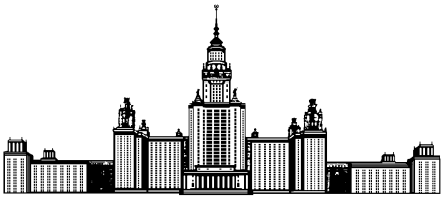

Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Полушин Владимир Владимирович

# Тематические модели для ранжирования рекомендаций текстового контента

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Научный руководитель:**
д.ф-м.н., доцент
Воронцов Константин Вячеславович

Москва, 2017

# Содержание

**Аннотация**

Задачи ранжирования текстов, информационного поиска и классификации текстового контента становятся все более актуальными в растущем мире информации. Для решения этих задач применяется вероятностное тематическое моделирование. Одним из важнейшим показателем качества тематической модели является *интерпретируемость тем*, то есть возможность сопоставить темам наиболее частотные слова из словаря коллекции так, чтобы человек смог понять о чем эта тема и дать ей адекватное название. Для повышения интерпретируемости тем в качестве элементов словаря данной коллекции можно использовать не только отдельные слова, но и словосочетания, которые также являются терминами предметных областей. Целью данной работы было предоставить эффективный алгоритм нахождения терминов в текстовых коллекциях.

В работе рассмотрено три главных подхода к фильтрации терминов (синтаксический, статистический и тематический). Эти подходы были скомбинированы для получения лучшего качества, и итоговый алгоритм был протестирован на текстовой коллекции. Эксперимент показал, что можно построить эффективный алгоритм предсказания терминов на основе трех основных методов фильтрации, позволяющий существенно сократить множество рассмативаемых словосочетаний, при этом оставив большую часть терминов.

# 1 Введение

Задачи ранжирования текстов, информационного поиска и классификации текстового контента становятся все более актуальными в растущем мире информации. Для решения этих задач применяется вероятностное тематическое моделирование. Тематическая модель — это модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции и какие слова наиболее ярко описывают каждую тему. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов, на выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем, и для каждой темы выдается вектор оценок принадлежности слова этой теме. По этим векторам оценок для тем можно составить список наиболее частотных слов и документов, с помощью которых человек сможет понять, о чём эта тема, дать ей адекватное название, определить более общие, более частные или близкие темы. Это важное свойство называется *интерпретируемостью тем*, оно позволяет систематизировать, визуализировать, объяснять результаты, выдаваемые пользователю информационной системы.

Интерпретируемость тем — важнейший показатель качества тематической модели. Для повышения интерпретируемости тем в качестве элементов словаря данной коллекции можно использовать не только отдельные слова, но и словосочетания, которые также являются терминами предметных областей. Поиск таких терминов в тексте является нетривиальной и трудоемкой задачей. Решение многих проблем, связанных с обработкой информации, элементарно для человека, но поиск терминов не входит в их число, так как при её решении необходимо не только различать синтаксис языка, что не является затруднительным для человека, но и понимать контекст, в котором был написан текст, и обладать большим словарным запасом в нужной предметной области. Таким образом, естественным путём встает вопрос об автоматизации извлечения терминов из текстов (ATE, automatic term extraction).

## 1.1 Автоматическое выделение терминов

Цель работы — предоставить эффективный алгоритм нахождения словосочетаний, являющихся терминами рассматриваемой предметной области, в автоматическом режиме. Для обработки таких словосочетаний требуется понять какими свойствами должен обладать термин и какими методами возможно определить эти свойства на основе текстовой коллекции.

Каждый термин в предметной области должен:

1. *быть синтаксически связным*, то есть является грамматически корректным словосочетанием. (“машина опорных векторов”, “рассмотреть частный случай”)

2. *обладать высокой частотностью*, много раз встречается в коллекции. (“автомобиль марки Ауди”, но не “автомобиль моего соседа по комнате”)

3. *иметь совстречаемость слов*, он должен состоять из слов, неслучайно часто встречающихся вместе (“машина опорных векторов”, но не “интересный метод исследования”)

4. *быть полным*, то есть является максимальной по включению цепочкой слов (“поиск ближайщих соседей”, но не “поиск соседей” и “ближайщих соседей”)

5. *иметь тематичность*, термин должен имеет пиковую тему в тематической модели.

