Анализ тем на TED

Алексей Лобанов

6 февраля 2016 г.

Содержание

1 Введение

В данном исследование рассмотрим некоторые особенности категоризации роликов на TED, найдём наиболее интересные пользователям темы и ролики

2 Порядок работы

С помощью scrapy был написан парсер, который за несколько часов собрал информацию по всем роликам. Я получал следующие данные:

1. Название ролика.
2. Список тегов.
3. Численное значение (не в виде процентов, как на сайте) каждой из зрительских оценок.
4. Количество просмотров.

Как источник тем для видео возьмём теги. В самом деле, если эти видео уже категоризованы, то лучше этим воспользоваться Всего 2133 видео и 357 тегов.

2.1 Обзор тегов и их очистка

Построим зависимость числа тегов, которые встретились хотя бы n раз от n. Можно предположить, что будет гипербола. График 1 это подтверждает.

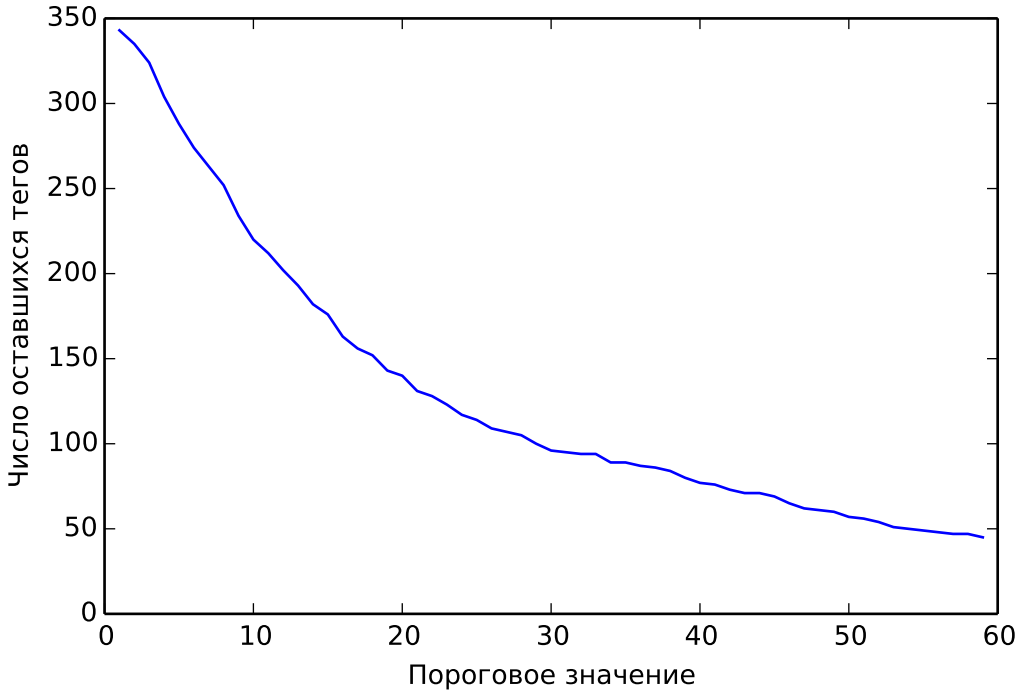

Рис. 1: Зависимость числа тегов от порогового значения

Давайте более детально проанализируем теги. Их достаточно мало, чтобы это можно было сделать вручную. Заметим, что среди них встречаются такие, которые не относятся к тематике ролика, например, *TED Conference* встретился 709 раз (33% от всех видео). По этим причинам, давайте удалим следующие теги: TED Conference, TEDx, happiness (слишком общее), TED Brain Trust, data (слишком общее), TED Prize, TED-Ed, TEDYouth, choice (слишком общее), materials (слишком общее), MacArthur grant, TED Books, TEDMED.

Заметим также, что существует значительное число тегов, которые встречаются слишком редко, чтобы можно было считать их значимыми. В качестве порога было выбрано число три. Если тег встречается реже, чем в *трёх* роликах, то мы его удаляем. Таким образом, будут удалены следующие теги: Moon, code, mobility, cloud, programming, capitalism, pandemic, urban, mining, TEDMED (этот тег редко встречается и не является темой), nuclear weapons, microsoft, Brand, testing, skateboarding, origami, vulnerability, evil, South America, glacier, cyborg, painting.

После удаления этих тегов, их осталось 323.

