

ВНУТРЕННИЕ КРИТЕРИИ КАЧЕСТВА ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

ПЕРПЛЕКСИЯ (PERPLEXITY)

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

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

$$n = \sum_{d \in D} \sum_{w \in d} n_{dw}$$

- Интерпретация перплексии:

- Если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$

ПЕРПЛЕКСИЯ (PERPLEXITY)

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

› Интерпретация перплексии:

- ▶ Если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- ▶ Мера различности или неопределенности слов в тексте
- ▶ Коэффициент ветвления (branching factor) текста

ПЕРПЛЕКСИЯ (PERPLEXITY)

- Перплексия тестовой коллекции D' (hold-out perplexity):

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

$$n'' = \sum_{d \in D'} \sum_{w \in d''} n_{dw}$$

- Параметры ϕ_{wt} оцениваются по обучающей коллекции D

ПЕРПЛЕКСИЯ (PERPLEXITY)

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

- » $d = d' \sqcup d'$ — случайное разбиение тестового документа на две половины равной длины
- » Параметры θ_{td} оцениваются по первой половине d'
- » Перплексия вычисляется по второй половине d''

МЕРЫ ИНТЕРПРЕТИРУЕМОСТИ ТЕМ

- Тема интерпретируема, если по топовым словам темы эксперт может определить, о чём эта тема, и дать её название
- Экспертные оценки:
 - ▶ Интерпретируемость по 2- или 5-балльной шкале
 - ▶ Каждую тему оценивают несколько экспертов

МЕРЫ ИНТЕРПРЕТИРУЕМОСТИ ТЕМ

- Тема интерпретируема, если по топовым словам темы эксперт может определить, о чём эта тема, и дать её название
- Метод интрузий (intrusion):
 - ▶ В список топовых слов внедряется лишнее слово
 - ▶ Измеряется доля ошибок экспертов при его определении

КОГЕРЕНТНОСТЬ (СОГЛАСОВАННОСТЬ)

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

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

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

- › $\text{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация

КОГЕРЕНТНОСТЬ (СОГЛАСОВАННОСТЬ)

- » $\text{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — 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 встретился хотя бы один раз

РЕЗЮМЕ

- Типы оценок качества тематических моделей:
 - ▶ Внутренние (intrinsic): оценивается сама модель
 - ▶ Внешние (extrinsic): оценивается качество решения прикладной задачи, например, поиска или классификации

РЕЗЮМЕ

- › Перплексия и когерентность — часто используемые внутренние оценки качества тематических моделей
- › Экспертные оценки обычно собирают с помощью краудсорсинга

ВНЕШНИЕ КРИТЕРИИ КАЧЕСТВА ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ И КАТЕГОРИЗАЦИИ ДОКУМЕНТОВ

- Определить жанр документа:
 - ▶ Художественный, научный, учебный, рекламный, ...
- Определить тематику новости:
 - ▶ Политика, экономика, наука, здоровье, спорт, ...

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ И КАТЕГОРИЗАЦИИ ДОКУМЕНТОВ

- Определить тематику новости:
 - ▶ Политика, экономика, наука, здоровье, спорт, ...
- Определить категорию (рубрику):
 - ▶ Наука / физика / большой адронный коллайдер
 - ▶ Спорт / футбол / чемпионат мира по футболу

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ И КАТЕГОРИЗАЦИИ ДОКУМЕНТОВ

- Две модальности: термины $w \in W$ и классы $c \in C$
- Этап обучения:
 - ▶ Вход: коллекция документов — матрицы $(n_{dw}), (n_{dc})$
 - ▶ Выход: модель коллекции $p(w|t), p(c|t)$

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ И КАТЕГОРИЗАЦИИ ДОКУМЕНТОВ

- Две модальности: термины $w \in W$ и классы $c \in C$
- Этап классификации:
 - ▶ Вход: документ $d : (n_{dw})$ модель коллекции $p(w|t), p(c|t)$
 - ▶ Выход: тематика документа $p(t|d)$, модель классификации:

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

КРИТЕРИИ КАЧЕСТВА КЛАССИФИКАЦИИ ИЛИ КАТЕГОРИЗАЦИИ

- › Число ошибок классификации
- › Чувствительность и специфичность
- › AUC: площадь под кривой
чувствительность-специфичность
- › Точность и полнота
- › AUC-PR: площадь под кривой
точность-полнота

ТОЧНОСТЬ И ПОЛНОТА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

- › TP_c — верные положительные
- › FP_c — ложные положительные
- › FN_c — ложные отрицательные
- › Точность и полнота для класса c :

$$\text{Precision: } P_c = \frac{TP_c}{TP_c + FP_c}$$

$$\text{Recall: } R_c = \frac{TP_c}{TP_c + FN_c}$$

ТОЧНОСТЬ И ПОЛНОТА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

- › TP_c — верные положительные
- › FP_c — ложные положительные
- › FN_c — ложные отрицательные
- › Точность и полнота с микроусреднением:
 - ▶ Precision: $P = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c)}$
 - ▶ Recall: $R = \frac{\sum_c TP_c}{\sum_c (TP_c + FN_c)}$