Для каждого из этих пунктов предлагается использовать отдельный инструмент, который бы осуществлял требуемую фильтрацию:

1. *Автоматический синтаксический анализ* текста и выделение из него связных словесных конструкций для пункта 1. Он позволяет найти связи в предложении и сопоставить словам метки частей речи и членов предложения. Используя эту информацию, можно пытаться выделять синтаксически корректные словосочетания из предложений, делать отбор по частям речи и членам предложения слов, входящих в термин.

2. *Статистический анализ* для пунктов 2, 3, 4, использующий информацию о частоте и совстречаемости слов в каждом документе коллекции. Этот анализ

позволяет находить высокочастотные слова, штрафовать словосочетания, которые входят как подмножество в другие, и выделять последовательности слов, которые неслучайно имеют такой порядок.

3. Построение *тематической модели* над текстовой коллекцией для пункта 5. В ней информацию о принадлежности слова каждой теме можно использовать для определения пиковых слов, то есть таких слов, которые принадлежат только одной или двум темам, тем самым, указывая на то, что они является важным компонентом в описании отдельной темы и всего текста.

Для достижения наилучшего качества поиска словосочетаний требуется сравнить несколько альтернативных стратегий оценки и отбора терминов на каждом этапе. Каждый из множества таких фильтров не может гарантировать высокую точность, но может дополнять другие фильтры, поэтому ставится задача объединения трех подходов отбора терминов в единую технологию. Она подразумевает решение задачи о том, какие фильтры стоит использовать, а какие нет, как строить комбинацию подходов и оценивать итоговое качество.

Оценку качества отсеивания и настройку комбинированного алгоритма отбора терминов предлагается проводить с помощью ручной разметки, то есть генерировать из большого множества словосочетаний *случайную* меньшую подвыборку ($\sim 1\%$), которую возможно разметить за приемлемое время, в рамках предположения о том, процент терминов в этой малой выборке будет соответствовать значению для большой. Таким образом, на малой выборке можно проводить приближенную настройку или проверку качества.

# 2 Синтаксический анализ

## 2.1 SyntaxNet

Синтаксический подход к отбору терминов предлагается проводить с помощью SyntaxNet [1] [2] — предобученной нейросети для распознавания синтаксиса и разметки слов в предложениях, которая поддерживает 40 языков, включая английский

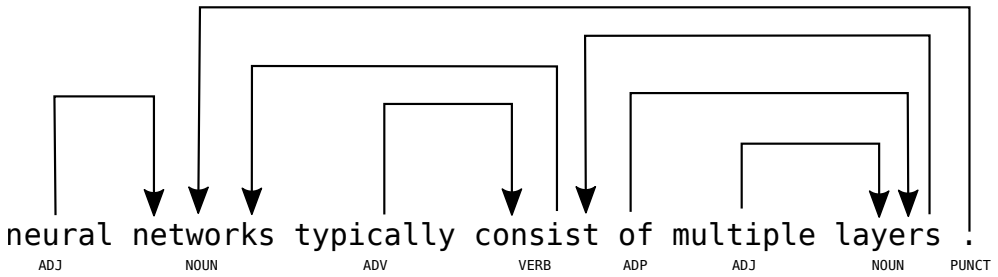


Рис. 1: Пример работы SyntaxNet

и русский. На вход ей подается список предложений, на выходе для каждого слова в предложении вычисляется:

- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- часть речи: `NOUN`, `VERB`, `CONJ`, `ADP`, ...
- член предложения: `nsubj`, `dobj`, `conj`, `cc`, `nmod`, ...

Пример работы `SyntaxNet` показан на схеме 1. У каждого слова кроме главного есть единственный предок, ребро от которого обозначает зависимую связь. Также на схеме под словами показаны части речи, выставленные в ходе обработки.

