Лаборатория интернет исследований

научный руководитель:

канд. физ.-мат. наук, доцент Департамента прикладной математики и бизнес-информатики Санкт-Петербургской школы экономики и менеджмента НИУ ВШЭ

Кольцов Сергей Николаевич

студент:

Агальцова Татьяна Александровна

Оптимизация тематического моделирования за счет изменения функции плотности в алгоритме семплирования Гиббса

Санкт-Петербург 2015

Тематическое моделирование

Тематическое моделирование - это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.

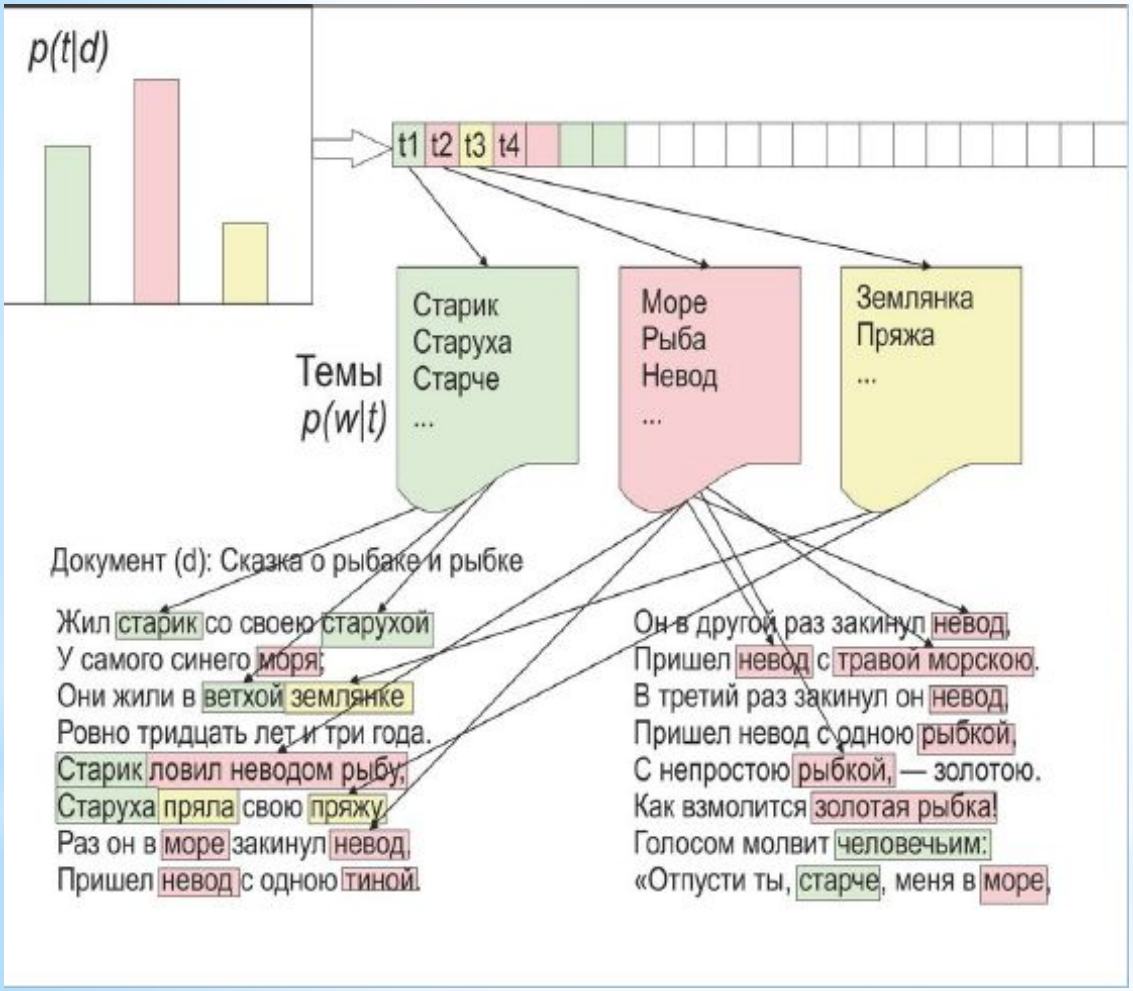
Тематическая модель (topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

Отвечает на вопросы:

1.Как выявлять смысл или тематику документов по их содержимому?