ТОЧНОСТЬ И ПОЛНОТА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

- › TP_c — верные положительные
- › FP_c — ложные положительные
- › FN_c — ложные отрицательные
- › Точность и полнота с макроусреднением:

$$\text{Precision: } P = \frac{1}{|C|} \sum_c P_c$$

$$\text{Recall: } R = \frac{1}{|C|} \sum_c R_c$$

ТЕМАТИЧЕСКИЙ ПОИСК

➤ Этап обучения:

- ▶ Вход: коллекция документов — матрица (n_{dw})
- ▶ Выход: модель коллекции $p(t|d)$

ТЕМАТИЧЕСКИЙ ПОИСК

› Этап обучения:

- ▶ Вход: коллекция документов — матрица (n_{dw})
- ▶ Выход: модель коллекции $p(t|d)$

› Этап поиска:

- ▶ Вход: запрос $q : (n_{qw})$, модель коллекции $p(t|d)$
- ▶ Выход: модель запроса $p(t|q)$ и документы коллекции d , ранжированные по близости к запросу q

СПОСОБЫ ОЦЕНИВАНИЯ БЛИЗОСТИ ЗАПРОСА q И ДОКУМЕНТА d

- Косинусная мера (чем больше, тем ближе):

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

- 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):

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

КРИТЕРИИ КАЧЕСТВА ТЕМАТИЧЕСКОГО ПОИСКА

- › Точность первых k позиций выдачи
(требуется экспертная оценка
релевантности)
- › Средняя позиция при поиске документа по
его аннотации
- › Средняя точность по релевантным позициям
(MAP, Mean Average Precision) при поиске
фрагментов документа по другим
фрагментам

КРИТЕРИИ КАЧЕСТВА ТЕМАТИЧЕСКОГО ПОИСКА

- Средняя позиция при поиске документа по его аннотации
- Средняя точность по релевантным позициям (MAP, Mean Average Precision) при поиске фрагментов документа по другим фрагментам
- Средняя позиция при поиске документа по его переводу на другой язык (при кросс-язычном поиске)

РЕЗЮМЕ

- Типы оценок качества тематических моделей:
 - ▶ Внутренние (intrinsic): оценивается сама модель;
 - ▶ Внешние (extrinsic): оценивается качество решения прикладной задачи
- Классификация и поиск — самые частные приложения

