Тематическое моделирование

Виталий Радченко

Kyiv Kaggle Trainings

Введение

Что такое тема

- **Тема** — условное распределение на множестве терминов, $p(w|t)$ - вероятность (частота) термина w в теме t
- **Тематика документа** — условное распределение, $p(w|t)$ - вероятность (частота) термина w в теме t
- **Тематическая модель** — автоматически выявляет латентные темы по наблюдаемым частотам терминов в документах $p(w|t)$

Введение

Задачи

1. Классификация и категоризация документов
2. Автоматическое аннотирование документов
3. Автоматическая суммаризация коллекций
4. Тематическая сегментация документов

Введение

Цели

1. Семантический поиск информации
2. Визуализация тематической структуры коллекции
3. Анализ динамики развития тем
4. Тематический мониторинг новых поступлений
5. Рекомендации новых документов пользователям
6. Поиск научной информации
7. Побдор экспертов, рецензентов, исполнителей проектов
8. Агрегирование новосных потоков
9. Аннотация генома и другие задачи биоинформатики

Постановка задачи и подготовка данных

Подготовка данных

Предварительная очистка текстов:

- Удаление форматирование переносов
- Удаление обрывочной и нетекстовой информации
- Исправление опечаток
- Слияние слишком коротких текстов

Постановка задачи и подготовка данных

Подготовка данных

Форматирование словаря:

- Приведение слов к нормальной форме
- Выделение терминов
- Удаление стоп-слов и слишком редких слов

Постановка задачи и подготовка данных

Базовые предположения

- Порядок документов не важен
- Порядок терминов не важен
- Каждая пара (d, w) связана с некоторой темой $t \in T$
- Гипотеза условной независимости:

$$p(w|t,d) = p(w|t)$$

Постановка задачи и подготовка данных

Вероятностный процесс

Документ d - это смесь распределений p(w|t) с весами p(t|d):

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

Постановка задачи и подготовка данных

Постановка задачи

Дано:

- W - словарь терминов(слов или сочетаний)
- D - коллекция текстовых документов$d \subset D$
- n_{dw} - сколько раз термин w встретился в документе d

Найти:

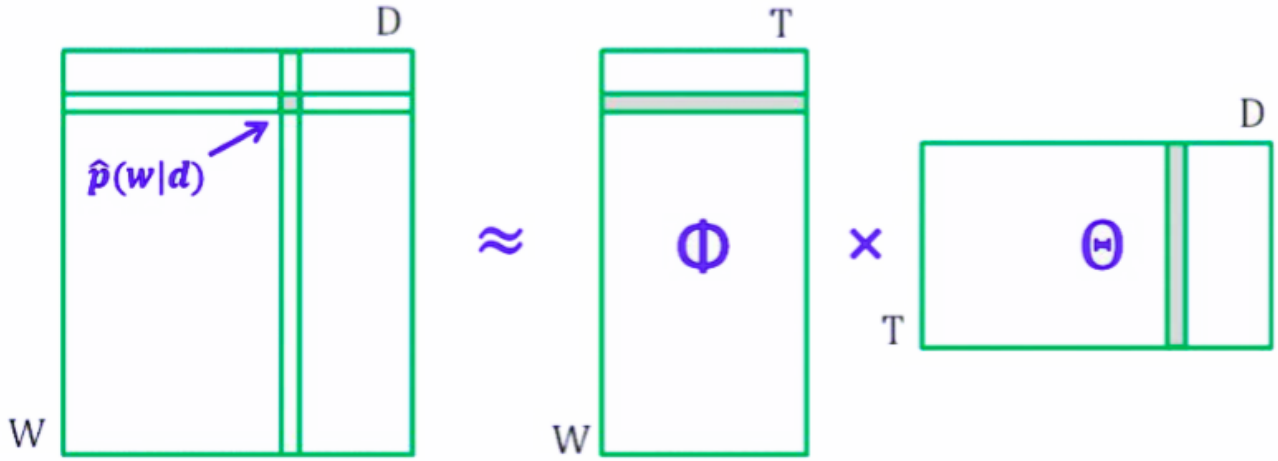
- Параметры вероятностной тематической модели

