Урок 8

Тематическое моделирование-2

Оценки качества делятся на две большие категории: внутренние критерии качества и внешние критерии качества. Внутренние критерии — такие критерии, которые позволяют оценить качество модели по матрицам Φ и Θ, которые модель дала на выходе. Внешние критерии измеряют качество модели, глядя на то, как она решает ту конечную прикладную задачу, ради которой она и была создана.

8.1. Внутренние критерии качества тематических моделей

8.1.1. Перплексия (Perplexity)

Перплексия — известная в вычислительной лингвистике мера качества модели языка. В данном случае моделью языка является условное распределение слов в документах.

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

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

Эта мера качества тесно связана с правдоподобием. По сути дела это экспонента от усредненного по всем словам всех документов значения логарифма правдоподобия.

Значение перплексии можно интерпретировать следующим образом. Если подставить в качестве распределения слов в документах равномерное распределение:

$$p(w|d) = \frac{1}{|W|},$$

получится, что перплексия равна мощности словаря:

$$P = |W|.$$

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 распределение. Еще можно сказать, что перплексия — коэффициент ветвления текста, то есть количество ожидаемых в среднем различных слов после каждого слова в документе.

8.1.2. Hold-out perplexity

Перплексия может быть вычислена по самой коллекции, по которой построена тематическая модель, но тогда существует риск переобучения. Чтобы избежать этого и получить несмещенную оценку, вычисляют перплексию тестовой (отложенной) коллекции D' (hold-out perplexity):

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

В отличие от предыдущего случая, параметры ϕ_{wt} оцениваются по обучающей коллекции D. Для каждого документа $d \in D$ строится случайное разбиение $d = d' \sqcup d''$ на две половины равной длины, причем параметры θ_{td} оцениваются по d', а перплексия вычисляется по d''.

Эксперименты на больших коллекциях показывают, что различие между перплексией на обучающей и на тестовой выборках, как правило, не существенно для цели сравнения разных моделей. Поэтому рекомендуется на очень больших выборках не считать hold-out perplexity, а считать обычную перплексию по основным данным.

8.1.3. Меры интерпретируемости тем

Перплексия ничего не говорит об интерпретируемости тем, а только говорит о том, насколько хорошо построилось матричное разложение. То есть ничего не говорит о том, насколько построенная модель будет полезна для конечных приложений. Поэтому были придуманы меры качества, которые измеряют, насколько темы хороши и понятны. Такую оценку можно сделать только при помощи экспертов.

Например, экспертам предлагается рассмотреть темы как последовательности слов, упорядоченные по вероятности встретить слово в данной теме, и оценить интерпретируемость темы по некоторой шкале (2 или 5-бальной). Обычно экспертам дают такую инструкцию: если эту тему можно кратко озаглавить или по этим словам можно сформулировать поисковый запрос и получить релевантную поисковую выдачу, то такую тему следует считать интерпретируемой. Такие оценки, естественно, являются субъективными, поэтому каждую тему оценивают несколько экспертов, чтобы понять, насколько непротиворечивы их оценки.

Другой метод, так называемый метод интрузий (intrusion), заключается в том, что в список топовых (имеющих большую вероятность) слов темы внедряется лишнее слово, которое заведомо этой теме не принадлежит. Экспертам предлагается определить, какое слово из списка лишнее, и измеряется доля ошибок при его определении. Такой способ оценки интерпретируемости существенно проще для экспертов, а, следовательно, эксперт может оценить больше тем в единицу времени.

8.1.4. Когерентность (Согласованность)

В ходе экспериментов было выявлено, что экспертные оценки хорошо коррелируют с такой мерой качества как когерентность, которая может быть вычислена полностью автоматически без участия человека. Когерентность темы — мера, которая показывает, насколько слова, встречающиеся рядом в текстах, оказываются в топах одних и тех же тем.

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-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.

