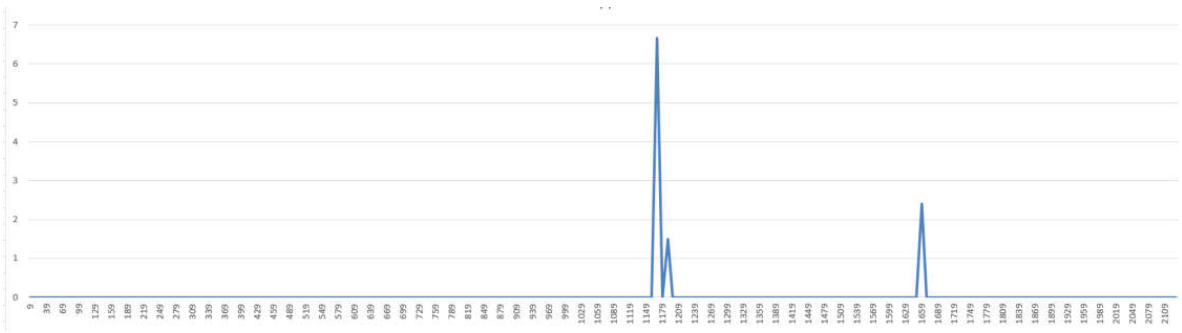

Рис.4. Совместная встречаемость героев «Карабас Барабас», «Тортила» и «Дуремар».

Самый большой всплеск на графике рисунка 4 отражает разговор Карабаса Барабаса с Дуремаром о тайне, которую рассказала черепаха Тортила:

> *...Синьор Карабас Барабас посопел трубкой и ответил: «Есть только одна тайна на свете, которую я хочу знать. На всё остальное я плевал и чихал». «Синьор, – опять сказал Дуремар, – я знаю великую тайну, её сообщила мне черепаха Тортила».*
> *При этих словах Карабас Барабас выпучил глаза, вскочил, запутался в бороде, полетел прямо на испуганного Дуремара, прижал его к животу и заревел, как бык: «Любезнейший Дуремар, драгоценнейший Дуремар, говори, говори скорее, что тебе сообщила черепаха Тортила!»*
> *Тогда Дуремар рассказал ему следующую историю: ...*

Вторым анализируемым произведением стал роман Михаила Афанасьевича Булгакова «Мастер и Маргарита», работа над которым началась в конце 1920-х годов и продолжалась вплоть до смерти писателя. Произведение обладает большим объемом, поэтому для исследования был взят не весь текст, а первые его четыре главы. В таблице 2 приведены частоты появления героев романа.

Таблица 2

Персонаж	**Количество упоминаний**	**Персонаж**	**Количество упоминаний**
Понтий Пилат	90	Крысобой	12
Берлиоз	78	Аннушка	11
Бездомный	26	Марк	8
Иван Николаевич	20	Михаил Александрович	7
Иешуа Га-Ноцри	18	Кант	5
Иосиф Каифа	17	Гестас	4
Вар-раввана	14	Дисмас	4

Иисус	14		

Для начала анализа были выбраны два персонажа: «Берлиоз», «Бездомный. На рисунках 5-6 визуализированы частоты их появлений в тексте произведения.

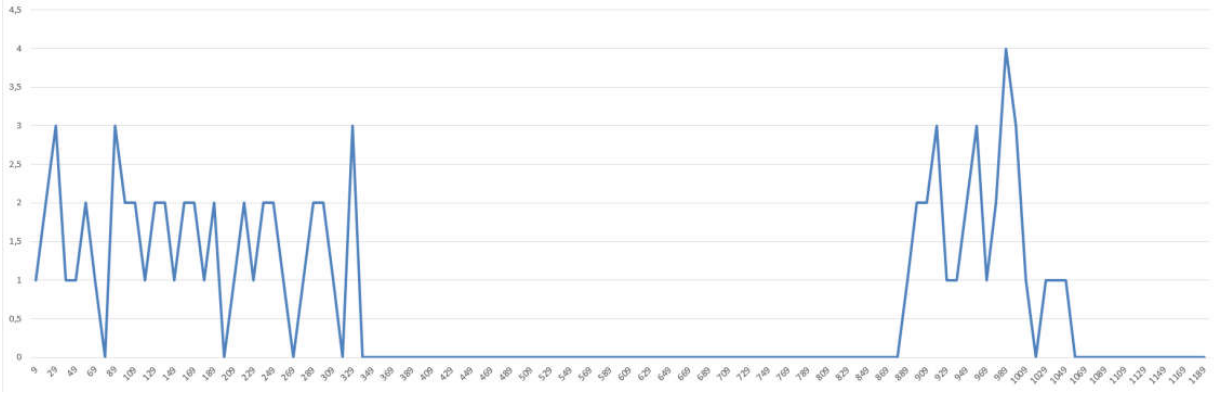

Рис. 5. Частота упоминания героя «Берлиоз».

