Тематическое моделирование

… и как сделать свой Brand Analytics

Задача

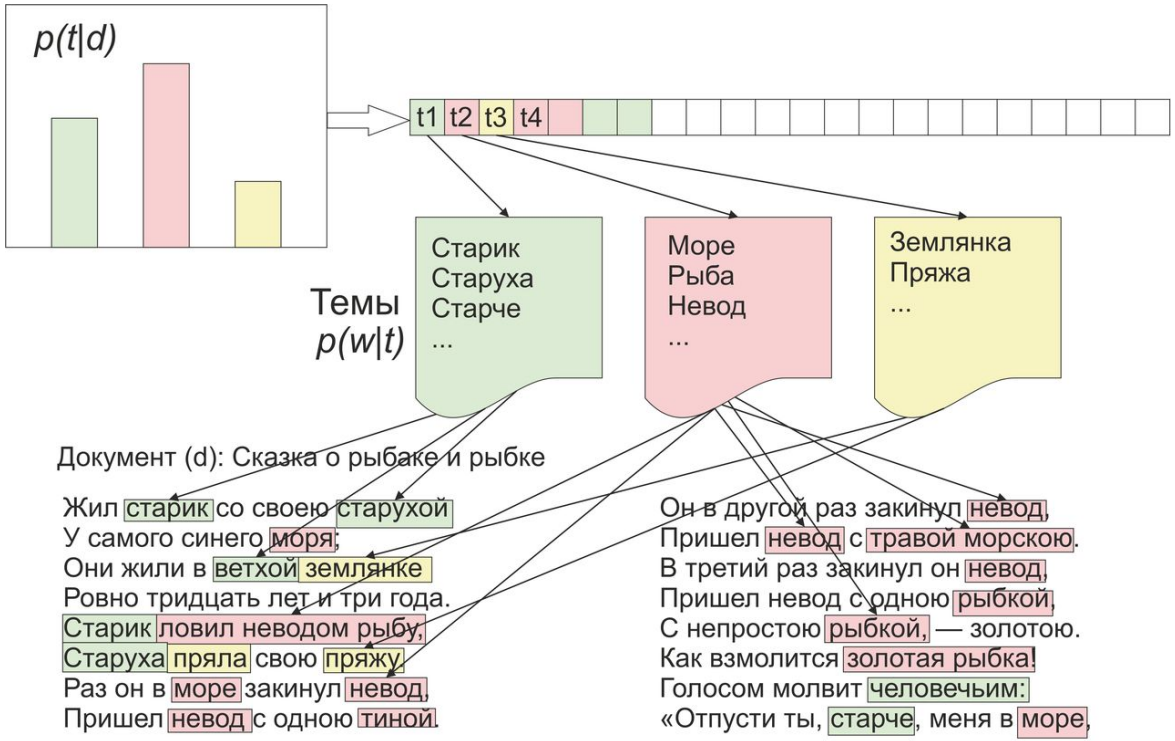
- Выделить в коллекции документов темы
- Мы заранее не знаем, сколько этих тем и какие они
- Приятный бонус - векторная модель (об этом попозже)

Что мы хотим?

Цель

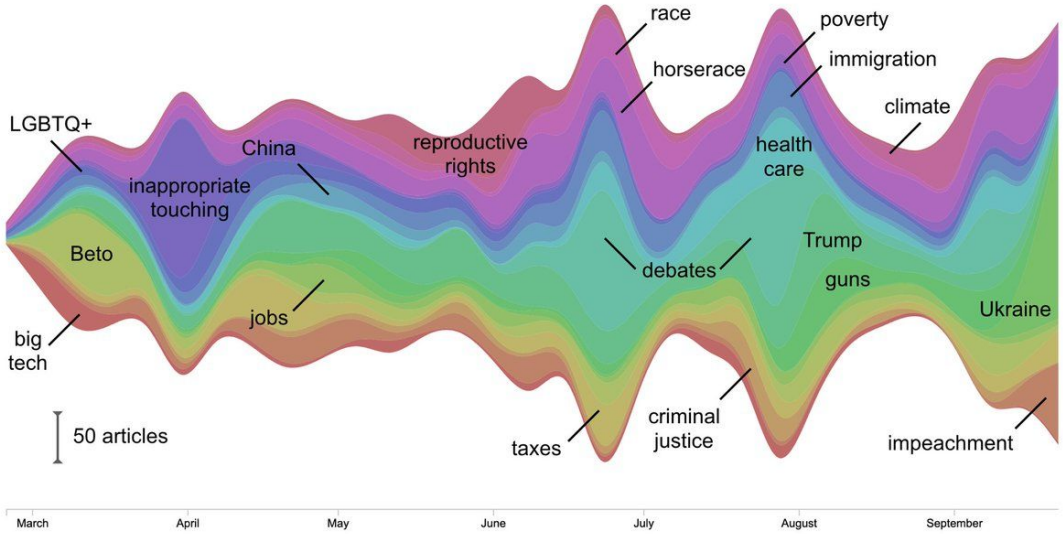
- Визуализация
- Эксплоративный анализ
- Кластеризация
- И снова векторная модель

