*Марголина Анастасия*

*БФЛ-192*

Корпус «О чем говорят персонажи фильмов братьев Коэн?»

Выборка данных

Представленный корпус содержит в себе субтитры 16 фильмов режиссерского дуэта Джоэла и Итана Коэн. В качестве данных, входящих в корпус, были отобраны только авторские работы, т.е. фильмы, в которых Коэн выступали и в качестве режиссеров, и в качестве продюсеров, и в качестве сценаристов. Список начинается с «Перекрестка Миллера (1990)» несмотря на то, что до этого фильма были сняты два «авторских». Это объясняется тем, что именно «Перекресток Миллера» можно считать отправной, стилистической точкой в творчестве Коэн. Корпус заканчивается на самой последней совместной работе «Баллада Бастера Скраггса (2018)».

Подготовка данных

Как было отмечено ранее, данные в этом корпусе – субтитры, т.е. реплики героев. Формат файла .srt предполагает наличие не только текста, но и отметок времени, на котором какая-то фраза будет показана (например, 00:00:03,220 --> 00:00:05,970). Так, было решено написать простую функцию на языке Python, которая удаляет ненужные числовые значения с помощью регулярных выражений, а затем сохраняет «чистый» текст в формат .txt. На рис.1 представлена полученная функция.

Тем не менее, этого недостаточно для начала работы с корпусом, поскольку в текстах присутствовали лишние данные, которые стоило убрать: песни (обозначались символом вопроса ‘? some lyrics? ’), ремарки (‘(SCREAMING)’), знак ‘-’ перед каждой новой репликой. Важно отметить, что в некоторых случаях в субтитрах обозначался говорящий, например ‘MAN: ’. Хоть эти данные и могли быть полезны для определенного анализа, для этого исследования они нерелевантные. Кроме того, больше половины субтитров не предоставляли таких значений (=> корпус был бы не сбалансирован).

С помощью регулярных выражений \?[^)]+\? и \([^)]+\) в Notepad ++ была удалена вся лишняя информация. Они убирают весь текст, заключенный в определенных символах (‘?’ и ‘)’ соответственно), а также сами символы. Такие знаки, как ‘-’ или кавычки перед некоторыми репликами были убраны обычной заменой символа на ‘ ’.



Рисунок 1

Полученный корпус

Корпус «О чем говорят персонажи фильмов братьев Коэн» состоит из 16 документов, 137786 слов, и которых 11175 – уникальные словоформы. Самый длинный документ – субтитры к фильму «Железная хватка», самый короткий – «Большой Лебовски». Интересно, что логичная, на первый взгляд, гипотеза - объем реплик персонажей зависит от длины самого фильма, в данном случае (на материале этого корпуса) не подтверждается. Например, «Да здравствует Цезарь!» (1час 45 минут) на втором месте по количеству слов (почти 11 тыс), когда «Баллада Бастера Скраггса» (2 часа 12 минут) входит в топ «the shortest» (всего лишь 7600).

Так как весь корпус англоязычный, стоп слова размечены достаточно хорошо. Как видно из рисунка 2, на котором представлено ‘word cloud’из Voyant Tools, наиболее частые слова ‘know’, ‘yeah’, ‘i’m’, ‘it’s’, ‘man’. Такой результат не удивителен: корпус состоит из устной речи персонажей, для которой характерны сленг и фонетическая редукция.