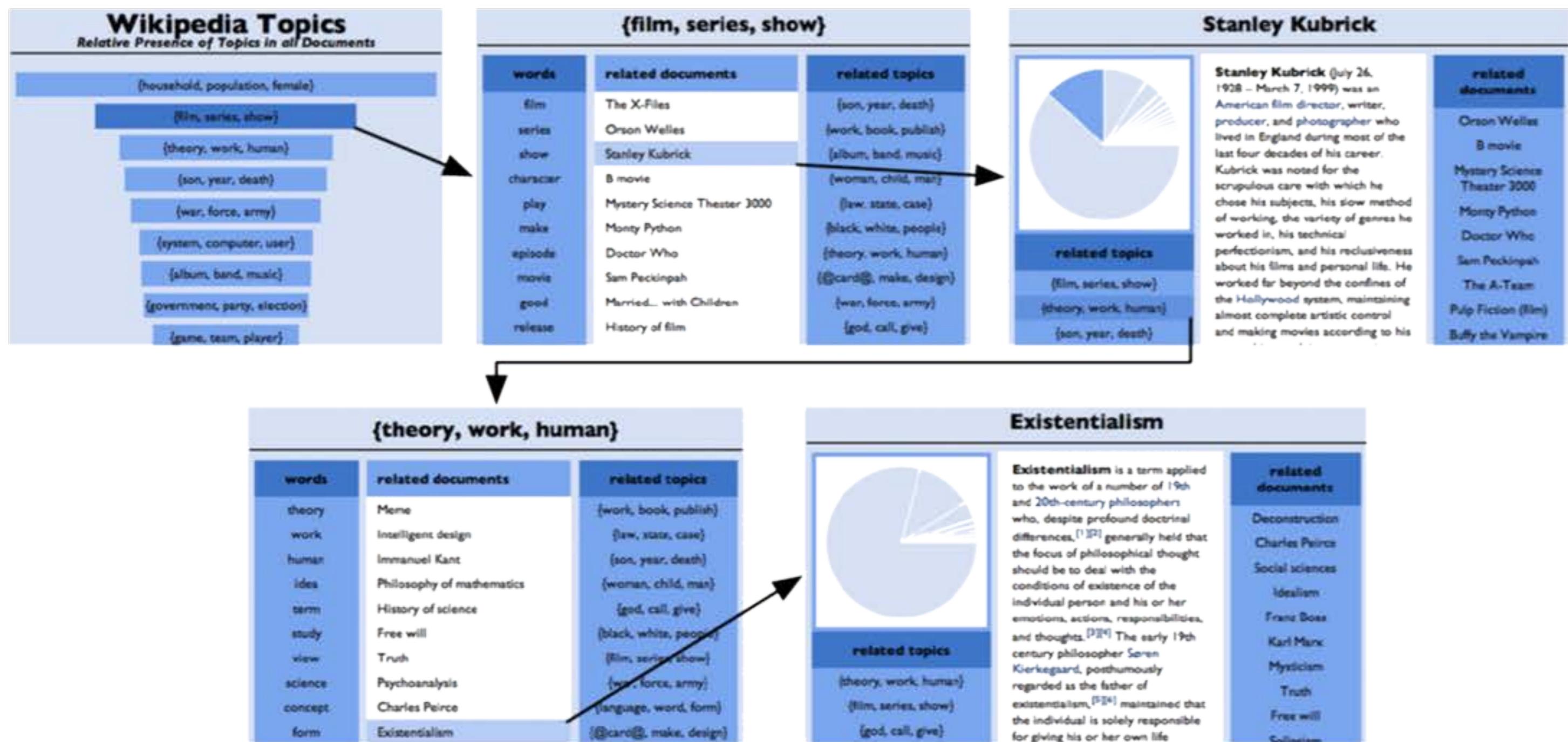
РЕЗЮМЕ

- Классификация и поиск — самые частные приложения
- Другие приложения тематического моделирования:
 - ▶ Сегментация
 - ▶ Аннотирование
 - ▶ Суммаризация

ВИЗУАЛИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

СИСТЕМА TMVE — TOPIC MODEL VISUALIZATION ENGINE

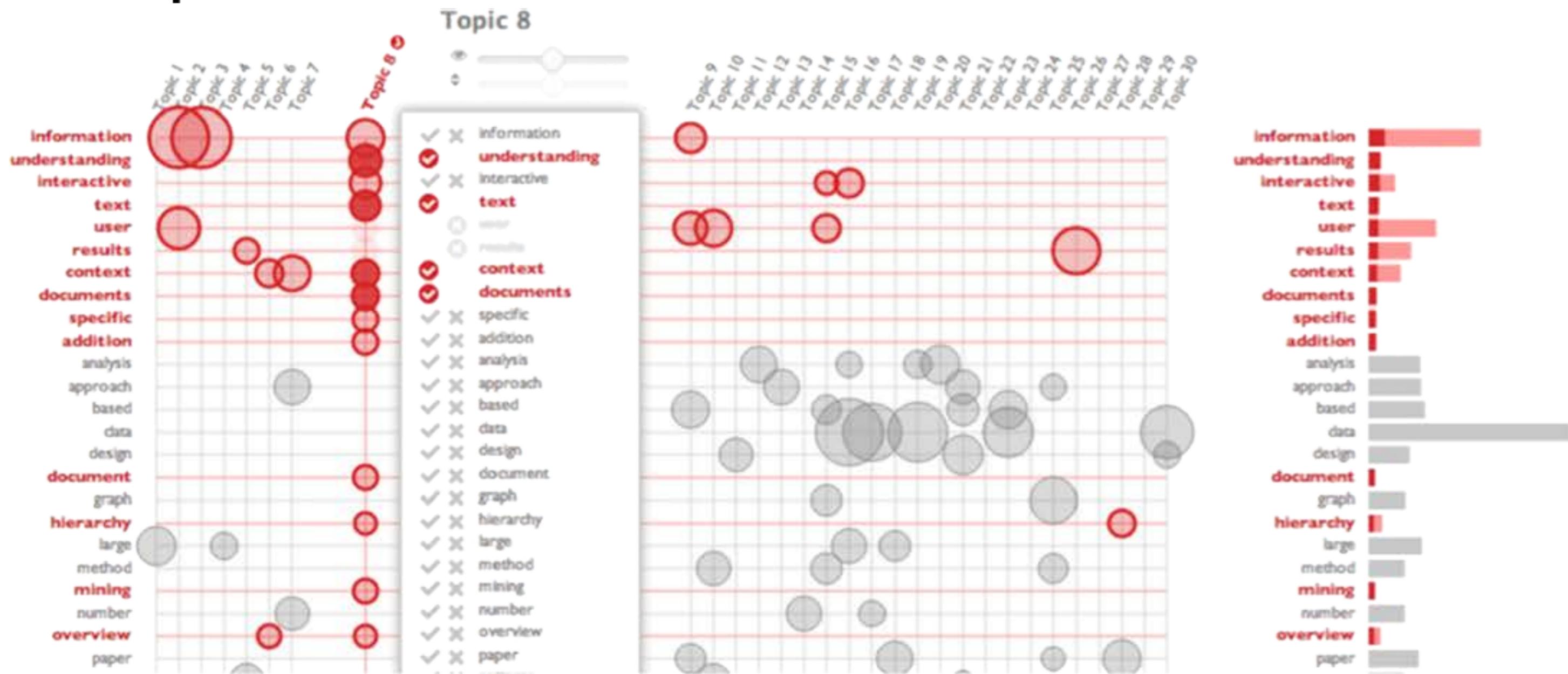
➤ Тематический навигатор с веб-интерфейсом:



Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77-84.

