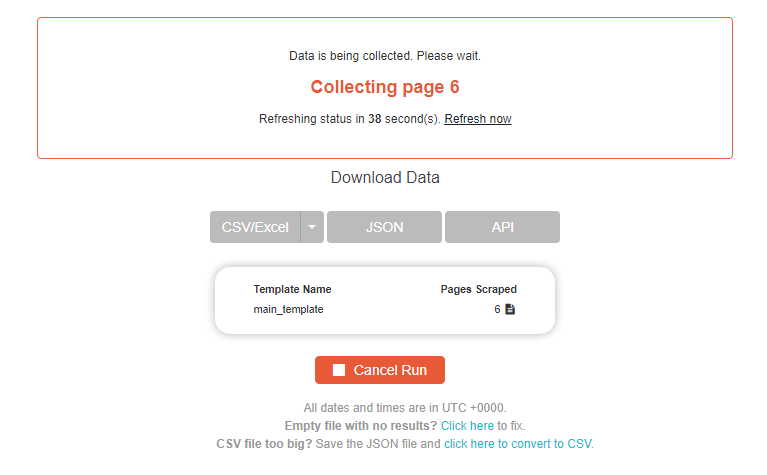
**Уровень 1**

**Задание 1**

**Номер 1**

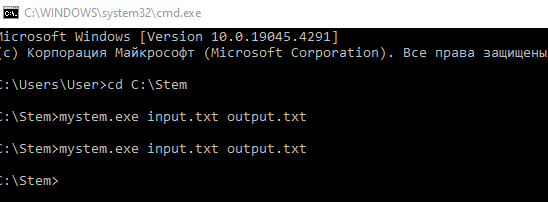
**Возьмем сайт РИА новости**

****

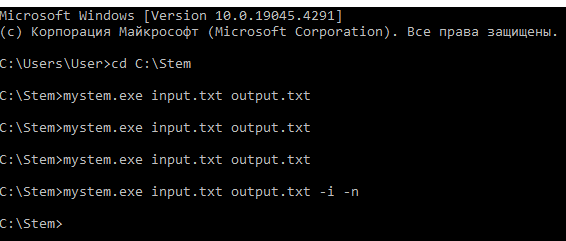
****

Номер 2

Stem

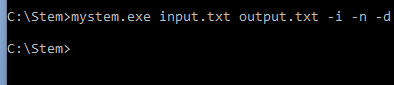


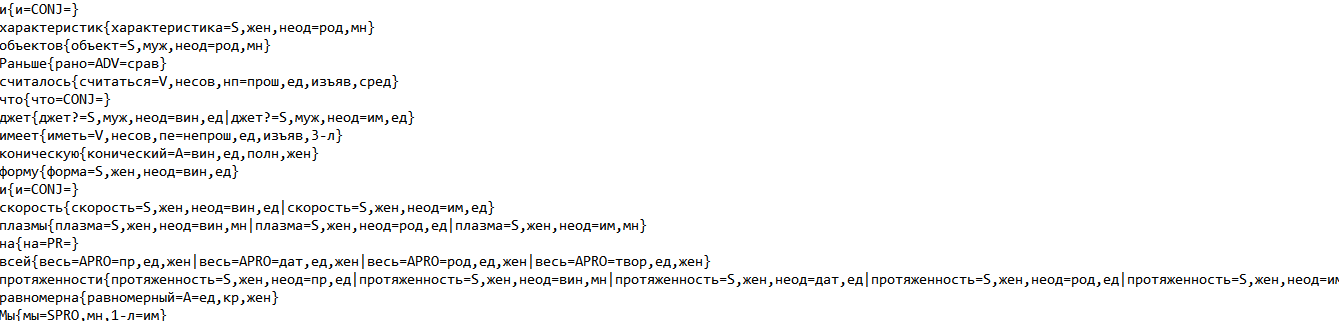
Каждое слово должно быть напечатано с новой строки.





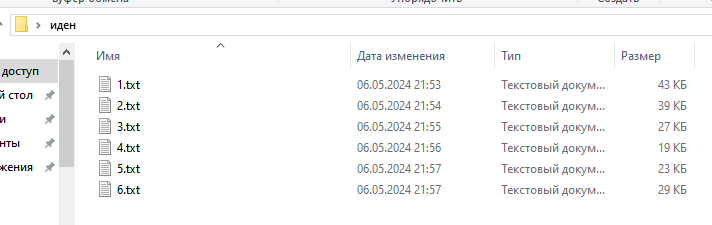
Чтобы получить более точные результаты.

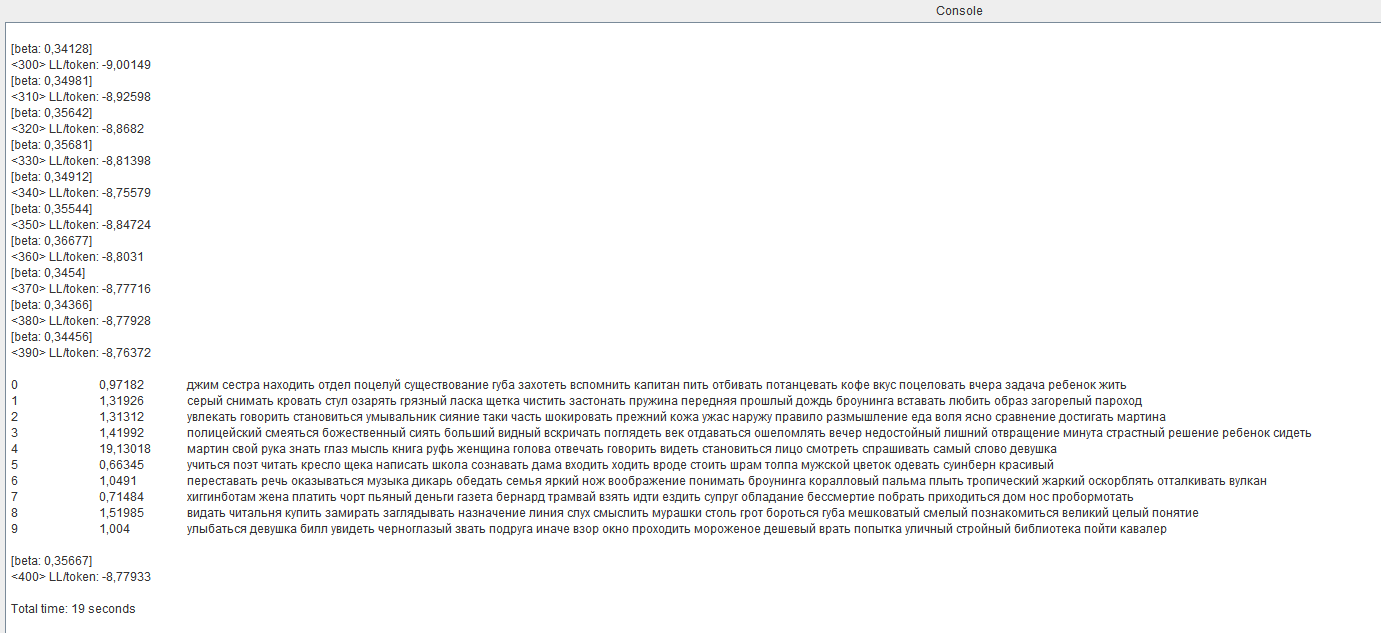


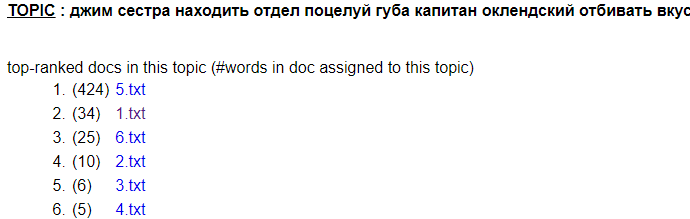


Номер 4

Берем первые 6 глав из Джек Лондон «Мартин Иден» и с помощью mystem проводим лемматизацию.