8.2. Внешние критерии качества тематических моделей

Если внутренние критерии качества оценивают матрицы Φ и Θ построенной тематической модели, то внешние критерии качества основаны на измерении полезности построенной тематической модели для исходной прикладной задачи.

8.2.1. Примеры задач классификации и категоризации документов

Задача классификации документов является задачей машинного обучения с учителем. Типичное применение задачи классификации:

- Определение жанра документа (художественный, научный, учебный, рекламный и так далее). Это может быть необходимо для анализа потока документов, полученных поисковыми роботами из интернета.
- Определение тематики сообщения в новостном потоке (политика, экономика, наука, здоровье, спорт и так далее)
- Определение категории (рубрики), к которой относится документ (например, Наука/Физика/Большой Адронный Коллайдер или Спорт/Футбол/Чемпионат мира по футболу). Основная особенность — наличие иерархической структуры.

Обучающая информация, которая используется для построения модели классификации, является результатом ручной классификации документов экспертами. То есть имеется обучающая выборка, в которой эксперты указали, к какой категории/жанру/тематике принадлежит тот или иной документ. Эта информация может быть использована для оценивания качества тематической модели.

8.2.2. Задача классификации

Безусловно, можно применить механизм мультимодальных моделей и использовать две модальности: термины $w \in W$ и классы $c \in C$.

На этапе обучения на вход подается коллекция документов, то есть матрицы (n_{dw}) (словарный состав документов) и (n_{dc}) (принадлежность документов классам, причем документ может быть отнесен сразу к нескольким классам), а на выходе — модель коллекции $p(w|t)$, $p(c|t)$. По полученной тематической модели можно будет классифицировать документы, для которых никаких классов не известно.

На этапе классификации на вход подается документ $d : (n_{dw})$ и известна модель классификации, построенная на этапе обучения. В первую очередь определяется тематика данного документа $p(t|d)$, а затем по модели, построенной на этапе обучения, определяются условные вероятности каждого класса для данного документа:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d).$$

На основании этого распределения определяется принадлежность документа классу или классам.

8.2.3. Критерии качества классификации или категоризации

Оценивать качество работы алгоритма, возвращающего условные вероятности классов для заданного документа, можно через:

- Число ошибок классификации (считается, что документ относится к классу, если соответствующая условная вероятность больше некоторого порогового значения). Такой способ является не очень хорошим, поскольку классы могут иметь разную важность и разную мощность.
- Критерий чувствительности и специфичности. Этот критерий подходит для работы с несбалансированными выборками.
- Критерий площади под кривой чувствительность–специфичность (AUC). Такой критерий не зависит от выбранного порога для классификации.
- Критерий точности и полноты.
- Критерий площади под кривой точность–полнота (AUC-PR).

8.2.4. Точность и полнота многоклассовой классификации

Пусть c — некоторый класс, TP_c — количество документов, которые были верно отнесены алгоритмом к классу c, FP_c — ошибочно отнесенных к классу c, FN_c — ошибочно неотнесенных к классу c. Точность (precision) — доля релевантных документов среди всех отнесенных алгоритмом к классу c, полнота (recall) — доля документов, отнесенных алгоритмом к классу c, среди всех релевантных. Точность P_c и полнота R_c для класса c выражаются так:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \qquad R_c = \frac{TP_c}{TP_c + FN_c}.$$

Если классов много, существуют две стратегии:

- Точность и полнота с микроусреднением:

$$P = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c)}, \qquad R = \frac{\sum_c TP_c}{\sum_c (TP_c + FN_c)}.$$

- Точность и полнота с макроусреднением:

$$P = \frac{1}{|C|} \sum_c P_c, \qquad R = \frac{1}{|C|} \sum_c R_c.$$

Принципиальная разница состоит в том, что макроусреднение более чувствительно к качеству классификации маломощных классов. При микроусреднении качество работы на маломощных классах слабо влияет на полное качество.

8.2.5. Тематический поиск

Тематический поиск — еще одна задача, которая часто используется для оценивания качества тематических моделей.

