

# Проект по расстановке тегов и выделению ключевых слов в корпусе текстов

Выполнил: Андрей Орлов

Курс ДПО по направлению  
«Компьютерная лингвистика», НИУ ВШЭ

Москва, 2023 г.

## Задачи проекта:

- сформировать корпус текстов;
- выполнить тематическое моделирование корпуса;
- каждому тексту присвоить соответствующие теги тем, имен и названий;
- выделить ключевые слова в тексте.

Материал для работы – новостные публикации на сайте [SecurityLab.ru](https://www.securitylab.ru)

Объем корпуса: 8000 текстов

## Перспективы использования:

- рубрикация текстов;
- исправление ошибок в расстановке тегов;
- дополнение поисковой системы сайта;
- тематический анализ публикаций.

Общая задача – облегчение  
пользовательской навигации на сайте

## Positive Technologies разъясняет указ президента №250

13:48 / 7 сентября, 2022



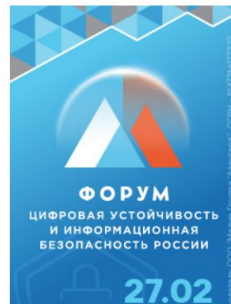
Вебинар состоится 13 сентября в 14.00.

Майский указ президента РФ №250 «О дополнительных мерах по обеспечению информационной безопасности» установил кратчайшие сроки для создания или преобразования отделов информационной безопасности, назначения ответственных за ИБ лиц и перехода на отечественные средства защиты информации. При этом под действие этого нормативного документа попадают, по экспертным оценкам, до 500 тысяч российских компаний.

Positive Technologies проведет 13 сентября **открытый вебинар**, чтобы разъяснить всем заинтересованным лицам, как выполнять указ №250 и выстраивать процессы кибербезопасности в новой цифровой реальности.

Positive Technologies

SOC



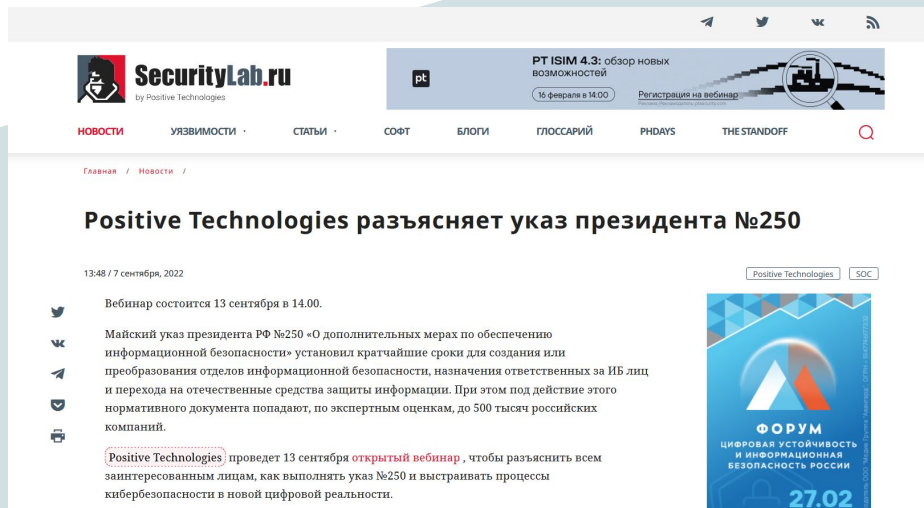
Пример:

анонс вебинара никак не отделен от новостных статей

Выделен тематический тег: 'события и мероприятия'

Ключевые слова: 'информационный', 'безопасность', 'майский', 'установить', 'назначение', 'президент', 'дополнительный', 'мера', 'краткий', 'отдел'

ссылка на публикацию



## Решение задачи: теория

Тематическое моделирование выполняется с помощью латентного разложения Дирихле (LDA).

Результаты оцениваются по показателю Topic Coherence.