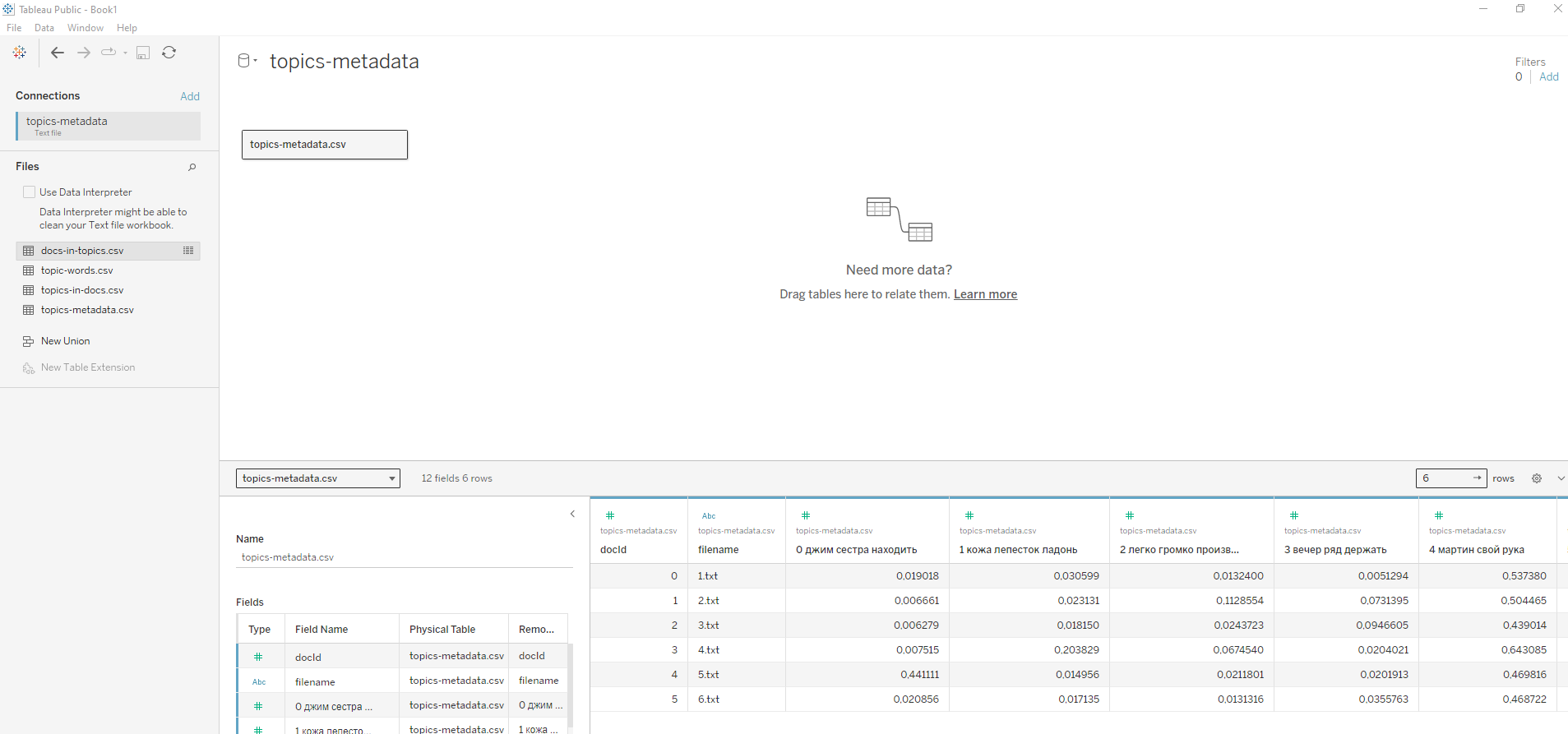


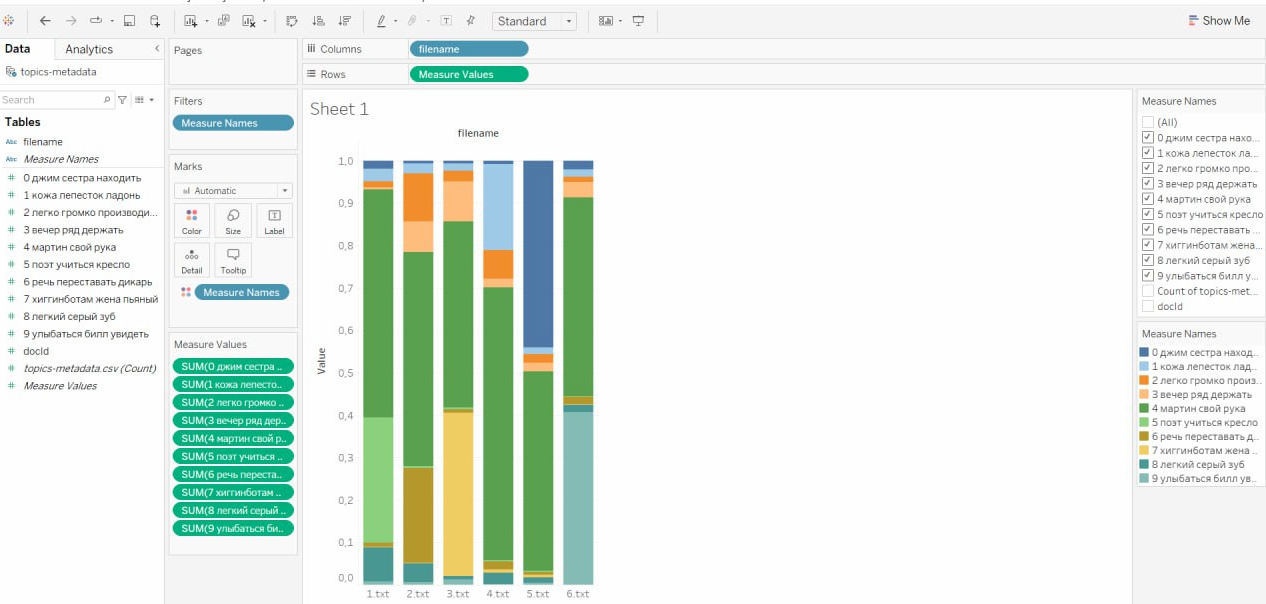


На протяжении первых 6 глав можно увидеть употребление имен некоторых второстепенных героев. В первой части шести глав идет описание жизни главного героя(что вокруг него, интерьер и т.д.), небольшой рассказ о путешествиях Мартина. В середине и в конце шести глав продолжается некоторое описание занятий Идена.

Номер 11

**Tableau**

****

****

**Уровень 2**

Номер 7

После загрузки стоп-слов

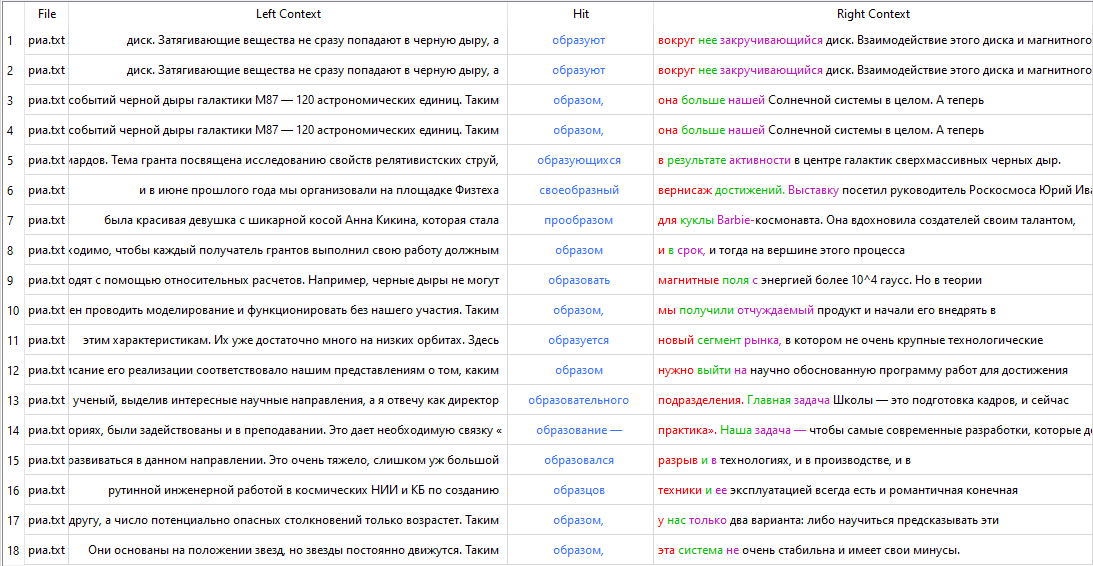
1.

****

**2. Контекстные употребления самых частотных слов**

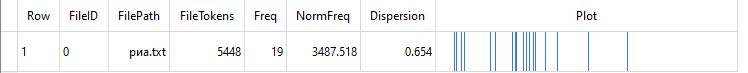
****

****

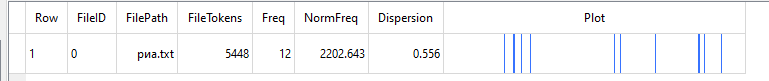
****

**3.Plot**

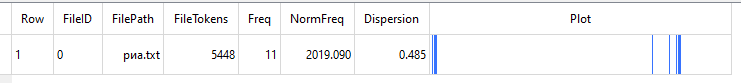
**космических**

****

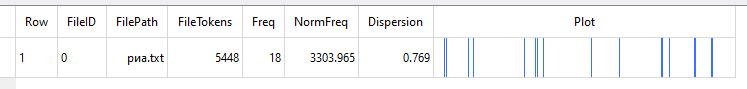
**например**

****

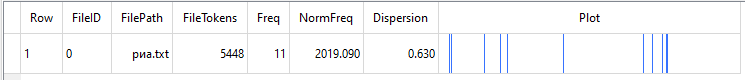
**джета**

****

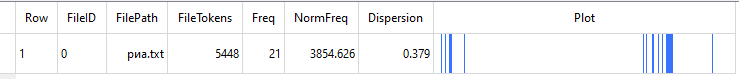
**Нашей**

****

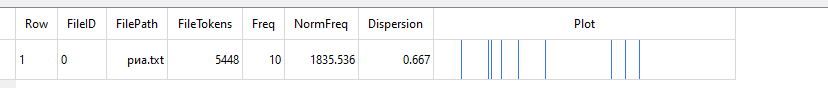
**Системы**

****

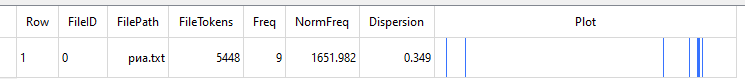
**Галактик**

****

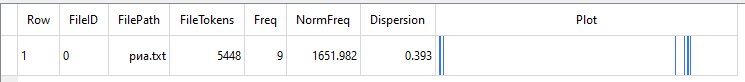
**Новых**

****

**Дыры**

****

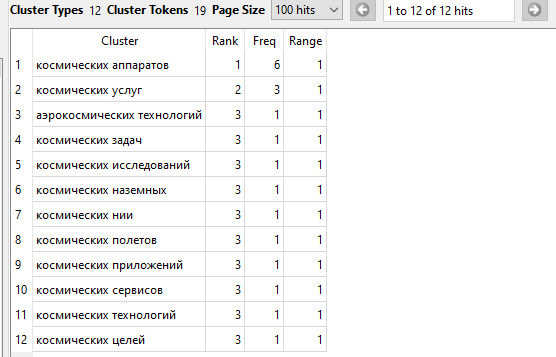
**Поля**

****

**Развития**

****

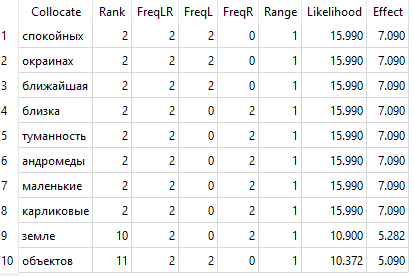
**4.Кластеры**

****

****

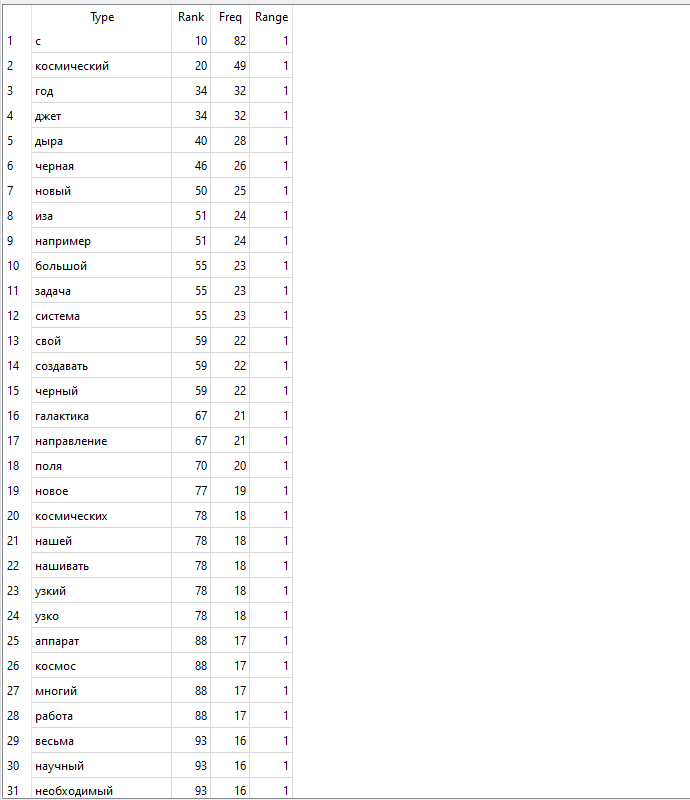
****

**5. топ 10 коллокаций со словом галактика**

****

**6.**

**6.1.**

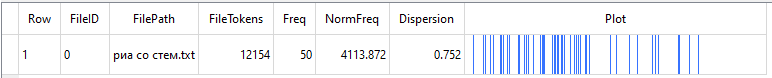
****

**6.2.**

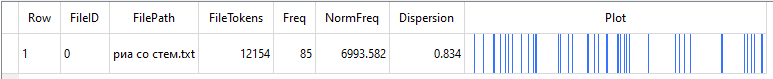
****

**6.3.**

**Космический**

****

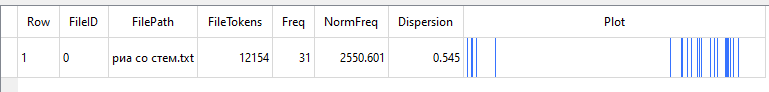
**Год**

****

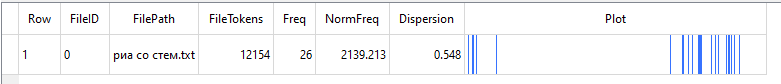
**Джет**

****

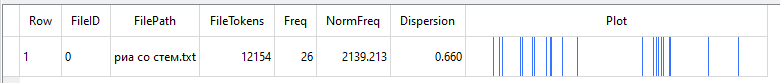
**Дыра**

****

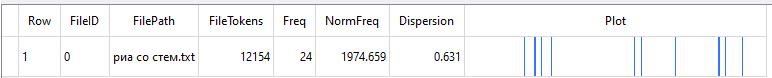
**Черная**

****

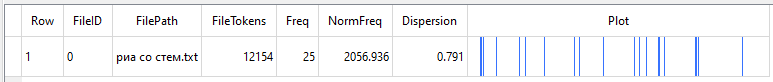
**Новый**

****

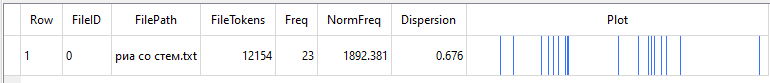
**Например**

****

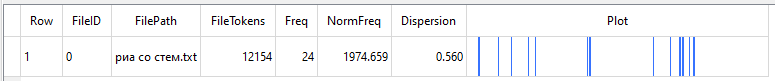
**Большой**

****

**Задача**

****

**Система**

****

**6.4**

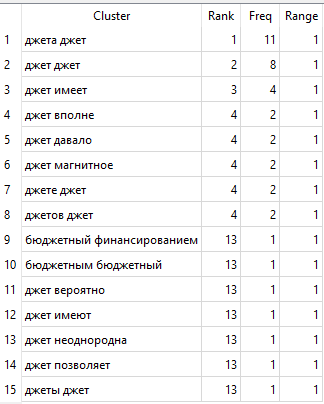
**Космический**

****

**Год**

****

**Джет**

****

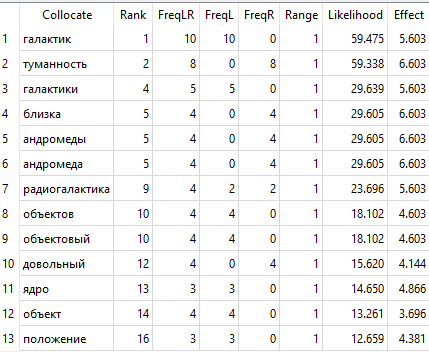
**Дыра**

****

**Черная**

****

**6.5. топ 10 коллокаций со словом галактика**

****

**Вывод: после лемматизации стало легче работать с корпусом, результаты поиска стали более точными и полными. Некоторые списки изменились(например, список частотности).В результате, мы видим, что с корпусом лучше и результативнее работать после лемматизации.**