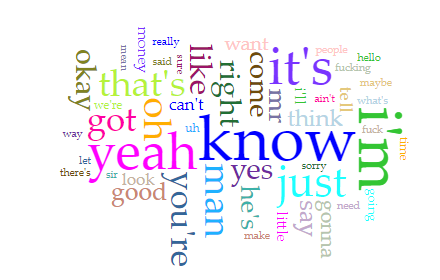


Рисунок 2

Визуализация корпуса с помощью Python

Библиотеки языка Python (pandas, seaborn, matplotlib,pyplot) предоставляют возможность визуализировать некоторые зависимости в корпусе. В качестве входных данных была выбрана таблица Excel, документация по корпусу, скачанная из Voyant Tools. В нее также вручную были добавлены столбцы author, language и genre.[[1]](#footnote-1)

На рисунке 3 представлен график зависимости жанр/количество слов. Для него были выбраны 4 самых частотных жанра. Тэги ‘black comedy’, ‘romantic comedy’, ‘comedy’ и ‘western’, ‘neo-western’ были выделены в два отдельных тэга – ‘all comedy’, ‘all western’, соответственно.

Построение графика было сделано с целью проверить гипотезу, что жанр ‘everyday life’, т.е. повседневность, будет демонстрировать наибольший объем слов. Как видно из рисунка 3 это суждение было опровергнуто: ‘all comedy’ самый объемный жанр по репликам. Безусловно, такой результат не может считаться точным, поскольку график построен на неравномерных данных: комедия - самый популярный жанр фильмов братьев Коэн (12 фильмов). Тем не менее, в жанре вестерн только 4 фильма, в то время как повседневность – 7 и, тем не менее, вестерн показывает большее значение. Кроме того, в графике не были учтены влияние одного жанра на другой: зачастую в фильме присутствовали теги, как и комедия, так и криминал. Впрочем, этот алгоритм может быть реализован в дальнейшем, на большем материале.

