MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

АТП кафедра Алгоритмов и Технологий Программирования

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Автоматическая обработка текстов
Тематическое моделирование

Лекция 3. Часть 1

Емельянов А. А.
login-const@mail.ru

2018

https://github.com/king-menin/nlp-hse-winter2018

Тематическое моделирование

- **Тематическое моделирование** — современное направление статистического анализа текстов. Методы тематического моделирования призваны ответить на вопрос, какой теме посвящена большая коллекция текстовых документов.

Что такое тема?

- **Тема** – специальная терминология предметной области.

- **Тема** – набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Что такое тема?

Более формально:

- **Тема** – это условное распределение на множестве терминов, $p(w|t)$ – вероятность (частота) термина w в теме t.

- **Тематика документа** – условное распределение, $p(t|d)$ – вероятность (частота) темы t в документе d.

- **Тематическая модель** должна автоматически выявлять латентные темы по наблюдаемым частотам терминов в документе $p(w|d)$.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

- **Задачи:**
  - Классификация и категоризация документов.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

- **Задачи:**
  - Классификация и категоризация документов.
  - Автоматическое аннотирование документов.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

- **Задачи:**
  - Классификация и категоризация документов.
  - Автоматическое аннотирование документов.
  - Автоматическая суммаризация коллекций.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

- **Задачи:**
  - Классификация и категоризация документов.
  - Автоматическое аннотирование документов.
  - Автоматическая суммаризация коллекций.
  - Тематическая сегментация документов.

Цели и задачи тематического моделирования

- **Цель** – автоматический анализ текста.

- **Задачи:**
  - Классификация и категоризация документов.
  - Автоматическое аннотирование документов.
  - Автоматическая суммаризация коллекций.
  - Тематическая сегментация документов.

- **Идея решения:**
  - Использовать признаковые описания документов $p(t|d)$.

Цели и задачи тематического моделирования

- **Цель** – систематизация больших объёмов информации.

Цели и задачи тематического моделирования

- **Цель** – систематизация больших объёмов информации.

- **Задачи:**
  - Семантический (разведочный) поиск информации.

Цели и задачи тематического моделирования

- **Цель** – систематизация больших объёмов информации.

- **Задачи:**
  - Семантический (разведочный) поиск информации.
  - Визуализация тематической структуры коллекции.

Цели и задачи тематического моделирования

- **Цель** – систематизация больших объёмов информации.

- **Задачи:**
  - Семантический (разведочный) поиск информации.
  - Визуализация тематической структуры коллекции.
  - Анализ динамики развития тем.

Цели и задачи тематического моделирования

- **Цель** – систематизация больших объёмов информации.

- **Задачи:**
  - Семантический (разведочный) поиск информации.
  - Визуализация тематической структуры коллекции.
  - Анализ динамики развития тем.
  - Рекомендация документов пользователям.

Приложения тематического моделирования

- Поиск научной информации, трендов, фронта исследований.

Приложения тематического моделирования

- Поиск научной информации, трендов, фронта исследований.

- Подбор экспертов, рецензентов, исполнителей проектов.

Приложения тематического моделирования

- Поиск научной информации, трендов, фронта исследований.
- Подбор экспертов, рецензентов, исполнителей проектов.
- Агрегирование новостных потоков.

Приложения тематического моделирования

- Поиск научной информации, трендов, фронта исследований.
- Подбор экспертов, рецензентов, исполнителей проектов.
- Агрегирование новостных потоков.
- Аннотирование и поиск изображений.

Приложения тематического моделирования

- Поиск научной информации, трендов, фронта исследований.
- Подбор экспертов, рецензентов, исполнителей проектов.
- Агрегирование новостных потоков.
- Аннотирование и поиск изображений.
- Анализ видеопоследовательностей, аннотация генома и другие задачи биоинформатики, анализ дискретизированных биомедицинских сигналов, мониторинг состояния технических систем.

Примеры использования тематического моделирования

- Мультиязычная модель Википедии.
  - 216 175 русско- английских пар статей Википедии и собрано 400 тем.
- Первые 10 слов (с их вероятностями $p(w|t)$ в %) в каждой из представленных тем.

тема 68				тема 79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

- Модель выявляет двуязычные темы без выравнивания, без словарей, даже когда тексты не являются точными переводами.
- В этом эксперименте независимый эксперт оценил 396 тем из 400 как хорошо интерпретируемые.