На этапе обучения на вход подается коллекция документов, то есть матрица слов в документах (n_{dw}), а на выходе строится тематическая модель коллекции $p(t|d)$. На этапе поиска по поисковому запросу (документу произвольной длины) $q : (n_{qw})$, используя построенную на этапе обучения модель коллекции, строится модель запроса $p(t|q)$. Также выводятся документы коллекции, ранжированные по близости к запросу.

Для этого необходимо уметь сравнивать условные распределения тем у запроса $p(t|q)$ и у документа d из коллекции $p(t|d)$.

8.2.6. Оценивание близости запроса и документа

Для сравнения вероятностных распределений тем в информационном поиске принято использовать несколько мер (которые показали себя наиболее эффективными в экспериментах на реальных задачах):

- Косинусная мера (чем больше, тем ближе):

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

- Расстояние Хеллингера (чем меньше, тем ближе):

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

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

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

Следует отметить, что KL-дивергенция является несимметричной мерой и имеет смысл того, насколько запрос подходит под данный документ. KL-дивергенцию имеет смысл использовать в тех случаях, когда запрос короче документов.

Еще один момент заключается в том, что на косинусную меру и расстояние Хеллингера может влиять «обрезание хвостов распределений». Этот прием используется, чтобы отбросить редкие темы, которые обычно мало интересуют. Это позволяет иногда повысить качество поиска и его скорость.

8.2.7. Критерии качества тематического поиска

Качество тематической модели по результату тематического поиска можно оценить по:

- точности первых k позиций поисковой выдачи (требуется экспертная оценка релевантности),
- средней позиции некоторого документа в поисковой выдаче при поиске по его аннотации,
- средней точности по релевантным позициям (MAP = Mean Average Precision) при поиске фрагментов документа по другим фрагментам,
- средней позиции документа при поиске по его переводу на другой язык (при кросс-язычном поиске).

8.3. Визуализация тематических моделей

Тематические модели, как правило, создаются для упрощения понимания и обеспечения навигации по большим текстовым коллекциям, поэтому важной задачей является визуализация тематических моделей. В последние годы было создано достаточно много средств визуализации, многие из которых находятся в свободном доступе. Большинство из этих инструментов ориентированы на то, чтобы визуализировать текстовые коллекции через web-интерфейсы.

8.3.1. Система TMVE

Система Topic Model Visualization Engine является одним из канонических примеров тематического навигатора с web-интерфейсом.

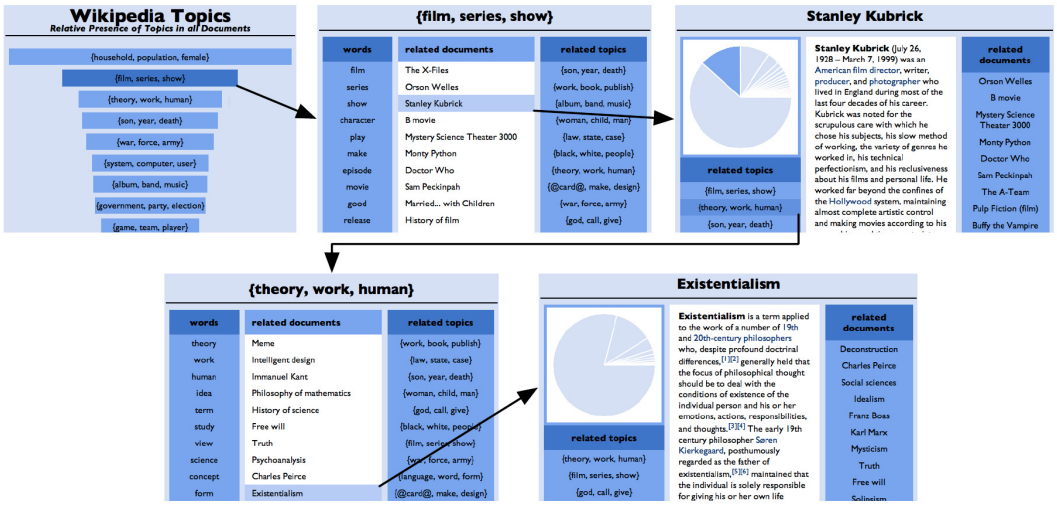


Рис. 8.1: «Wikipedia Topics», демонстрационный пример, представленный авторами TMVE.