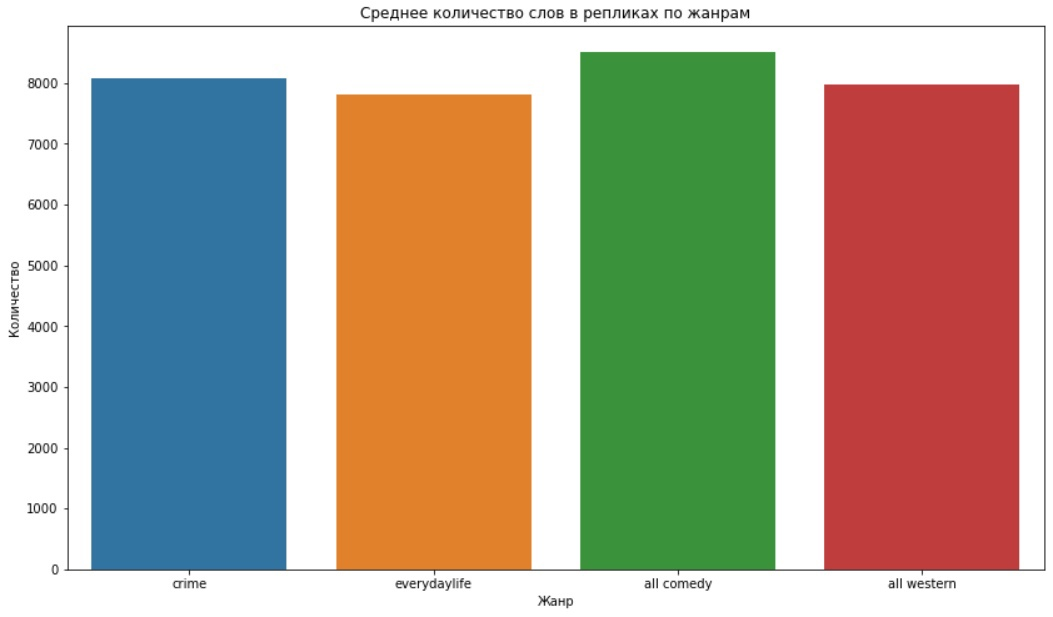


Рисунок 3

Следующие два графика (рис. 4 и рис.5) показывают соотношение количества слов, уникальных словоформ (соответственно) к году [выпуска фильма]. На них можно заметить определенную закономерность: раз в несколько лет объем слов, как и уникальность словоформ резко падают. Братья Коэн плодотворные режиссеры, как можно заметить из таблицы в документации (см. Documentation.xlsx); зачастую года выпусков их фильмов идут подряд. Можно сделать тривиальное предположение о ‘творческом кризисе’, т.е. о сложности генерировать большие, ‘разнообразные’ работы каждый раз. Такая гипотеза может быть основной междисциплинарного исследования на стыке филологии, социологии и психологии, данные которого могут быть полезны в разных сферах: от бизнеса и литературоведения/киноиндустрии до разработки системы против выгорания. Однако, все, безусловно, зависит от полученных результатов.



Рисунок 4



Рисунок 5

Интерпретация с помощью веб-инструментов Voyant Tools

Область Digital Humanities предоставляет много инструментов для филологов для интерпретации какого-либо художественного корпуса. В этой части эссе будут рассмотрены один такой Voyant Tools, который позволяет визуализировать какие-либо статистические данные по корпусу без применения языков программирования (R, Python).

Мандала

Мандала – это еще один инструмент для визуализации в Voyant Tools. Она представляет собой концептуализированную схему отношений между документами и терминами (terms).

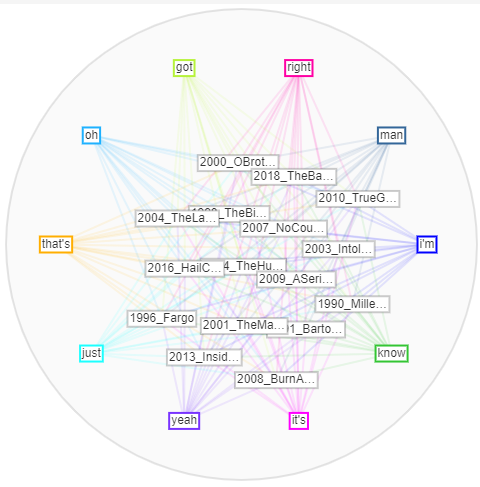


Рисунок 6

На рисунке 6 применяется такая функция к корпусу субтитров фильмов братьев Коэн. Этот инструмент удобен для поиска семантических связей между разными фильмами. Чем ближе фильм к какому-либо слову, тем чаще оно там встречается. Например, как видно из рисунка 6, слово ‘man’ наиболее часто употребляется в трех фильмах: ‘O Brother where art thou? ’, ‘The Ballad of Buster Scruggs’, ’True Grit’. Интересно, что два последних – вестерны и не комедии. Первый упомянутый фильм отчасти тоже можно назвать вестерном: это ‘road movie’, события которого происходят в южных штатах времен Великой депрессии. Такой результат можно связать с двумя идеями: ‘man’ часто используется в качестве сленгового обращения (‘look, man,...’) или выражения сожаления (‘oh man’), такой просторечный язык характерен для сеттинга вестерна; ‘man’ буквально – человек, так как зачастую главная тема фильмов вестерн – какой-либо человек (например, ‘охота за головами’). Стоит отметить, инструмент мандала может быть очень полезен для нахождения закономерностей как в фильмах, так и в книгах. Так, выбрать интересующих двух авторов и проследить схожести их языка (например, выбрать пару писателей, где первый художественно повлиял на другого).

Loom

Инструмент Loom показывает, как распределены самые частотные слова во всем корпусе. Он удобен тем, что предоставляет наглядную статистику. На рисунке 7 отчетливо видно шесть кривых линий, выбивающихся из ритма всего корпуса.