Примеры использования тематического моделирования

- Биграммная модель термины – словосочетания.
  - Несколько тем, построенные по 850 статьям конференций ММРО, ИОИ на русском языке.

распознавание образов в биоинформатике	
unigrams	bigrams
объект	задача распознавания
задача	множество мотивов
множество	система масок
мотив	вторичная структура
разрешимость	структура белка
выборка	распознавание вторичной
маска	состояние объекта
распознавание	обучающая выборка
информативность	оценка информативности
состояние	множество объектов
закономерность	разрешимость задачи
система	критерий разрешимости
структура	информативность мотива
значение	первичная структура
регулярность	тупиковое множество

теория вычислительной сложности	
unigrams	bigrams
задача	разделять множества
множество	конечное множество
подмножество	условие задачи
условие	задача о покрытии
класс	покрытие множества
решение	сильный смысл
конечный	разделяющий комитет
число	минимальный аффинный
аффинный	аффинный комитет
случай	аффинный разделяющий
покрытие	общее положение
общий	множество точек
пространство	случай задачи
схема	общий случай
комитет	задача MASC

Подготовка данных для тематического моделирования

- Удаление форматирования и переносов.
- Удаление обрывочной и нетекстовой информации.
- Исправление опечаток.
- Слияние слишком коротких текстов.
- Выделение терминов (term extraction).
- Приведение слов к нормальной форме.
- Удаление стоп-слов и слишком редких слов

Базовые предположения простых тематических моделей

- **Порядок документов в коллекции не важен.**

Базовые предположения простых тематических моделей

- **Порядок документов в коллекции не важен.**

- **Порядок терминов в документе не важен (bag of words)**: переставив в документе слова или выделенные словосочетания, все равно можно определить его тематику.

Базовые предположения простых тематических моделей

- **Порядок документов в коллекции не важен.**

- **Порядок терминов в документе не важен (bag of words)**: переставив в документе слова или выделенные словосочетания, все равно можно определить его тематику.

- **Употребление каждого слова в каждом документе связано с некоторой темой**, то есть каждая пара (d, w) связана с некоторой темой $t \in T$.

Базовые предположения простых тематических моделей

- **Порядок документов в коллекции не важен.**
- **Порядок терминов в документе не важен (bag of words)**: переставив в документе слова или выделенные словосочетания, все равно можно определить его тематику.
- **Употребление каждого слова в каждом документе связано с некоторой темой**, то есть каждая пара (d, w) связана с некоторой темой $t \in T$.
- **Гипотеза условной независимости**: $p(w|t, d) = p(w|t)$ заключается в том, что вероятность слова документах определяется только темой, а не самим документом. **Следовательно**, коллекция документов представляет собой последовательность троек (d, w, t), в которой темы являются латентными: *они не видны и для их определения как раз используется тематическая модель*.

Дополнительные предположения простых тематических моделей

- Дополнительные предположения разреженности:
  - Предположение, что документ относится к небольшому числу тем.

Дополнительные предположения простых тематических моделей

- Дополнительные предположения разреженности:
  - Предположение, что документ относится к небольшому числу тем.
  - Предположение, что тема состоит из небольшого числа терминов, лексического ядра, которое существенно отличает эту тему от остальных.

