Московская культурная жизнь

Целью данной работы является анализ культурных событий Москвы по данным сайтов ‘Культура Москвы’[1] и ‘2go2go’[2].

I. Данные

Мы использовали архивы культурных событий выше упомянутых сайтов. Сайт ‘Культура Москвы’ предоставляет меньше метаданных о событии, чем сайт ‘2go2go’, однако данные сайта ‘2go2go’ менее последовательны (в частности, очень много пересекающихся категорий, которые отнюдь не всегда приписываются закономерно). Помимо этого, на сайте ‘Культура Москвы’ заметно больше событий, правда на обоих сайтах события неравномерно распределены по годам (см. информацию ниже). Большая часть анализа основывается на данных сайта ‘Культура Москвы’.

Метаданные сайта ‘Культура Москвы’:

- ➔ название события
- ➔ место проведения мероприятия (например, ‘Центральный дом художника’)
- ➔ адрес проведения мероприятия (например, ‘Крымский Вал, 10’)
- ➔ даты
- ➔ категории
- ➔ стоимость билета
- ➔ текст объявлени

Всего: 5634 объявления (1934 мероприятия -- 2014 год, 3735 -- 2015 год, 402 -- январь-май 2016 года).

Метаданные сайта ‘2go2go’:

- ➔ название события
- ➔ место проведения мероприятия
- ➔ адрес проведения мероприятия
- ➔ даты
- ➔ категории
- ➔ стоимость билета
- ➔ количество просмотров объявления
- ➔ количество комментариев к объявлению
- ➔ текст объявления

Всего: 1024 объявления (44 мероприятия -- 2013 год, 140 -- 2014 год, 462 -- 2015 год, 974 -- январь-май 2016 года).

---

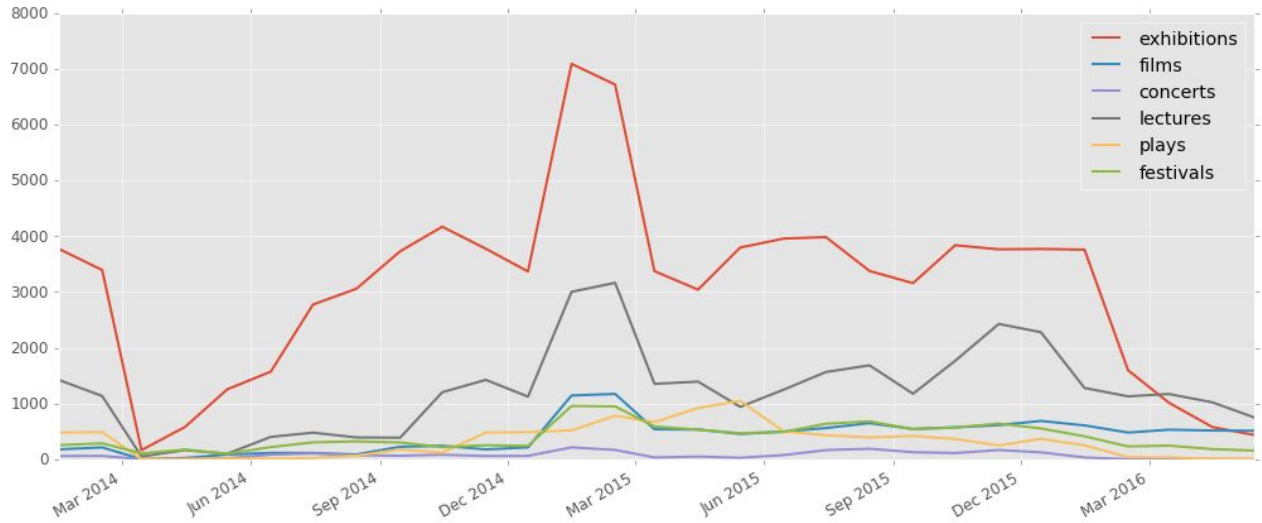
[1] http://cult.mos.ru/

[2] http://www.2do2go.ru/msk

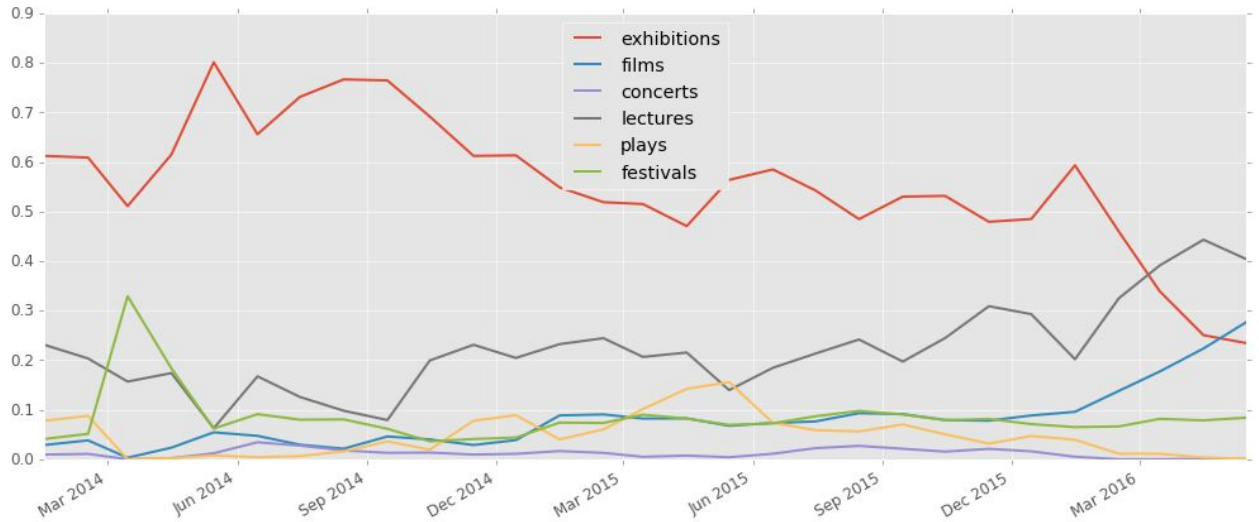
II. Анализ данных

1) Распределение типов культурных событий по месяцам

Для данного распределения используем данные сайта ‘Культура Москвы’. Доступный временной период: январь 2014 г. - май 2016 г. Посчитав, сколько событий какой категории было в каждом месяце, и построив соответствующий график, замечаем, что количество мероприятий неравномерно распределено по месяцам. Отчасти это может объясняться человеческим факторов: причиной того, что, например, в определенный период (апрель 2014 г.) количество мероприятий близко к 0, может являться не отсутствие большого числа культурных событий в этот период, а отсутствие соответствующих данных на сайте.

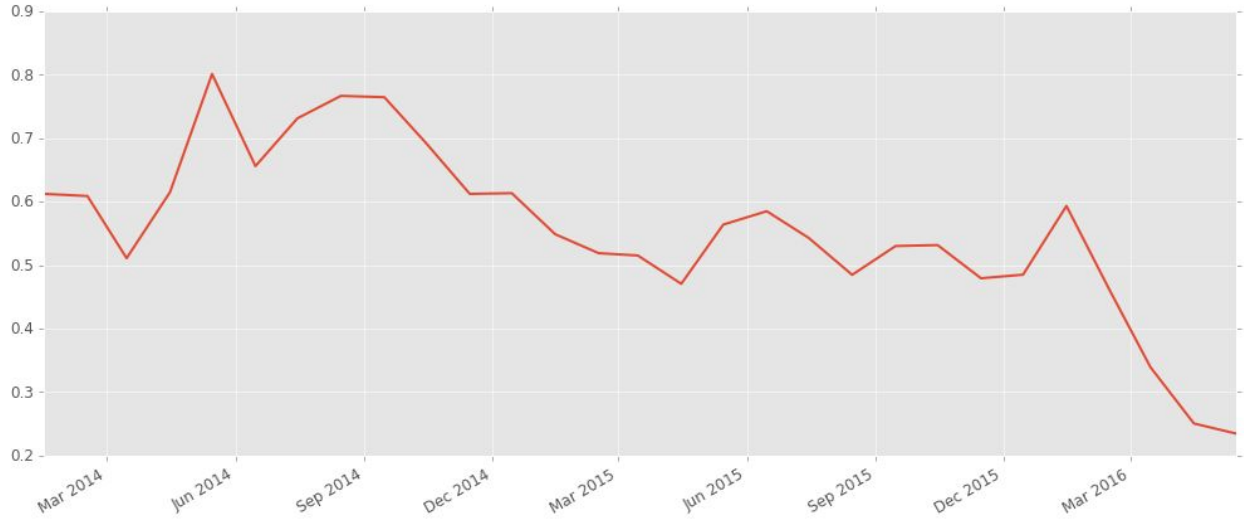


Нормализуем данные по доле мероприятий данной категории в данном месяце. В целом в месяц из всех культурных событий большее всего проходит выставок, и меньше всего -- концертов.