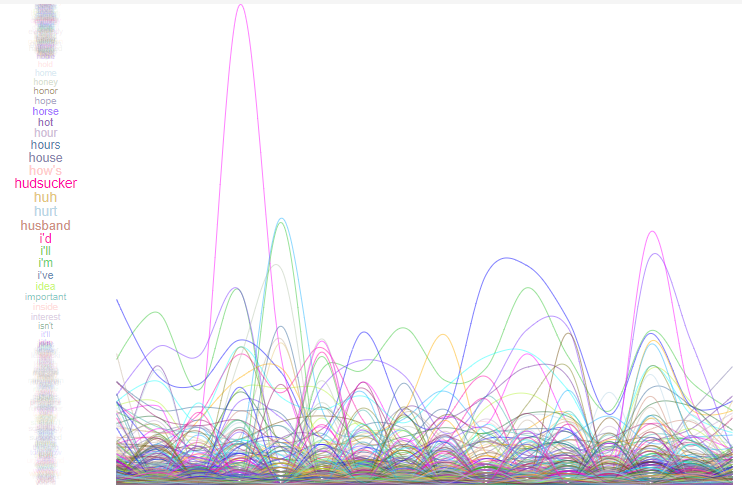


Рисунок 7

Две розовые линии – ‘yeah’, синяя – ‘know’, голубая – ‘it’s’. Однако, интересна другая закономерность: около пика розовой линии видны две почти идентичные кривые – ‘man’, ‘dude’. То есть, в одном документе эти два синонимичных слова употребляются с одинаковой частотой. Вероятно, это тот случай, о котором было сказано ранее: ‘man’ в сленговой парадигме.

В целом, такой инструмент удобен для визуализации необычных, доминирующих распределений. ‘Loom’ может помочь сделать выводы о корпусе: единичный ли случай такого частотного употребления какого-либо слова или это наблюдаемая тенденция через весь массив текста.

Terms Radio

Этот инструмент похож на ‘Loom’, он также показывает тенденции по частоте употребления слова на всем корпусе в виде кривых. Тем не менее, его функционал менее абстрактен, ‘Terms Radio’ отображает на оси X названия фильмов, а на оси Y частотность слова. Кроме того, можно выбрать желаемые слова (рис.8).

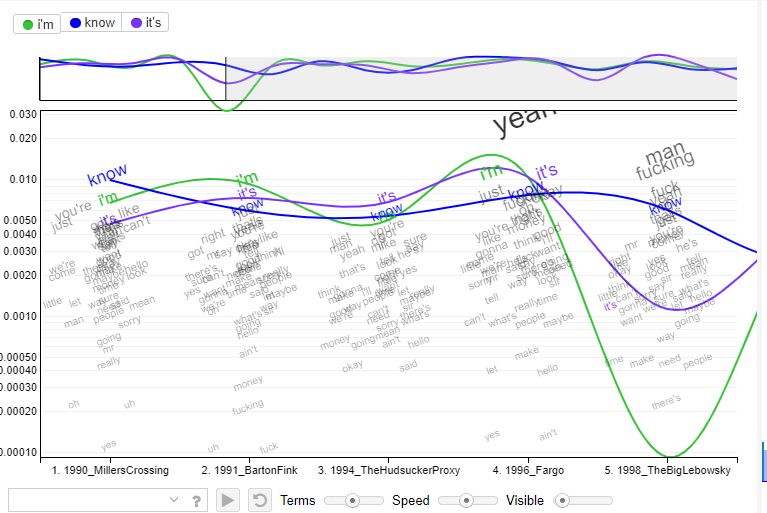


Рисунок 8

На рисунке 8 можно заметить три трэка: ‘I’m’, ‘it’s’, ‘know’. Глагол знать равномерно проходит через весь корпус, на рисунке 1 видно, что это самое популярное слово для корпуса. Интересно, что на фильме ‘Большой Лебовски’ происходит резкий спад ‘I’m’ и ‘it’s’. Действительно, это можно связать с субтитрами: вероятно, в этом фильме ‘I’m’ написано, как ‘I am’, без редукции. Тем не менее, это не объяснение конфликтует с наличием it’s’. Причину такого падения можно найти в сюжете фильма: главный герой во- первых мало говорит, во-вторых, не зациклен на себе, как и, в целом, персонажи этого фильма.