Зачем мы это хотим?


p(t|d)
t1 t2 t3 t4
Темы
p(w|t)
Старик
Старуха
Старче
...
Море
Рыба
Невод
...
Землянка
Пряжа
...
Документ (d): Сказка о рыбаке и рыбке
Жил старик со своею старухой
У самого синего моря;
Они жили в ветхой землянке
Ровно тридцать лет и три года.
Старик ловил неводом рыбу,
Старуха пряла свою пряжу.
Раз он в море закинул невод,
Пришел невод с одною тиной.
Он в другой раз закинул невод,
Пришел невод с травой морскою.
В третий раз закинул он невод,
Пришел невод с одною рыбкой,
С непростою рыбкой, — золотою.
Как взмолится золотая рыбка!
Голосом молвит человечьим:
«Отпусти ты, старче, меня в море,

How media are setting the 2020 agenda

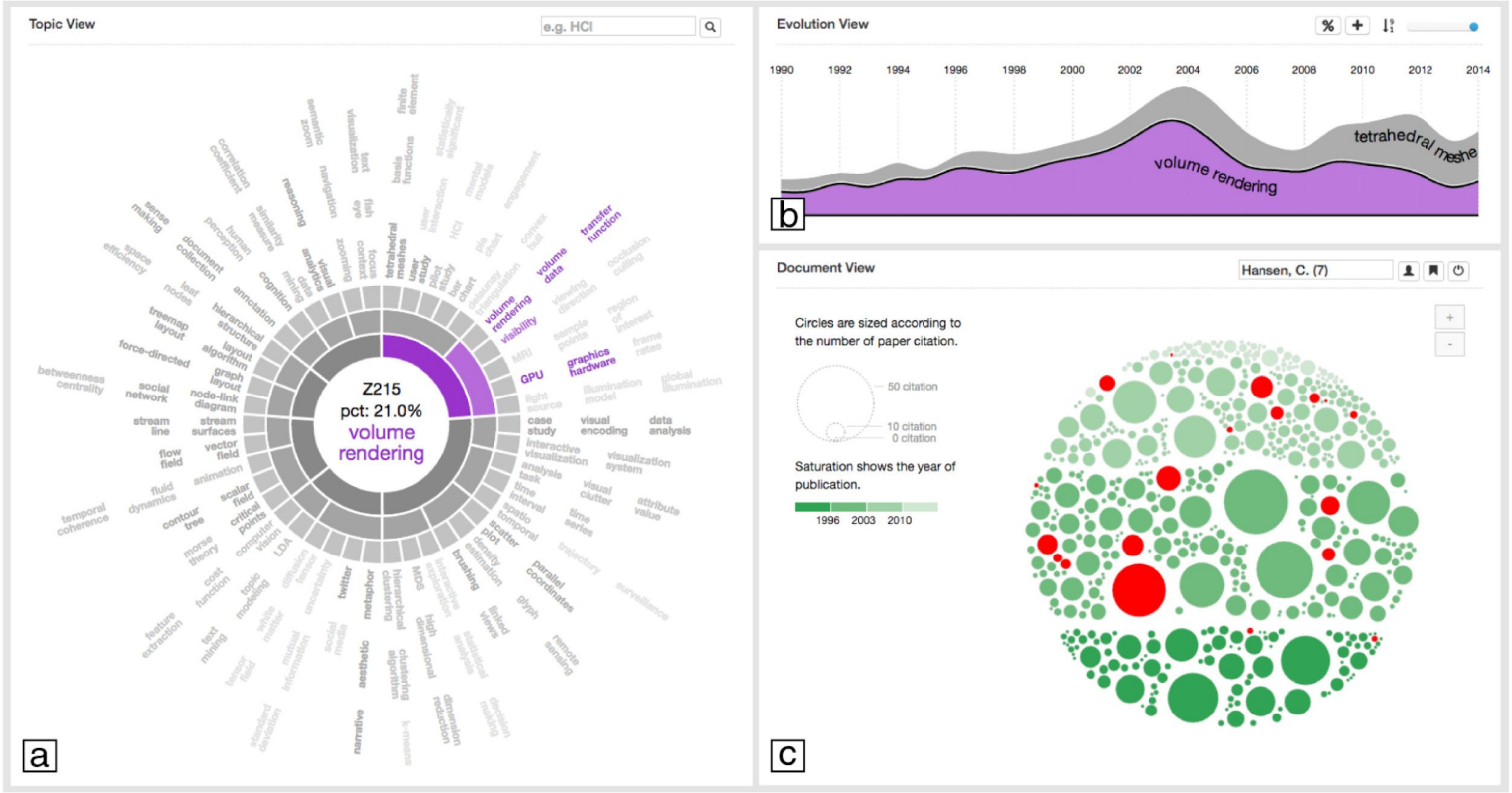
A topic analysis of news articles published by 28 outlets since March 2019 mentioning Joe Biden, Bernie Sanders, Elizabeth Warren, Kamala Harris, Pete Buttigieg, Beto O'Rourke, Cory Booker, Kirsten Gillibrand, Amy Klobuchar, or Tulsi Gabbard