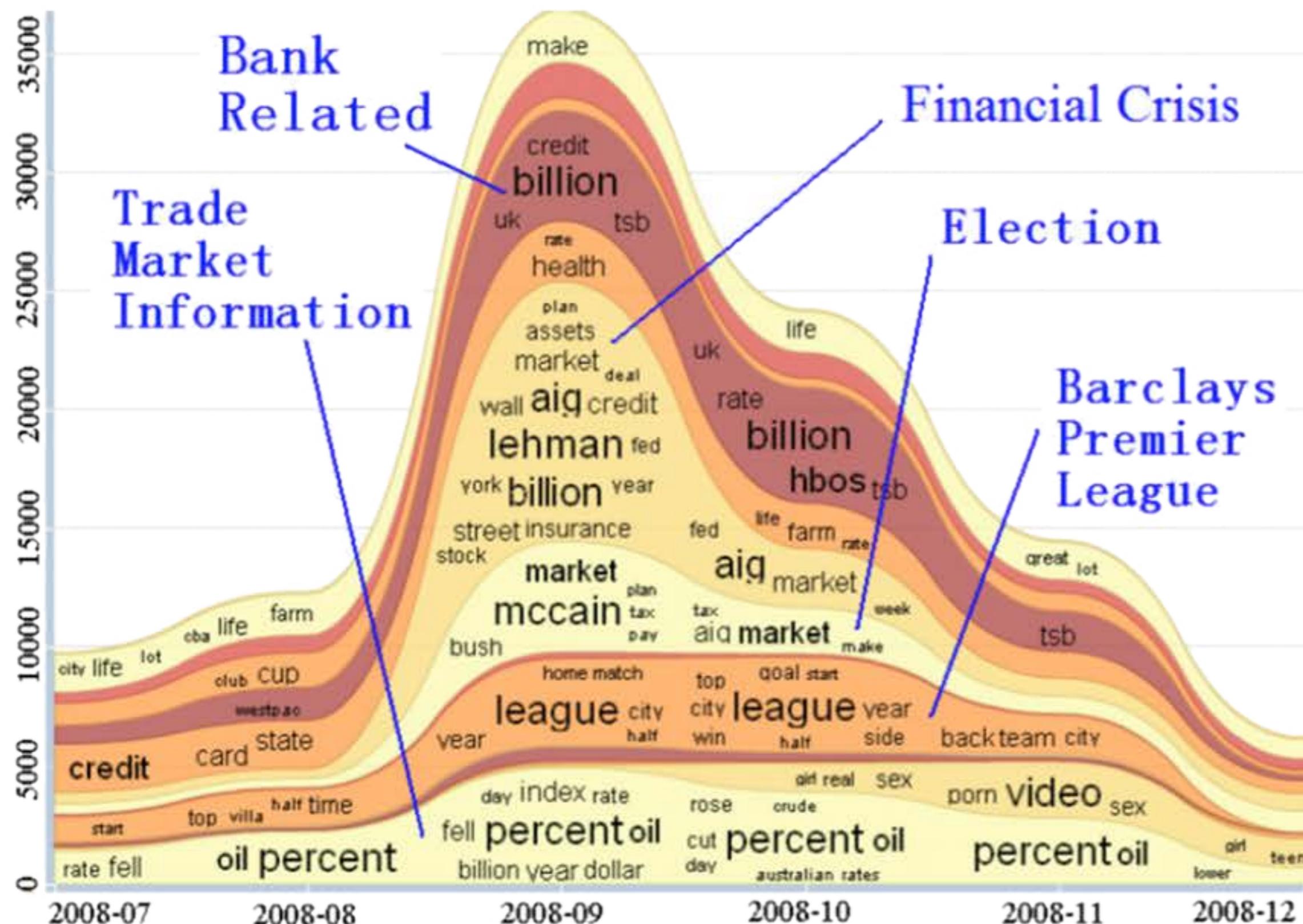
СИСТЕМА TERMITE

› Интерактивная визуализация матрицы Φ и сравнение тем:



Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models // International Working Conference on Advanced Visual Interfaces, 2012. ACM. pp. 74-77