Данная технология является свободной (open-source) и высоко расширяемой на другие языки и синтаксисы. В ней не используются вручную созданные правила и зависимости, а происходит стандартный метод обучения на размеченных данных. С другой стороны это влечет за собой снижение качества распознавания по сравнению с аналогами, где часть правил обработки задана с помощью специалистов, некоторые специфичные случаи (имена собственные, редкие склонения слов) `SyntaxNet` может

обработать неправильно. Но в случае определения связей в простой структуре термина детальное распознавание не нужно, достаточно только базовых синтаксических зависимостей между словами.

### 2.2 Синтаксический отбор

Синтаксический отбор N-грамм, то есть N подряд идущих слов в тексте, основан на предположении о том, что термины должны образовывать полное поддерево в дереве связей предложения. Это означается, что каждое ребро, исходящее из слов в выбранном подмножестве предложения, должно либо вести в какую-то одну фиксированную для N-граммы вершину, не принадлежащую N-грамме, либо вести в слово внутри N-граммы. С другой стороны N-грамма должна образовывать строгую последовательность, не имея разрывов и других слов внутри. Для примера на схеме 1 есть 4 подходящие би-граммы `neural networks`, `typically consist`, `of multiple`, `multiple layers`, 3 из них являются словом с дополнением, то есть предок одного слова есть второе, а в `of multiple` предком обоих слов является `layers`, что тоже подходит под критерий.

Также предлагается фильтровать некоторые части речи в предложении, в таблице 1 показаны части речи и примеры слов, которые не должны присуствовать в термине. (*В данной работе рассмотрены тексты только на английском, но подобную информацию по любому языку можно найти на сайте документации для Universal Dependencies [3]*).

## 3 Статистический анализ

Статистический подход к отбору терминов включает в себя несколько критериев, которые позволяют оценивать неслучайность последовательности слов в N-грамме, штрафовать словосочетание за то, что оно входит в другие и учитывать частоту и совстречаемость слов.

Частоту N-граммы $w$ в коллекции, состоящей из слов $w^1, w^2, \ldots, w^N$ обозначим $f(w^1, w^2, \ldots, w^N)$, оценку вероятности встретить N-грамму обозначим $p(w^1, w^2, \ldots, w^N)$.

| Метка части речи | Описание и примеры |
|---|---|
| ADJ | Прилагательные |
| ADP | Предлоги |
| ADV | Наречия |
| NOUN | Существительные |
| NUM | Числительные |
| PRON | Местоимение |
| PROPN | Имена собственные |
| VERB | Глаголы |
| CONJ | Только 7 слов, которые могут быть помечены так: and, but, for, nor, or, so, yet. Обозначает связь между другими словами |
| AUX | Разные формы глаголов be, have, do, get, используемые для построения грамматических конструкций (например разные времена глаголов) |
| SCONJ | Слова для связи частей предложения: that, whether, if, when, since, before и т.д. |
| PUNCT | Пунктуация |
| DET | Слова the, my, this, some, twenty, each, any и т.д, используемые перед существительными |
| PART | Части слов: 's, ', not, to (как обозначение инфинитива) |
| INTJ | Междометия |
| SYM | Различные символы, отличные от пунктуации |

Таблица 1: Части речи, получаемые в процессе обработки `SyntaxNet`. Сверху обозначены части речи, которые встречаются в терминах, снизу те, которые встречаться не должны

- **PMI** - Pointwise mutual information. Для каждой N-граммы вычисляется функция:

$$\log \frac{p(w^1, w^2, \ldots, w^N)}{p(w^1)p(w^2)\ldots p(w^N)}$$

Она показывает насколько совместная вероятность N-граммы отличается от произведения вероятностей каждого из слов. Неслучайные сочетания слов будут иметь существенное различия в этих двух величинах, а в случайных N-граммах каждое из слов встречается независимо от других, поэтому данное отношение будет примерно равно 1.

- **C-value** [4] - Статистика вычисляемая как:

$$\text{C-value} = \begin{cases} \log_2(N) f(w^1, \ldots, w^N), & \text{если } T_w = \emptyset \\ \log_2(N) \Big[ f(w^1, \ldots, w^N) - \\ -\dfrac{1}{|T_w|} \displaystyle\sum_{\hat{w} \in T_w} f(\hat{w}^1, \ldots, \hat{w}^K) \Big], K > N, & \text{в ином случае} \end{cases}$$

где $T_w$ обозначает все N-граммы, которые включают в себя как подмножество данную N-грамму $w$.