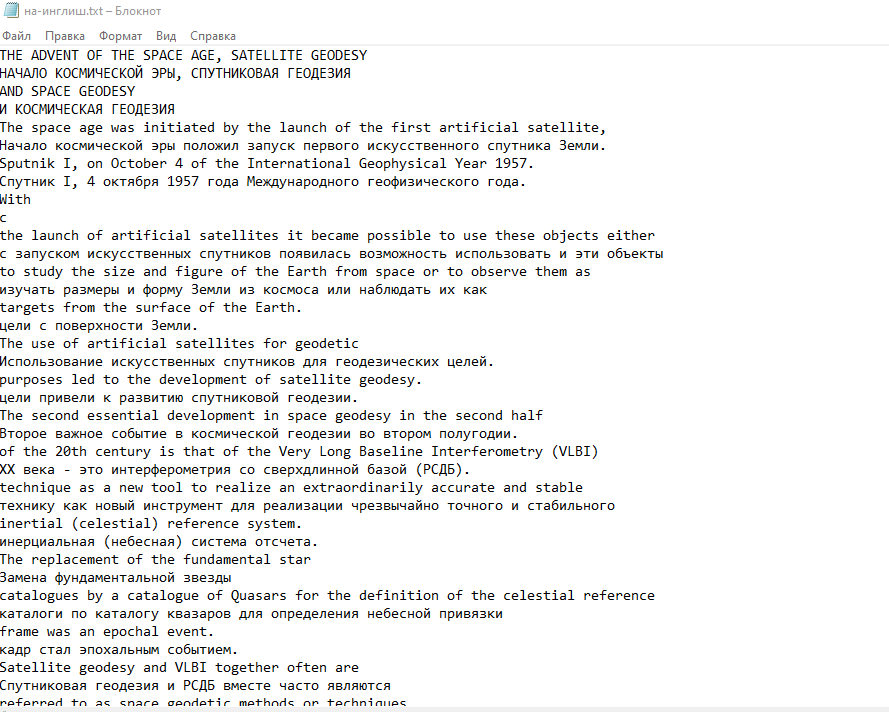
**Номер 8**

****

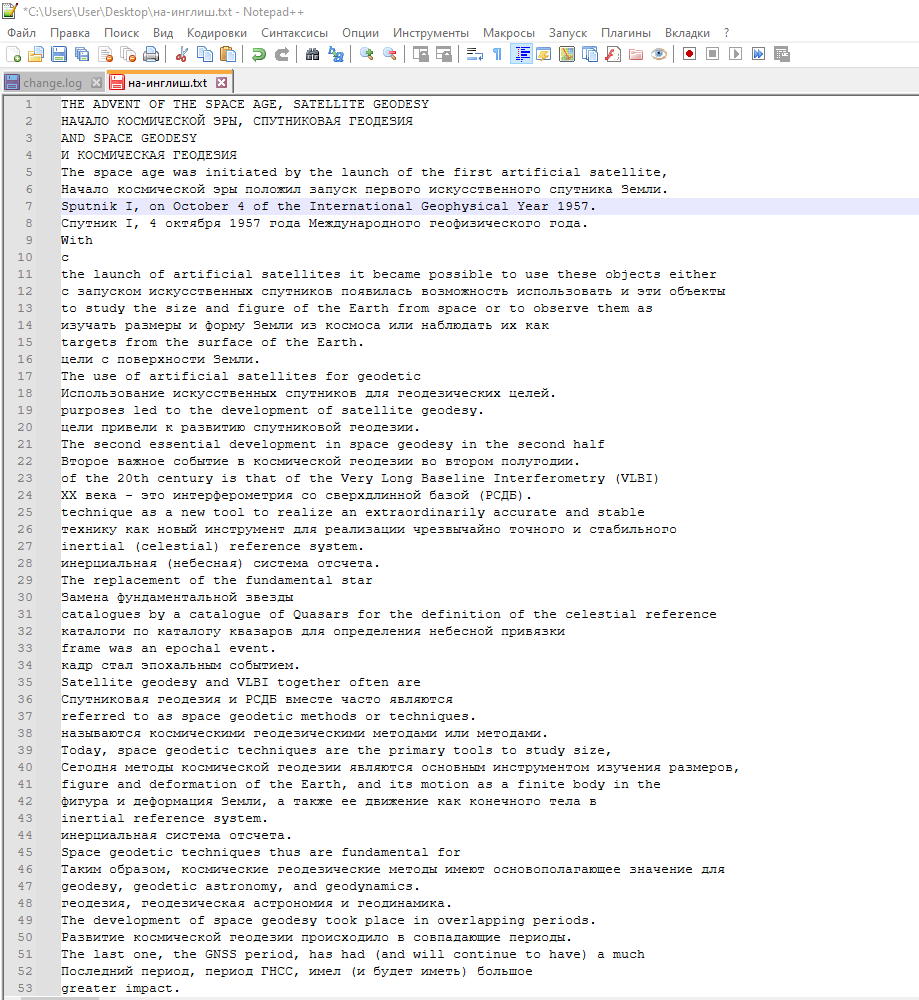
Экспортируем данные

****

Конвертируем

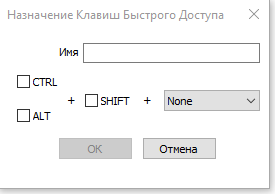
****

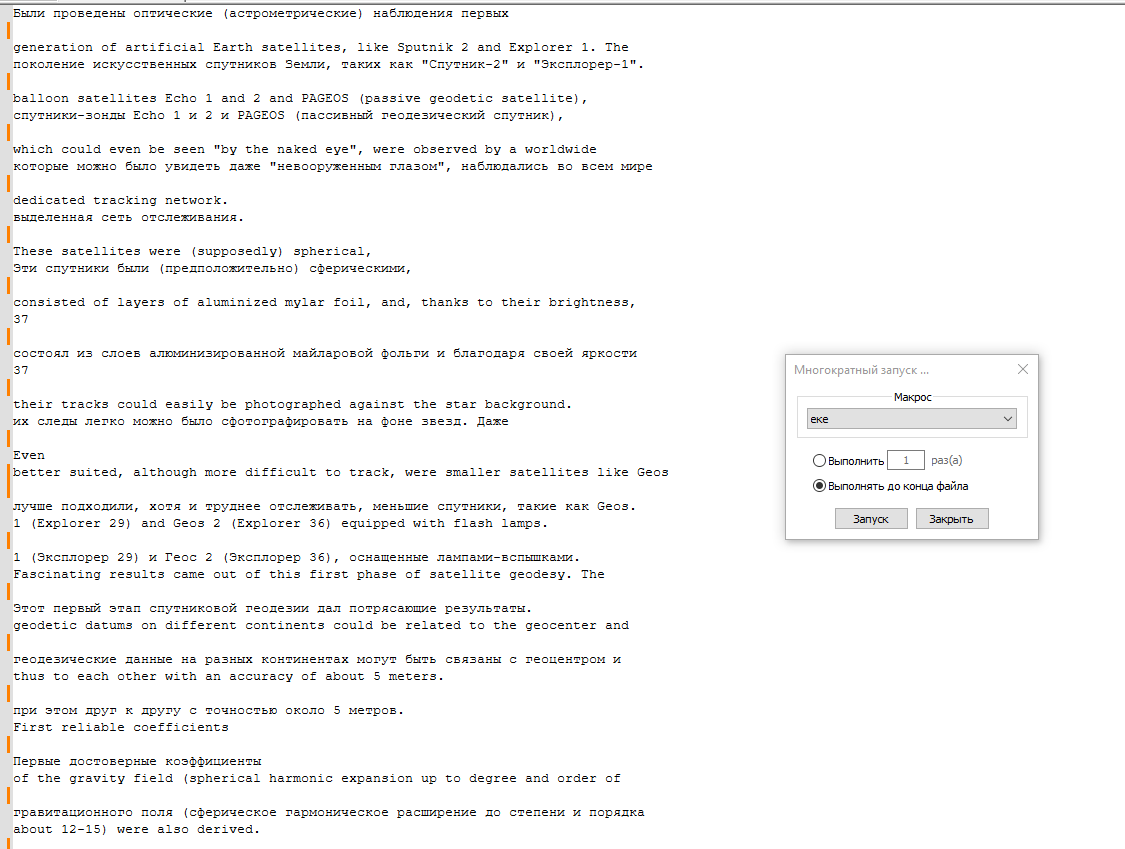
Уберем пустые строки

****

**Делаем макросы**

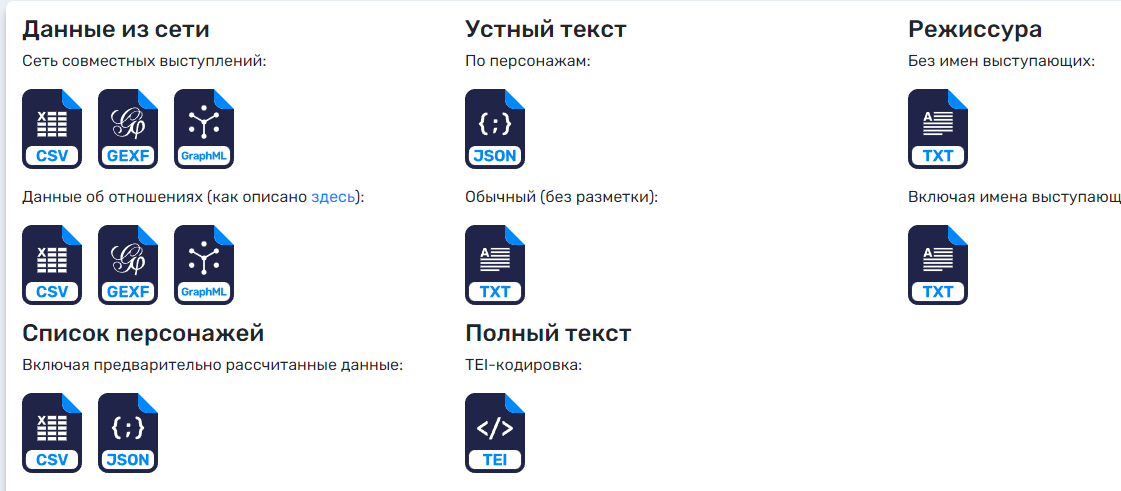
**Сохраняем**

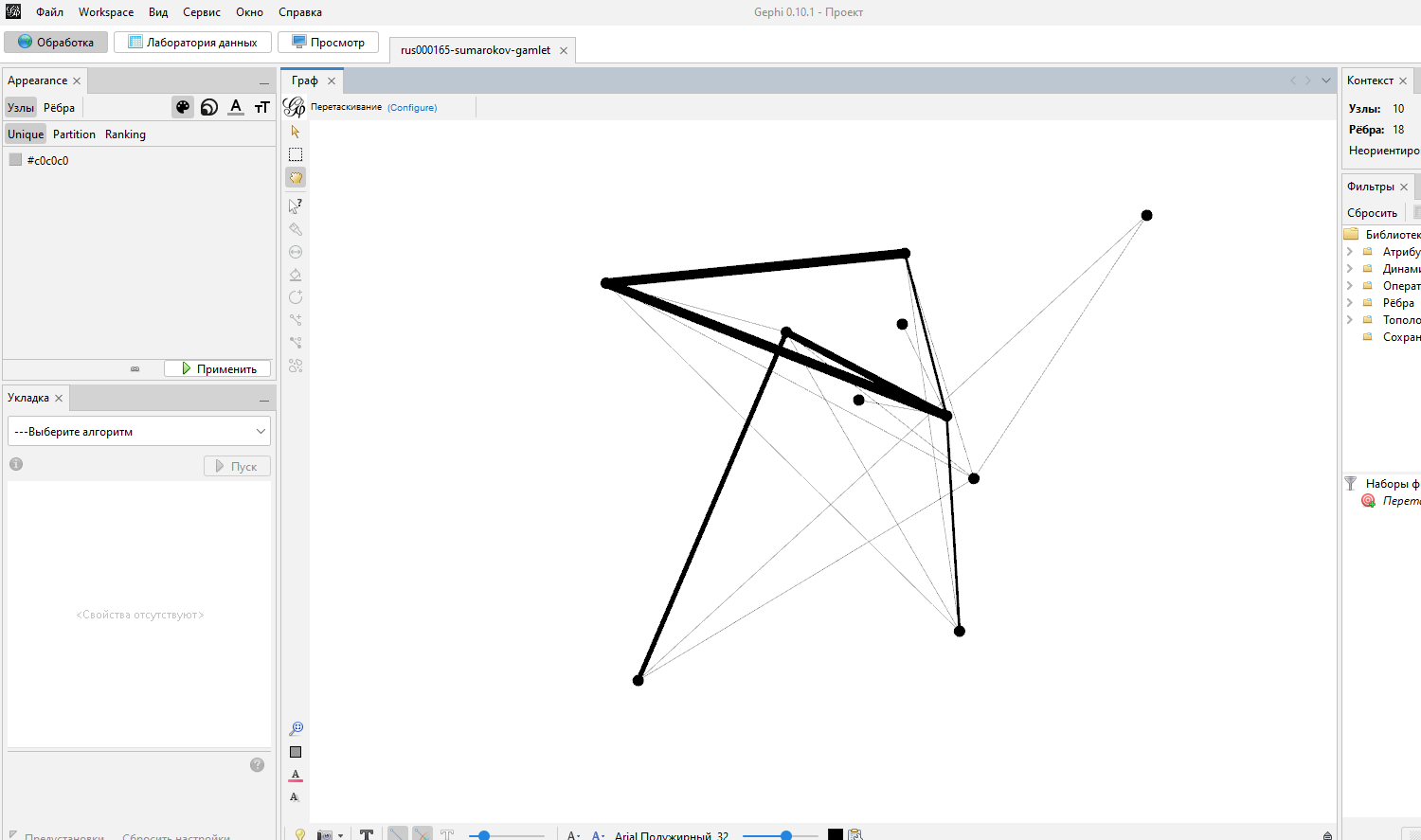
****

****

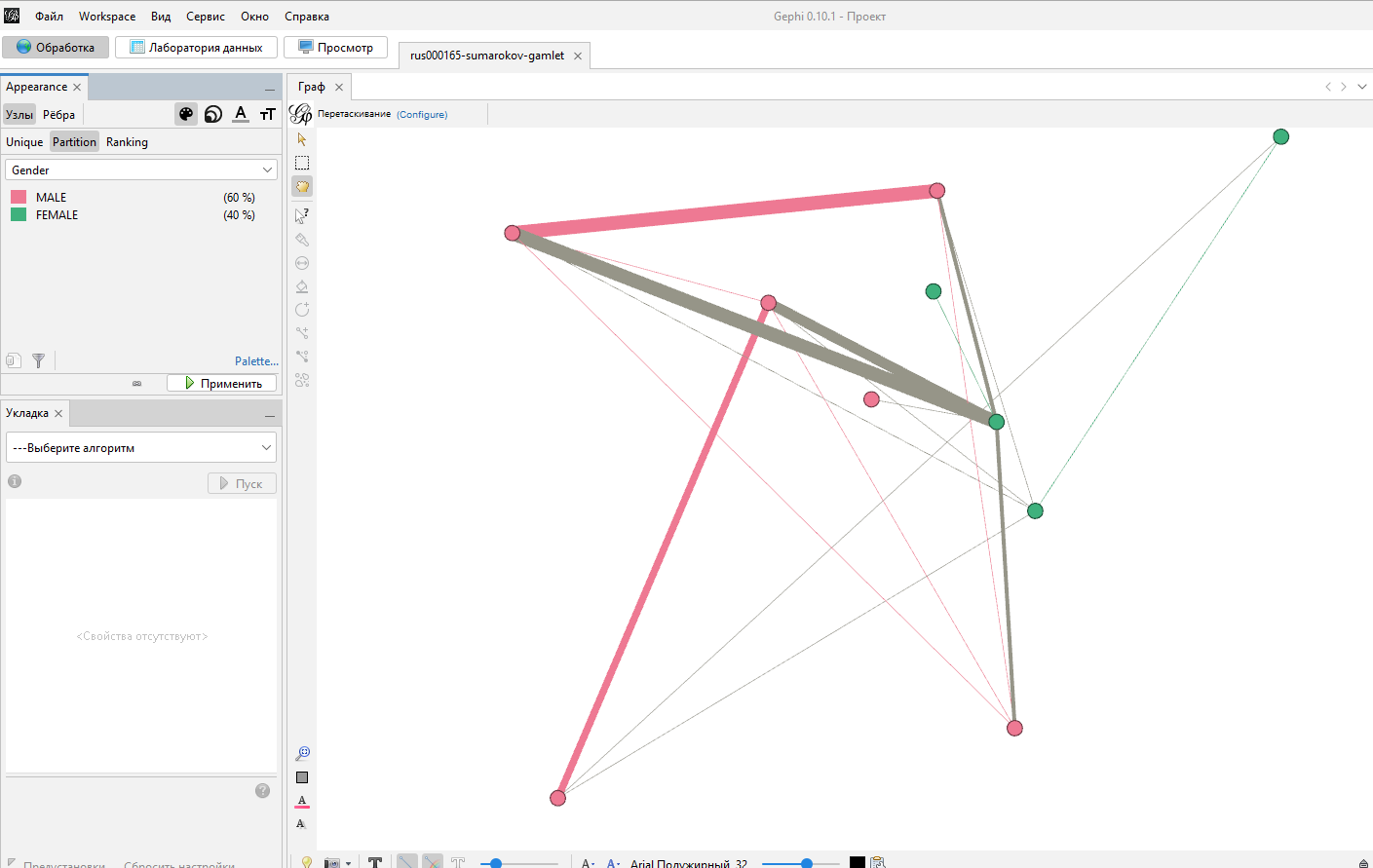
**Номер 10**

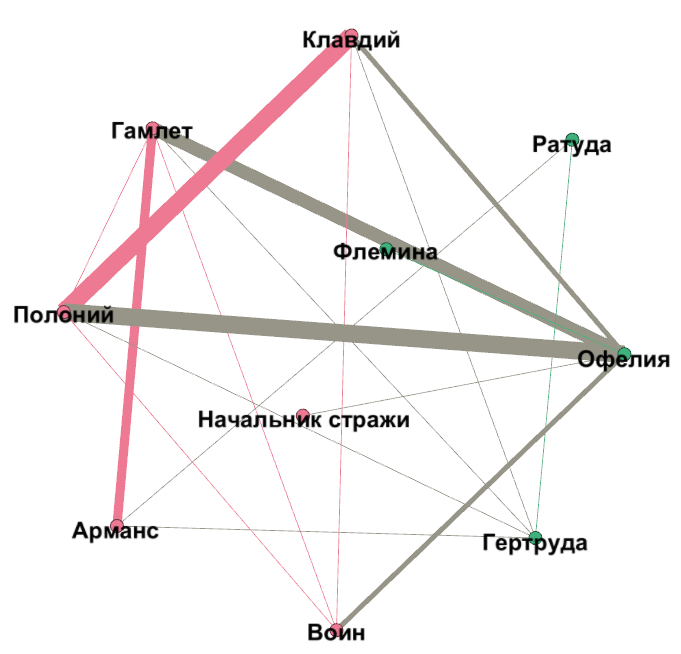
**В.Шекспир «Гамлет»**

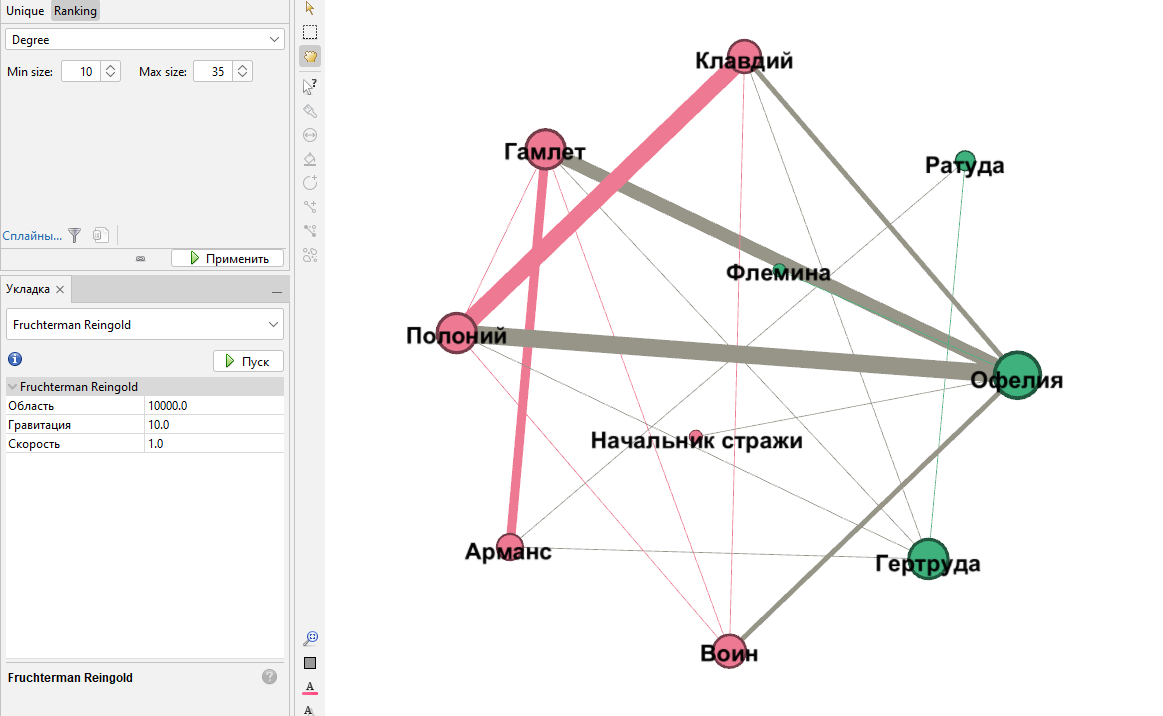