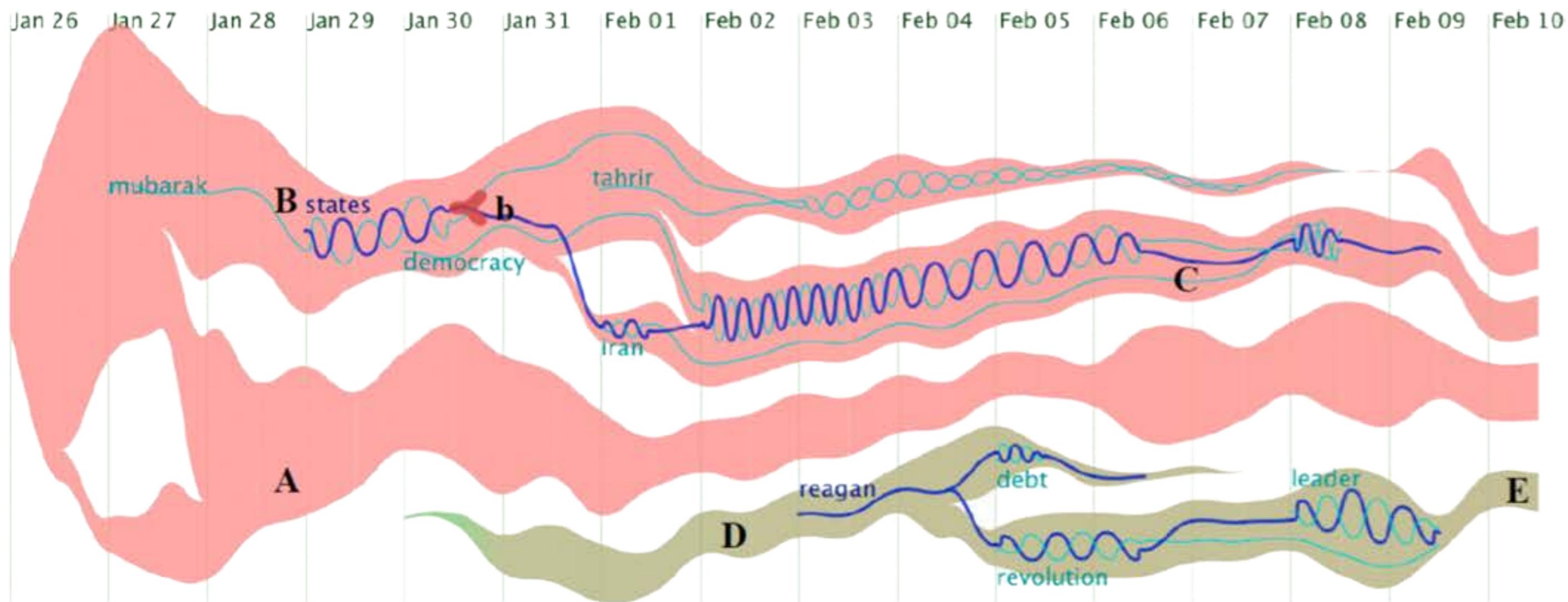
ДИНАМИЧЕСКИЕ МОДЕЛИ, УЧИТЫВАЮЩИЕ ВРЕМЯ



Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu.

Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25-28, 2010

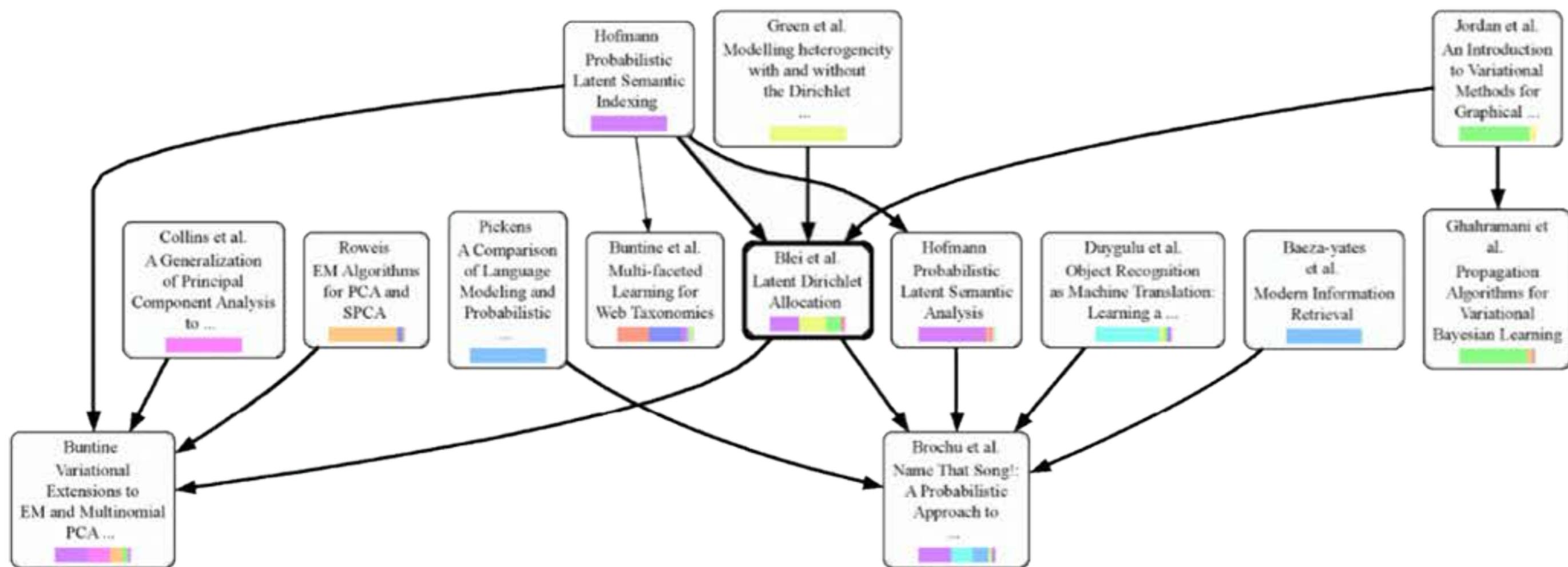
ДИНАМИЧЕСКИЕ МОДЕЛИ ЭВОЛЮЦИИ ТЕМ



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu. TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12, December 2011.