На главной странице системы находится список тем, по каждой теме можно просмотреть документы и термины этой темы. Таким образом реализуется возможность навигации пользователя по коллекции.

8.3.2. Система TERMITE

Система TERMITE позволяет интерактивно визуализировать матрицу Φ и больше подходит для разработчиков тематической модели.

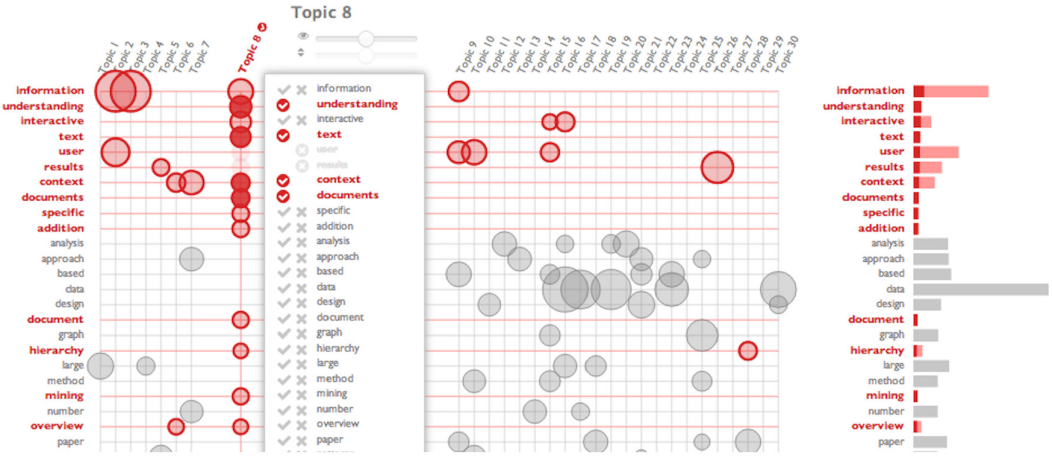


Рис. 8.2: Система TERMITE

8.3.3. Динамические модели, учитывающие время

Есть огромное количество средств визуализации для тематических моделей потоков новостей, научных статей или любых других коллекций, где каждому документу приписывается метка времени. Тогда строить тематические модели очень удобно, визуализируя их в виде графиков, на которых отображено, как развивались темы во времени, в какие моменты темы набирали популярность.