2.2 Объединение тегов

Рассмотрим две функции:

- Расстояние Жаккара:

$$\rho_J(A, B) = 1 - \frac{|A \cup B|}{|A \cap B|}$$

  Это классическое расстояние поможет определить максимально близкие теги.

- Мера Шимкевича-Симпсона:

$$\rho'(A, B) = 1 - \max\left(\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}\right)$$

  Такая функция расстояния удобна для определения того, что какое-то множество *почти* включает другое

Мне кажется разумным выдвинуть следующие гипотезы по тегам:

1. Есть очень много *похожих* тегов, которые можно будет объединить.
2. Некоторые теги могут содержать другие почти полностью, то есть может существовать некая иерархия.

Начать проверку можно с первой. В таком случае будет удобно рассматривать расстояние Жаккара. Тогда есть лишь одна пара тегов, для которой данное расстояние меньше 0.5, это пара *meditation* и *mindfulness*. Каждому из этих тегов принадлежат 4 одинаковых видео. Таким образом, можно удалить один из них, например, mindfulness. Это гипотеза неверна.

Рассмотрим вторую гипотезу. Есть 29 пар тегов, для которых один из них полностью содержит другой, например global issues и population. Построим граф 2, показывающей иерархию тегов (рассматриваются с расстоянием Жаккара не более 0.15). Видно, что получилось несколько больших групп с центрами в тегах