Извлечение ключевых слов с помощью алгоритма YAKE

Особенность: YAKE рассматривает каждый текст отдельно

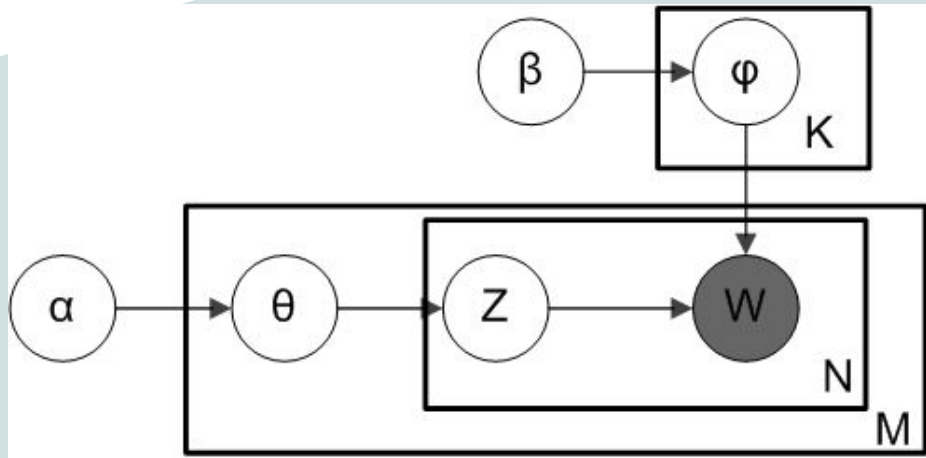
## Распределение Дирихле (LDA):

предсказывает, какими именно могут быть  
распределения вероятностей встретить тему в  
документе

Подробнее про вероятностное  
тематическое моделирование

## Распределение Дирихле (LDA):

На основе распределения Дирихле  $\text{Dir}(\alpha)$  выбирается распределение тем в документе —  $\theta$  (тета). На основе распределения тем  $\theta$  выбирается тема  $Z$ . На основе другого распределения Дирихле —  $\text{Dir}(\beta)$  — выбирается распределение слов в теме  $Z$  —  $\phi$  (фи). Из  $\phi$  выбирается слово  $W$ .





## Решение задачи: практика

Для сортировки текстов по темам используется модель LDA в библиотеке Gensim.

Именованные сущности извлекаются с помощью библиотеки Spacy.

Ключевые слова выделяются с помощью программной реализации алгоритма YAKE на языке Python

## Порядок работы:

- разработка парсера;
- создание текстового корпуса;
- токенизация и лемматизация текстов;
- тематическое моделирование корпуса;
- выделение и нормализация названий и имен;
- извлечение и нормализация ключевых слов.