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

- $\phi_{wt} = p(w|t)$ - вероятность терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ - вероятность темы t в документе d

Матричное разложение

Принцип максимального правдоподобия



Наблюдаемые частоты терминов в документах:

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}$$

Вероятнотсная тематическая модель:

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

Матричное разложение

Принцип максимального правдоподобия

Максимизируется логарифм правдоподобия:

$$\begin{cases} \sum\limits_{d\in D}\sum\limits_{w\in d} n_{dw} \ln \sum\limits_{t\in T} \phi_{wt}\cdot\theta_{td} \to \max\limits_{\phi,\theta} \\ \sum\limits_{w\in W} \phi_{wt} = 1 \quad \phi_{wt} \geq 0 \\ \sum\limits_{t\in T} \theta_{td} = 1 \quad \theta_{td} \geq 0 \end{cases}$$

Матричное разложение

Принцип максимального правдоподобия

Что бы из множества решений выбрать наиболее подходящее, вводится критерий регуляризации $R(\Phi, \Theta)$

$$\begin{cases} \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw} \ln \sum\limits_{t \in T} \phi_{wt} \cdot \theta_{td} + R(\Phi, \Theta) \to \max\limits_{\Phi, \Theta} \\ \sum\limits_{w \in W} \phi_{wt} = 1 \quad \phi_{wt} \geq 0 \\ \sum\limits_{t \in T} \theta_{td} = 1 \quad \theta_{td} \geq 0 \end{cases}$$

Регуляризированный EM-алгоритм

Общее представление

$$\begin{cases} p_{tdw} = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}(\sum\limits_{d \in D} n_{dw} \cdot p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}) \\ \theta_{td} = \underset{t \in T}{\text{norm}}(\sum\limits_{w \in d} n_{dw} \cdot p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}) \end{cases}$$

Операция нормировки вектора

$$\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum\limits_{s \in T} \max\{x_s, 0\}}$$

Регуляризированный EM-алгоритм

Примеры

1. PLSA, вероятностный латентный семантический анализ

$$R(\Phi,\Theta) = 0$$

2. LDA, латентное размещение Дирихле

$$R(\Phi,\Theta) = \sum_{t,w}(\beta_w - 1)\cdot \ln \phi_{wt} + \sum_{d,t}(\alpha_t - 1)\cdot \ln \theta_{td}$$

где $\beta_w > 0, \alpha_t > 0$ - параметры регуляризатора

Регуляризация тематических моделей

Аддитивная регуляризация тематических моделей

Максимизация правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi\Theta}$$

где τ_i - коэффициенты регуляризации

Типы регуляризаторов:

- для учета дополнительных данных
- для получения решения Φ, Θ с заданными свойствами

Регуляризация тематических моделей

Разделение тем на предметные и фоновые

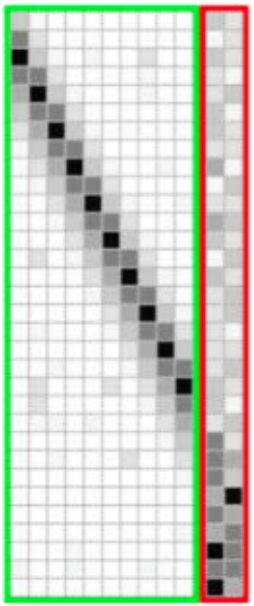
- Предметные темы S содержат термины предметной области
- Фоновые темы B содержат слова общей лексики

Регуляризация тематических моделей

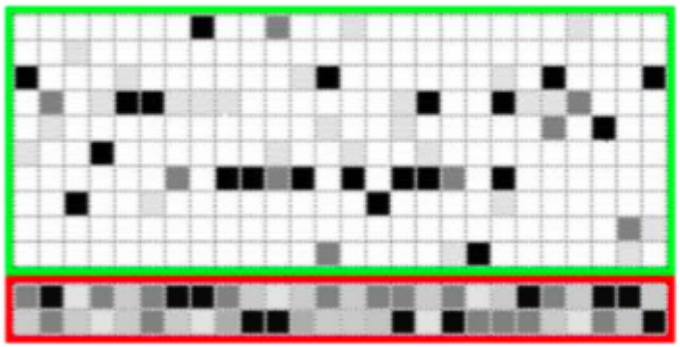
Разделение тем на предметные и фоновые

- Предметные темы S - разреженные, существенно различные
- Фоновые темы B - существенно отличные от нуля