Вероятностный процесс порождения текстовой коллекции

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

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

Постановка задачи тематического моделирования

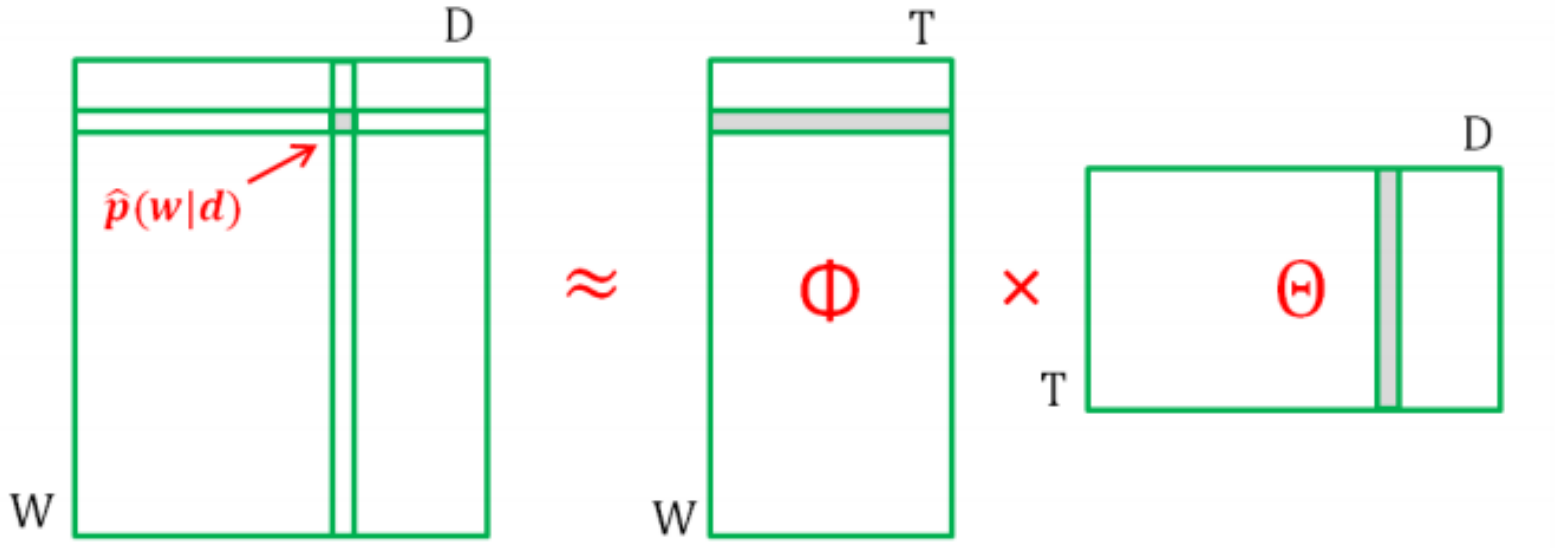
- **Дано**:
  - W – словарь терминов (слов или словосочетаний),
  - D – коллекция текстовых документов $d \subset W$,
  - n_{dw} – сколько раз термин w встретился в документе d,
  - n_d – длина документа.
- **Найти**:
  - Параметры вероятностной тематической модели

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

  - $\phi_{wt} = p(w|t)$,
  - $\theta_{td} = p(t|d)$.

Постановка задачи тематического моделирования

- Порождающая модель описывает процесс построения коллекции по ϕ_{wt} и θ_{td}. Тематическое моделирование представляет собой обратную задачу: по наблюдаемой коллекции необходимо понять, какими распределениями ϕ_{wt} и θ_{td} она могла бы быть получена.