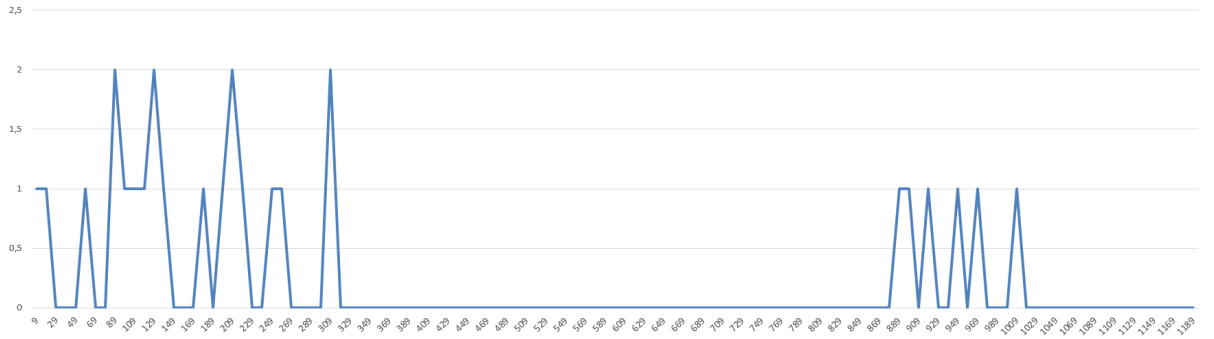

Рис. 6. Частота упоминания героя «Бездомный».

Если визуально разделить график рисунка 5 на четыре части, то можно сделать вывод, что герой «Берлиоз» фигурирует в первой и в начале четвертой главы произведения.

На графике рисунка 6 мы видим всплески приблизительно в тех же местах, следовательно, эти два персонажа фигурируют в тексте совместно. Анализируя роман, можно убедиться, что действительно оба персонажа встречаются в тексте вместе и они ведут тесный диалог в первой и начале четвертой главах произведения. Применив формулу (1), построим график совместного присутствия этих героев, который отображен на рисунке 7.

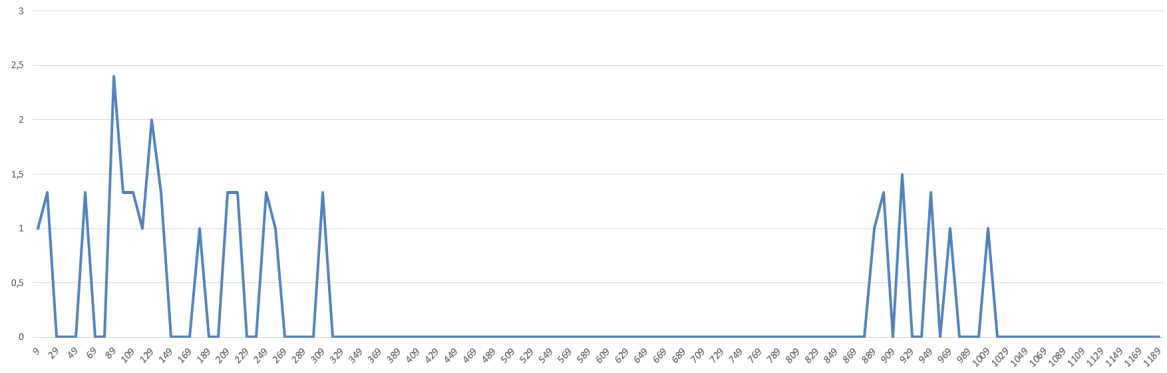

Рис.7. График встречаемости героев «Бездомный» и «Берлиоз».