****

****

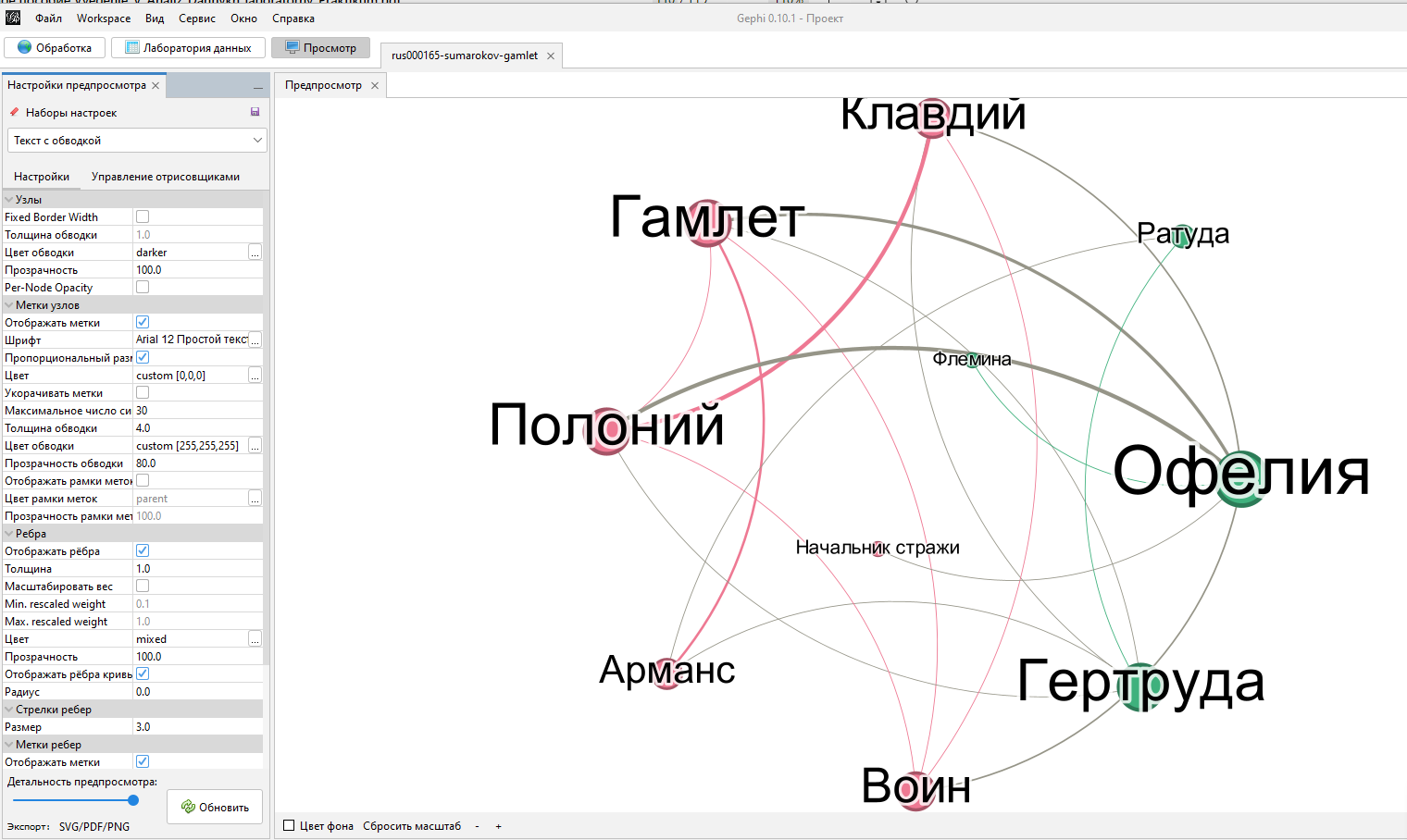
**Поменяем цвет ребер и узлов в зависимости от пола героя**

****

**Подпишем узлы **

**Настроим размер узлов **

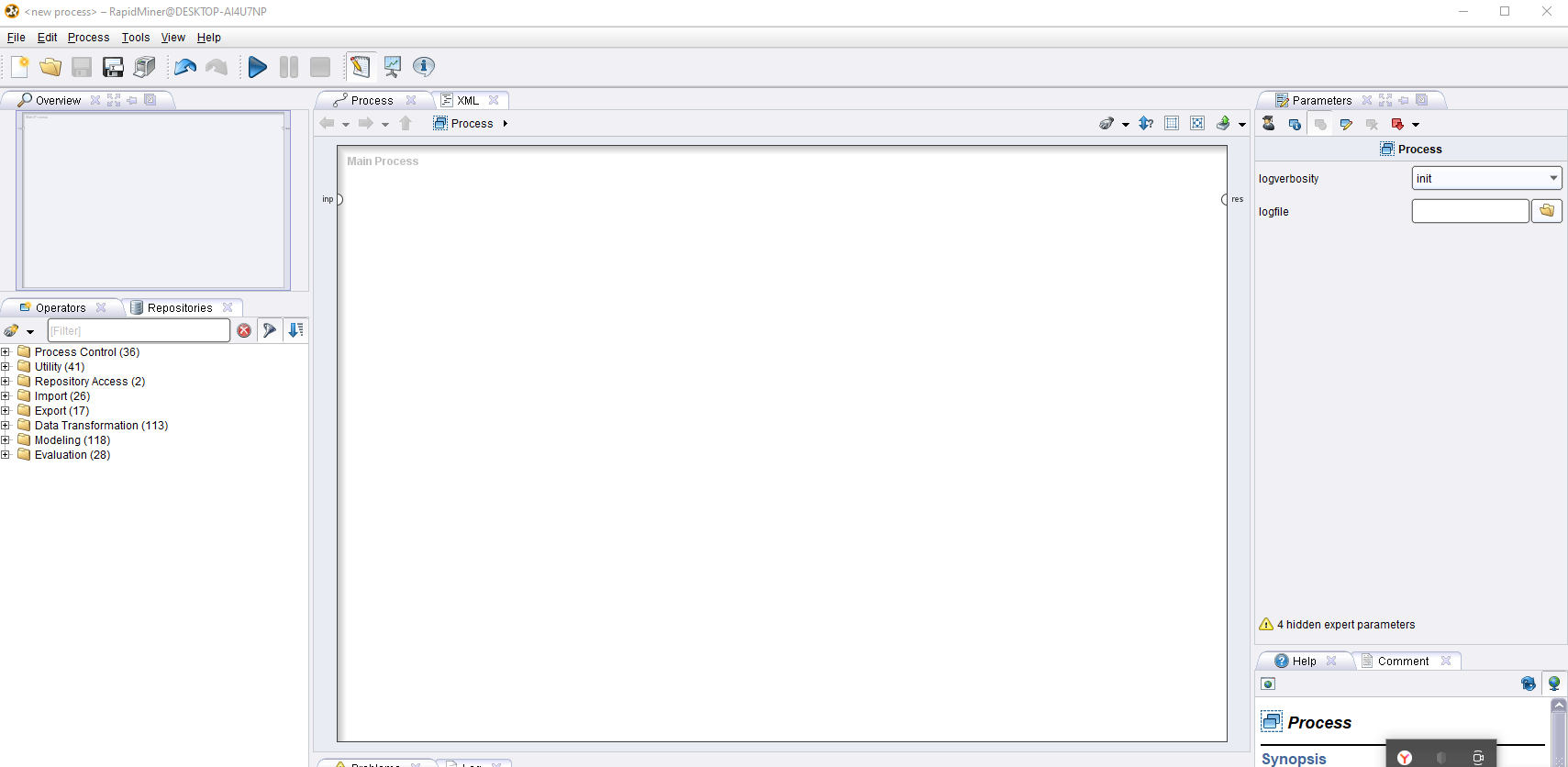
**В конце получается данный граф**

****

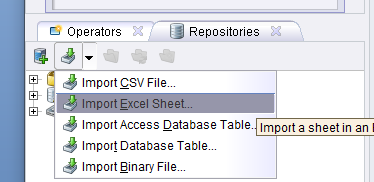
**В результате анализа, стало ясно, что в произведении преобладают мужские персонажи. Ключевые герои: Гамлет, Офелия, Гертруда, Полоний. Больше всего контактов с другими у Офелии, Гамлета, Полония. Неочевидные связи, видные из графа: начальник стражи контактировал в произведении только с Офелией.**

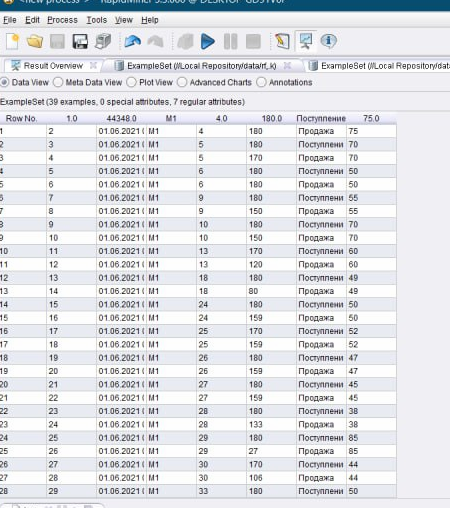