Данный критерий поощряет частые N-граммы, но штрафует за вхождения в другие более длинные термины. Также положительно учитывается длина N-граммы, так как длинные термины наиболее редки и важны в текстовых коллекциях.

- **TopMine** из статьи [El-Kishky, 2014][5]. Это алгоритм, который итеративно сливает слова и фразы в предложении, рассчитывая для каждого слияния *значимость* и останавливаясь, когда для всех возможных слияний значимость меньше данного порога. В качестве функции значимости используется следующее выражение:

$$\text{SignificanceScore}(W_1, W_2) = \frac{f(W_1 \oplus W_2) - f(W_1)f(W_2)/L}{\sqrt{f(W_1 \oplus W_2)}}$$

где $W_1, W_2$ — сливаемые N-граммы, $\oplus$ обозначает конкатенацию и $L$ — длина коллекции.

Точное описание алгоритма приведено на схеме Алгоритм 1, пример работы на Рис. 2.

**Вход:** Частоты N-грамм $f(\dots)$, порог $\alpha$

**Выход:** Разбиение на N-граммы и значения SignificanceScore для каждой

H ← Куча по возрастанию

Поместить все рядом стоящие пары слов вместе с их SignificanceScore в H

**пока** размер H $> 1$

  Best ← максимум по SignificanceScore из H

  **если** Best $\geq \alpha$ **то**

    New ← Слить(Best)

    Удалить Best из H

    Обновить SignificanceScore для New с помощью фраз, находящихся слева и справа

  **иначе**

    выйти из цикла

**Алгоритм 1.** TopMine

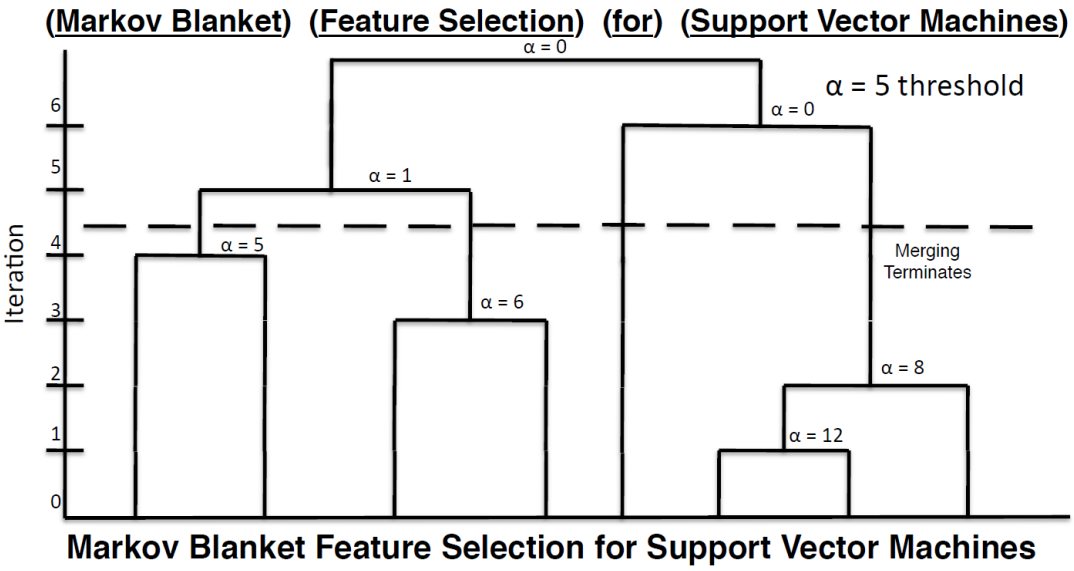


Рис. 2: Пример работы алгоритма TopMine

Данный подход к выделению терминов сочетает в себе качества двух предыдущих критериев: неслучайность фразы оценивается с помощью статистики

*SignificanceScore*, и полнота термина достигается благодаря итеративному слиянию с порогом.