Самый большой всплеск отражает реакцию Бездомного и Берлиоза на появление иностранца на аллее, который, пройдя мимо, сел на соседнюю скамейку и рассматривал все вокруг, а они продолжили разговор, который заинтересовал гостя, и он направился к ним:

«Немец», – подумал Берлиоз.
«Англичанин, – подумал Бездомный, – ишь, и не жарко ему в перчатках».
А иностранец окинул взглядом высокие дома, квадратом окаймлявшие пруд, причем заметно стало, что видит это место он впервые и что оно его заинтересовало.
Он остановил свой взор на верхних этажах, ослепительно отражающих в стеклах изломанное и навсегда уходящее от Михаила Александровича солнце, затем перевел его вниз, где стекла начали предвечерне темнеть, чему-то снисходительно усмехнулся, прищурился, руки положил на набалдашник, а подбородок на руки.
– Ты, Иван, – говорил Берлиоз, – очень хорошо и сатирически изобразил, например, рождение Иисуса, сына божия, но соль-то в том, что еще до Иисуса родился еще ряд сынов божиих, как, скажем, фригийский Аттис, коротко же говоря, ни один из них не рождался и никого не было, в том числе и Иисуса, и необходимо, чтобы ты, вместо рождения и, скажем, прихода волхвов, описал нелепые слухи об этом рождении... А то выходит по твоему рассказу, что он действительно родился!..
Тут Бездомный сделал попытку прекратить замучившую его икоту, задержав дыхание, отчего икнул мучительнее и громче, и в этот же момент Берлиоз прервал свою речь, потому что иностранец вдруг поднялся и направился к писателям.

В анализируемых четырех главах романа присутствует еще один интересный персонаж - «Понтий Пилат». График изменения частоты упоминания этого героя отражен на рисунке 8.

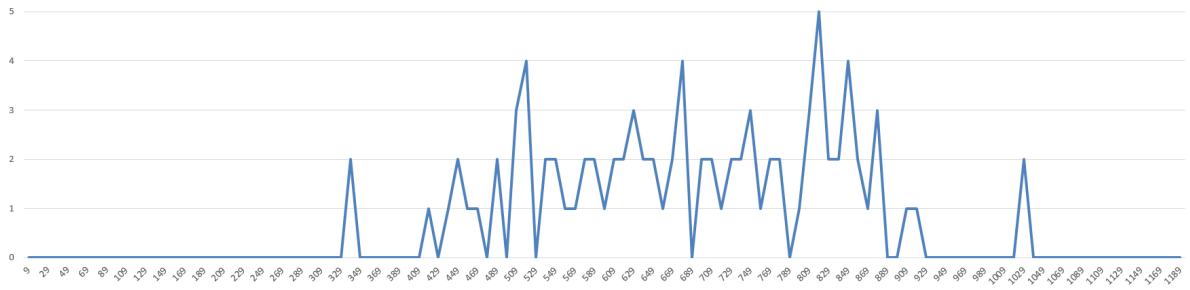

Рис. 8.Частота упоминания героя «Понтий Пилат».

График свидетельствует о том, что данный персонаж активно фигурирует во второй и третьей главах, и это действительно так. В этих главах «Иешуа» предстает перед судом прокуратора «Понтия Пилата», причем это происходит отдельно от предыдущих героев, об этом и свидетельствует следующий график:

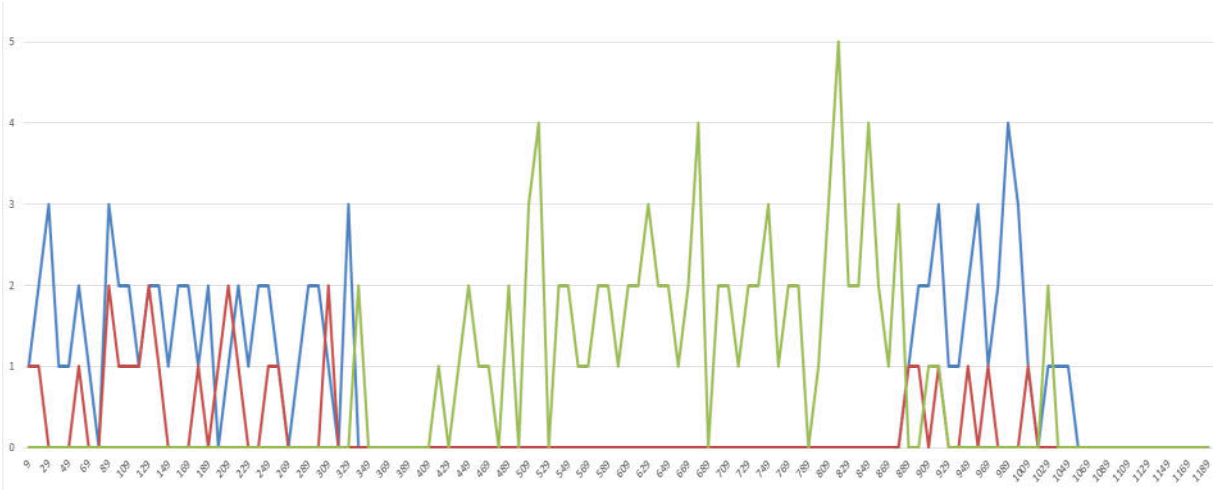

Рис. 9. График персонажей «Берлиоз», «Бездомный» и «Понтий Пилат».