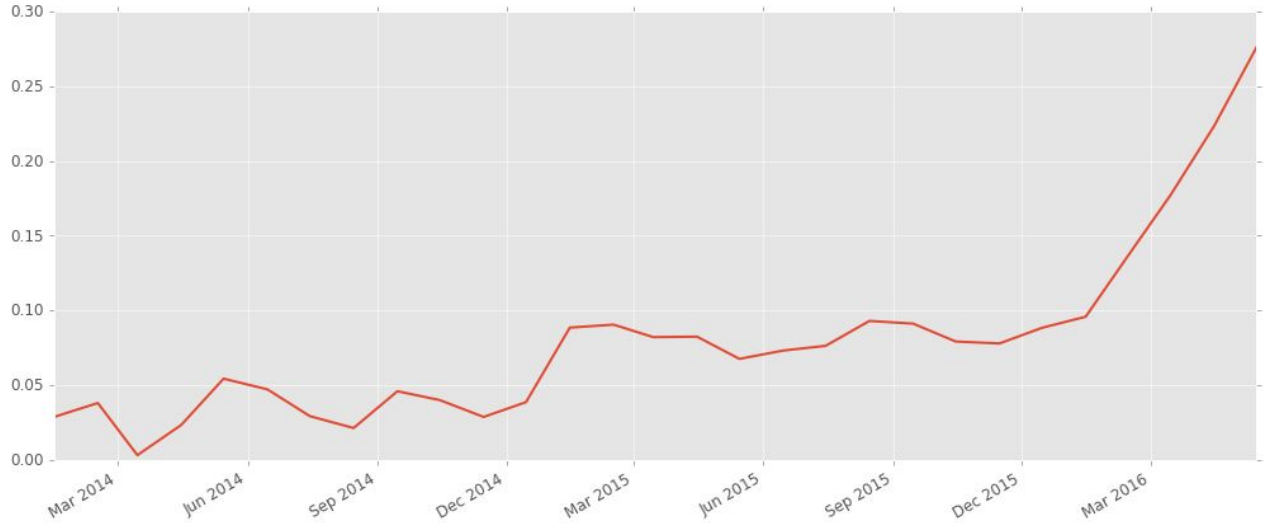


Можно заметить, что доля выставок с февраля 2016 г. стала снижаться, а доля лекций, мастер-классов, экскурсий и кинопоказов -- увеличиваться.

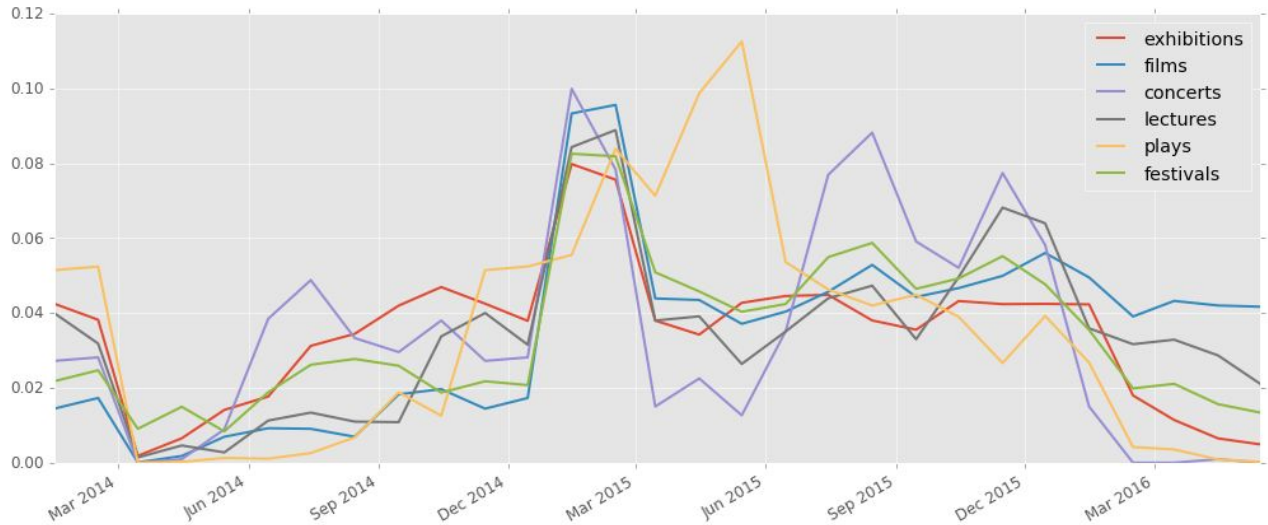
Доля выставок:



Доля кинопоказов:



Теперь нормализуем данные по доле мероприятий данной категории в данном месяце относительно общего количества мероприятий данной категории.



Как видно, в мае - июне 2015 г. было неожиданно много спектаклей; общее увеличение событий в декабре 2014 г. - январе 2015 г. должно быть связано с пред- и посленовогоднем временем.

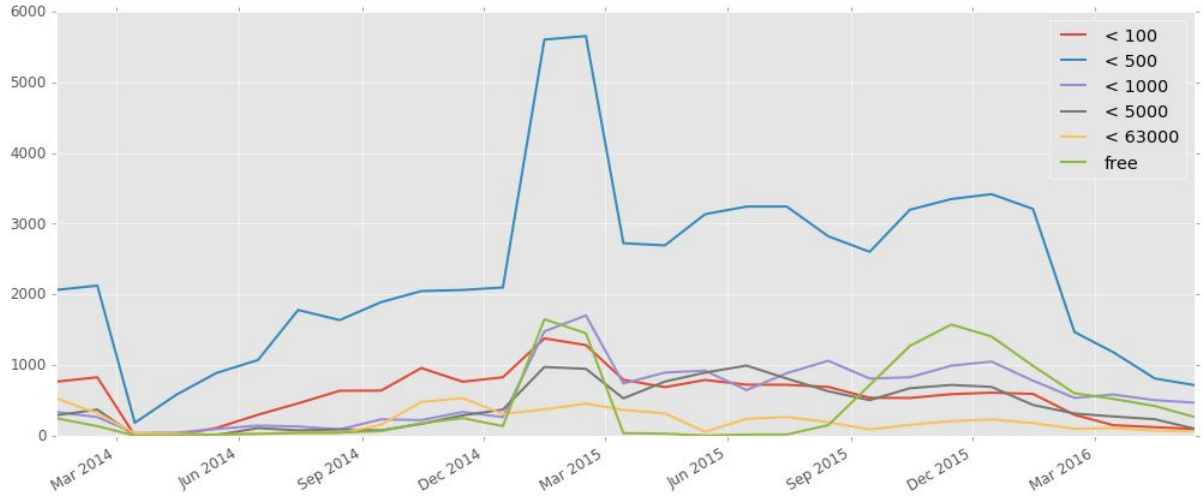
*

festivals = фестивали
concerts = концерты
lectures = лекции, мастер-классы и экскурсии
plays = спектакли
films = кинопоказы
exhibitions = выставки

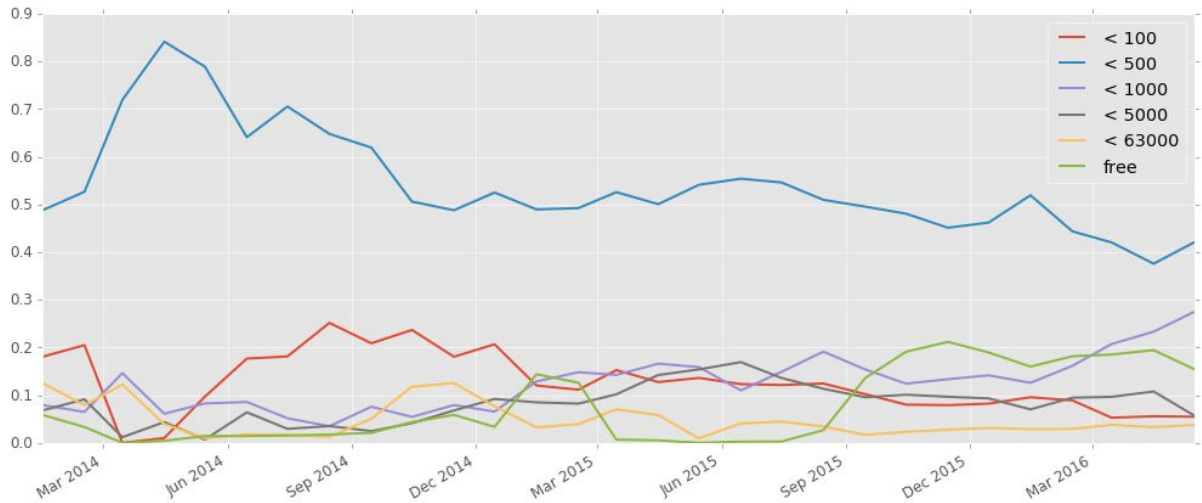
2) Распределение стоимости билета на культурные мероприятия по месяцам

Строим аналогичные рассмотренным выше графикам, только теперь на графиках вместо шести различных категорий культурных событий представлены шесть ценовых групп, отражающих стоимость билета на мероприятия: 1) бесплатно, 2) стоимость билета составляет менее 100 рублей, 3) от 100 до 500 рублей, 4) от 500 до 1000 рублей, 5) от 1000 до 5000 рублей, 6) более 5000 рублей (63000 - это максимальная цена билета за мероприятие).

Стоит сказать, что это очень условное приближение реальной ценовой ситуации, потому что в качестве цены бралось среднее значение от всех цен указанных в объявлении. А в объявлениях говорится и о абонементах, и о льготных билетах, и о цене за билет при покупке до определенного числа, и о ценах на путевки и продолжительные (недели, месяцы) курсы… Кроме того, здесь не учитывались те случаи, когда в тексте объявления прямо не сказано, что мероприятие бесплатно, но требуется регистрация или, например, клубная карточка, и когда мероприятие (чаще всего выставка) бесплатна для посещения при предъявлении билета, например, в музей.