Алгоритм в отличие от двух предыдущих критериев производит отбор N-грамм, отбрасывая большую часть рассматриваемых словосочетаний, но для итоговых терминов считается статистика *значимости*, которую можно использовать на подобии других критериев.

# 4 Тематическая модель

Пусть $D$ — коллекция текстовых документов, $W$ — множество употребляемых в них N-грамм, каждый документ $d \in D$ представляет собой последовательность входящих в него $n_d$ N-грамм $(w_1, \ldots, w_{n_d})$ из словаря $W$. Приняв гипотезу "мешка N-грамм" о том, что порядок N-грамм в документе не важен для определения тематики текста, можно перейти к более компактному представлению документа как подмножества $d \subset W$, в котором каждой N-грамме $w \in d$ поставлено в соответствие число $n_{dw}$ ее вхождений в документ $d$. Тематическая модель описывает вероятности появления N-грамм $w$ в документах $d$ при предположении условной независимости $p(w|d,t) = p(w|t)p(w|d)$,

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td},$$

где условные вероятности $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ не известны и являются параметрами модели. Будем записывать их в виде двух матриц: матрицы терминов тем $\Phi = (\phi_{wt})_{W \times T}$ и матрицы тем документов $\Theta = (\theta_{td})_{T \times D}$. Для оценивания параметров тематической модели $\Phi$ и $\Theta$ по коллекции документов $D$ максимизируется логарифм правдоподобия

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w \in W} \phi_{wt} = 1, \ \phi_{wt} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1, \ \theta_{td} \geqslant 0.$$

### 4.1 Тематический отбор

Подход к отбору терминов, основанному на тематических моделях, состоит в поиске *пиковых* N-грамм, то есть таких N-грамм $w$, для которых вероятность принадлежности теме $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ высока только для некоторых $t$ из малого подмножества всех тем $T$. Это означает, что вероятностное распределение тем при заданной N-грамме $p(t|w)$, $t = 1 \dots T$ должно быть максимально отдалено от равномерного $p_0(t) = 1/T$, $t = 1 \dots T$, что можно измерить с помощь нескольких методов:

- **Дивергенция Кульбака-Лейблера**:

$$\mathsf{KL}(w) = \mathsf{KL}(p_0 \| p) = \sum_{t \in T} \frac{1}{|T|} \ln \frac{\frac{1}{|T|}}{p(t|w)} \to \max$$

- **Дивергенция Йенсена-Шеннона** (метрика, не имеет проблем с нулевыми вероятностями), где $\bar{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|})$:

$$\mathsf{JS}(w) = \frac{1}{2}\mathsf{KL}(p_0 \| \bar{p}) + \frac{1}{2}\mathsf{KL}(p \| \bar{p}) \to \max$$

- **Сумма степенных функций**, $\gamma > 1$:

$$\mathsf{Тематичность}(w) = g(w, \gamma) = \sum_{t \in T} p(t|w)^{\gamma} \to \max$$

Большие значения этих функционалов для N-граммы свидетельствуют о большем расстоянии до равномерного распределения, позволяя проводить еще одну часть отбора.

## 5 Признаковое описание

Три описанных подхода по отдельности не могут гарантировать высокую точность, поэтому требуется скомбинировать их для получения лучшего результата. В каждом из них есть несколько стратегий, по которым можно отбирать и оценивать термины. Данные стратегии генерируют для каждой N-граммы оценку, и чем выше эта оценка, тем уверенее можно сказать, что рассматриваемое словосочетание действительно термин. Таким образом, из этих оценок можно составить признаковое описание N-грамм, по которому предлагается настраивать финальный алгоритм отбора.

Опишем признаки для каждой стратегии фильтрации:

1. *Синтаксический анализ*

   Синтаксический подход к фильтрации позволяет делать только отбор N-грамм без промежуточных значений, поэтому было рассмотрено два индикатора:

   - Не отброшена ли N-грамма при фильтрации поддеревьев с помощью `SyntaxNet`.
   - В дополнении к предыдущему условию, не отброшена ли N-грамм при фильтрации по частям речи, описанным в Таблице 1.