n = 5,850

Data: Media Cloud

https://www.trendsmap.com/twitter/tweet/1194233155642941440



https://www.sciencedirect.com/science/article/pii/S2468502X17300074

Тематические модели: много трехбуквенных аббревиатур

- **LSA** (Latent semantic analysis, катентно-семантический анализ)
- **pLSA** (probabilistic latent semantic analysis,вероятностный латентно-семантический анализ)
- **LDA** (Latent Dirichlet Allocation, латентное размещение Дирихле)
- **BigARTM** (ARTM — Additive Regularization for Topic Modeling)

LSA

- Самая ранняя модель
- Предполагает наличие скрытого (латентного) параметра - темы
- Наиболее популярная реализация - сингулярное разложение (SVD) матрицы термы-документы

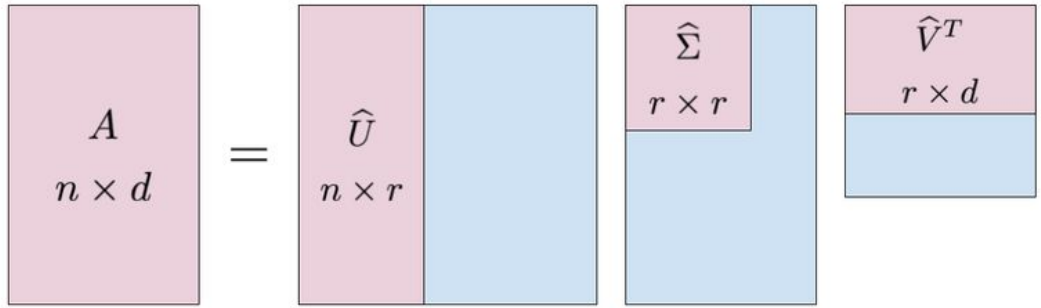
Матрица термы-документы

1. Медведев прибыл с официальным визитом во Вьетнам.
2. В Крыму объявили экстренное предупреждение из-за шторма.
3. В Пушкинской галерее пройдет выставка Рафаэля.
4. Премьер-министр России встретился с коллегами на саммите во Вьетнаме.
5. На презентации Apple продемонстрировали новый iPad Pro.
6. Работы Рафаэля впервые привезут в Москву.
7. В понедельник в Крыму ожидаются сильные дожди и штормовой ветер.
8. В Калифорнии стартовала ежегодная презентация компании Apple.