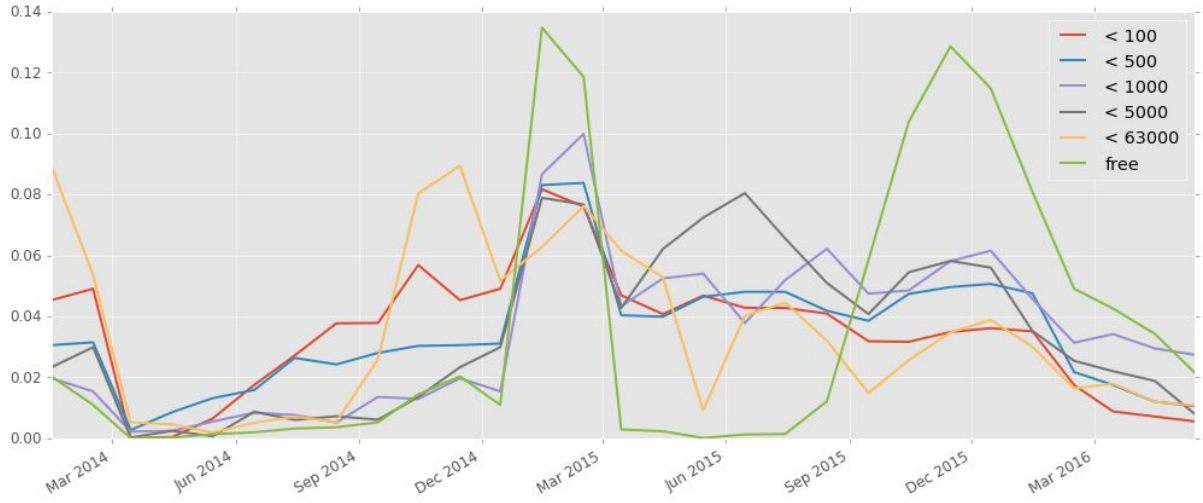


Нормализуем данные по доле данной ценовой категории в данном месяце.



Доля событий со стоимостью билета в 100 и менее рублей постепенно снижается (если не принимать во внимание апрель 2014 г., когда данных в принципе практически не было). Также постепенно снижается доля мероприятий со стоимостью билета от 100 до 500 рублей и увеличение доли мероприятий со стоимостью билета от 500 до 1000 рублей. С осени 2015 г. наблюдается установление доли бесплатных событий на уровне 18% в среднем (хотя колебания доли бесплатных мероприятий может быть связано с неполнотой данных сайта или с тем, что не все бесплатные мероприятия распознаются как бесплатные).

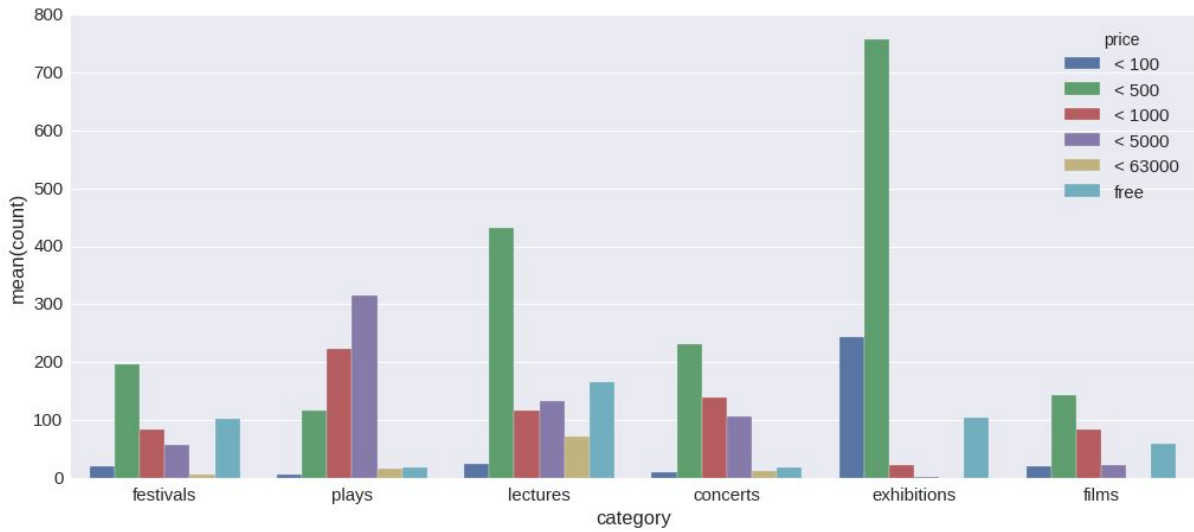
Нормализуем данные по доле данной ценовой группы в определенном месяце относительно суммарного количества событий данной ценовой группы. Замечаем, что бесплатных мероприятий больше всего в периоды зимних праздников, а увеличение событий ценовой категории от 5000 рублей часто приходится на дни школьных каникул и праздников.



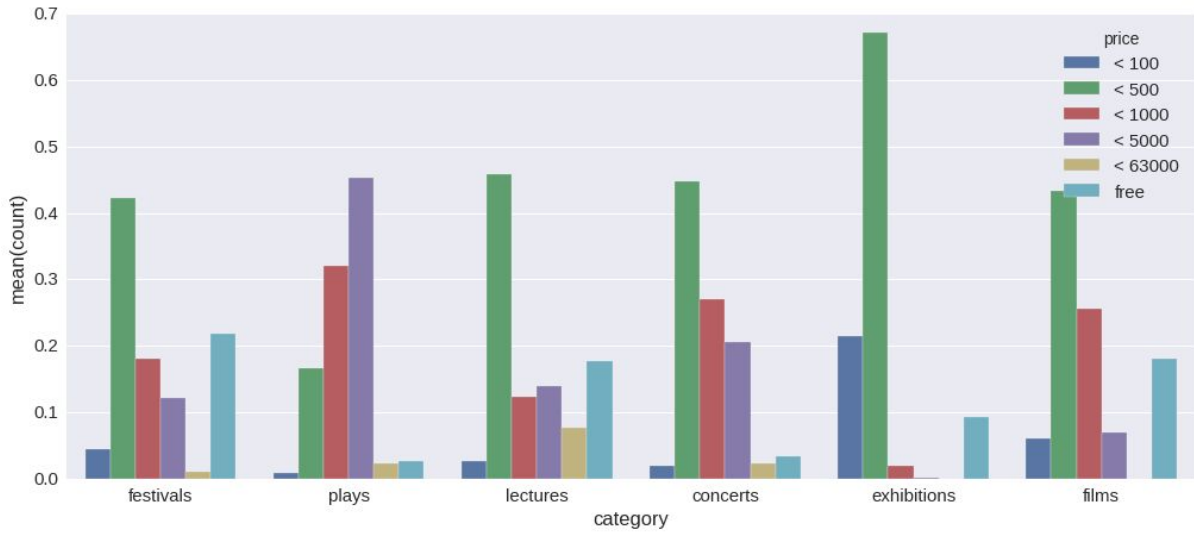
3) Стоимость билетов vs тип культурных событий

Следующие столбиковые диаграммы показывают связь между категорией и ценовой группой мероприятия.

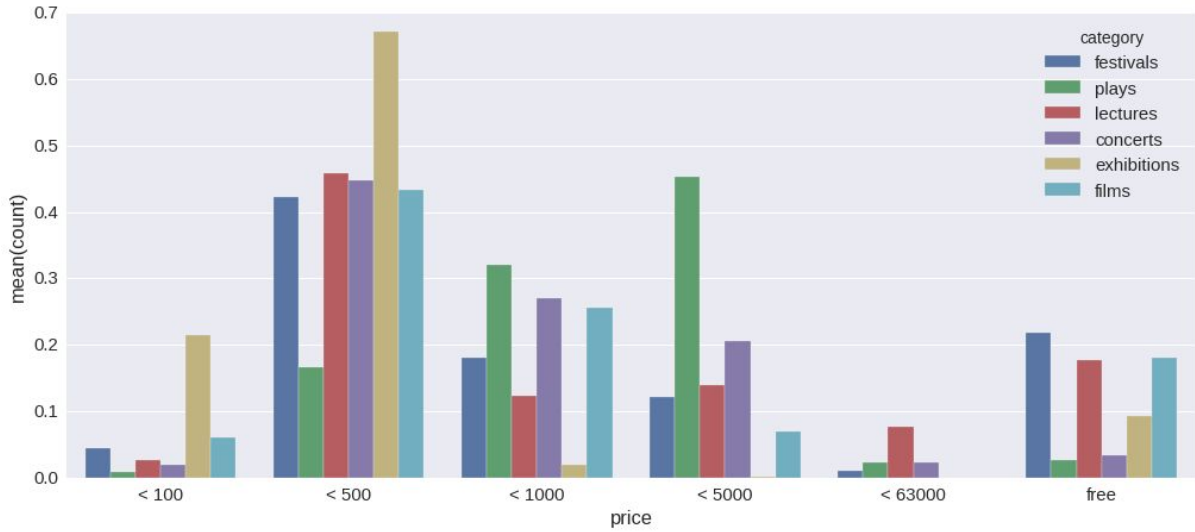
“Сырые” данные показывают, что среди всех событий больше всего выставок со стоимостью билета от 100 до 500 рублей, а выставок и фильмов с билетом более, чем в 5000 рублей, вообще не случалось.



Нормализованные данные говорят, что среди лекций, мастер-классов и экскурсий, фестивалей и кинопоказов наблюдается наибольшая доля бесплатных событий -- 20(±2)%, а наименьшая -- у спектаклей и концертов, что вполне соответствует ожиданиям. Наибольшая доля самых дешевых билетов у выставок -- 21%, наибольшая доля самых дорогих билетов (более 5000 рублей) у лекций, мастер-классов и экскурсий -- 8% (что, вероятно, связно с высокой ценой на различные курсы и тренинги). Доля событий с ценой билета от 100 до 500 рублей превалирует во всех категориях, кроме спектаклей, где наибольшая доля принадлежит мероприятиям с ценой билета от 1000 до 5000 рублей.



В дополнение приведем столбиковую диаграмму, “обратную” представленной раннее, на которой можно проследить те же самые закономерности.