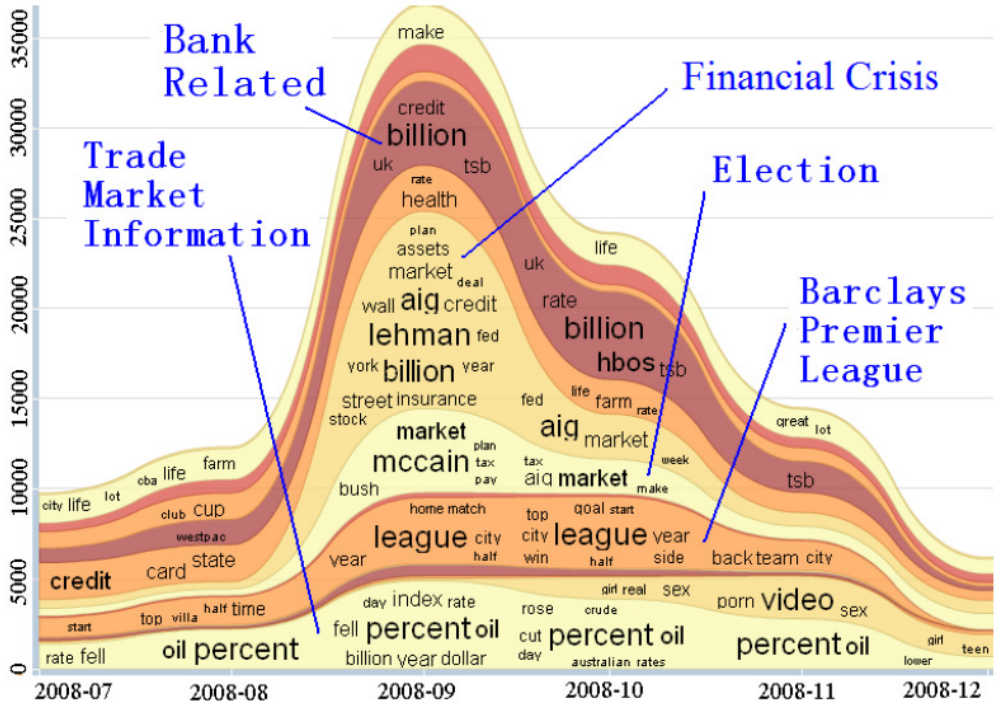


Рис. 8.3: На этом графике видны темы, которые возникли в связи с финансовым кризисом 2008.

На таких графиках можно изучать предвестники, последствия и связанные темы.

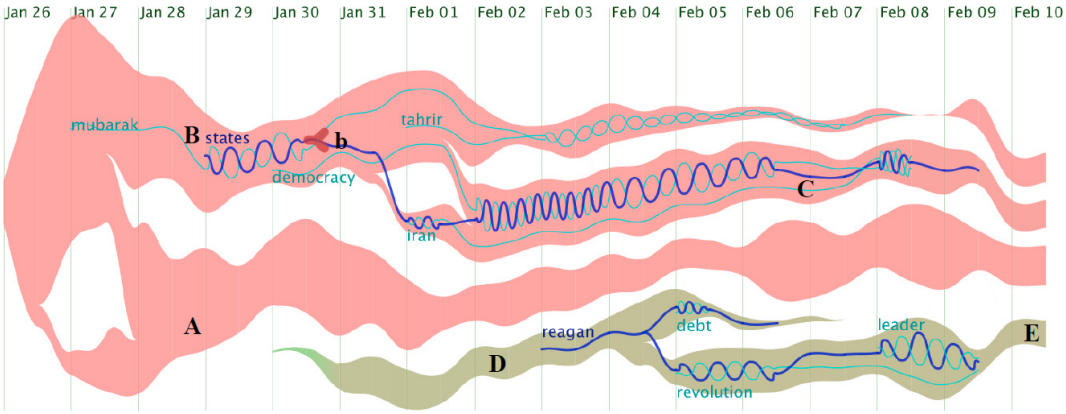
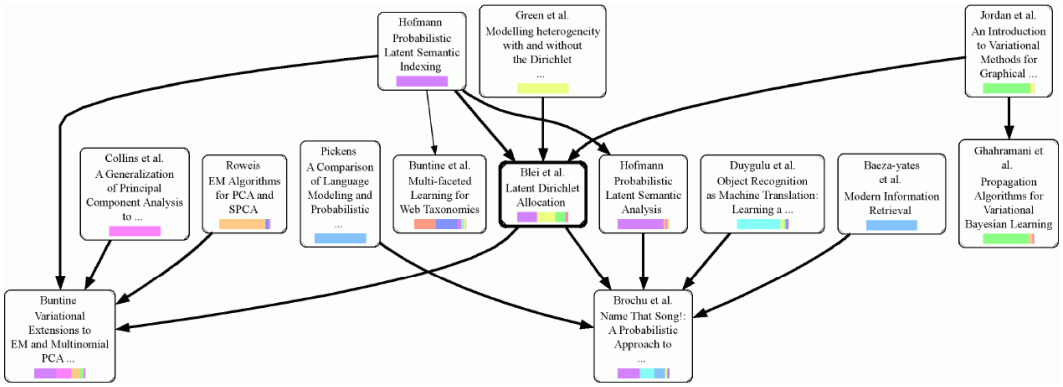


Рис. 8.4: Визуализация «река тем»: изображены моменты зарождения тем, исчезновения. Волнами изображаются траектории отдельных слов, причем часто колеблющаяся волна значит, что слово использовалось часто.