**Задание 3**

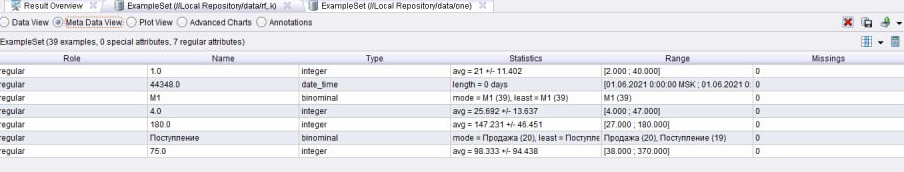
**RapidMiner 5**

****

Импортируем таблицу Excel

****

****

****

****

RapidMiner:

- Описание: RapidMiner — мощный инструмент для анализа данных, предоставляющий широкий спектр функций для машинного обучения и аналитики.

- Плюсы:

- Обширные возможности для работы с данными и построения моделей.

- Поддерживает различные методы машинного обучения.

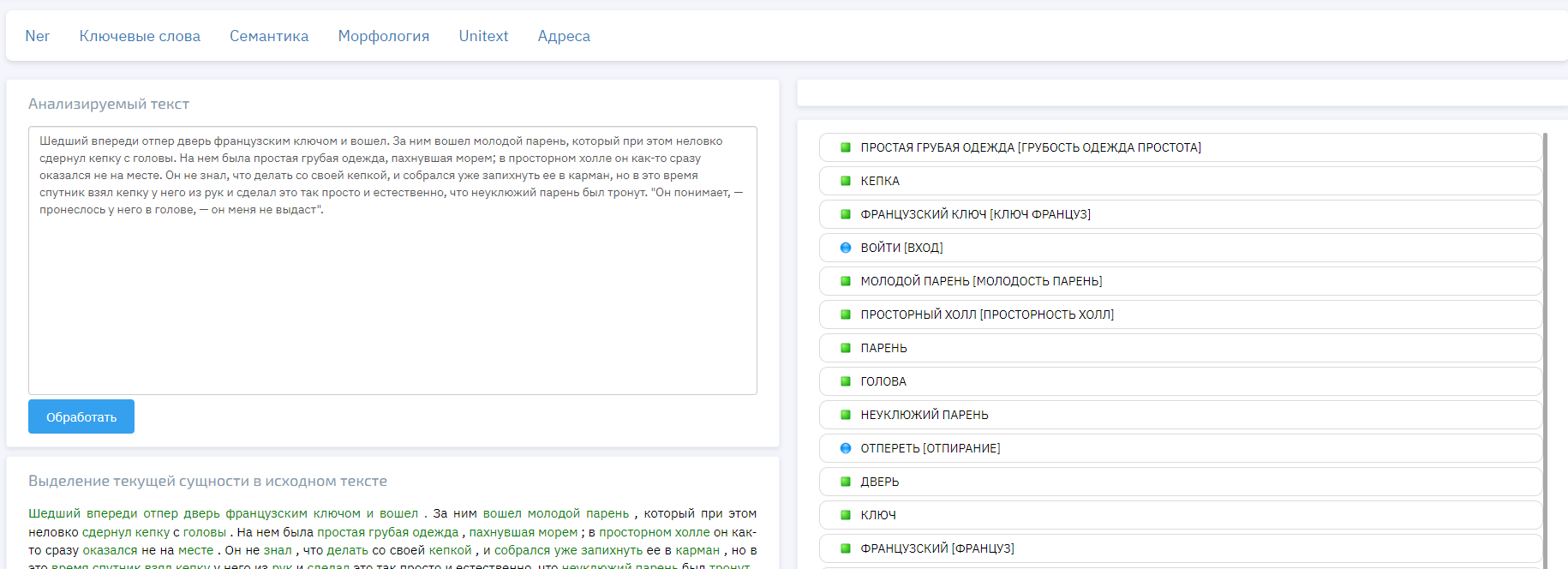
- Имеет большое сообщество пользователей.