4) Частотные слова в виде облака тегов

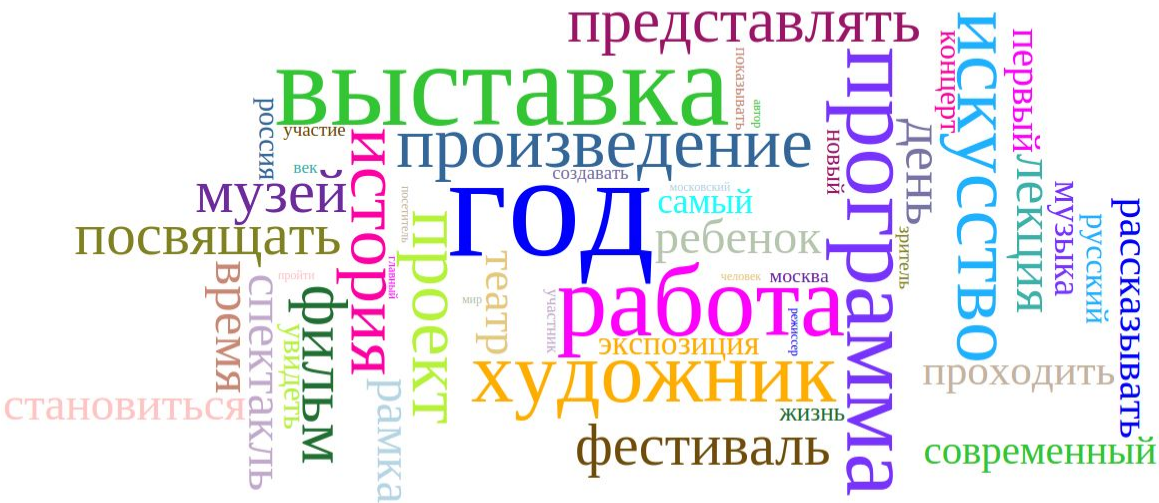
Все тексты объявлений были предобработаны (токенизация, лемматизация, удаление стоп-слов -- служебных частей речи) и сгруппированы в подкорпуса по категориям культурных событий. Далее с помощью ресурса Voyant Tools[3] были построены облака тегов для каждого подкорпуса.

Voyant Tools позволяет делать поиск по подкорпусам, различные визуализации, просматривать коллокации определенных слов и многое другое. Несмотря на огромные возможности этого ресурса, мы не используем его в нашей работе (за исключением иллюстрации частотных слов), но предлагаем заинтересовавшимся загрузить наш корпус текстов объявлений в Voyant Tools и поэкспериментировать[4]. Ещё одним интересным ресурсом является Lexos, который помимо предобработки, статистики и визуализации предлагает также различные алгоритмы кластеризации[5].

Что касается категорий, то данные сайта ‘2go2go’ содержат 136 уникальный комбинаций рубрик (например, 'Развлечения Игры Игры для компаний Квесты', 'Экстремальные развлечения Распродажи', 'Экстремальный спорт Водный спорт Фестивали Распродажи'), которые, как отмечалось ранее, не организованы иерархически в отличие от категорий сайта ‘Культура Москвы’, где имеется 6 категорий “верхнего уровня”: ‘Концерты’, ‘Фестивали’, ‘Выставки’, ‘Лекции, мастер-классы и экскурсии’, ‘Кинопоказы’, ‘Спектакли’. Поэтому все тексты сайта ‘2go2go’ были сгруппированы в более общие категории (такое разбиение не идеально, но что-то лучше придумать не удалось): ‘Выставки’, ‘Спектакли’, ‘Развлечения Концерты Шоу Фестивали Игры Ярамарки и др.’, ‘Лекции Курсы Обучение Тренинги Мастер-классы Встречи и др.’, ‘Спорт Квесты Прогулки Экскурсии и др.’.

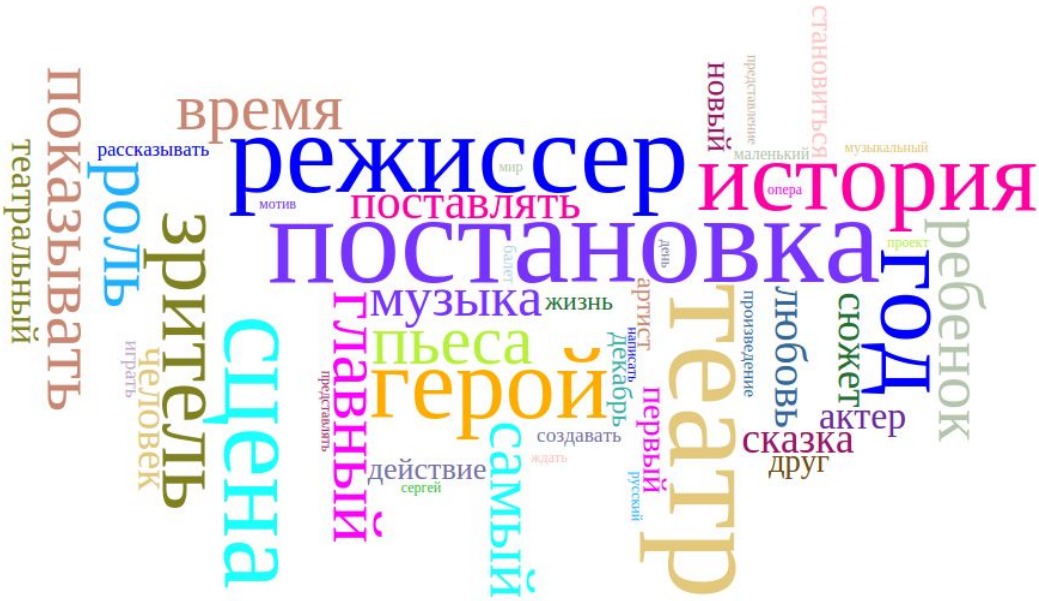
Сайт ‘Культура Москвы’

- Выставки

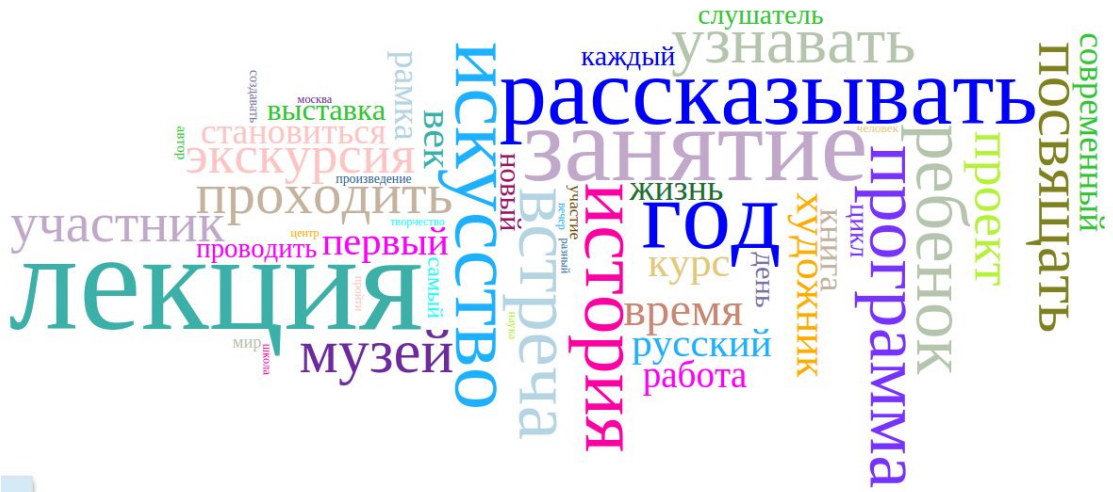


[3] http://voyant-tools.org/
[4] Инструкцию можно найти по ссылке http://voyant-tools.org/docs/#!/guide/about
[5] http://lexos.wheatoncollege.edu/upload