$\Phi_{W \times T}$



$\Theta_{T \times D}$



Регуляризация тематических моделей

Регуляризатор сглаживания фоновых тем

- Распределения ϕ_{wt} близки к заданому распределению β_w
- Распределения θ_{td} близки к заданому распределению α_t

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B} \sum_{w\in W} \beta_w \ln \phi_{wt} + \\ + \alpha_0 \sum_{d\in D} \sum_{t\in B} \alpha_t \ln \theta_{td} \to \max$$

где α_0, β_0 - коэффициенты регуляризации

Регуляризация тематических моделей

Регуляризатор сглаживания фоновых тем

- Распределения ϕ_{wt} **далеки** от заданого распределения β_w
- Распределения θ_{td} **далеки** от заданого распределения α_t

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \to \max$$

где α_0, β_0 - коэффициенты регуляризации

Регуляризатор декоррелирования тем

Лексическое ядро

- **Лексическое ядро темы** - множество терминов, отличающее ее от других тем.
- Минимизируем ковариации между вектор-столбцами ϕ_t :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \neq s \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws} \to \max$$

где τ - коэффициент регуляризации

Регуляризатор декоррелирования тем

Регуляризатор для отбора тем

- Разреживаем распределение

$$p(t) = \sum_d p(d)\theta_{td}$$

максимизируя KL-дивергенцию между $p(t)$ и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max$$

где τ - коэффициент регуляризации

Регуляризатор декоррелирования тем

Дивергенция Кульбака-Лейблера

- Расстояние между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

$$KL(P||Q) = \sum_i p_i \ln \frac{p_i}{q_i}$$

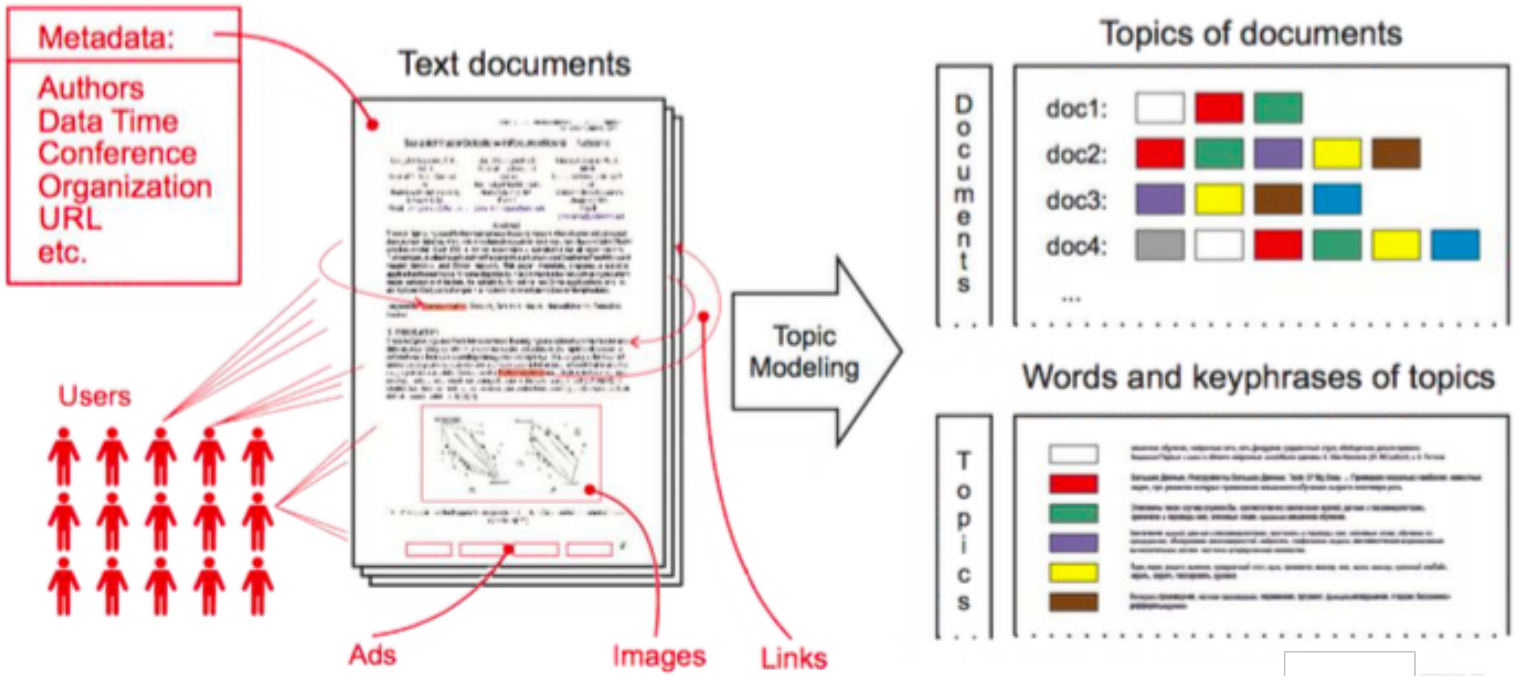
- Связь с принципом максимума правдоподобия