2. *Статистический анализ*

   Было рассмотрено несколько статистик для N-грамм и алгоритм **TopMine**, значения полученные в этих критериях возьмем как признаки:

   - PMI
   - C-value
   - Индикатор того, что алгоритм **TopMine** взял N-грамму
   - Для взятых в ходе **TopMine** N-грамм значение *SignificanceScore*, иначе 0
   - Частота N-граммы

3. *Тематическая модель*

   В качестве признаков возьмем оценку тематичности $g(w, \gamma)$ при разных $\gamma$:

   - $g(w, \gamma = 2)$
   - $g(w, \gamma = 5)$

В итоге для каждой N-граммы конструировалось 9 признаков, разбитых на три группы.

В дополнению к базовому описанию строились дополнительные признаки, улучшающие качество линейной модели:

- Отрицательный индикатор того, что $g(k, w) = 1$, это часто говорит о том, что N-грамма имеет вырожденное распределение по темам, где только при одном $t$ $p(t|w) = 1$ и остальные условные вероятности равны 0.

- Отрицательный индикатор того, что c-value$(w) = 0$, так как это свидетельствует о том, что N-грамма всегда входит в K-грамму большего размера.

- Логарифм частоты N-граммы, который не так сильно как изначальное значение влияет на результат предсказания в линейной модели.

Таким образом, при построении линейной модели комбинирования подходов к фильтрации использовалось 12 признаков.

# 6 Вычислительные эксперименты

## 6.1 NIPS

Цель эксперимента состояла в проверки предположения о том, что комбинация описанных выше подходов к отбору терминов может выдавать достаточно хорошее качество отсеивания в сравнении с отдельными техниками. Также проверялось то, что каждый из трех подходов существенно влияет на итоговый результат.

Эксперимент проводился на коллекции аннотаций с конференции NIPS (Neural Information Processing Systems). Она содержала около 3200 аннотаций, в которых было около 500000 слов. Рассматривались би-граммы, 3-граммы и 4-граммы, среди которых было большое количество N-грамм с частотой, равной 1. График частот для би-грамм показан на Рис. 3, на нем видно, что более половины из всех би-грамм имеют единичную частоту. Такие N-граммы несут недостаточную информацию для статистических и тематических критериев, поэтому они были удалены из рассмотрения в отборе терминов, но информация о них все же использовалась при анализе остальных. В итоге осталось 45391 уникальных би-грамм, 50539 3-грамм, 26644 4-грамм.

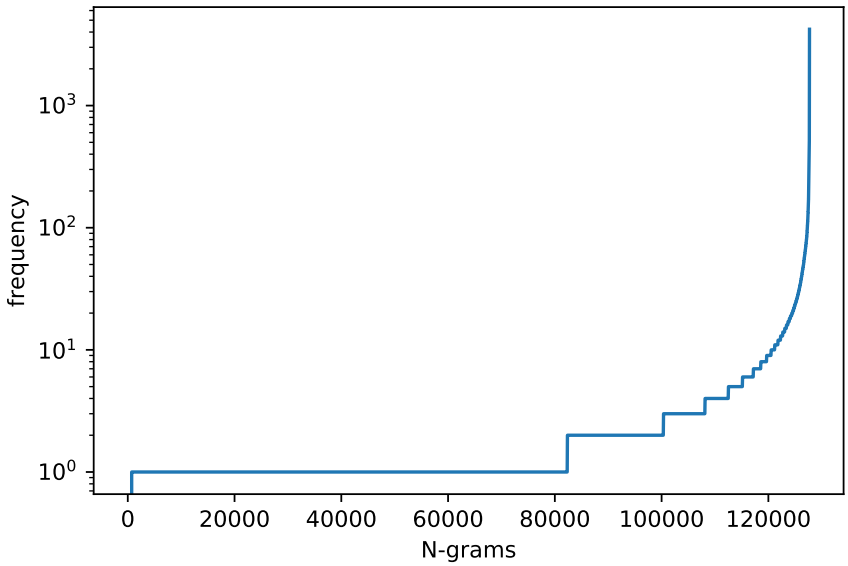


Рис. 3: Отсортированные значения частот для биграмм из коллекции аннотаций NIPS.