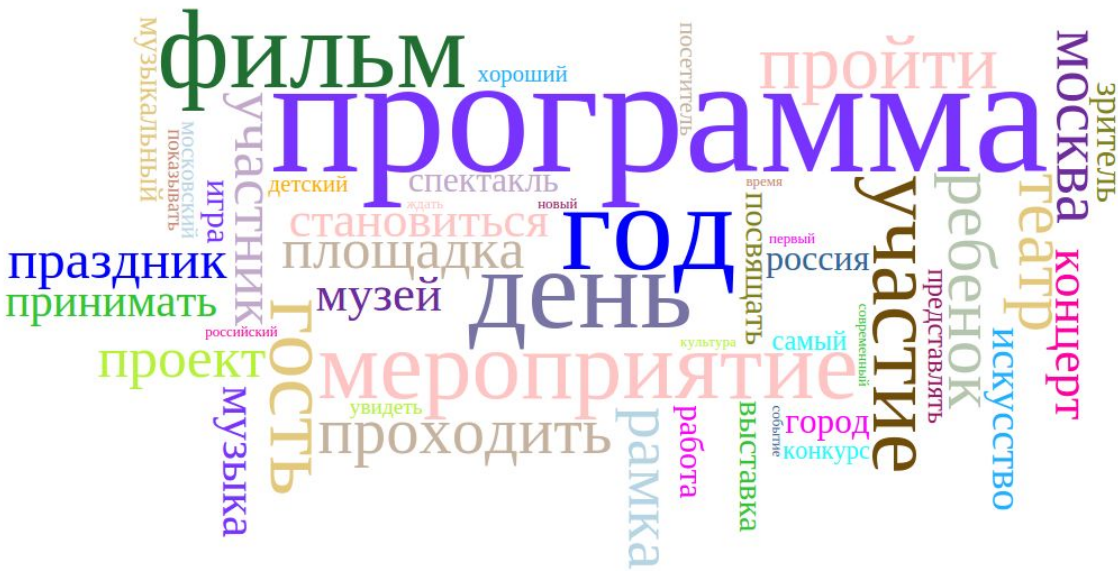
- Спектакли



- Лекции, мастер-классы и экскурсии

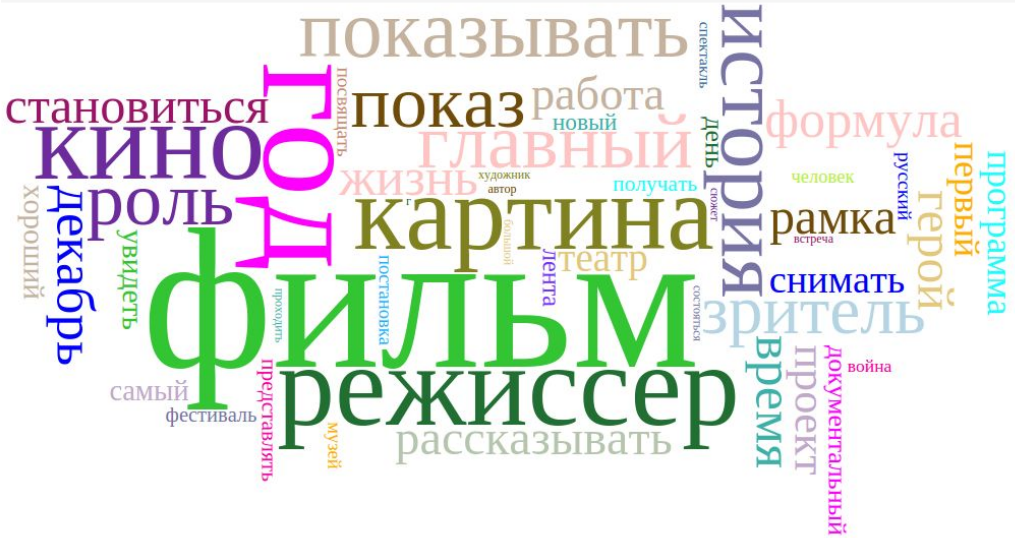


- Фестивали



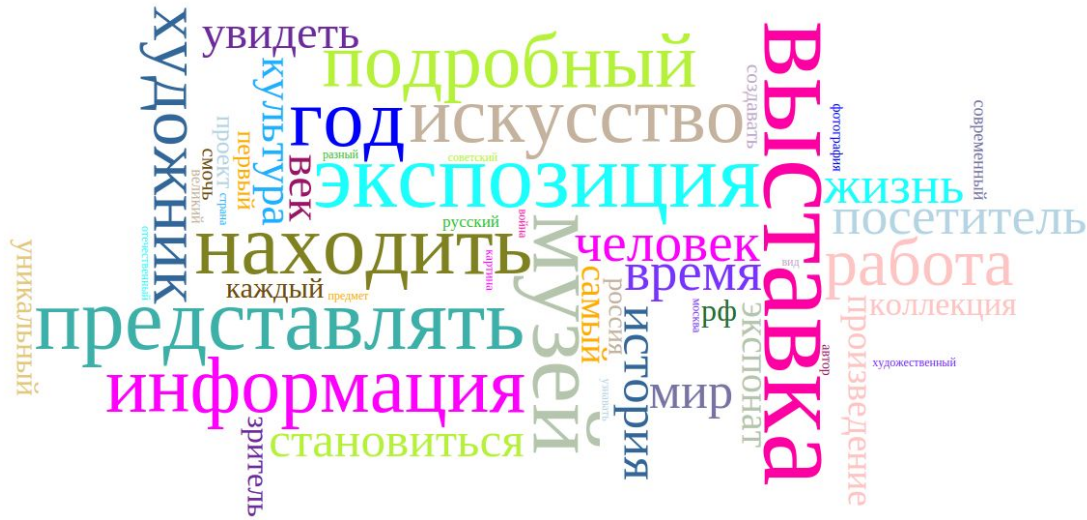
- Концерты

- Кинопоказы

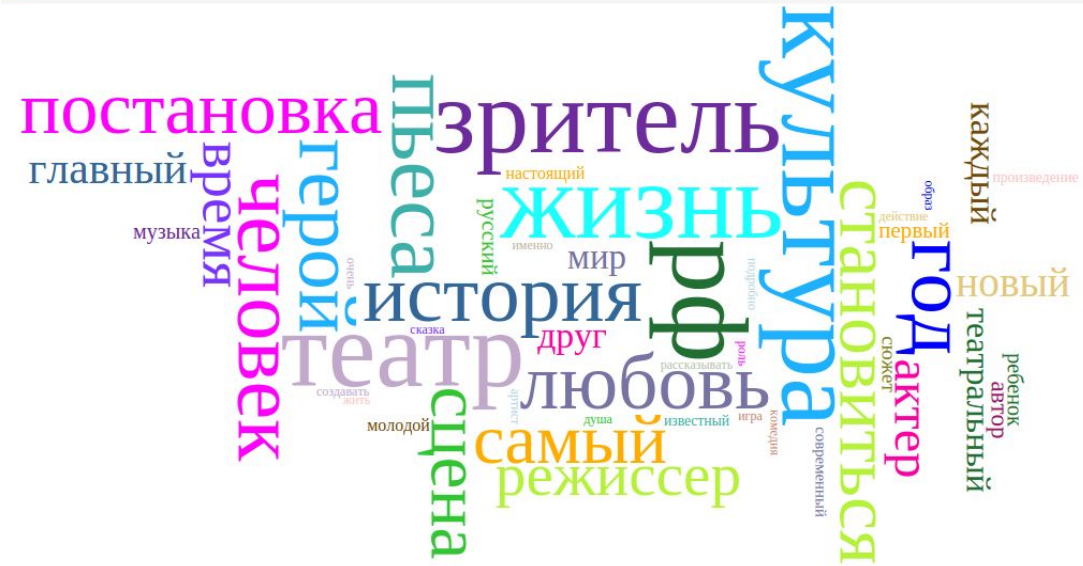


Сайт ‘2go2go’