$$\sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \Leftrightarrow \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha$$

Мультимодальная тематическая модель

Выделение модальностей

Выявляет тематику документов $p(t|d)$, терминов $p(t|w)$ и токенов других модальностей:
$p(t|author), p(t|time), p(t|URL), p(t|user)...$



Внутренние критерии качества тематических моделей

Перплексия

Перплексия коллекции D для языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(d) = exp\Big(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \Big)$$

Внутренние критерии качества тематических моделей

Меры интерпритируемости тем

Тема интерпритируема, если по топовым словам темы эксперт может определить, о чем эта тема, и дать ее название

- Метод интрузий
  - В список топоповых слов внедряется лишнее слово
  - Измеряется доля ошибок экспертов при его определении

Внутренние критерии качества тематических моделей

Когерентность

- Когерентность темы t — средняя поточечная взаимная информация топ-слов темы (pointwise mutual information, PMI):

$$PMI_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} PMI(w_i, w_j)$$

где w_i — i-термин в порядке убывания ϕ_{wt}, $k = 10$

- $PMI(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ – поточечная взаимная информация
- N_{uv} — число документов, в которых термины u,v хотя бы один раз встречаются рядом (в окне 10 слов)
- N_u — число документов, в которых u встречается хотя бы один раз

Внешние критерии качества тематических моделей

Способы оценивания близости запроса q и документа d

- Косинусная мера(чем больше, тем ближе):

$$\cos(q,d) = \frac{\sum_t p(t|q)p(t|d)}{\left(\sum_t p(t|q)^2\right)^{1/2}\left(\sum_t p(t|d)^2\right)^{1/2}}$$

- Расстояние Хеллингера (чем меньше, тем ближе):

$$H^2(q,d) = \frac{1}{2}\sum_t(\sqrt{p(t|d)} - \sqrt{p(t|q)})^2$$

- KL-дивергенция (чем меньше, тем ближе):

$$KL(q,d) = \sum_t p(t|q)\ln\frac{p(t|q)}{p(t|d)}$$

Сравнение методов

kaggle

LDA	ARTM
Очень популярный	Молодой
Множество модифицкаций для разных задач	Мощный аппарат регуляризаторов для модифицирования модели
Для каждого усложнения нужно искать реализацию	Одна реализация для разных задач
Нужно настраивать гиперпараметры	Нужно настраивать параметры регуляризации

Реализация в Python

kaggle

gensim для LDA	**BigARTM** для ARTM
Есть функционал для решения разных задач анализа текстов	Специализированная библиотека для тематического моделирования
Проще в использовании	Больше возможностей, но чуть-чуть больше кода
Дольше обучается	Быстрее обучается
Больше форматов данных, самый понятный — UCI Bag of Words	Можно импортировать данные в формате UCI Bag of Words, но vowpal wabbit формат проще

Материалы

kaggle

1. Курс "Поиск структруры в данных на курсере"
2. Вероятностные тематические модели (курс лекций, К.В.Воронцов)

Спасибо за внимание!