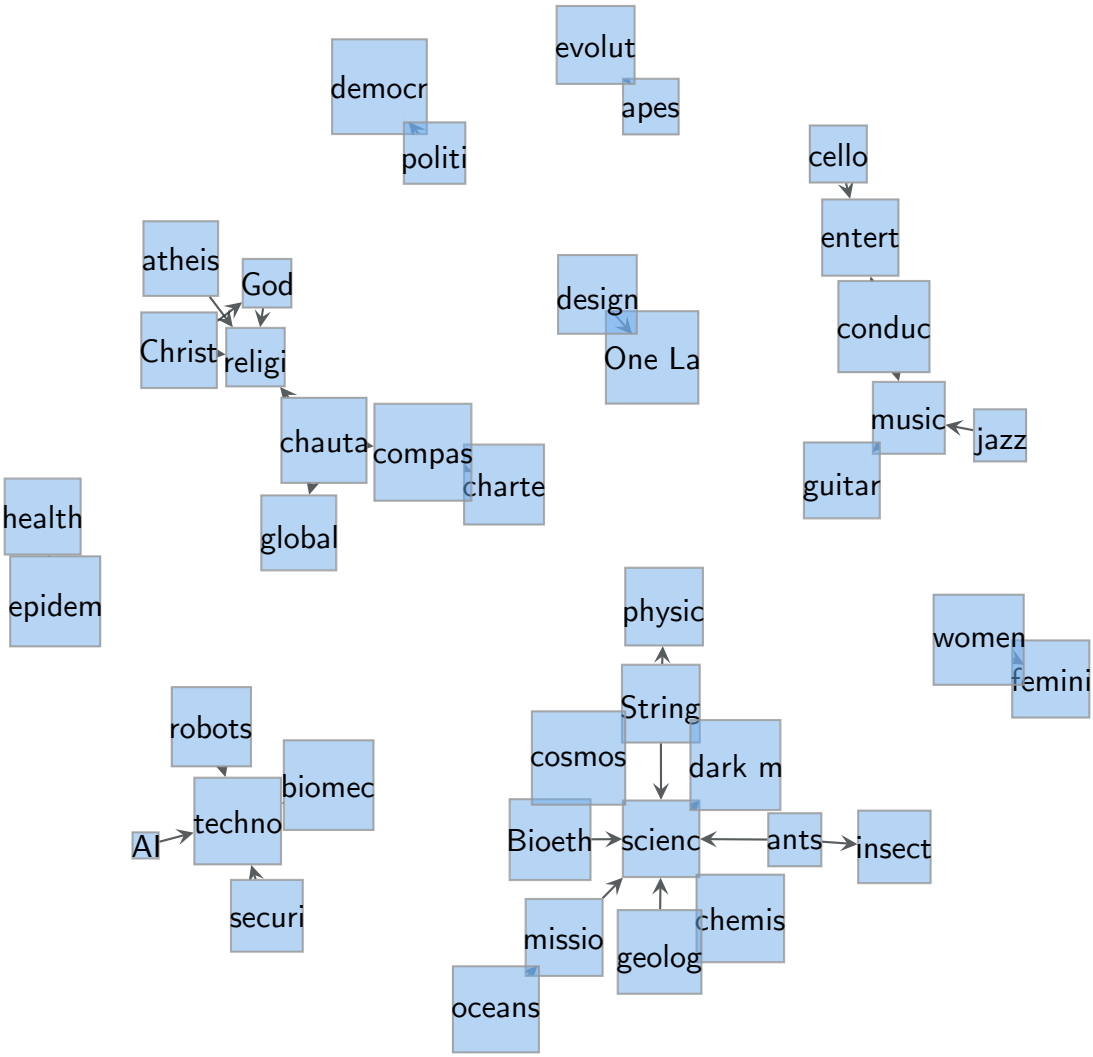

Рис. 2: Иерархия тегов

technology, science, music. Таким образом, можно считать эту гипотезу верной.

2.3 Анализ просмотров

Среднее число просмотров 1433164, медиана 973111. Построим распределение числа просмотров 3. Самые популярные ролики:

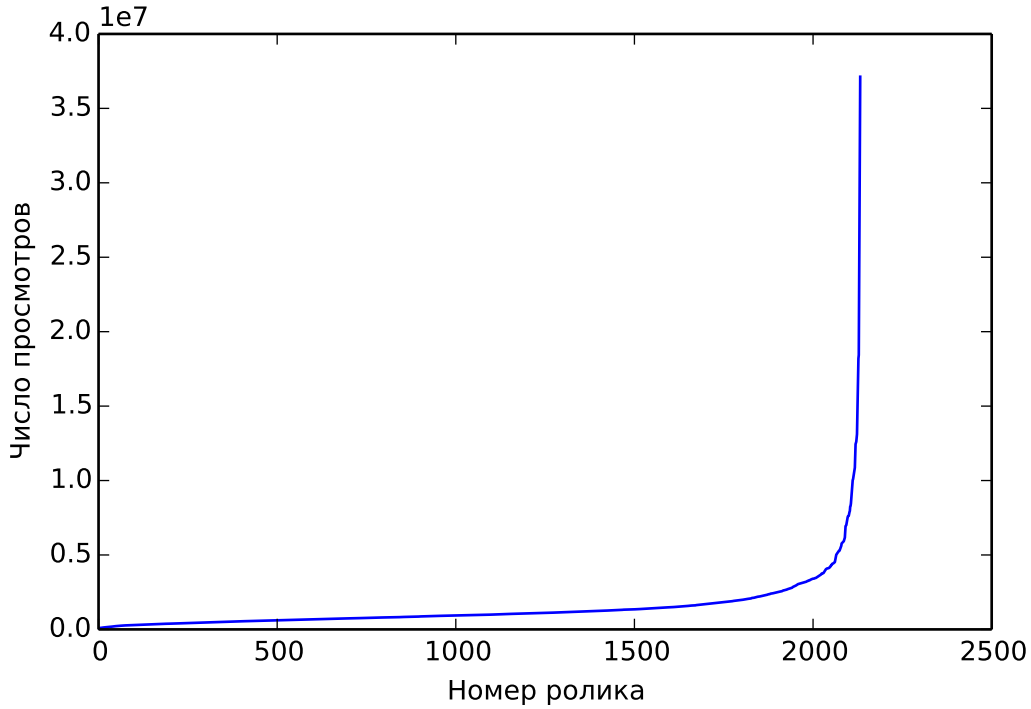

Рис. 3: Распределение числа просмотров

1. Do schools kill creativity? и 37124197 просмотров
2. Your body language shapes who you are и 31154001 просмотров
3. How great leaders inspire action и 25449477 просмотров
4. The power of vulnerability и 23416955 просмотров
5. My stroke of insight и 18430876 просмотров
6. 10 things you didn't know about orgasm и 18195966 просмотров
7. Why we do what we do и 16146117 просмотров
8. The puzzle of motivation и 15014711 просмотров
9. The thrilling potential of SixthSense technology и 14464629 просмотров
10. Looks aren't everything. Believe me, I'm a model. и 13116493 просмотров

Видно, что разница между первым и десятым роликами по числу просмотров — 2.8 раза.

Самые популярные темы, по среднему числу просмотров:

1. succ и 6590998 просмотров
2. body language и 6118118 просмотров
3. Buddhism и 4036186 просмотров
4. goal-setting и 3788322 просмотров
5. dance и 3635241 просмотров
6. psychology и 3499243 просмотров
7. magic и 3473116 просмотров
8. motivation и 3473039 просмотров
9. fear и 3352476 просмотров
10. self и 3344291 просмотров

Видно, что разница между первым и десятым роликами по числу просмотров — 2 раза.