- Минусы:

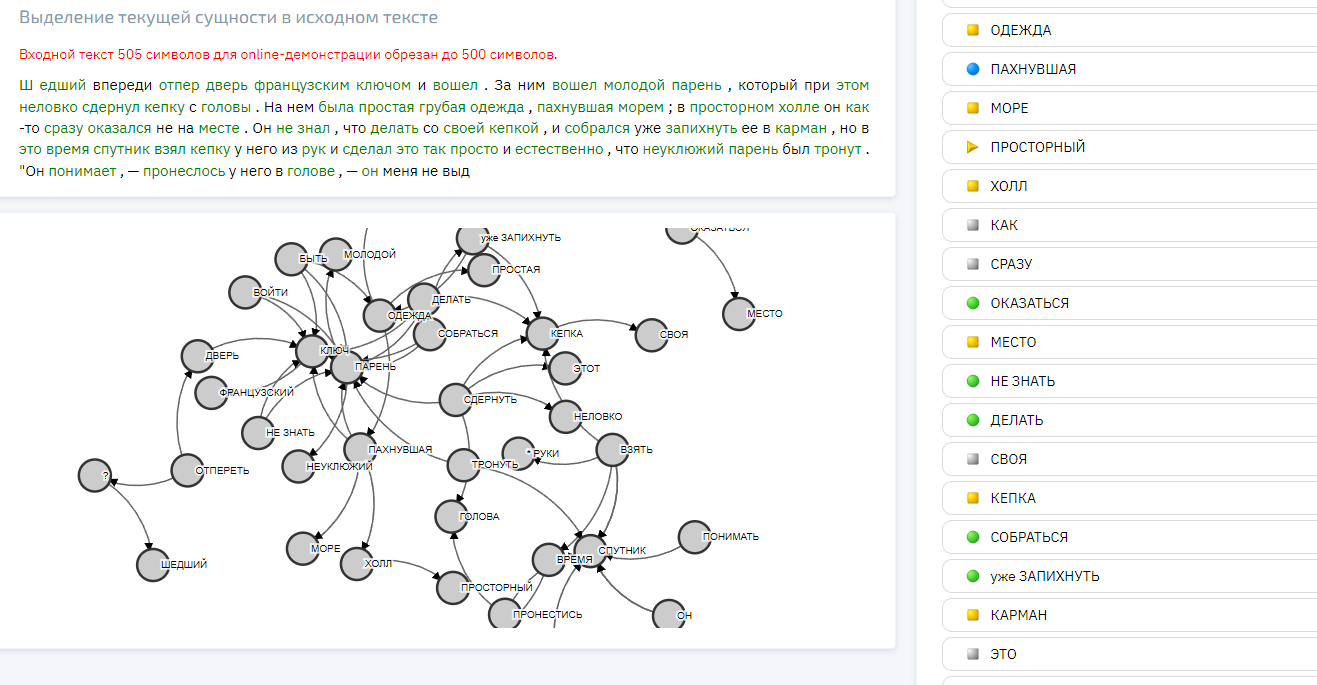
- Не всегда прост в использовании для начинающих.

- Платный доступ к расширенным функциям.

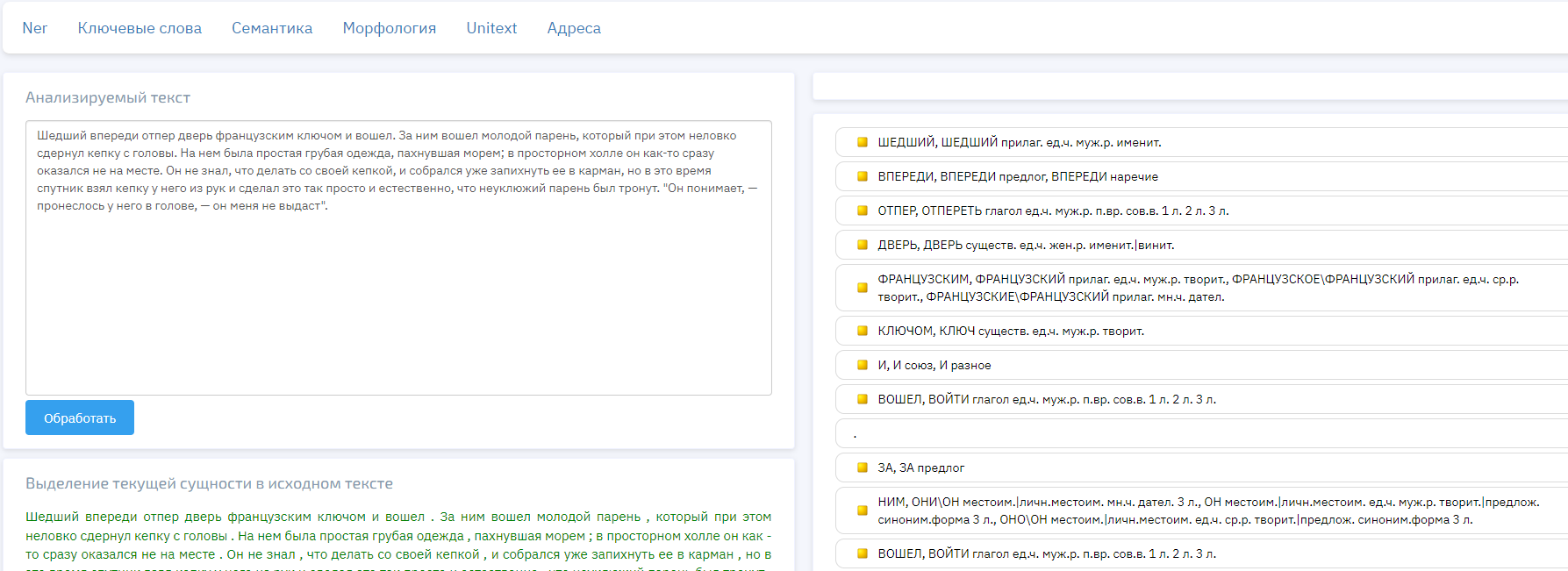
2.Pullenti

Ключевые слова 

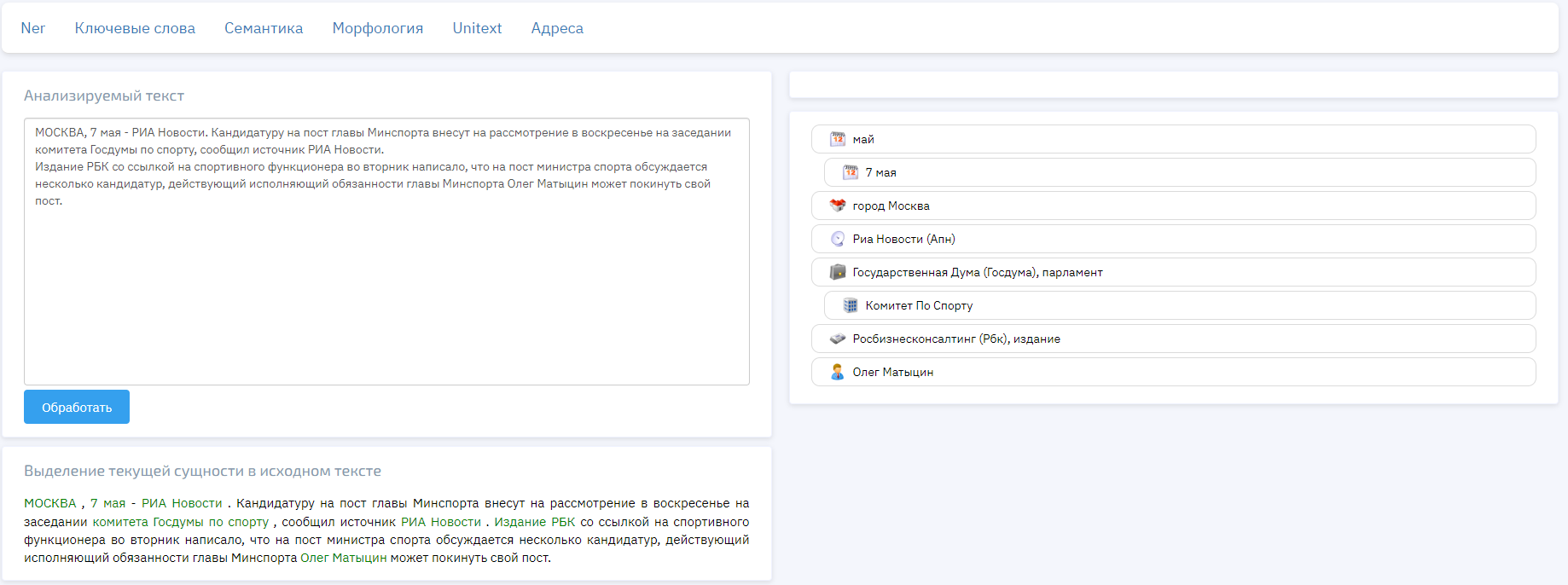
Семантика



**Морфология**

****

**Именованные сущности**

****

**569**

**Уровень 3**

АОЕЯ

Морфологический разбор всех слов

****

Этот{этот=APRO=вин,ед,муж,неод|этот=APRO=им,ед,муж}бумажный{бумажный=A=вин,ед,полн,муж,неод|бумажный=A=им,ед,полн,муж}фонарик{фонарик=S,муж,неод=вин,ед|фонарик=S,муж,неод=им,ед}мой{мой=APRO=вин,ед,муж,неод|мой=APRO=им,ед,муж|мыть=V,несов,пе=ед,пов,2-л}бойфренд{бойфренд=S,муж,од=им,ед}прислал{присылать=V=прош,ед,изъяв,муж,сов}к{к=PR=|к=S,сокр=пр,мн|к=S,сокр=пр,ед|к=S,сокр=вин,мн|к=S,сокр=вин,ед|к=S,сокр=дат,мн|к=S,сокр=дат,ед|к=S,сокр=род,мн|к=S,сокр=род,ед|к=S,сокр=твор,мн|к=S,сокр=твор,ед|к=S,сокр=им,мн|к=S,сокр=им,ед}празднику{праздник=S,муж,неод=дат,ед}из{из=PR=|иза=S,имя,жен,од=вин,мн|иза=S,имя,жен,од=род,мн}Пекина{пекин=S,гео,муж,неод=род,ед}где{где=ADVPRO=}он{он=SPRO,ед,3-л,муж=им}проходит{проходить=V=непрош,ед,изъяв,3-л,сов,пе|проходить=V=непрош,ед,изъяв,3-л,несов,пе|проходить=V,нп=непрош,ед,изъяв,3-л,несов|проходить=V,нп=непрош,ед,изъяв,3-л,сов}стажировку{стажировка=S,жен,неод=вин,ед}в{в=PR=|в=S,сокр=пр,мн|в=S,сокр=пр,ед|в=S,сокр=вин,мн|в=S,сокр=вин,ед|в=S,сокр=дат,мн|в=S,сокр=дат,ед|в=S,сокр=род,мн|в=S,сокр=род,ед|в=S,сокр=твор,мн|в=S,сокр=твор,ед|в=S,сокр=им,мн|в=S,сокр=им,ед}большой{большой=A=вин,ед,полн,муж,неод|большой=A=им,ед,полн,муж|большой=A=пр,ед,полн,жен|большой=A=дат,ед,полн,жен|большой=A=род,ед,полн,жен|большой=A=твор,ед,полн,жен}фирме{фирма=S,жен,неод=пр,ед|фирма=S,жен,неод=дат,ед}

Красным выделены слова с несколькими вариантами разбора

Уберем омонимию

Этот{этот}бумажный{бумажный}фонарик{фонарик}мой{мой}бойфренд{бойфренд}прислал{присылать}к{к}празднику{праздник}из{из}Пекина{пекин}где{где}он{он}проходит{проходить}стажировку{стажировка}в{в}большой{большой}фирме{фирма}



1. Омонимия:

- "Фонарик" может быть понят как предмет освещения или как человек, который приносит свет или надежду.