	D1	D2	D3	D4	D5	D6	D7	D8
apple	0	0	0	0	1	0	0	1
ipad	0	0	0	0	1	0	0	1
pro	0	0	0	0	1	0	0	1
визит	1	0	0	0	0	0	0	0
вьетнам	1	0	0	1	0	0	0	0
галерея	0	0	1	0	0	0	0	0
ежегодный	0	0	0	1	0	0	0	1
крым	0	1	0	0	0	0	1	0
...	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**

Матрицу можно заполнить tf-idf весами!

SVD



Исходная матрица термы-документы

U
$n \times n$

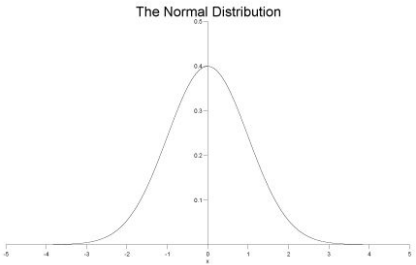
Σ
$n \times d$

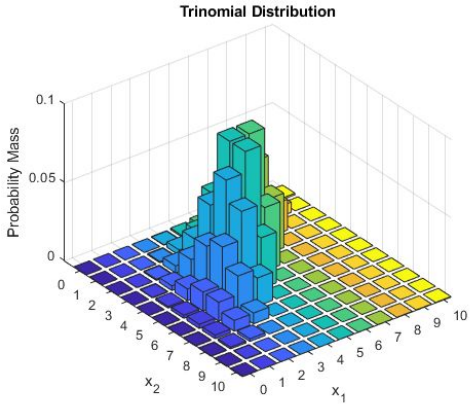
V^T
$d \times d$

[Наглядная гифка]

pLSA

- LSA предполагет, что распределение слов и документов нормальное, pLSA - мультиномиальное
- Совместная встречаемость терма и документа (w, d) моделируется как сочетание независимых мультиномиальных распределений, где c — тема. Количество тем — это гиперпараметр, который выбирается до начала анализа





LDA

- генеративная вероятностная модель
- документ - это набор случайных скрытых тем, где каждая тема определяется распределением слов, при этом каждое слово в конкретном документе можно отнести к одной из его тем
- в качестве априорного распределения для тем используется распределение Дирихле

А что там с векторной моделью?

- Тематическое моделирование - еще один способ (помимо подходов дистрибутивной семантики) получить векторное представление слов и документов, при этом полученные вектора тем будут **интерпретируемыми**

Topic 1		Topic 2		Topic 3	
term	**weight**	**term**	**weight**	**term**	**weight**
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

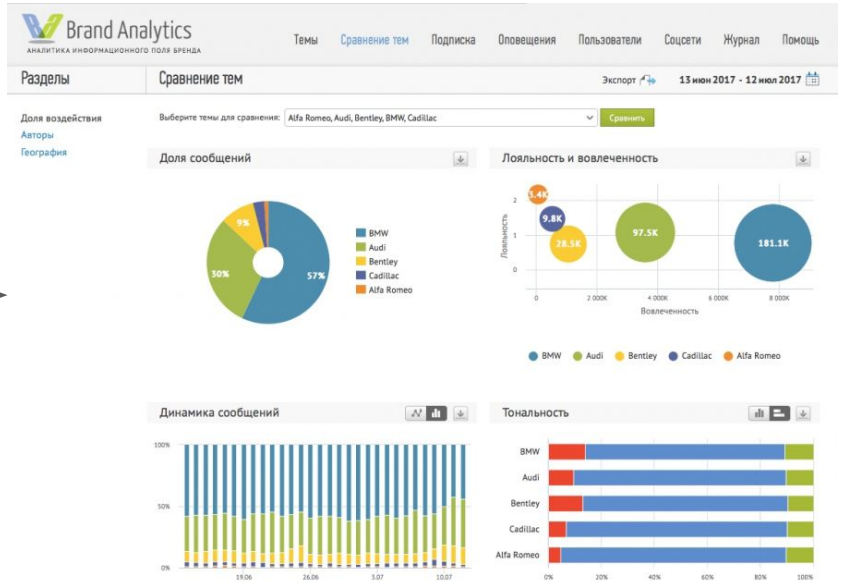
Что на практике

1. Выбираем количество тем N
2. Строим тематическую модель с N темами
3. Оцениваем качество (в этом, в том числе, помогает визуализация)
4. Пробуем другое число тем N

Перерыв!