2.4 Анализ пользовательских рейтингов

Пользователь может оценить каждый из роликов одним из следующих рейтингов (можно выбрать до трёх, если выбран один, то он учитывается трижды):

1. beautiful
2. inspiring
3. informative
4. fascinating
5. OK
6. persuasive
7. jaw-dropping
8. ingenious
9. courageous
10. confusing

11. obnoxious
12. longwinded

Отдельный анализ по всем этим рейтингам малополезен, так как такие, как OK не несут практически никакой информации. Тем не менее, некоторые из них вполне достойны некоторого анализа.

Будем ранжировать ролики по параметру R/V, где R и V это количество данных рейтингов и количество просмотров соответственно. Для упорядочивания рейтингов воспользуемся параметром $\left(\sum_{R_i \in R} R_i\right) / \left(\sum_{V_i \in V} V_i\right)$, где R и V это множества данных рейтингов и просмотров роликов данного тега соответственно.

2.4.1 Анализ по рейтингу beautiful

Начнём с 5 роликов с самым большим параметром:

1. Kounandi, у которого 70851 просмотров
2. How many lives can you live?, у которого 514237 просмотров
3. The secret life of plankton, у которого 182096 просмотров
4. M'Bifo, у которого 264956 просмотров
5. A message to gay teens: It gets better, у которого 264170 просмотров

Далее, 5 тегов с самым большим значением данного параметра:

1. Foreign Policy
2. conducting
3. guitar
4. violin
5. cello

2.4.2 Анализ по рейтингу inspiring

Начнём с 5 роликов с самым большим параметром:

1. Transplant cells, not organs, у которого 591904 просмотров
2. A message to gay teens: It gets better, у которого 264170 просмотров
3. When a reporter becomes the story, у которого 118258 просмотров
4. Which country does the most good for the world?, у которого 2194074 просмотров

5. Why we all need to practice emotional first aid, у которого 3105779 просмотров

Далее, 5 тегов с самым большим значением данного параметра:

1. Foreign Policy
2. chautauqua
3. Latin America
4. Buddhism
5. death

2.4.3 Анализ по рейтингу informative

Начнём с 5 роликов с самым большим параметром:

1. The early birdwatchers, у которого 108989 просмотров
2. Just how small is an atom?, у которого 349159 просмотров
3. What happens when an NGO admits failure, у которого 202608 просмотров
4. Inventing is the easy part. Marketing takes work, у которого 179460 просмотров
5. The sea we've hardly seen, у которого 135665 просмотров

Далее, 5 тегов с самым большим значением данного параметра:

1. oil
2. privacy
3. human origins
4. wind energy
5. presentation

2.4.4 Анализ по рейтингу obnoxious

Это пример рейтинга, который, как мне кажется, следует считать негативным. Начнём с 5 роликов с самым большим параметром:

1. 17 words of architectural inspiration, у которого 680410 просмотров
2. I believe we evolved from aquatic apes, у которого 987158 просмотров
3. Protecting the brain against concussion, у которого 394909 просмотров

4. Put the financial aid in the bag, у которого 164513 просмотров
5. Gotta share!, у которого 349869 просмотров

Далее, 5 тегов с самым большим значением данного параметра:

1. apes
2. Christianity
3. oil
4. atheism
5. God

Кажется весьма логичным, что такие спорные категории, как Christianity, atheism и God попали в этот список.

2.4.5 Анализ активности

Давайте введём параметр активность. Для ролика определим его так: $\sum_{R_i \in R} / V$, где где R и V это множество с рейтингами и количество просмотров данного ролика соответственно. Тогда, по аналогии, на тег это можно обобщить как:

$$\left(\sum_{R^k \in R} \sum_{R_i \in R^k} R_i \right) / \left(\sum_{V_i \in V} V_i \right)$$

Здесь R это множество множеств рейтингов каждого из роликов, а V множество с просмотрами каждого из роликов.

Начнём с 5 роликов с с самым большим параметром:

1. A message to gay teens: It gets better, у которого 264170 просмотров
2. Transplant cells, not organs, у которого 591904 просмотров
3. The case for same-sex marriage, у которого 284271 просмотров
4. "Kounandi у которого 70851 просмотров
5. What happens when an NGO admits failure, у которого 202608 просмотров

Далее, 5 тегов с самым большим значением данного параметра:

1. oil
2. Foreign Policy
3. illness

4. apes

5. consciousness

Видно, что у этих роликов достаточно мало просмотров. Давайте посмотрим 4 на значение этого параметра, в зависимости от числа просмотров ролика. Легко заметить, что график достаточно ровный, за исключением роликов с небольшим числом просмотров. Таким образом, можно сделать вывод, что эта характеристика малоинформативна.