- "Бойфренд" может быть понят как парень или друг, а также как устройство для поддержания темпа в музыке.

2. Многозначность:

- "Фирма" может означать как компанию, так и устойчивое выражение "в хорошей форме".

- "Прислал" может означать отправку по почте или передачу информации.

3. Отсутствие словоформ в базовом словаре:

- В зависимости от словаря или алгоритма морфологического анализа, некоторые словоформы могут отсутствовать.

4. Неоднозначность:

- Словосочетание "из Пекина" может вызывать неоднозначность в понимании, так как можно интерпретировать как место происхождения фонарика или место, из которого он был отправлен.

Типология вариантности в данном предложении включает омонимию, многозначность, возможное отсутствие словоформ в базовом словаре и неоднозначность, что делает его интересным примером для анализа в контексте лингвистических исследований и обработки естественного языка.

В современной русской морфологии неоднозначность встречается довольно часто из-за особенностей языка. Вот несколько теоретических соображений и примеров конкретного анализа:

1. Флексия и морфологическая неоднозначность:

- Например, слово "белый" может быть прилагательным (белый цвет) или существительным (белая птица). В данном случае, контекст помогает определить, какая часть речи используется.

2. Омонимия и синтаксическая неоднозначность:

- Например, в предложении "Он видел звезду на небе", слово "на" может быть предлогом или частью глагола "видел". Контекст и синтаксический анализ помогут разрешить эту неоднозначность.

3. Многозначность глаголов:

- Некоторые глаголы имеют несколько значений, что может вызывать неоднозначность. Например, глагол "делать" может означать выполнять действие или создавать что-то.

Для улучшения программы-парсера и снятия обнаруженной неоднозначности можно предложить следующие направления дополнения:

1. Использование контекста:

- Разработка алгоритмов, которые учитывают контекст предложения для более точного определения значения слова или фразы.

2. Семантический анализ:

- Внедрение методов семантического анализа для определения значения слова на основе его значения в данном контексте.

3. Машинное обучение:

- Использование методов машинного обучения для обучения парсера на больших корпусах текста с размеченными данными, чтобы улучшить способность разрешать неоднозначности.

4. Использование синтаксических и семантических шаблонов:

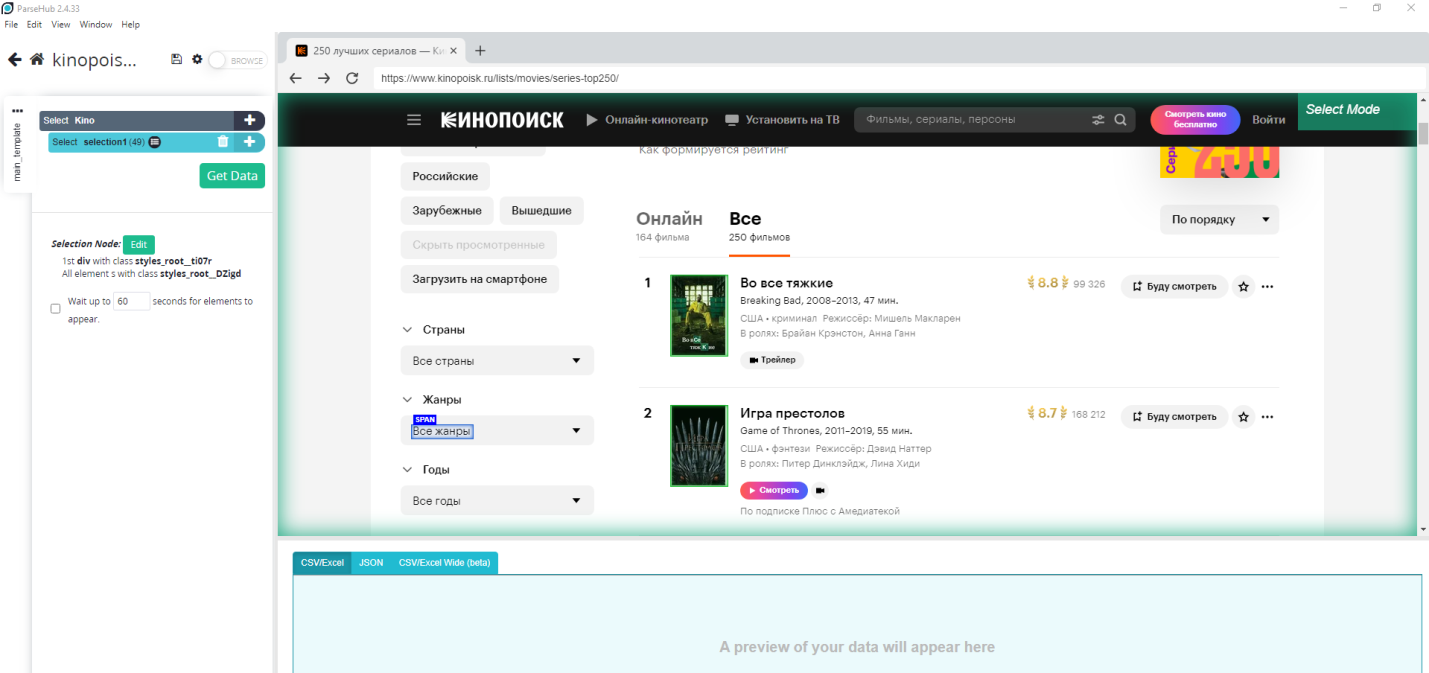
- Разработка и использование синтаксических и семантических шаблонов для анализа предложений и выявления возможных вариантов разбора.

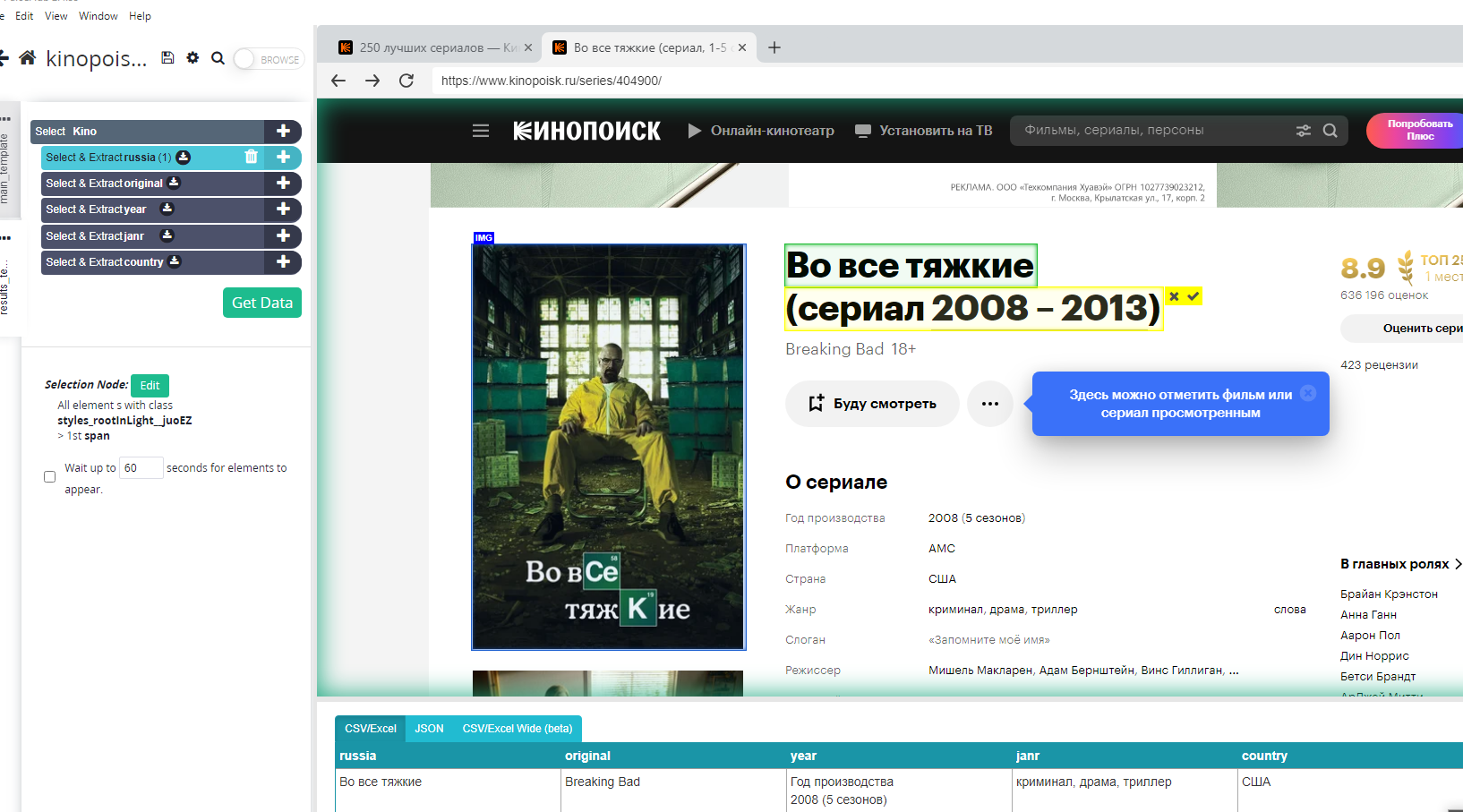
Эти направления могут помочь улучшить работу программы-парсера и более эффективно справляться с неоднозначностями, характерными для современной русской морфологии.