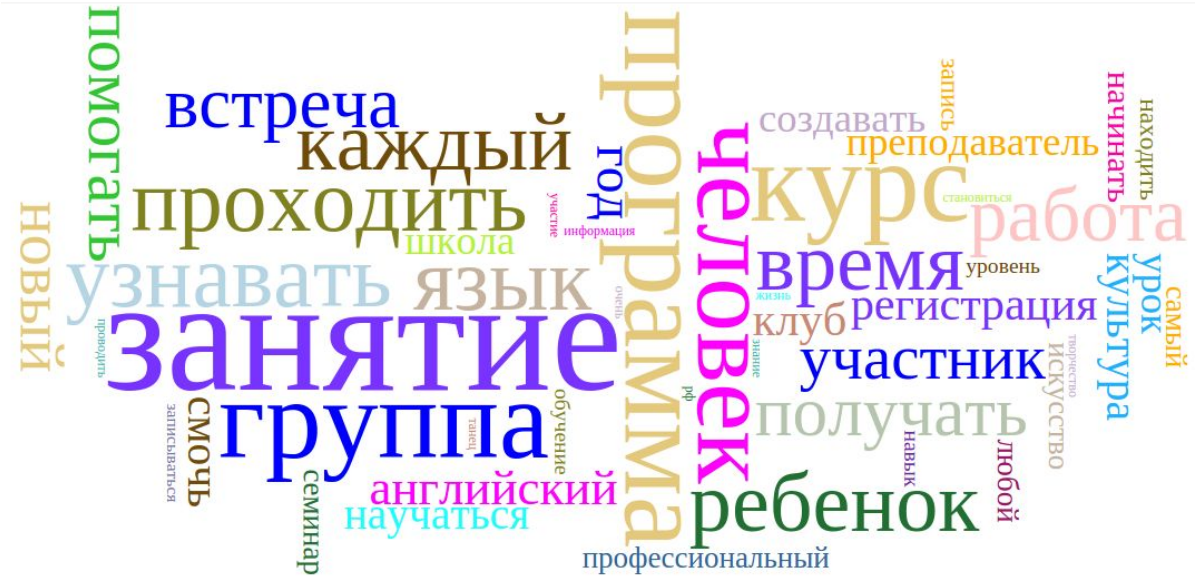
- Выставки



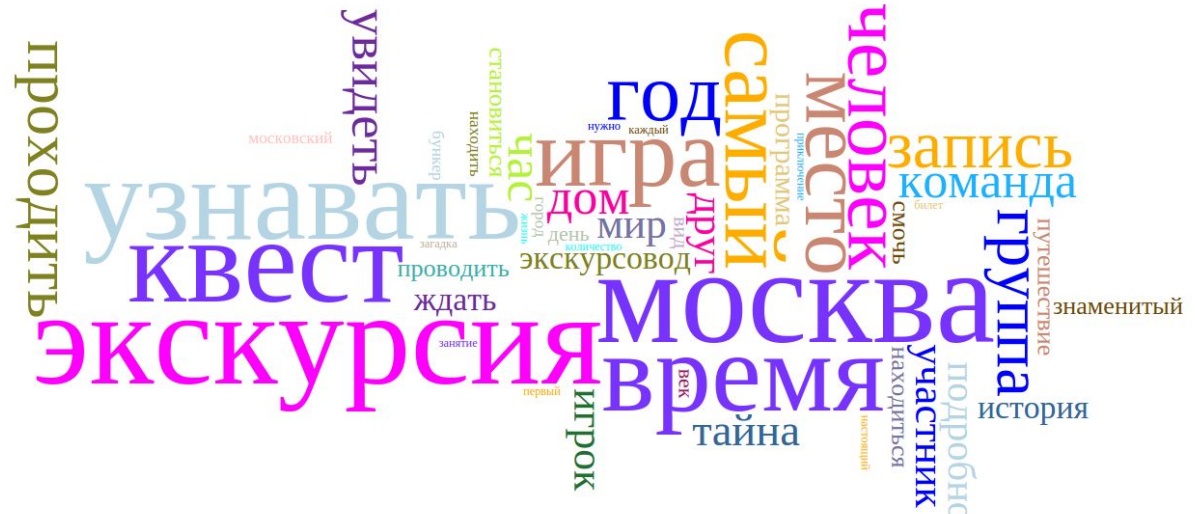
- Спектакли



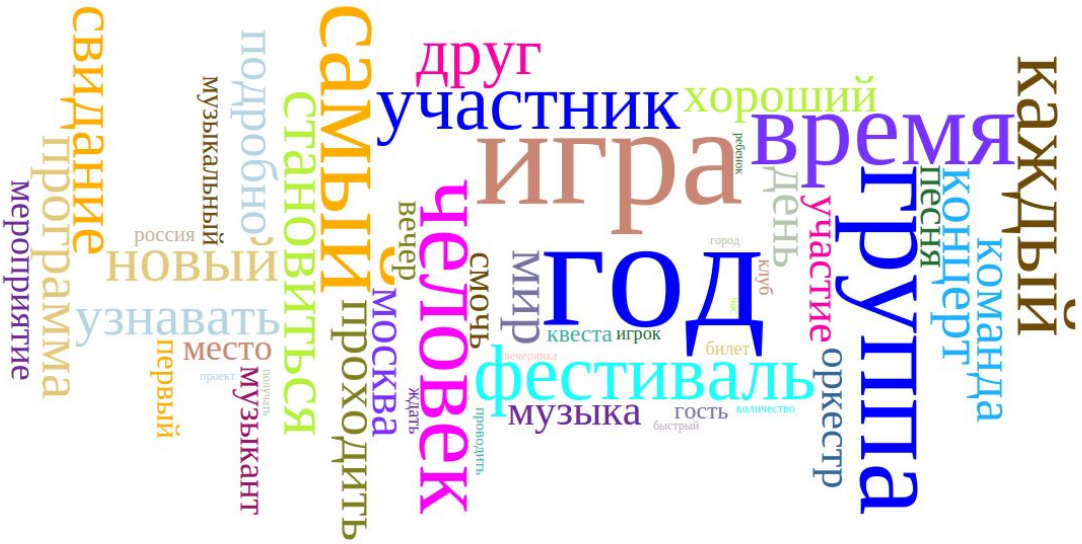
- Лекции Курсы Обучение Тренинги Мастер-классы и др.



- Спорт Квесты Прогулки Экскурсии и др.



- Развлечения Концерты Шоу Фестивали Игры Ярмарки и др.



Высокую частотность слова *год* можно объяснить тем, что в описании событий очень часто указывается, что что-то произошло в таком-то году.

5) Кластеризация текстов объявлений

Для задачи кластеризации использовались предварительно предобработанные тексты объявлений сайта “Культура Москвы”.

В компьютерной лингвистике в задачах обучения без учителя на текстовых данных есть два принципиально разных подхода. Во-первых, собственно кластеризация (k-means и другие алгоритмы), которая предлагает отнесение каждого документа к ровно одному из кластеров. Во-вторых, LDA[6]-подобные алгоритмы, которые каждому документу соотносят набор кластеров (или, как принято их называть, топиков) с их вероятностями. Идея модели LDA исходит из того, что документ может содержать несколько тем (и поэтому каждому документу соответствует не один кластер, а набор топиков).

Рассмотрим сначала первый подход. Здесь, как и при LDA анализе, требуется задать число кластеров (при LDA -- топиков). Будем использовать реализацию k-means кластеризации библиотеки scikit-learn[7]. Мы знаем, что все тексты объявлений сайта ‘Культура Москвы’ распределены на шесть категорий первого уровня. Проверим, как данные тексты автоматически распределяться на 6 категорий (т.е. кластеризуются на 6 кластеров).

Мы провели несколько экспериментов с параметрами модели, в результате получилось, что самое удачное распределение на кластеры происходит при использовании униграмм при подсчете матрицы tf-idf, а не би- и триграмм, как это бывает для более длинных текстов (новостных статей и др.).

Кроме того, было посчитано, сколько текстов какой категории содержится в каждом из кластеров. Результаты представлены ниже в таблице и на столбиковой диаграмме.

В целом получившиеся кластеры отражают исходные категории объявлений. Единственным исключением является категория “Фестивали”, которая, как оказалась, очень близка к категории “Лекции”. Кроме того, поскольку на каждом этапе работы алгоритма k-means есть элемент случайности, при обучении модели иногда получается, что мы имеем два похожих кластера. Списки топ-5 слов этих двух кластеров могут быть, например, такими:

- ➔ лекция занятие курс искусство рассказывать
- ➔ лекция ребенок фестиваль день программа.

Заметим, что категория “Фестивали” занимает “второе” место и в доле кластеров, соответствующим категориям “Концерты” и “Кинопоказы”, что вполне закономерно: в Москве часто проходят различные фестивали кино и музыки.

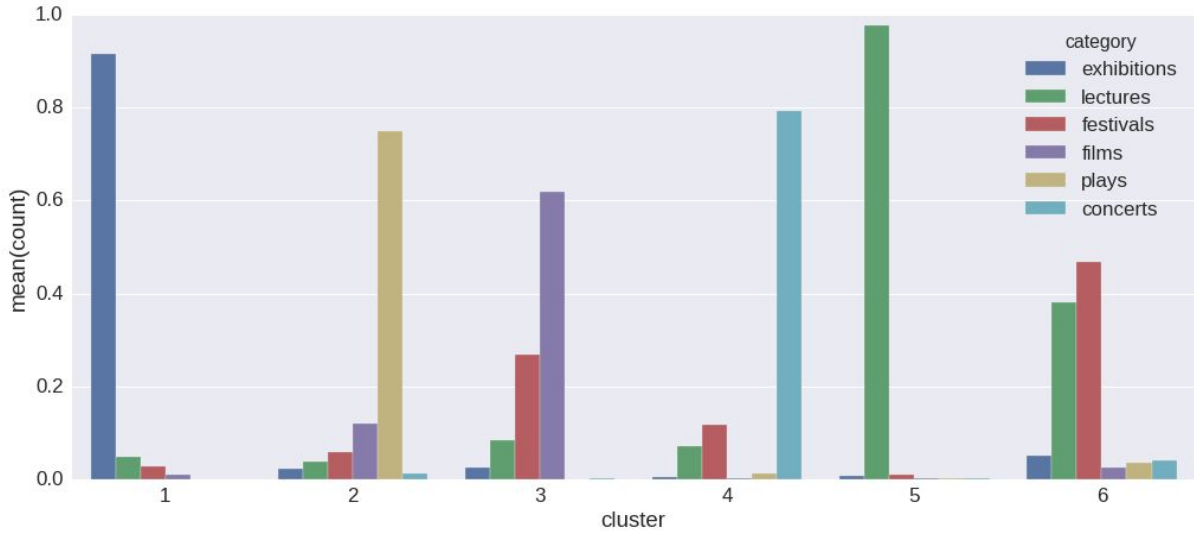
Самый “чистый” кластер -- это кластер, соответствующий категории “Выставки”.