8.3.4. Динамические модели, учитывающие ссылки

Если тематическая модель учитывает не только слова, но также связи между документами, например связи цитирования между научными статьями, то можно ставить очень интересные задачи.

Например, можно попытаться ответить на вопрос, какие предшествующие работы действительно существенно повлияли на данную статью. В статье часто десятки ссылок, многие из которых чисто формальные, или дань вежливости, или же незначительные моменты, которые для данной статьи не важны. Оказывается, что это можно сделать с помощью тематической модели: выявить тематику статьи и выбрать из списка литературы те статьи, которые тоже соответствуют этой тематике.



С другой стороны, использование ссылок и цитат позволяет уточнить саму тематическую модель. Для этого предполагается, что если статья ссылается на другую, то у них есть общая тематика, и это учитывается с помощью регуляризатора.

8.3.5. Выявление взаимосвязей между темами

Оказывается, что можно выявлять связи между темами. Это особенно хорошо получается на коллекциях научных текстов. Так, например, в статье про археологию скорее появится термин из геологии, чем из генетики. Выявление таких связей между отраслями знаний представляет отдельный прикладной интерес.

Если каждую тему изображать в виде вершины графа, а ребро проводить только в том случае, когда соответствующие две темы часто появлялись в документах одновременно, то получившаяся тематическая модель будет называться кореллированной тематической моделью.

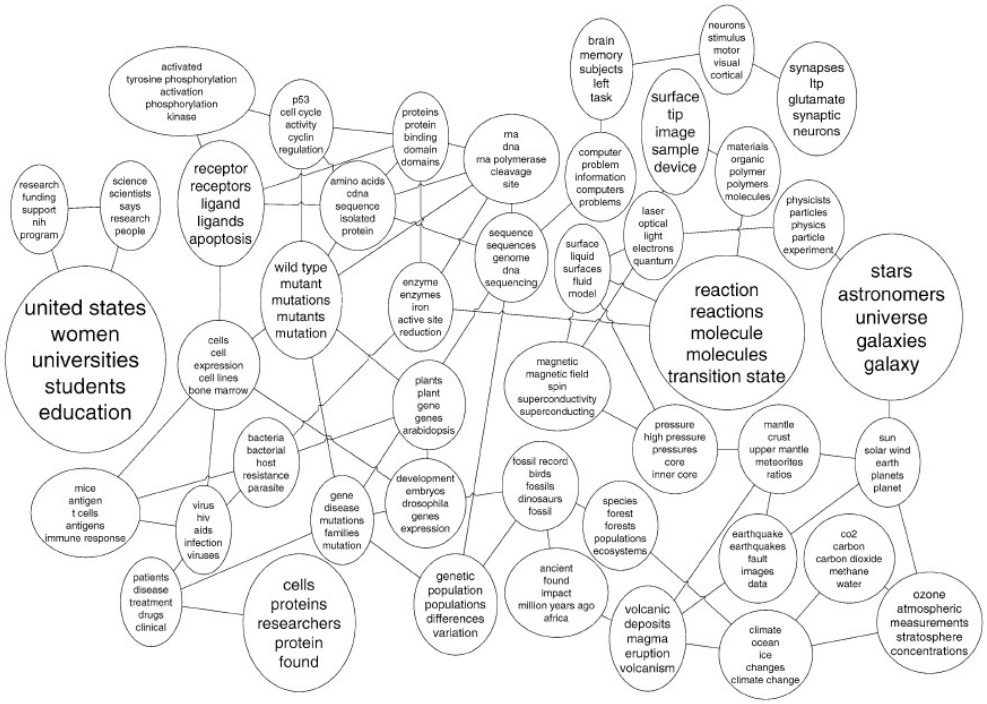


Рис. 8.5: Кореллированная тематическая модель, построенная на текстах из журнала Science.