Полученные данные в виде тегов и ключевых слов присваиваются текстам

Тексты сохраняются в файл CSV

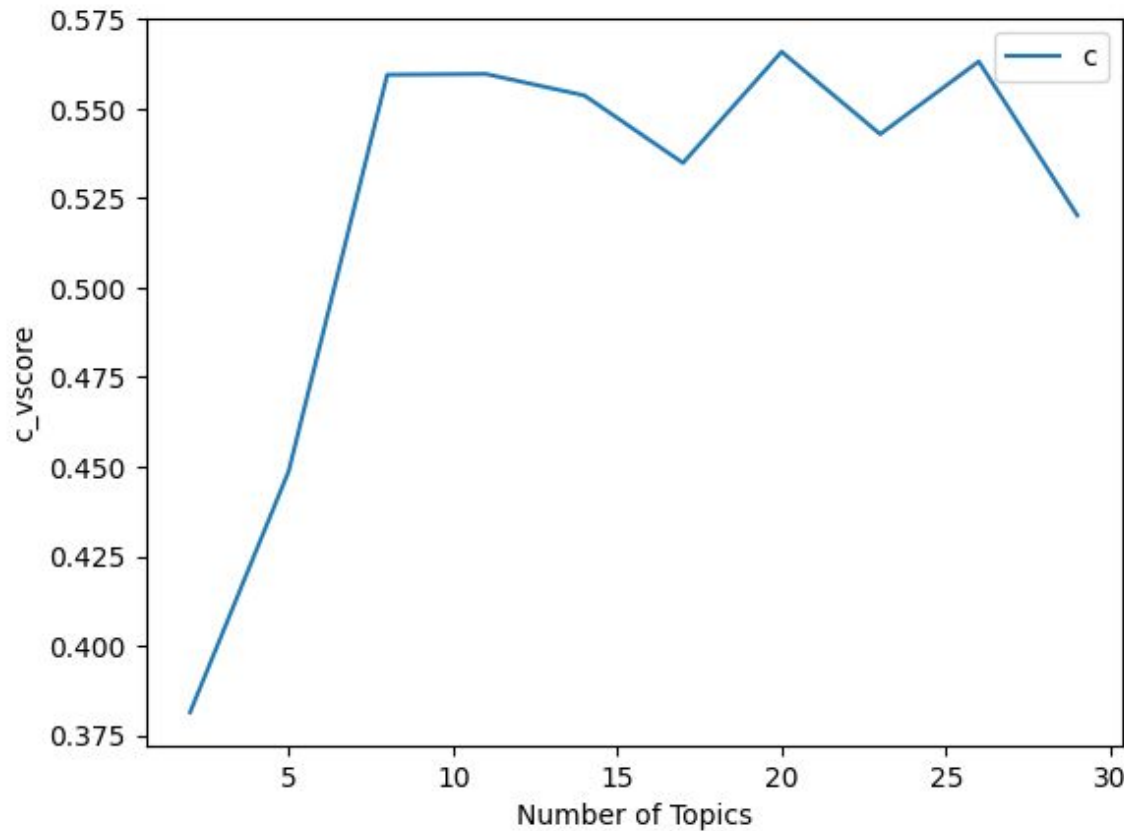
**Ссылки по теме:**

Объяснение тематического моделирования

Сравнительное описание алгоритмов для  
выделения ключевых слов

**Пример:**

Тематическое моделирование  
интернет-издания «Нож»



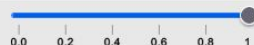
Изменение параметра Topic Coherence по отношению к числу выделяемых в корпусе тем

Out[11]:

Selected Topic:

Slide to adjust relevance metric ( $\alpha$ )