Функционал и применения этого инструмента, как было сказано ранее, очень схожи с ‘Loom’. Однако, ‘Radio Terms’ позволяют рассматривать статистику на конкретных примерах.

Topics

Анализируя далее фразу ‘I’m’, мною было решено использовать ‘Topics’, который генерирует кластеры слов (темы).

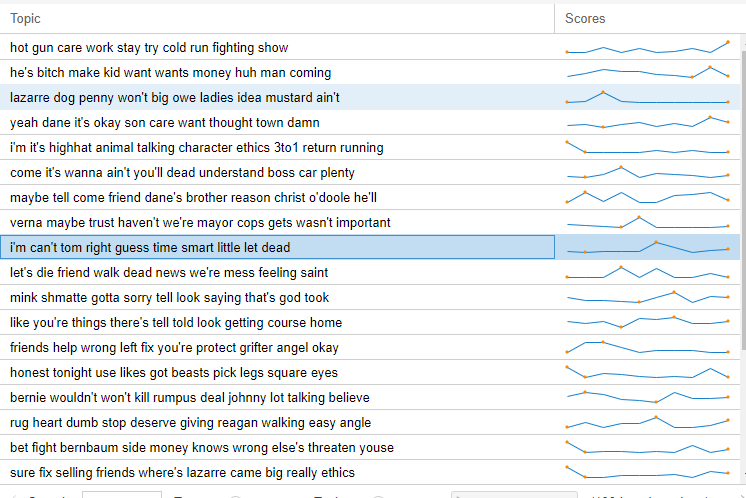


Рисунок 9

На рисунке 9 выделена одна тема, начинающаяся с ‘I’m’. Примечательно, что ее пик – фильм ‘O Brother Where Art Thou’. В отличии от главного героя ‘The Big Lebowsky’, герой этого фильма – эгоистичен и самовлюблён, что, вероятно, и связано с частым употреблением ‘I’m’. Кроме того, выделенная тема характерна еще для одного вестерна ‘No Country for Old Men’.

Этот инструмент – наиболее подходит для задачи этого корпуса, проанализировать, о чем в основном говорят персонажи братьев Коэн. Из рисунка 9 видно, что зачастую темы связаны с криминалом, убийством, оружием и деньгами.

Trends

Предыдущее утверждение также доказывается с помощью Trends – раздела, который в виде линии позволяет сравнить какие-либо слова по их частоте использования.

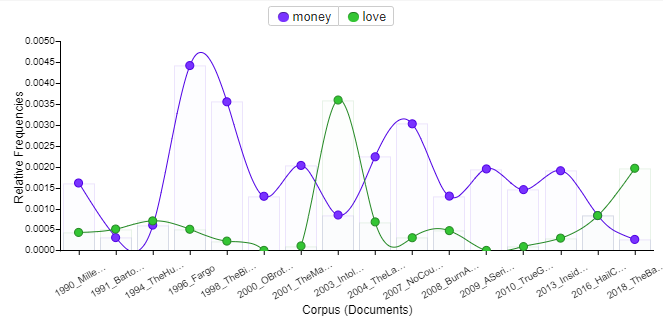


Рисунок 10

На рисунке 10 показано распределение ‘money’ и ‘love’. Можно заметить, что деньгам персонажи фильмов Коэн уделяют больше внимания, чем любви. Интересно, что в единственной романтической комедии авторов наблюдается обратная пропорция: ‘love’ произносят чаще, чем ‘money’.

Context

Важный инструмент для работы с любым корпусом – контексты. В продолжении анализа ‘money’ мною были рассмотрены самые частые правые и левые контексты для этого слова (рис. 11 и рис. 12).