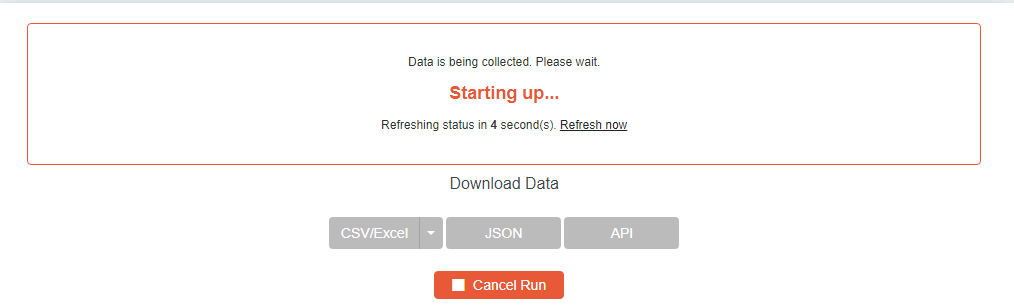
**Задание 1**

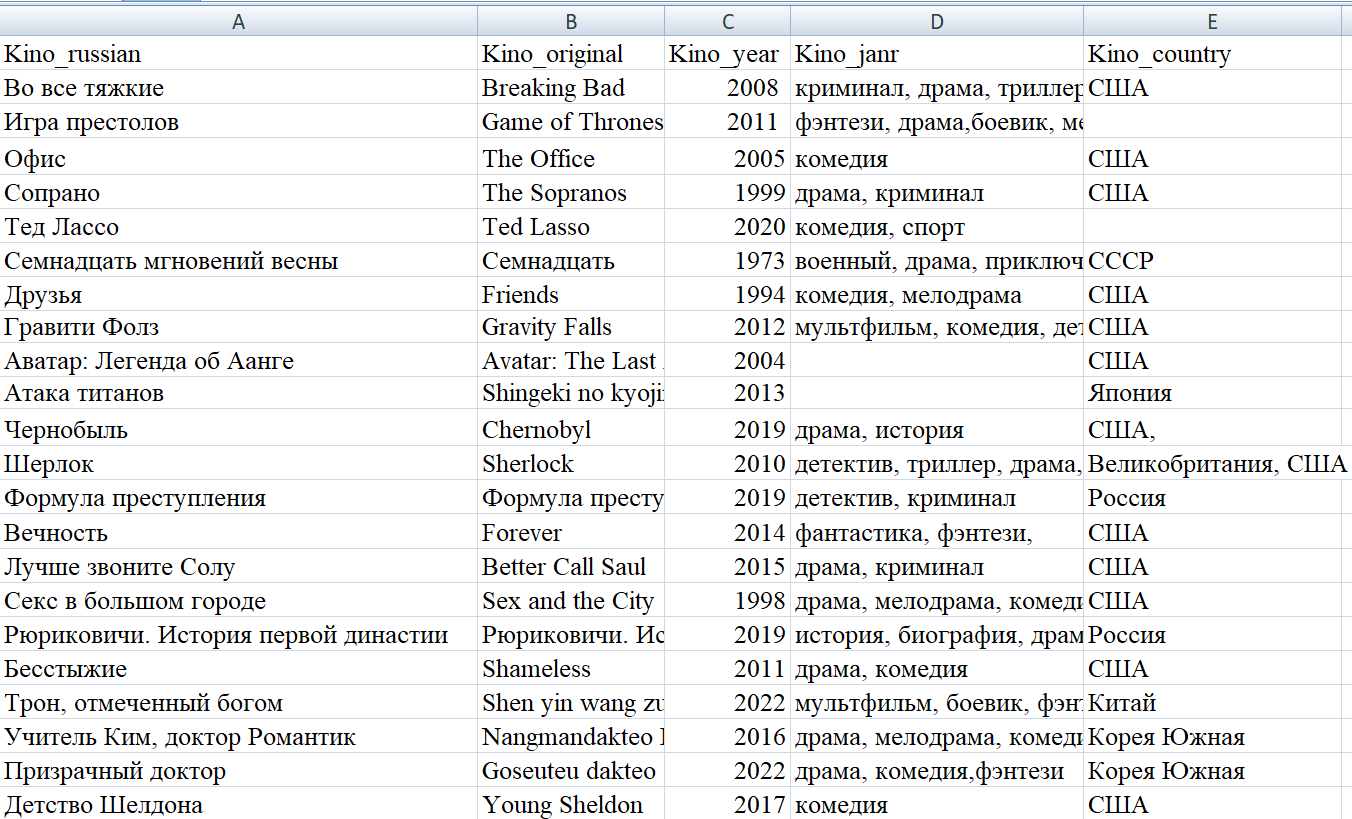
**Номер 5**

**Топ-250 сериалов на кинопоиске**

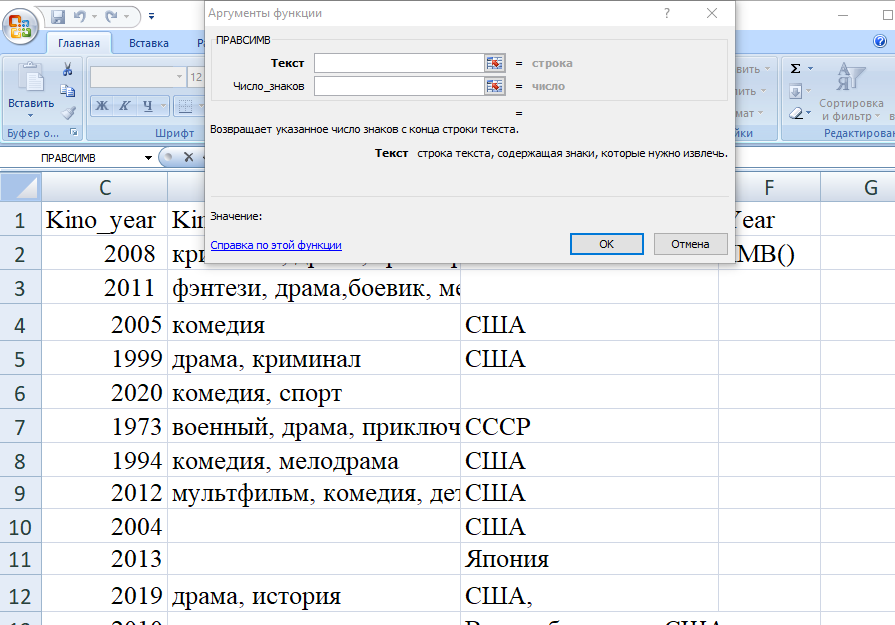
****

****

****

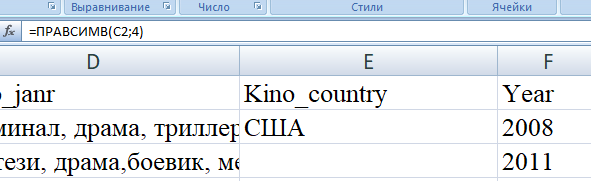
****

**Работаем с таблицей**

****

**С помощью ПРАВСИМВ**

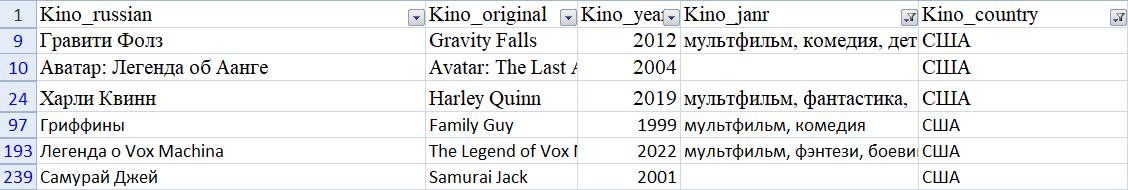
****

****

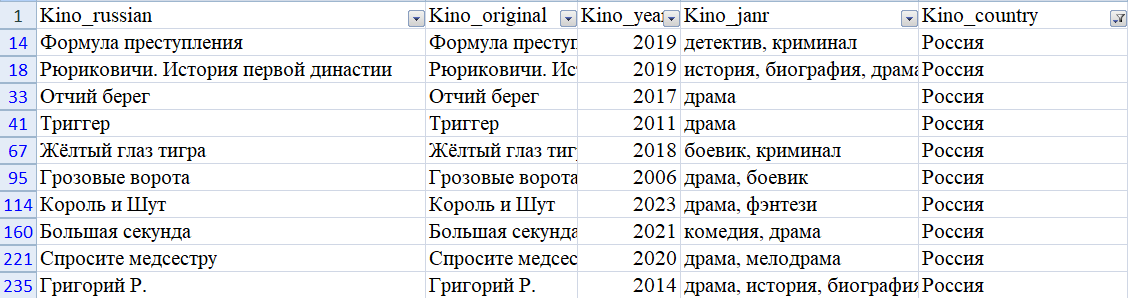
**Ищем нужные данные, используя фильтр**

****

**Мультсериалы из США**

****

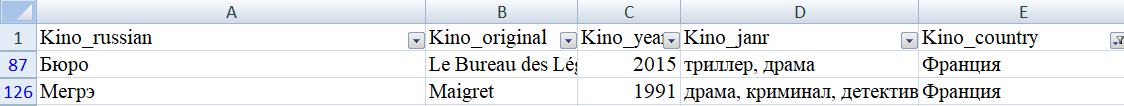
**Сериалы из России**

****

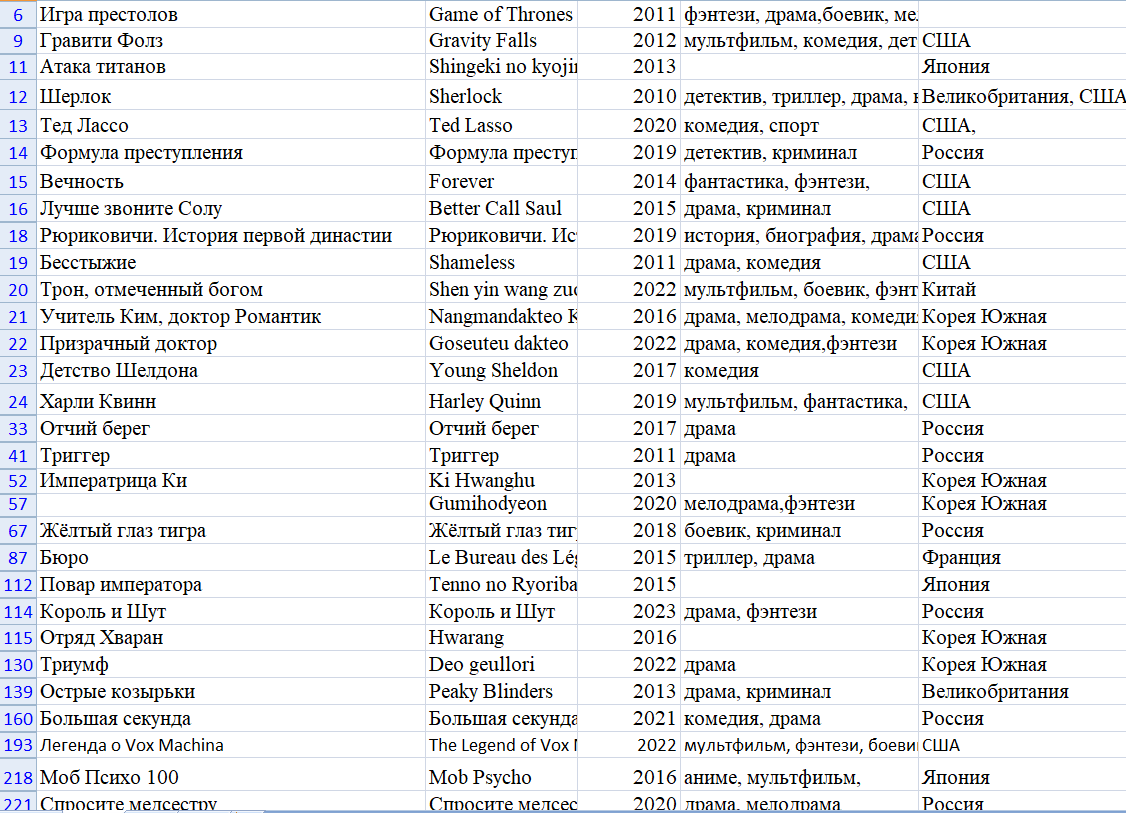
**Сериалы не из Европы**

****

**Сериалы из Франции и Германии**

****

**Сериалы с 2009 г.**

****

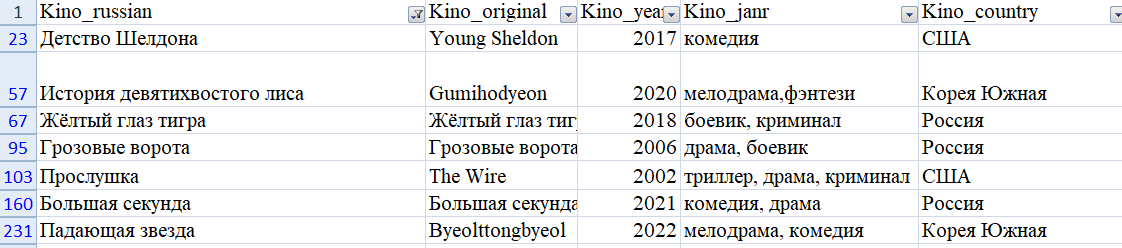
**Комедии**

****

**Сериалы из Великобритании в алфавитном порядке**

****

**Заканчивается на а**

****

**Номер 6**

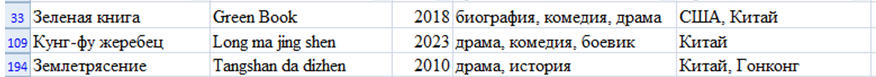
**Топ-250 фильмов на кинопоиск**

****

**Фильмы в жанре биография**

****

**Фильмы из Китая, начиная с 2000**

****

**Мюзиклы из США**

****

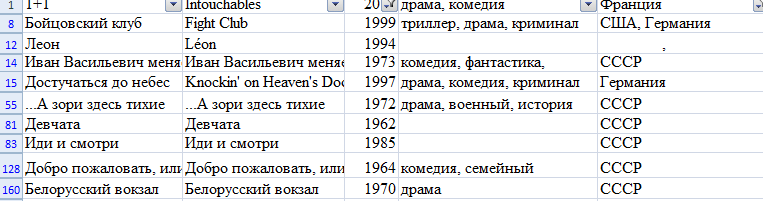
**Фильмы до 1990г**



Фильмы только из СССР



Фильмы, произведённые в странах Европы в период с 1960 по 2005 год в прямом алфавитном порядке



Фильмы, название которых начинается на букву «А»