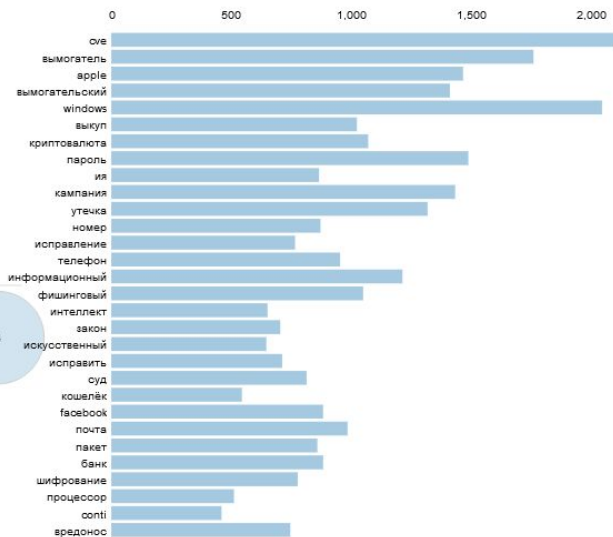
$\alpha = 1$



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms<sup>1</sup>



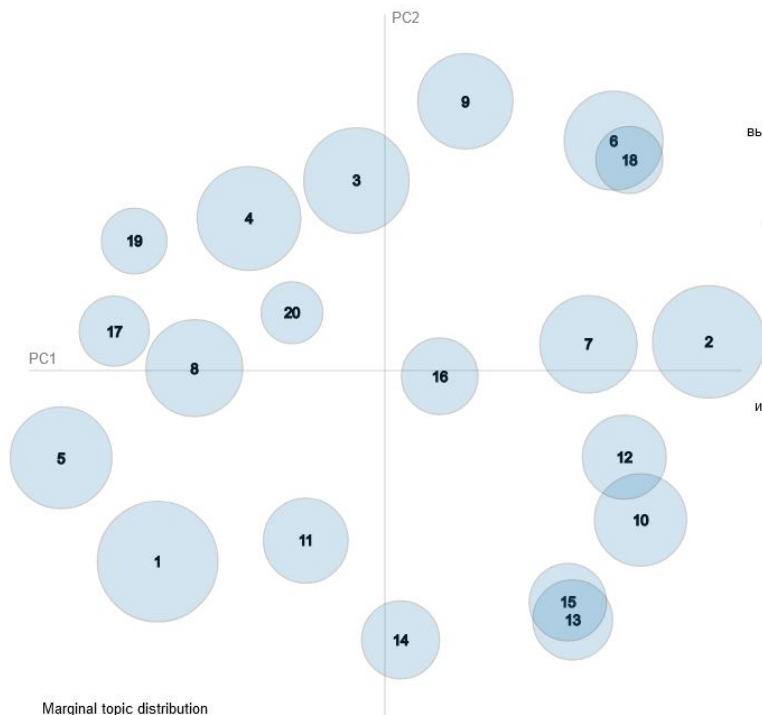
Overall term frequency

Estimated term frequency within the selected topic

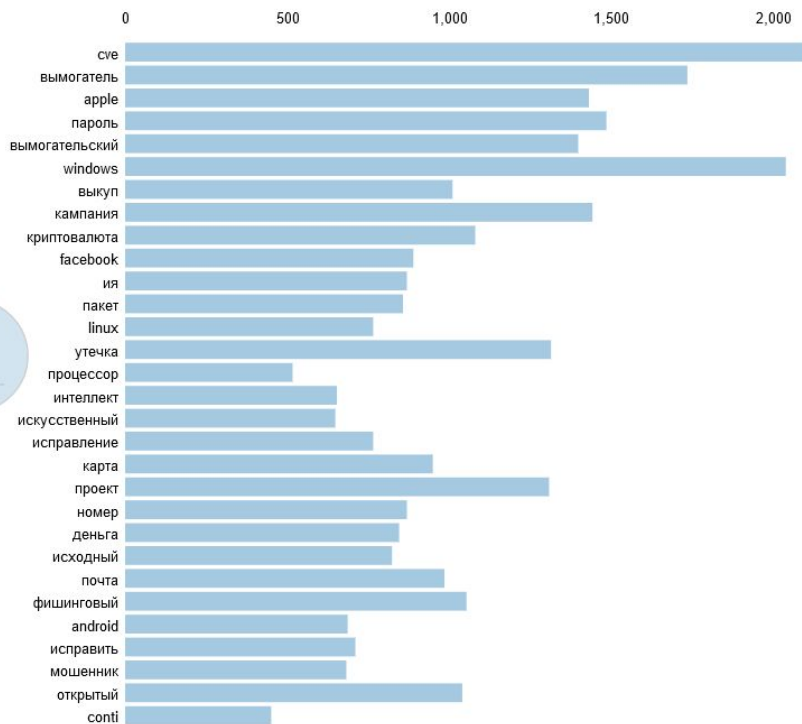
<sup>1</sup>  $salience(term, w) = frequency(w) * \sum_{t \in topics} p(t|w) * \log(p(t|w) / p(t))$  for topics  $t$  see Chuang et al. (2012)  
<sup>2</sup>  $relevance(term, w, topic_t) = \alpha * p(w|t) + (1 - \alpha) * p(w, t|w)$  see Slavet & Shirley (2014)

Распределение 12 тем,  $c\_v = 0.5990480012463285$

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms<sup>1</sup>

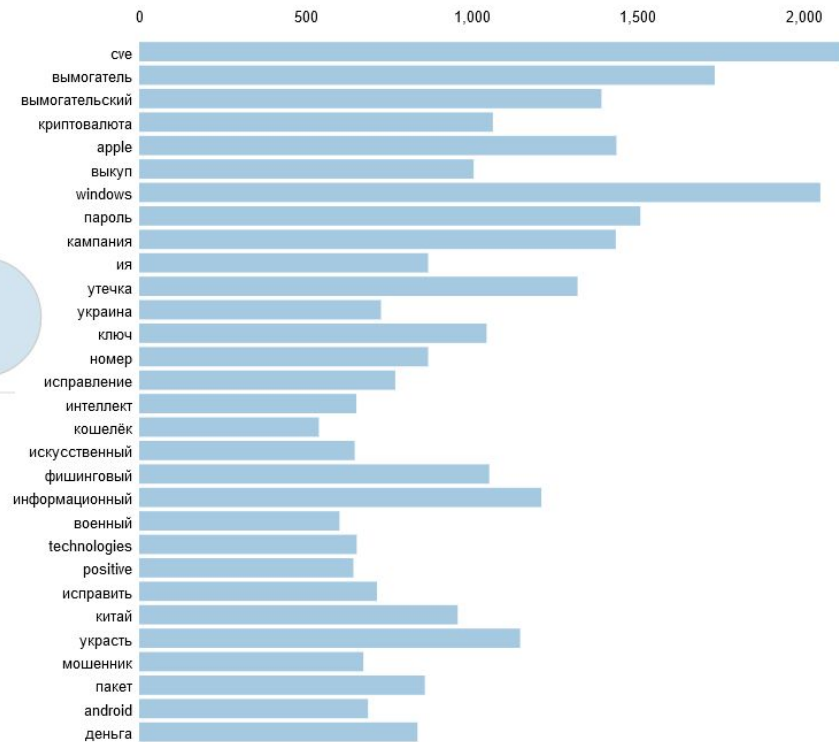


Распределение 20 тем,  $c\_v = 0.5972458367311672$

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms<sup>1</sup>



Распределение 15 тем,  $c\_v = 0.6041639125950901$

**Итоги: выделено 12 тем:**

"исследования", "разработка и техника", "финансы и бизнес", "госрегулирование",  
"вредоносы", "мошенничества", "законы и нарушения", "события и мероприятия",  
"программное обеспечение", "уязвимости",  
утечки и обновления", "кибербезопасность"

Выделены теги имен собственных

Ключевые слова помогают идентифицировать текст в результатах поиска



## Перспективы разработки:

- дальнейшая классификация текстов с выделением подтем;
- коррекция нормализации;
- выделение двусложных тегов;
- тематический анализ и выделение информационных трендов.

Ссылка на проект