---

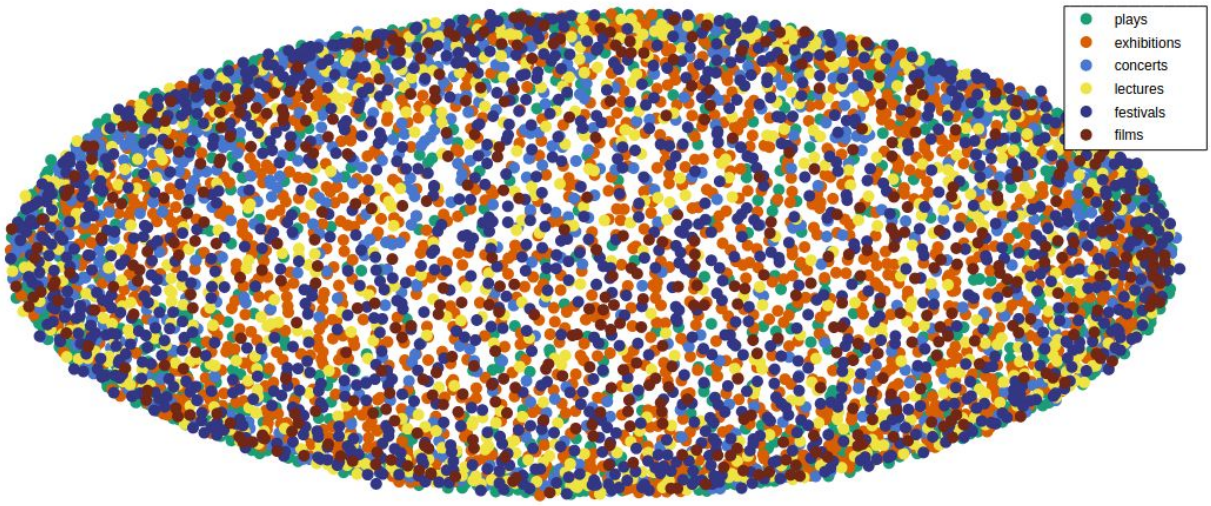
[6] Латентное размещение Дирихле

[7] http://scikit-learn.org/stable/

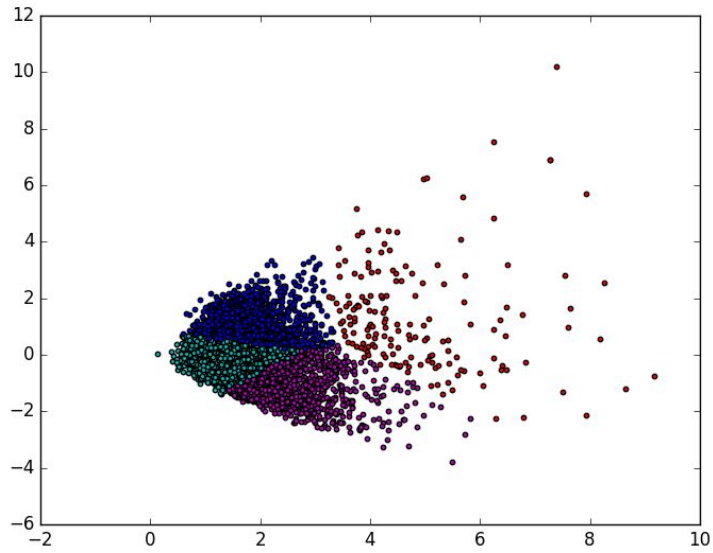
N	Top-10 близких к центру кластера слов	Категории кластера
1	выставка художник работа экспозиция искусство год представлять фотография музей произведение	выставки: 1388 лекции: 71 фестивали: 40 кинопоказы: 14 спектакли: 1 концерты: 1
2	спектакль театр постановка режиссер сцена герой роль пьеса зритель декабрь	спектакли: 694 кинопоказы: 110 фестивали: 55 лекции: 35 выставки: 20 концерты: 12
3	фильм кино картина показ режиссер год фестиваль документальный зритель лента	кинопоказы: 307 фестивали: 133 лекции: 41 выставки: 13 концерты: 1
4	концерт музыка музыкальный композитор произведение программа вечер исполнять оркестр прозвучать	концерты: 551 фестивали: 83 лекции: 51 спектакли: 9 выставки: 3 кинопоказы: 1
5	лекция занятие курс искусство рассказывать узнавать слушатель ребенок история цикл	лекции: 768 фестивали: 8 выставки: 6 кинопоказы: 2 спектакли: 1 концерты: 1
6	фестиваль день программа гость участие ребенок мероприятие книга праздник участник	фестивали: 568 лекции: 463 выставки: 60 концерты: 48 спектакли: 43 кинопоказы: 31



Далее визуализируем результаты кластеризации и исходные категории с помощью многомерного шкалирования (используя реализацию библиотеки scikit-learn, а также модуля mpld3 -- matplotlib оболочки для библиотеки D3.js).

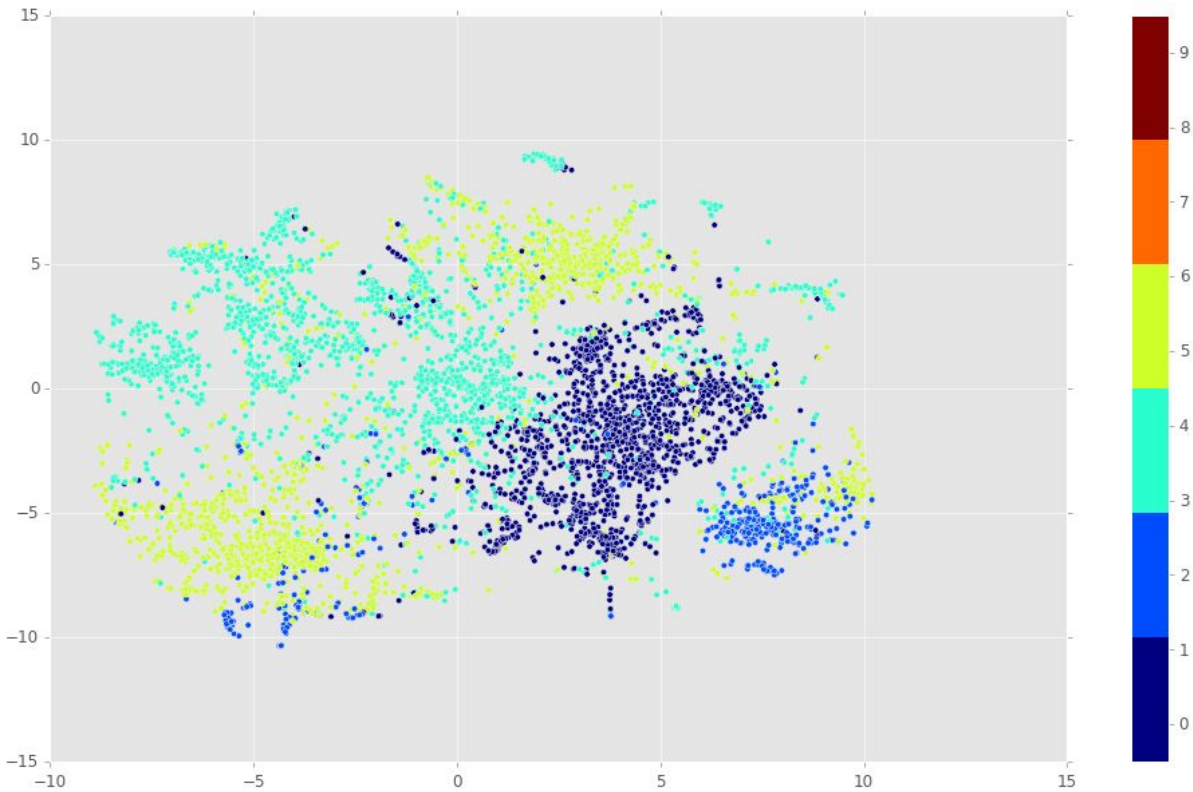


Представим себе, что мы бы не знали исходного количества категорий. Как выбрать число кластеров в таком случай? Существует несколько техник, одна из них -- это использование LSA (Латентно-семантический анализ) или, что то же самое, LSI (Латентно-семантическое индексирование). При таком подходе получилось, что наша коллекция документов содержит 4 кластера. Далее документы были кластеризованны всё тем же алгоритмом k-means. Как видно на визуализации, три из четырех кластеров выделяются достаточно четко.



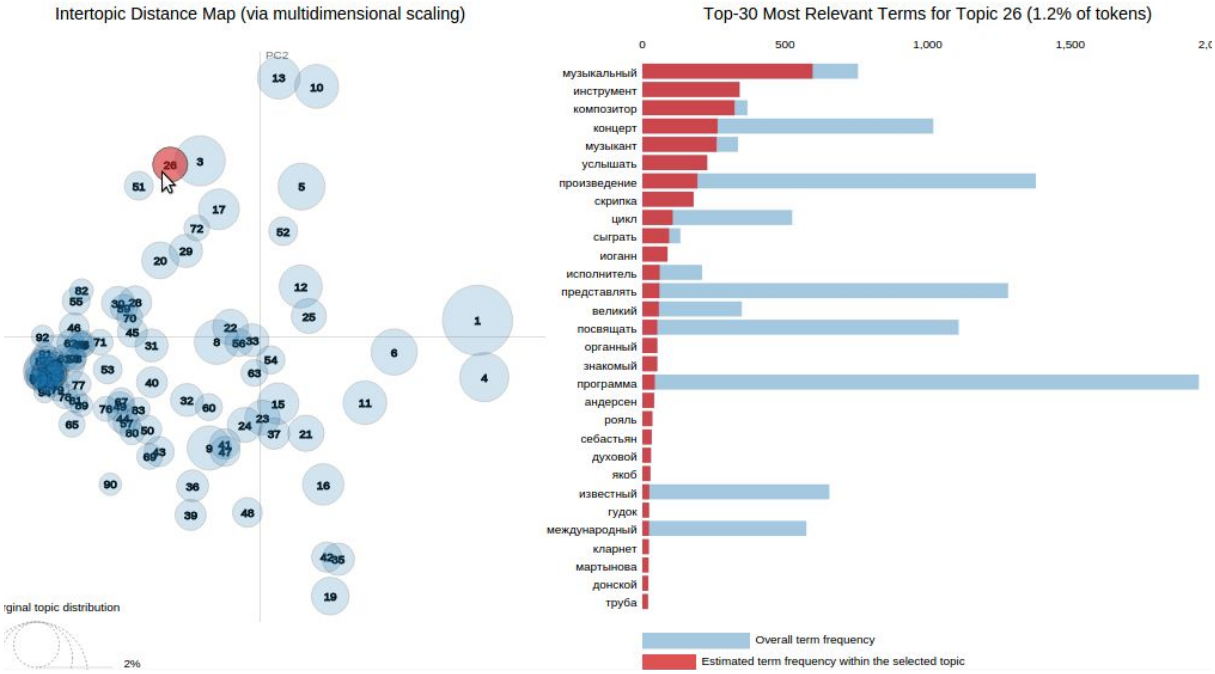
Если интересно провести LSA на данных корпуса объявлений сайта “Культура Москвы”, то предлагаем воспользоваться ранее упомянутым нами ресурсом Lexos. С помощью него для каждого документа Вы сможете просмотреть список наиболее близких к нему документов и соответствующие значения близости.

Мы также попробовали один из методов понижения размерности, а именно метод нелинейного снижения размерности t-SNE (t-distributed stochastic neighbor embedding), используя библиотеку scikit-learn. Результаты визуализации t-SNE подтверждают результаты нахождения количества кластеров с помощью LSA. Их снова получилось 4, хотя в качестве предполагаемый целевых кластеров задавали 6 исходных категорий.