Лишь в конце главы можно наблюдать пересечение героев: «Понтий Пилат», «Берлиоз» и «Бездомным» (рисунок 10).

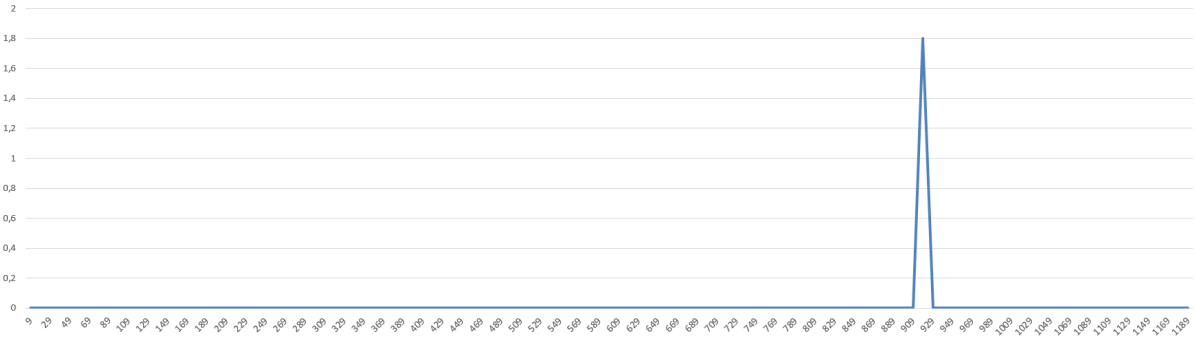

Рис. 10. Совместное упоминание персонажей «Берлиоз», «Бездомный» и «Понтий Пилат».

Они обсуждали рассказ иностранца о «Понтии Пилате» и «Иешуа» при этом несколько раз упомянули его имя. Ниже представлен фрагмент подтверждающий это.

> *Берлиоз тотчас сообразил, что следует делать. Откинувшись на спинку скамьи, он за спиною профессора замигал Бездомному, – не противоречь, мол, ему, – но растерявшийся поэт этих сигналов не понял.*
> *– Да, да, да, – возбужденно говорил Берлиоз, – впрочем, все это возможно! Даже очень возможно, и Понтий Пилат, и балкон, и тому подобное... А вы одни приехали или с супругой?*
> *– Один, один, я всегда один, – горько ответил профессор.*
> *– А где же ваши вещи, профессор? – вкрадчиво спрашивал Берлиоз, – в «Метрополе»? Вы где остановились?»*

Выводы

В ходе проделанной работы был достигнут результат в виде функционирующей программы, которая облегчит литературным критикам анализ произведений и позволит быстро и продуктивно, возвращаясь в ту или иную часть произведения, обращаясь к любому персонажу, а, при необходимости, сразу к нескольким, достигать более точных и конкретных результатов в своей работе, затрачивая на это в разы меньше времени, чем при классическом поиске в произведении.

Графики очень удобно и информативно отображают ситуацию в произведении и позволяют увидеть нюансы, которые не видны при классическом поиске по ключевым словам. Таким образом, исследователям литературных произведений не требуется много времени для многократного перечитывания одних и тех же его частей.

Литература

1. Алексеев А.А., Катасёв А.С., Кириллов А.Е., Кирпичников А.П.Классификация текстовых документов на основе технологии textmining// Вестник Казанского технологического университета, 2016, Т19, №18, с. 116 – 119.

2. Канунова Е.Е. Вопросы автоматизации музейного дела// Алгоритмы, методы и системы обработки данных. 2014. № 4 (29). С. 72-76.

3. Ленкин А.В., Баженов Р.И. Исследование систем для TextMining// Постулат. 2017. № 1 (15). С. 3.

4. Шарапова Е.В., Шарапов Р.В. Универсальная система проверки текстов на плагиат "автор.net"// Информатика и ее применения. 2012. Т. 6. № 3. С. 52-58.

5. Щербатов И.А., Беляев И.О. Многоагентная поисковая система: применение фибоначчиевх куч// Алгоритмы, методы и системы обработки данных. 2015. № 2 (31). С. 86-92.