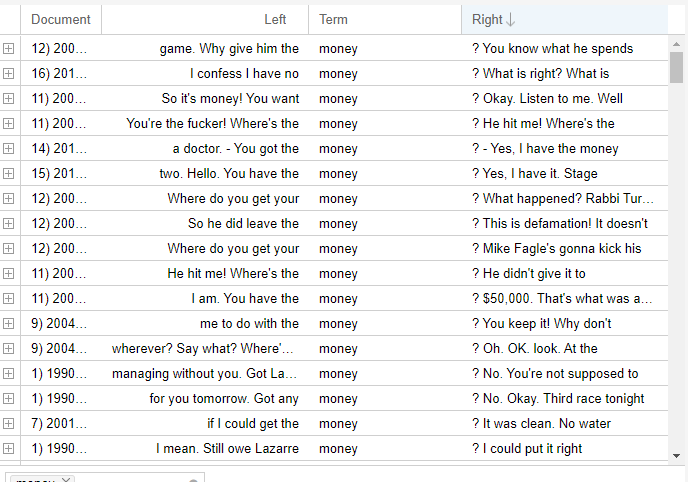


Рисунок 11



Рисунок 12

Как видно из приложенных таблиц, герои фильмов в основном задают вопросы, связанные с деньгами, а также предпочитают говорить о деньгах, принадлежащих другим людям (‘your money’).

Links

Последний примененный к корпусу инструмент – Links, который, как и Mandala рисует графы связей, но только между словами, без привязки к документу.

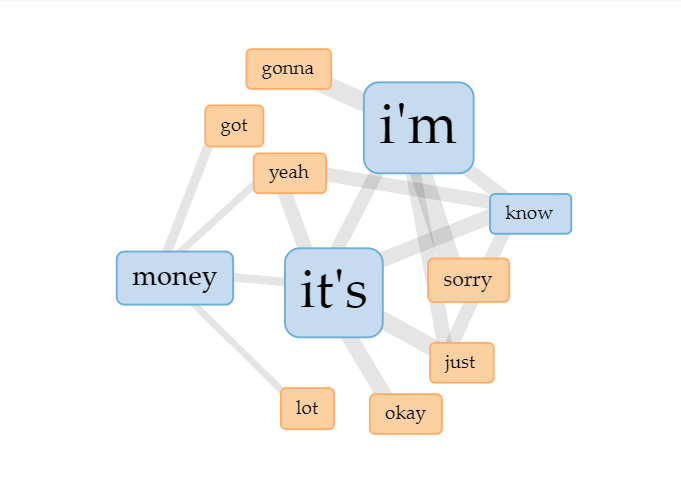


Рисунок 13

На рисунке 13 видно, что ‘money’ в основном связаны со словами ‘got’ и ‘lot’. Это, а также информация из предыдущего пункта опять отсылают к главной теме фильмов братьев Коэн – криминал, американская мечта и желание обогатиться.

Заключение

При исследовании корпуса «О чем говорят персонажи фильмов братьев Коэн?» были применены методы анализа как с помощью Python, так и с помощью готового приложения – Voyant Tools. Они оба имеют свои преимущества: язык Python позволяет интерпретировать и визуализировать любую входную информацию, в частости – метаданные (год выпуска фильма, количество слов и т.д.), в то время как Voyant Tools – инструмент для работы только с текстовой информацией. Удобство последнего метода заключается в его простоте и доступности любому человеку. Благодаря Voyant Tools можно найти интересные закономерности между лексикой корпуса и содержанием его документов.

Все примененные инструменты из Voyant Tools – визуализации, работающие со статистикой по корпусу. Они показывают заданные параметры и помогают сделать репрезентативное исследование с наглядными и понятными иллюстрациями.

В этом исследовании было также показано, что текстовая информация для корпуса может быть любой: не только художественные произведения, но и субтитры к фильмам. Так, инструментами для работы с корпусами digital humanities может пользоваться каждый гуманитарий: не только филолог, но и киновед, антрополог, культуролог.

1. Описание всех столбцов и код на Python можно найти в приложенной к письму ссылке Google Colab [↑](#footnote-ref-1)