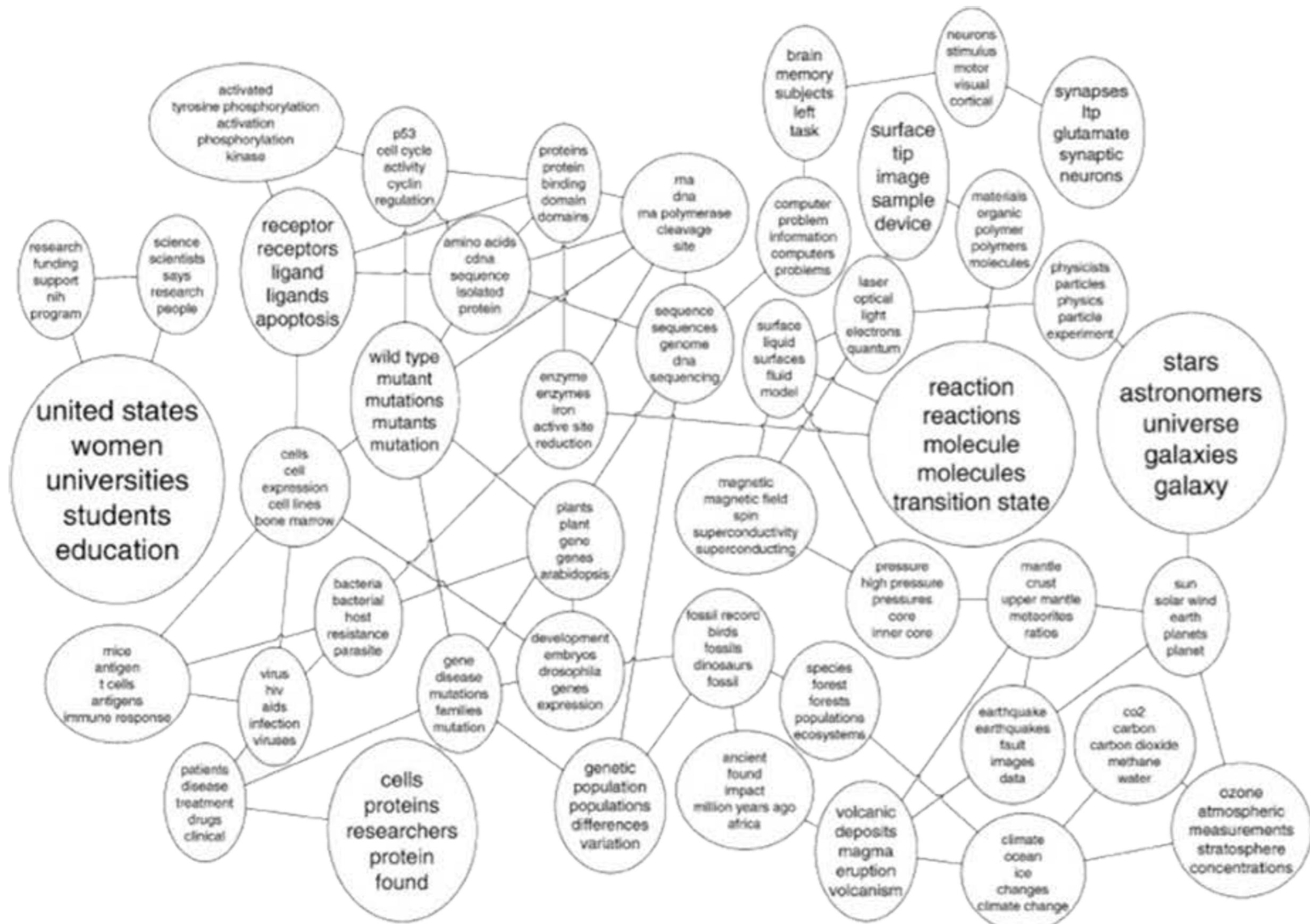
МОДЕЛИ, УЧИТЫВАЮЩИЕ ЦИТИРОВАНИЯ ИЛИ ГИПЕРССЫЛКИ

- › Учёт ссылок уточняет тематическую модель
 - › Тематическая модель выявляет самые влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction
of citation influences // ICML-2007, pp. 233-240

ВЫЯВЛЕНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ТЕМАМИ



David Blei, John Lafferty. A correlated topic model of Science // Annals of Applied Statistics, 2007. Vol. 1, pp. 17-35.

РЕЗЮМЕ

➤ Цели визуализации:

- ▶ *для конечного пользователя:*
изучение тем, навигация и поиск по коллекции

- ▶ *для разработчика моделей:*
анализ и тестирование тематической модели

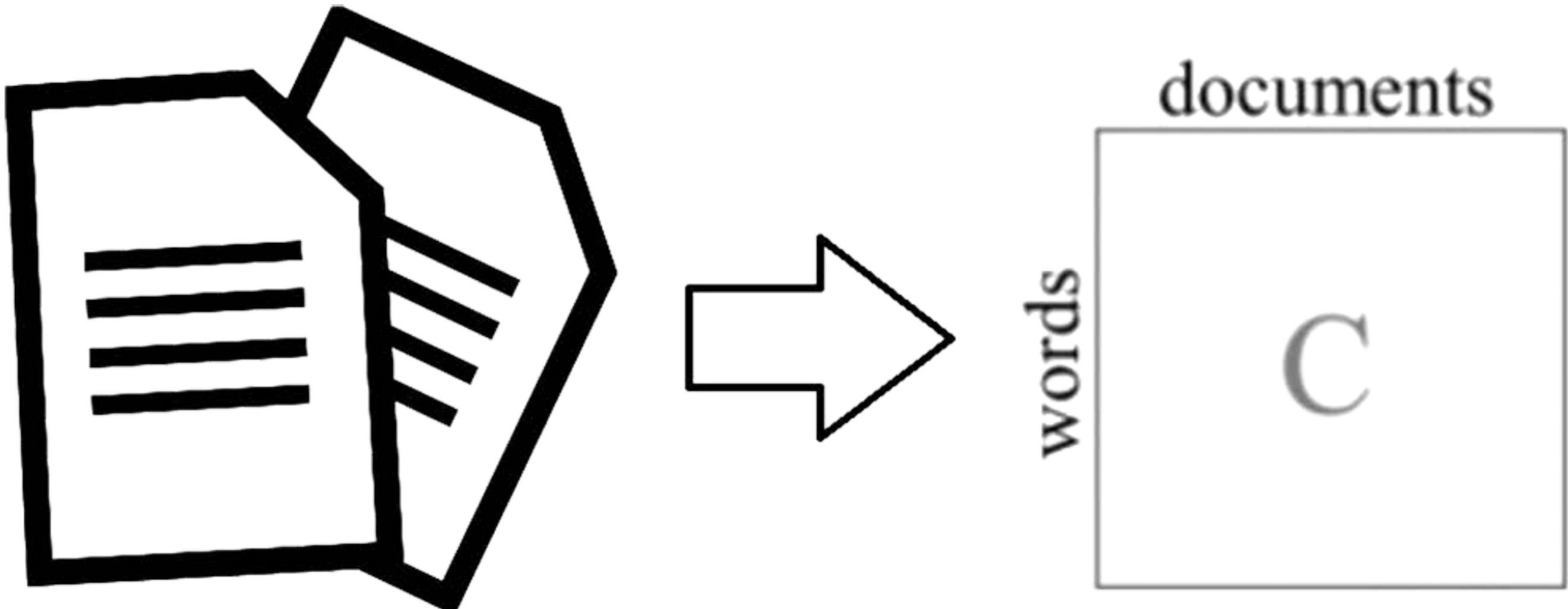
РЕЗЮМЕ

➤ Цели визуализации:

- ▶ для разработчика моделей:
анализ и тестирование тематической модели
- ▶ разметка и оценивание качества тематической модели
- ▶ разметка и дообучение тематической модели

ТЕМАТИЧЕСКИЕ МОДЕЛИ НА ПРАКТИКЕ

ТЕМАТИЧЕСКАЯ МОДЕЛЬ — МАТРИЧНОЕ РАЗЛОЖЕНИЕ

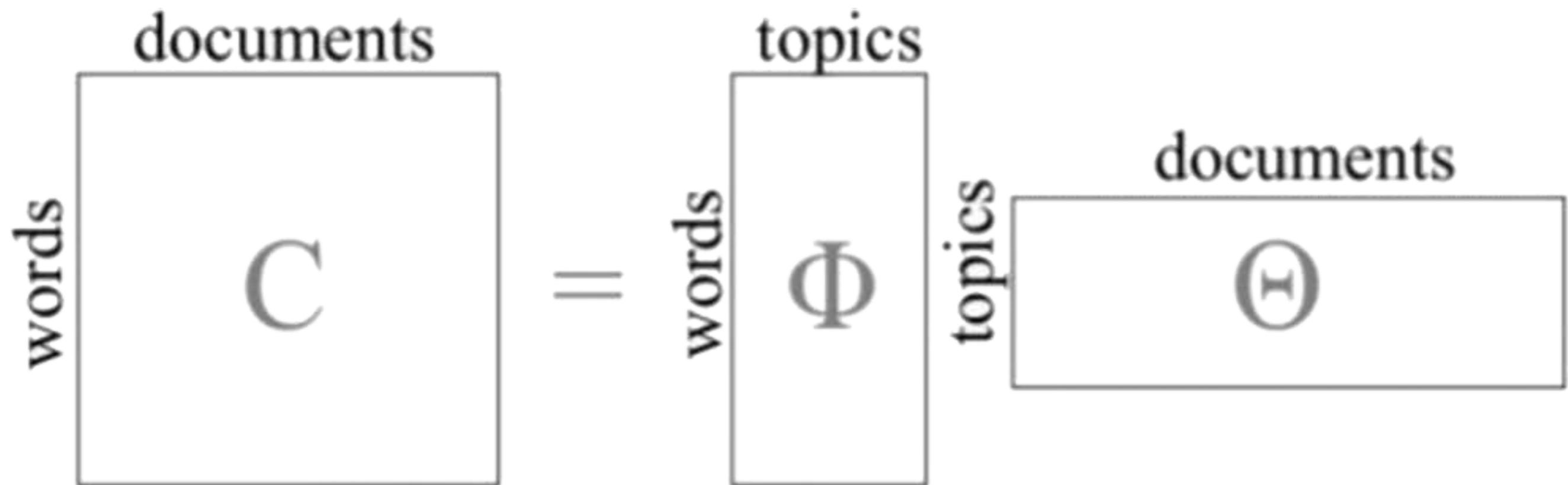


МАТРИЧНЫЕ РАЗЛОЖЕНИЯ — МОЩНЫЙ ИНСТРУМЕНТ

$$\begin{matrix} C \end{matrix} = \begin{matrix} \Phi \end{matrix} \begin{matrix} \Theta \end{matrix}$$