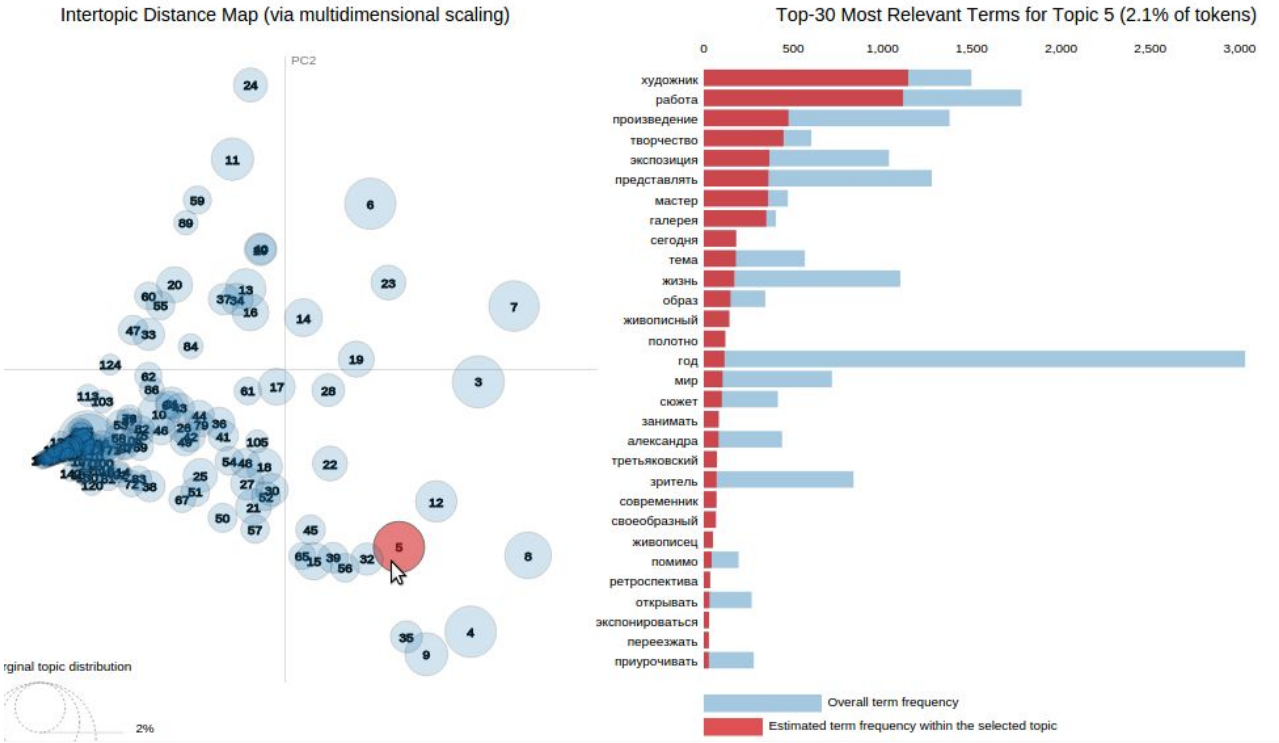
Перейдем к LDA. Основная сложность при построении модели LDA -- это задать оптимальное количество топиков, а также некоторые параметры модели так, чтобы перплексия (своеобразная мера качества модели, которая правда не всегда коррелирует с интерпретируемостью топиков). На данных текстов объявлений получалась такая закономерность: чем больше топиков, тем ниже перплексия. Однако при большом количестве топиков, появляется все больше топиков, содержащих, например, название месяцев, единиц измерения, город или других подобных объектов.

Лучшая модель (выбор делался на основании как значения перплексии, так и субъективной оценки получившихся топиков) имеет 100 топиков. Библиотека pyLDAvis позволяет изучить карту расстояния между топиками (строится с помощью многомерного шкалирования), а также проанализировать для каждого топика топ-30 слов (в сравнении с их частотой на всей коллекции документов).



Пример того, как меняется карта расстояния между топиками, при увеличении их числа до 150 ниже. При дальнейшем увеличении числа топиков карта расстояния между топиками все больше напоминает визуализацию результатов k-means кластеризации при четырех кластерах. Принимая во внимание визуализацию t-SNE и результаты k-means кластеризации при шести кластерах, можно сделать вывод, что две из шести исходных категорий неоднородны, и своим содержанием они отчасти напоминают остальные категории (т.е. не всегда их легко выделить и разграничить). Одной из таких категорий однозначно является категория “Фестивали” (это обсуждалось выше), второй категорией (но в меньшей степени “нечеткой”), кажется, является категория “Кинопоказы”. Действительно, описания кинопоказов в зависимости от содержания могут быть похожи на описания спектаклей или экскурсий

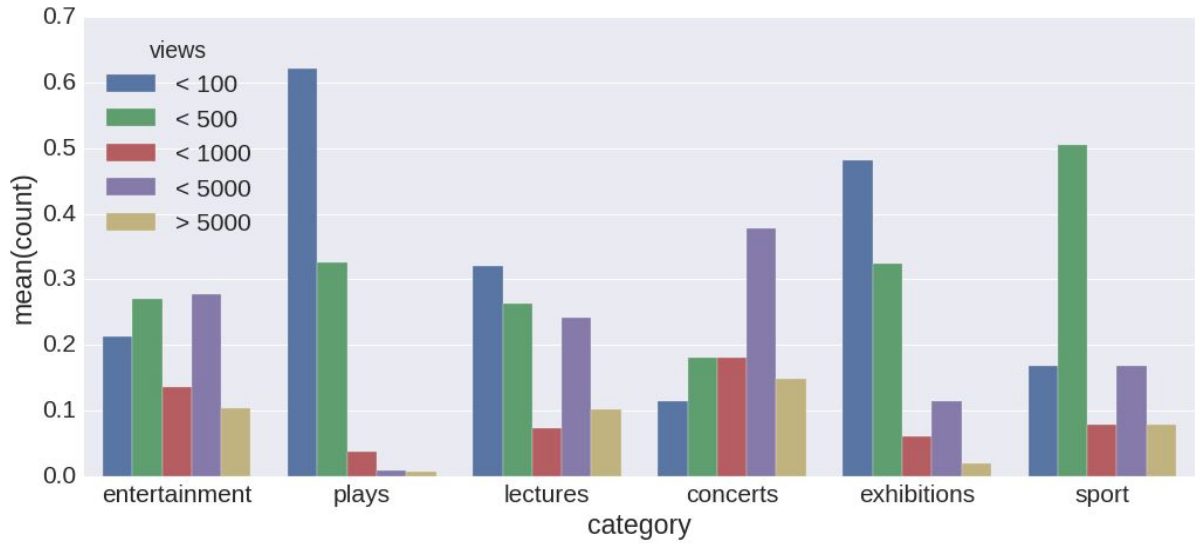
(последние относятся к категории “Лекции”). Кроме того, причиной того, что категория “Кинопоказы” несколько размыта, может служить небольшое количество соответствующих ей текстовых описаний.

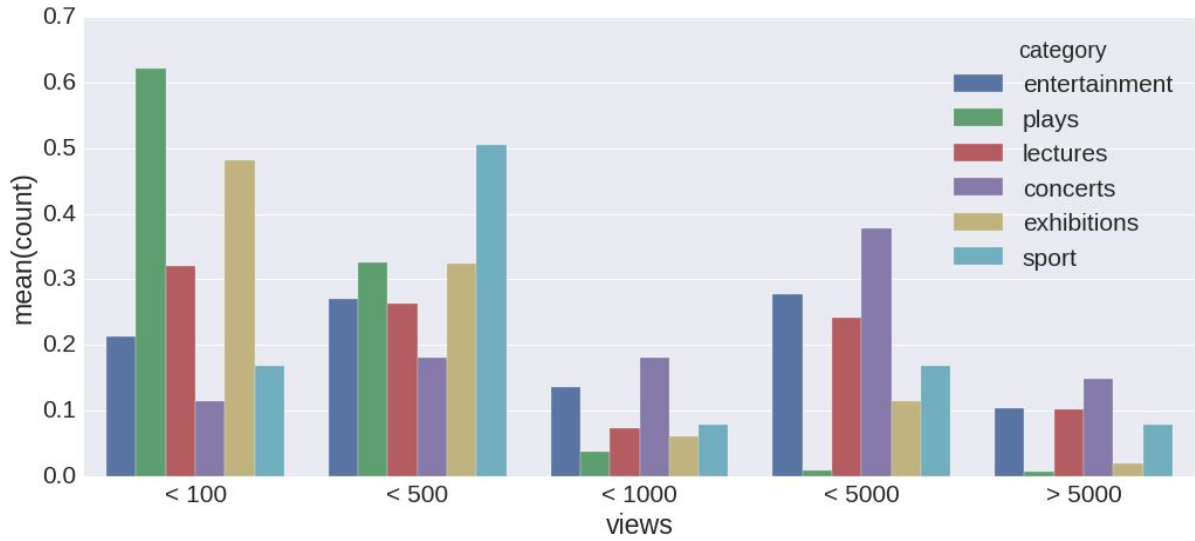


6) Популярные культурные мероприятия

На сайте ‘2go2go’ имеется доступ к информации о количестве просмотров каждого из объявлений. Эти значение были сгруппированы в пять категорий: 1) количество просмотров менее 100, 2) количество просмотров от 100 до 500, 3) количество просмотров от 500 до 1000, 4) количество просмотров от 1000 до 5000, 5) количество просмотров более 5000 (заметим, что в последнюю категорию попадают и весьма “популярные” мероприятия с количеством просмотров более 100000, но таких лишь несколько).

Нормализуем данные, строим столбиковую диаграмму и получаем следующие наблюдения: самая большая доля мало просматриваемых мероприятий приходится на спектакли -- 62% (мало просматриваемые мероприятия “лидируют” также в общей доле категорий “Лекции, мастер-классы, обучение, курсы” и “Выставки”, однако не с таким отрывом), самая большая доля “популярных” (много просматриваемых) событий -- среди объявлений категории “Концерты и фестивали” (несколько в меньшей степени “популярными” можно назвать мероприятий категорий “Развлечения, шоу, игры”, “Лекции, мастер-классы, обучение, курсы” и “Спорт, квесты, прогулки”).





III. Заключение

Анализировать культурную жизнь города, основываясь на данных того или иного веб-сайта, можно, но чтобы это исследование было надежным (в той степени, насколько об этом можно говорить), необходимо сопоставлять данные нескольких источников, тем самым сводя к минимуму “погрешности” того или иного источника.

Конечно, было бы интересно

1) посмотреть на более длительный временной промежуток;
2) сравнить, например, культурную жизнь Москвы и Санкт-Петербурга, или Москвы и Казани

но в настоящий момент в открытом доступе не имеется подходящих данных.