На основе заданного ранее признакового описания N-грамм были построены две модели: логистическая регрессия и градиентный бустинг `XGBoost` [6]. Логистическая регрессия как простая линейная модель строилась для того, чтобы по настроенным весам признаков легко проинтерпретировать значимость и вклад каждого. Градиентный бустинг наоборот является усложненной моделью и позволяет получить лучшее качество отбора терминов.

Настройка моделей и оценка качества алгоритмов проводились на двух *случайных* подвыборках размера 1000. В ходе настройки модели были удалены признаки, не меняющее качество фильтрации, веса и значимость оставленных признаков отображены в Таблице 2. Из нее видно, что каждая группа признаков вносит заметный вклад в настроенные модели. Признаки индикаторы не использовались при настройке `XGBoost`, для линейной модели существование таких индикаторов важно.

| Признак | Вес линейной модели | Значимость в градиентном бустинге |
|---|---|---|
| Синтаксический отбор с частями речи | 1.86 | 0.09 |
| C-value | -0.04 | 0.17 |
| $-\mathbb{I}[\text{c-value} = 0]$ | 1.67 | – |
| Логарифм частоты | 1.02 | 0.16 |
| $\mathbb{I}$[TopMine оставил N-грамму] | 1.06 | – |
| SignificanceScore для отобранных N-грамм, иначе 0 | 0.24 | 0.21 |
| Тематичность $g(w, \gamma = 2)$ | 2.85 | 0.38 |
| $-\mathbb{I}[g = 1]$ | 0.98 | – |

Таблица 2: Веса признаков в линейной модели и значимость (feature importance) в модели градиентного бустинга, построенных для эксперимента на коллекции аннотаций NIPS.

Оценки качества фильтрации проводились для различных комбинаций групп признаков для того, чтобы понять какие из этих подходов существенно влияют на конечный отбор признаков. Результаты тестирования показаны в Таблице 3. Линейная модель показывает результаты хуже градиентного бустинга, точность заметно меньше при одинаковой полноте, но это ожидаемый эффект от упрощения модели. Градиентный бустинг при полноте почти 100% показывает точность на уровне 40%. Итоговый лучший результат — 41% точности при полноте 99%. Каждая из групп признаков в моделях градиентного бустинга и логистической регрессии оказывает значимое влияние на финальное качество фильтрации. Также на Рис. 4 показана кривые Precision-Recall для лучших результатов двух моделей.

| Группа признаков | Линейная модель | | | Градиентный бустинг | | |
|---|---|---|---|---|---|---|
| | AUC-ROC | Точность | Полнота | AUC-ROC | Точность | Полнота |
| Только синтаксические признаки | 0.83 | 0.20 | 0.91 | 0.83 | 0.20 | 0.91 |
| Только статистические признаки | 0.71 | 0.09 | 0.94 | 0.73 | 0.11 | 0.90 |
| Только тематические признаки | 0.92 | 0.32 | 1.00 | 0.95 | 0.32 | 1.00 |
| Синтаксические + статистические | 0.88 | 0.22 | 0.91 | 0.88 | 0.24 | 0.91 |
| Синтаксические + тематические | 0.91 | 0.36 | 0.91 | 0.95 | 0.34 | 0.99 |
| Статистические + тематические | 0.93 | 0.29 | 0.94 | 0.98 | 0.34 | 1.00 |
| Все признаки | **0.95** | **0.38** | **0.91** | **0.97** | **0.41** | **0.99** |

Таблица 3: Значения AUC-ROC, точности и полноты для предсказаний логистической регрессии и градиентного бустинга на разных группах признаков для эксперимента на коллекции аннотаций NIPS.