Brand Analytics: от названия бренда до аналитических отчетов

Что умеет сервис?



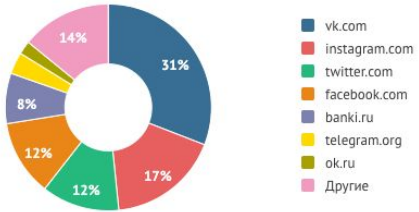
https://br-analytics.ru/

Что это значит на самом деле:

1. Используем открытые API соцсетей (твиттер, VK и т.д.) и медиа-порталов, чтобы вытаскивать все тексты с упоминанием бренда

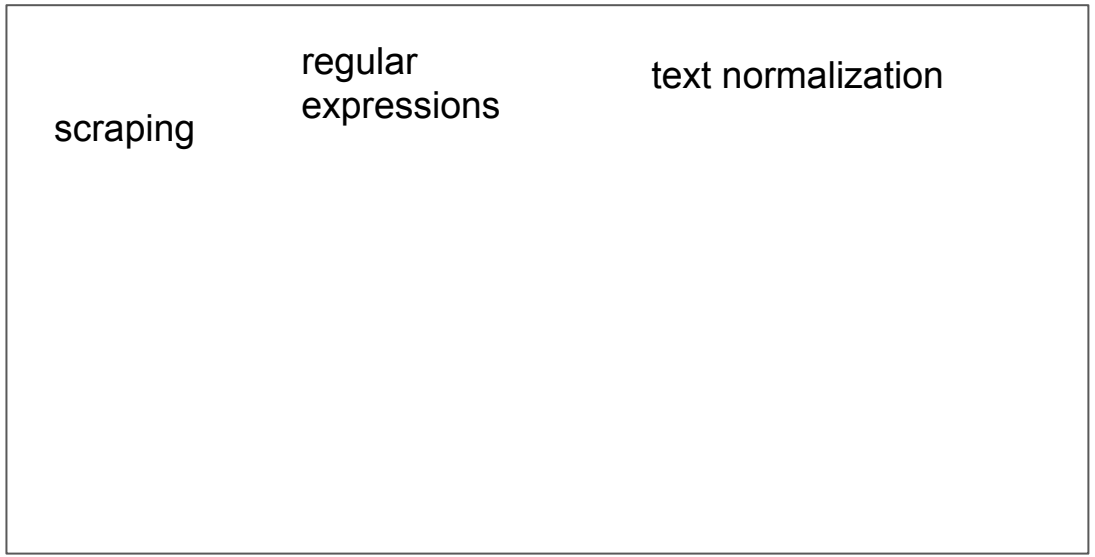
scraping

regular expressions



Что это значит на самом деле:

2. Предобрабатываем текст



Что это значит на самом деле:

3. Частотный анализ: количество упоминаний, ключевые слова и словосочетания, облака слов и т.д.

scraping

regular expressions

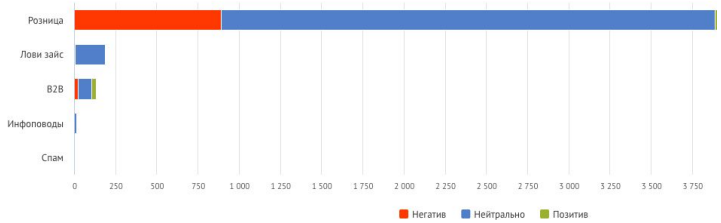
text normalization

tf-idf

keywords extraction

Что это значит на самом деле:

4. Строим тематические модели



scraping

regular expressions

text normalization

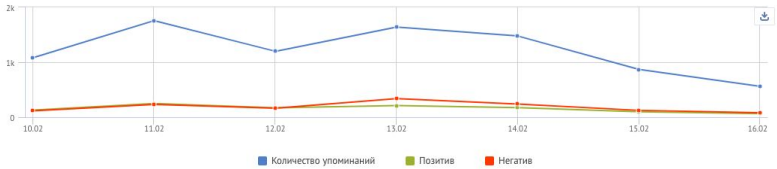
tf-idf

keywords extraction

topic modeling

Что это значит на самом деле:

5. Делаем анализ тональности



scraping

regular expressions

text normalization

tf-idf

keywords extraction

topic modeling

word embeddings

text classification