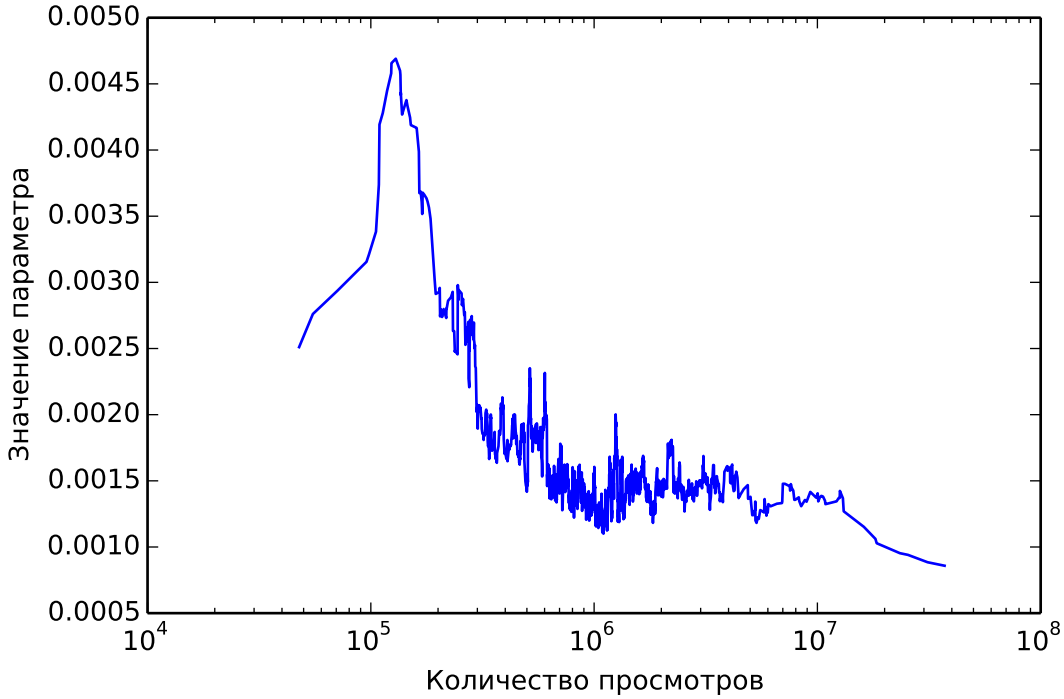

Рис. 4: Распределение значения параметра, после применения скользящего среднего

3 Инструменты

При исследовании были использованы следующие инструменты:

Python — основной язык программирования при исследовании.

Scrapy — библиотека для написания парсера.

PyLab — создание графиков.

Numpy — некоторые статистические функции.

Ipython — утилита для более удобного исследования данных.

Graph_tool — библиотека для работы с графами.

XeTeX — создания отчёта.

4 Выводы

4.1 Получение данных

Несмотря на то, что TED достаточно открыто делится со всем миром своими выводами, API был закрыт. Мне кажется, что если бы был API, то удалось бы больше усилий приложить непосредственно к исследованию.

Слишком частые запросы завершались сервером с ошибкой 329.

4.2 Теги

Теги являются самыми очевидными индикаторами темы выступления. Гипотеза о том, что их можно было объединять с тем, чтобы сократить их количество не оправдалась, хотя предполагалось, что это выполняется из-за большого их числа.

4.3 Работа с субтитрами

Как мне кажется, работа с субтитрами затруднена, применительно к данному исследованию. В самом деле, получать темы бессмысленно, так как они уже есть в тегах, получать же более специфические темы тоже бессмысленно, так анализ таких редких тем будет малополезен.

Единственное применение, которое я смог найти — анализ эмоциональной составляющей доклада. Но для этого нужна размеченная база со словами, и мне кажется, что это не совсем соотносится с темой исследования.

4.4 Направления для дальнейшей работы

1. Анализ эмоциональности докладов, по субтитрам
2. Добавление к данным авторов (есть авторы, которые выступали более раза). Поиск наиболее успешных авторов.
3. Добавление к данным длительности роликов. Поиск возможных зависимостей между длительностью и какими-либо рейтингами (например, между длительностью и рейтингом longwinded)