2. Как осуществлять классификацию документов на основе этих скрытых тематических закономерностей?

Тематическое моделирование



Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. 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.

Тематическое моделирование (Latent Dirichlet allocation)

Основное предположение тематической модели Latent Dirichlet Allocation состоит в том, что каждый документ с некоторой вероятностью может принадлежать множеству тематик. Тема - это совокупность слов, где каждое слово имеет некоторую вероятность принадлежности к данной тематике. Формально тема определяется как дискретное (мультиномиальное) вероятностное распределение в пространстве слов заданного словаря. Тематическим моделированием называется решение задачи, обратной классификации . Каждый документ в корпусе текстов рассматривается как наблюдаемая случайная независимая выборка слов (мешок слов), порождённая некоторым, скрытым (латентным) множеством тем. По этим данным требуется восстановить вероятностные распределения всех тем в корпусе и определить, каким именно подмножеством тем порождён каждый документ.
Тематическое моделирование основано на применении формулы Байеса, в которой распределение слов и тем выражено в виде смеси плотностей распределений слов и документов.

Тематическое моделирование

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Вероятностная модель порождения документа d:

$$p(w|d) = \sum_{t \in T} p(w|d, t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Дано $\hat{p}(w|d) \equiv n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Тематическое моделирование

Задача классификации заключается в расчете (оценке) апостериорной информации на основании априорной информации. Такая оценка может быть реализована при помощи формулы Байеса.

$$P(A|B) = \frac{p(\mathrm{A},\mathrm{B})}{P(A)} = \frac{p(A)\,\mathrm{P}(\mathrm{B}|A)}{P(A)}$$

$P(A|B)$ - Апостериорная вероятность

$p(A,\mathrm{B})$ - Априорная вероятность

Однако существует проблема оценивания априорной величины $p(A,\mathrm{B})$

Задача восстановления априорного распределения p(x,y)

Оценка функции p(x,y) может быть реализован при помощи трех методов.

1. Непараметрическое восстановление плотности основано на локальной аппроксимации плотности p(x) в окрестности классифицируемого объекта $x \in X$. Пример, Алгоритм Парзена-Розенблатта (метод парзеновского окна).

2. Параметрическое восстановление плотности основано на предположении, что плотность распределения известна с точностью до параметра, $p(x,y) = \phi(x; \theta)$, где ϕ фиксированная функция.

3. Восстановление смеси плотностей. Если функцию плотности p(x,y) не удаётся смоделировать параметрическим распределением, можно попытаться описать её смесью нескольких распределений:

Собственно именно третий метод является основой тематического моделирования.

Семплирование по Гиббсу

Семплирование по Гиббсу — алгоритм для генерации выборки совместного распределения множества случайных величин. Он используется для оценки совместного распределения и для вычисления интегралов методом Монте-Карло . Этот алгоритм является частным случаем алгоритма Метрополиса-Гастингса.

Семплирование по Гиббсу замечательно тем, что для него не требуется явно выраженное совместное распределение, а нужны лишь условные вероятности для каждой переменной, входящей в распределение. Алгоритм на каждом шаге берет одну случайную величину и выбирает ее значение при условии фиксированных остальных. Можно показать, что последовательность получаемых значений образуют возвратную цепь Маркова, устойчивое распределение которой является как раз искомым совместным распределением.

Применяется семплирование по Гиббсу в тех случаях, когда совместное распределение случайных величин очень велико или неизвестно явно, но условные вероятности известны и имеют простую форму.

Цели и задачи

Цель:

Оценить работу тематического моделирования при изменении структуры функции плотности, переходя от функции Дирихле к полетам Леви в алгоритме семплирования Гиббса.

Задачи:

1) Вычислить и запрограммировать полеты Леви.
2) Анализ полученных данных в topic maner.
3) Сравнение результатов, полученных из данной модели с результатами простой модели LDA.
4) Выявить преимущества и недостатки исследованной модели.