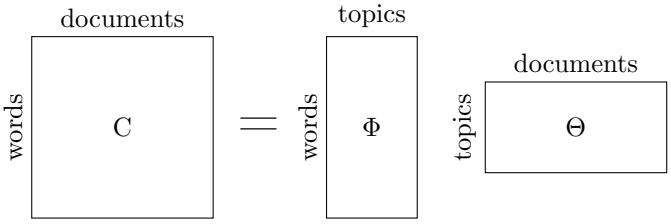
8.4. Тематические модели на практике

В этом разделе будет рассказано про практические аспекты применения тематических моделей.

8.4.1. Тематическая модель как матричное разложение

Тематические модели является еще одним видом матричного разложения, адаптированное для работу с текстом. Коллекция документов представляется в виде матрицы частот слов, причем эта матрица обычно сильно разрежена. Матричное разложение позволяет приблизить исходную матрицу в виде произведения двух матриц меньшего ранга. Причем, чем больше промежуточная размерность, тем лучше аппроксимация исходной матрицы. Разные виды матричного разложения отличаются ограничениями на матрицы и метрикой, с помощью которой оценивается мера схожести с исходной матрицей.

Тематические модели приближают исходную матрицу по псевдометрике, которая называется дивергенция Кульбака-Лейблера. Промежуточная размерность, то есть количество тем, обычно подбирается экспериментально, так как большее значение промежуточной размерности не всегда хорошо. Иногда количество тем диктуется поставленной задачей — это самый лучший вариант.



С помощью тематических моделей оказывается решенной одна из популярных проблем машинного обучения — интерпретируемость модели. Матрица Φ — матрица распределений слов в темах, а матрица Θ — матрица распределений тем в документах. По списку наиболее вероятных слов тем можно понять, о чем эта тема, и дать ей название. Тогда будет легко охарактеризовать тематический профиль документа.

8.4.2. Примеры задач тематического моделирования

Тематическое моделирование можно использовать для анализа огромного количества сообщений социальной сети. Например, коллекция документов — миллион сообщений из социальной сети Twitter. Если на этой коллекции построить тематическую модель, то можно будет смотреть не все твиты, а только на интересующую тему. Таким образом, тематическое моделирование позволяет сузить множество документов, которые необходимо просматривать. Использование других инструментов автоматического анализа текстов позволит автоматически построить навигатор по коллекции текстовых документов.

Тематические модели используются в рекомендательных системах, при медицинской диагностике и даже при распознавании изображений. На сайте habrahabr.ru доступна статья, в которой разработчики рассказывали о применении тематического моделирования для задачи рекомендации web-страниц. В качестве матрицы частот слов использовалась матрица популярности страниц для пользователей, то есть в качестве документов выступали пользователи, а в качестве слов — отдельные страницы. После построения тематической модели оказалось, что статьи, которые попали в одну тему, действительно семантически похожи друг на друга.

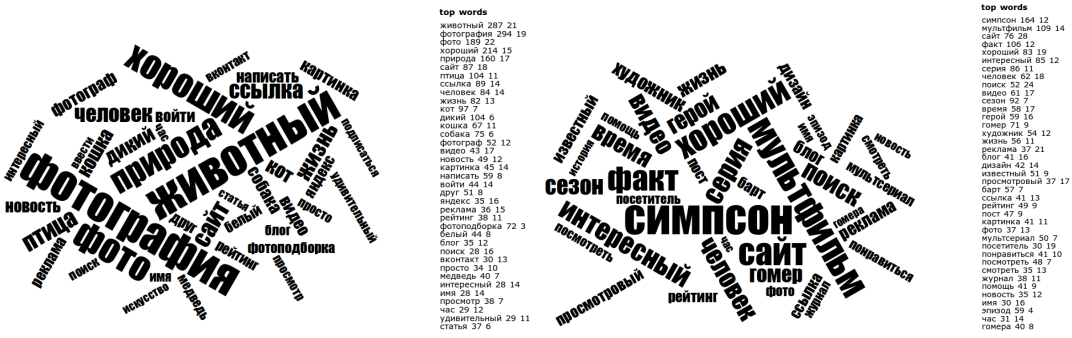


Рис. 8.6: Для статей из каждой темы были определены наиболее часто встречающиеся слова.