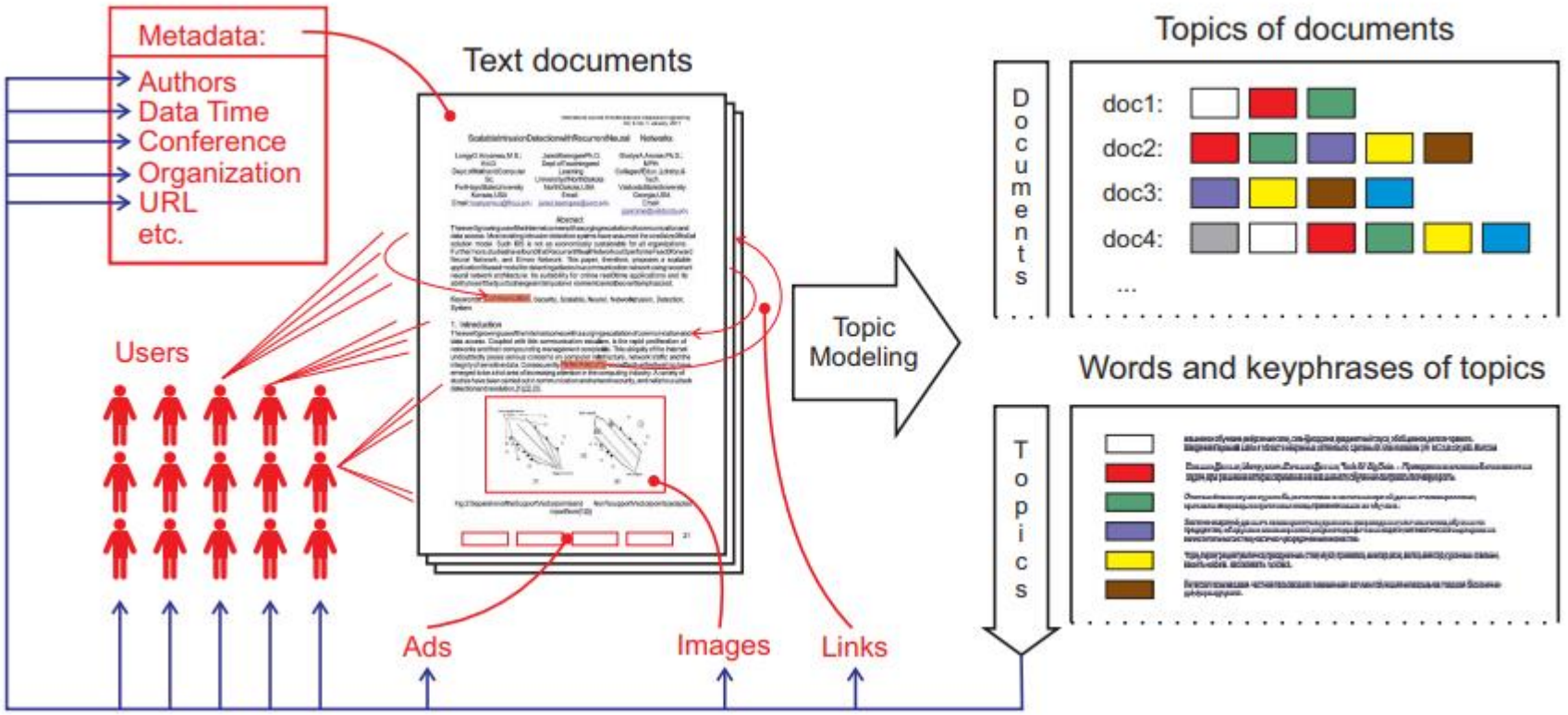


Решение задачи тематического моделирования

- EM-алгоритм – решает задачу в общем случае.
- Латентный семантический анализ (Latent Semantic Indexing).
- Латентное размещение Дирихле (Latent Dirichlet allocation) – частный случай EM алгоритма.

Мультимодальные тематические модели

- Примеры модальностей:
  - Авторы, моменты времени и так далее.
  - Элементы изображений.
  - Множество ссылок на другие документы.
  - Множество рекламных баннеров, которые появились на данной странице, а также множество пользователей, которые кликнули на данные баннеры.
  - Множество пользователей, сделавших определенное действие с документом (скачал, лайкнул, поставил оценку и так далее).



Меры качества тематической модели

- **Перплексия** (Perplexity)
  - Перплексия коллекции D для языковой модели $p(w|d)$:

$$P(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right), \qquad n = \sum_{d\in D}\sum_{w\in d} n_{dw}.$$

  - Можно сказать, что это мера неопределенности или различности слов в тексте. Если распределение слов неравномерно, то перплексия уменьшается по сравнению с тем значением, которое дает равномерное распре- деление. Еще можно сказать, что перплексия — коэффициент ветвления текста, то есть количество ожидаемых в среднем различных слов после каждого слова в документе.

Меры качества тематической модели

- **Когерентность** (Согласованность) 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-ый термин в порядке убывания ϕ_{wk}, $k = 10$.

Поточечная взаимная информация

$$PMI(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v},$$

где N_{uv} — число документов, в которых термины u и v хотя бы один раз встречаются рядом (в окне 10 слов), N_u — число документов, в которых термин u встретился хотя бы один раз. Чем выше величина поточечной взаимной информации, тем выше неслучайность того, что два слова стоят рядом.

Методы тематического моделирования

- Gensim.lda (Radim Řehůřek)

- BigARTM (Yandex)

LDA	ARTM
Очень популярный	Молодой
Множество модификации для различных 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

Пример применения в python

- topic modeling.ipynb

- (для тех, кто хочет выполнять семинары онлайн) topic modeling.ipynb

СПАСИБО ЗА ВНИМАНИЕ

https://github.com/king-menin/nlp-hse-winter2018