- › Приближение исходной матрицы в определенной метрике
- › Чем больше промежуточная размерность, тем лучше аппроксимация
- › Разные ограничения на матрицы => решения разных задач

МАТРИЧНОЕ РАЗЛОЖЕНИЕ ДЛЯ ТЕКСТОВ



- Сумма по столбцам равна единице => интерпретируемые числа в матрицах
- Дивергенция Кульбака-Лейблера <=> Правдоподобие данных
- Промежуточная размерность? Надо подбирать, больше не всегда лучше

ТЕМАТИЧЕСКИЕ МОДЕЛИ для упрощения доступа к документам

related topics

{work, book, publish}

{law, state, case}

{rate, high, increase}

{@card@, make, design}

{government, party, election}

{war, force, army}

{style, bgcolor, rowspan}

{ship, engine, design}

{black, white, people}

{country, population, people}

{car, race, vehicle}

{service, military, aircraft}

{day, year, event}

{group, member, jewish}

{game, team, player}

{island, water, area}

{film, series, show}

{theory, work, human}

{city, large, area}

{school, student, university}

{food, make, wine}

{woman, child, man}

{system, computer, user}

{son, year, death}

ТЕМАТИЧЕСКИЕ МОДЕЛИ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ ВЕБ-СТРАНИЦ

top words

животный 287 21
фотография 294 19
фото 189 22
хороший 214 15
природа 160 17
сайт 87 18
птица 104 11
ссылка 89 14
человек 84 14
жизнь 82 13
кот 97 7
дикий 104 6
кошка 67 11
собака 75 6
фотограф 52 12
видео 43 17
новость 49 12
картинка 45 14
написать 59 8
войти 44 14
друг 51 8
яндекс 35 16
реклама 36 15
рейтинг 38 11
фотоподборка 72 3
белый 44 8
блог 35 12
поиск 28 16
вконтакт 30 13
просто 34 10
медведь 40 7
интересный 28 14
имя 28 14
просмотр 38 7
час 29 12
удивительный 29 11
статья 37 6

top words

симпсон 164 12
мультфильм 109 14
сайт 76 28
факт 106 12
хороший 83 19
интересный 85 12
серия 86 11
человек 62 18
поиск 52 24
видео 61 17
сезон 92 7
время 58 17
герой 59 16
гомер 71 9
художник 54 12
жизнь 56 11
реклама 37 21
блог 41 16
дизайн 42 14
известный 51 9
просмотровый 37 17
барт 57 7
ссылка 41 13
рейтинг 49 9
пост 47 9
картина 41 11
фото 37 13
мультсериал 50 7
посетитель 30 19
понравиться 41 10
посмотреть 48 7
смотреть 35 13
журнал 38 11
помощь 41 9
новость 35 12
имя 30 16
эпизод 59 4
час 31 14
гомера 40 8



ТЕМАТИЧЕСКИЕ МОДЕЛИ В ПОИСКОВЫХ МАШИНАХ

Search Query: Pianist

Dropping his meeting notes at the door, he jiggled the keys into the lock but found it wouldn't budge.

Content A

Her hands mercilessly pounded the keys, notes cascading into the surrounding stairway.

Content B

Solution: Topic Modeling

As humans reading both sentences, we can infer that Content B is obviously about the musical instrument - a piano - and the woman playing it. But a search engine armed with only the methods we described above will struggle since both sentences use the words "keys" and "notes," some of the only clues to the puzzle.

NOTE: We were excited to see that our LDA modeling tool correctly scored B higher than A :-)

МЕТОДЫ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- › **PLSA** (Probabilistic Latent Semantic Analysis):
матричное разложение с ограничением
нормировки столбцов, обучение методом
максимального правдоподобия

МЕТОДЫ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- › **PLSA** (Probabilistic Latent Semantic Analysis):
матричное разложение с ограничением
нормировки столбцов, обучение методом
максимального правдоподобия
- › **LDA** (Latent Dirichlet Allocation):
вероятностный подход к той же задаче,
вместо матриц — распределения над
матрицами

МЕТОДЫ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- › **PLSA** (Probabilistic Latent Semantic Analysis):
матричное разложение с ограничением
нормировки столбцов, обучение методом
максимального правдоподобия
- › **LDA** (Latent Dirichlet Allocation):
вероятностный подход к той же задаче,
вместо матриц — распределения над
матрицами
- › **ARTM** (Additive Regularization of Topic Models): регуляризация PLSA с целью
получения лучших моделей

КАКОЙ МЕТОД ЛУЧШЕ?

LDA

Очень популярный

Множество модификаций для разных задач

Для каждого усложнения нужно искать реализацию

Нужно настраивать гиперпараметры

ARTM

Молодой

Мощный аппарат регуляризаторов для модификации модели

Одна реализация для разных задач

Нужно настраивать параметры регуляризации

РЕАЛИЗАЦИЯ В PYTHON

gensim для LDA

Есть функционал для решения разных задач анализа текстов

Проще в использовании

Дольше обучается

Больше форматов данных, самый понятный — UCI Bag of Words

BigARTM для ARTM

Специализированная библиотека для тематического моделирования

Больше возможностей, но чуть-чуть больше кода

Быстрее обучается

Можно импортировать данные в формате UCI Bag of Words, но vowpal wabbit формат проще