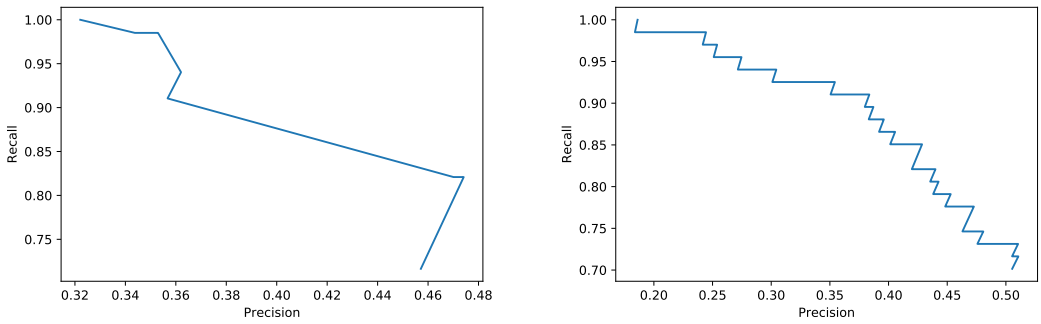


Рис. 4: Кривые Precision-Recall для предсказаний градиентного бустинга и логистической регрессии в эксперименте на коллекции аннотаций NIPS.

Для иллюстрации отбора терминов в Таблице 4 показаны N-граммы с наибольшими вероятностями в модели `XGBoost`. Все они отвечают рассматриваемой области конференции NIPS.

| N-грамма | Вероятность термина | Би-грамма | Вероятность термина |
|---|---|---|---|
| low rank matrix completion | 0.86 | conditional random field | 0.86 |
| multiple kernel learning | 0.86 | lifted inference | 0.86 |
| observable markov decision | 0.86 | multi armed bandit | 0.86 |
| stochastic block models | 0.86 | sparse inverse covariance estimators | 0.86 |
| retinal ganglion cells | 0.86 | working memory | 0.86 |
| latent dirichlet allocation | 0.86 | kernel based method | 0.86 |
| two sample testing | 0.86 | monte carlo method | 0.86 |
| gaussian mixture models | 0.86 | online convex optimization | 0.86 |

Таблица 4: N-граммы с наибольшими вероятностями в модели градиентного бустинга из эксперимента на коллекции аннотаций NIPS.

## 6.2 Выводы

Эксперимент на коллекции NIPS показал, что можно построить достаточно эффективный алгоритм предсказания терминов на основе трех основных групп признаков: синтаксической, статистическо и тематической. Каждая из групп вносит существенный вклад в итоговый ответ. Лучшее качество фильтрации, полученное в ходе комбинирования трех подходов к отбору терминов — 41% точности при полноте 99%.

# 7 Заключение

В рамках выпускной квалификационной работы был предложен алгоритм выделение терминов для улучшения качества интерпретируемости тематических моделей, состоящий из трех основных частей фильтрации (синтаксической, статистической, тематической), которые в свою очередь также включают несколько различных стратегий. Было показано, что предложенный алгоритм позволяет существенно сократить исходное множество словосочетаний, оставив в нем большинство терминов. Эксперимент на коллекции аннотаций NIPS (таблица 2 и 3) показывает, что, используя значения критериев N-грамм и результаты отбора терминов, при помощи приемлемой по размеру ручной разметки можно настроить эффективный алгоритм распознавания терминов. Улучшение качества фильтрации рассмотренным методом и применение новых техник на каждом из этапов требует дальнейшего изучения.

# Список литературы


[1] Announcing SyntaxNet: the world's most accurate parser goes open source. — https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html. — 2016.

[2] Globally normalized transition-based neural networks / Daniel Andor, Chris Alberti, David Weiss et al. // *arXiv preprint arXiv:1603.06042*. — 2016.

[3] Online documentation for Universal Dependencies, version 2. — http://universaldependencies.org. — 2016.

[4] *Frantzi Katerina, Ananiadou Sophia, Mima Hideki.* Automatic recognition of multi-word terms:. the c-value/nc-value method // *International journal on digital libraries*. — 2000. — Vol. 3, no. 2. — Pp. 115–130.

[5] Scalable Topical Phrase Mining from Text Corpora / Ahmed El-Kishky, Yanglei Song, Chi Wang et al. // *PVLDB*. — 2014. — Vol. 8, no. 3. — Pp. 305–316.

[6] *Chen Tianqi, Guestrin Carlos.* XGBoost: A Scalable Tree Boosting System // KDD 2016. — 2016.