Еще одно применение тематических моделей — использование в поисковых машинах. В частности, они используются для выбора документов, которые потенциально могут быть релевантны запросу.

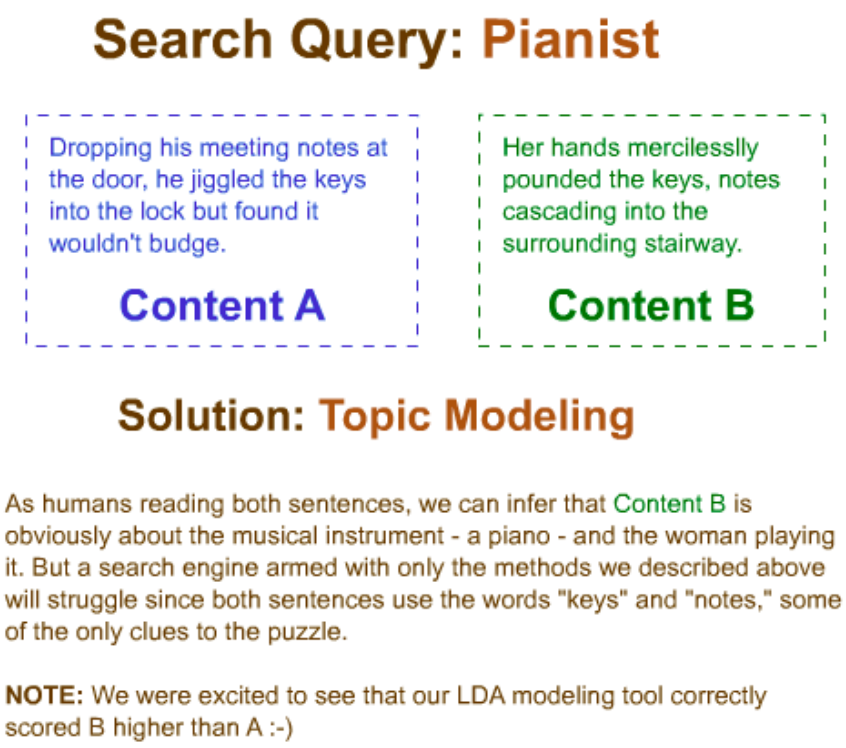


Рис. 8.7: Тематические модели позволяют отличить релевантный документ от нерелевантного за счет построения тематического профиля.

8.4.3. Методы тематического моделирования

Одним из самых первых подходов к построению тематических моделей, Probabilistic Latent Semantic Analysis (PLSA), был предложен в 1999 году. В этом методе ставилась задача матричного разложения с дополнительными условиями на нормировку столбцов матриц (столбцы матриц должны представлять собой вероятностное распределение), которая решалась методом максимального правдоподобия, стандартным методом в статистике.

В 2003 году эта же задача была рассмотрена в байесовской постановке, в которой вместо матриц строится вероятностное распределение над матрицами. Этот подход получил название Latent Dirichlet Allocation (LDA), или латентное размещение Дирихле. Стоит отметить, что LDA — самая изученная модель тематического моделирования.

Не так давно был разработан другой подход, Additive Regularization of Topic Models (ARTM), который заключался в регуляризации PLSA с целью получения лучших моделей. Для этого предполагается ввести дополнительные критерии как регуляризаторы в модель PLSA, за счет чего модель получается более гибкой и ее можно адаптировать к большему числу задач.

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