Тематическое моделирование

Квантитативный анализ текста

Кирилл Александрович Маслинский

25.04.2022 / 08

ЕУСПб

От слов к темам

Multinomial distribution over words

Исходный текст

Карл у Клары украл кораллы, Клара у Карла украла кларнет.

Словарь (распределение)

Карл	у	Клара	украсть	коралл	кларнет
Карл	у	Клара	украсть		
2	2	2	2	1	1
0.2	0.2	0.2	0.2	0.1	0.1

Корпус как смесь тем (распределений)

Темы — события

Тема							всего
	Карл	у	Клара	украсть	коралл	кларнет	
Карл	0,1	0,1	0,1	0,1	0,1	0	**0,5**
Клара	0,1	0,1	0,1	0,1	0	0,1	**0,5**

Темы — общее и различное

Тема							всего
	Карл	у	Клара	украсть	коралл	кларнет	
Общее	0,2	0,2	0,2	0,2	0	0	**0,8**
Разное	0	0	0	0	0,1	0,1	**0,2**

Probabilistic Latent semnatic analysis (pLSA)

Hoffman 1999

Тема:

- латентная переменная (объясняет распределение слов в корпусе)
- представляет собой мультиномиальное распределение

Порождающие вероятностные модели

1. Предположим, что тексты сгенерированы с помощью случайного процесса, у которого есть параметры.
2. Подберем значения параметров, которые наилучшим образом объясняют наши данные (предсказывают, что будут сгенерированы именно такие тексты).
3. Используем модель, чтобы предсказывать новые данные, основываясь на уже виденных.

Генеративная модель pLSA

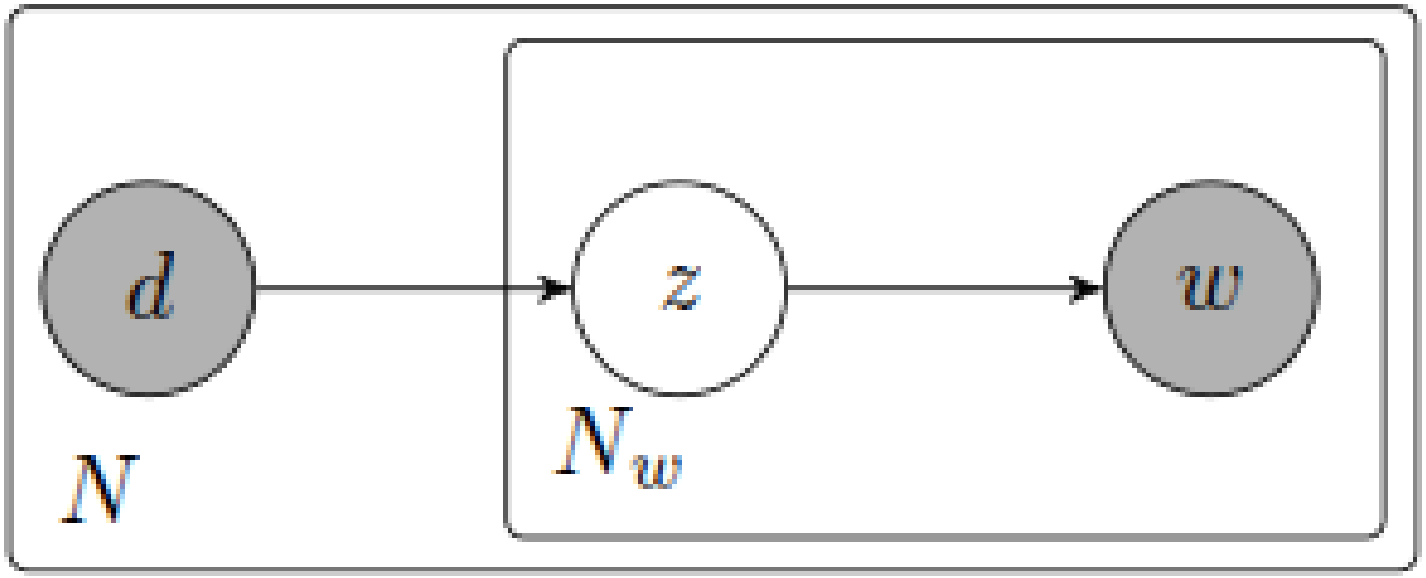
Для каждого слова в каждом документе:

1. Выбрать <случайным образом> тему z на основании распределения вероятностей тем в коллекции.
2. Выбрать <случайным образом> слово на основании распределения вероятностей слов в теме z.

Свойства:

- у каждого слова — одна тема;
- в документе может быть несколько тем;
- в коллекции — одно общее распределение вероятностей тем.

Графическая модель pLSA



LDA — Модель

Предположим:

- В коллекции представлено K тем.
- Каждый документ представляет собой смесь тем.
- «Тема» — мультиномиальное распределение слов. Каждое слово в словаре имеет определенный вес (вероятность) в каждой теме.

Свойства:

- У документа может быть несколько тем.
- У каждого слова — одна тема.
- У каждого документа в коллекции — свое распределение тем (в одном документе представлены только некоторые темы коллекции).

Порождающий процесс для LDA

для каждого документа d_d в корпусе D

Шаг: Выбрать $\theta_d \sim \text{Dirichlet}(\alpha)$

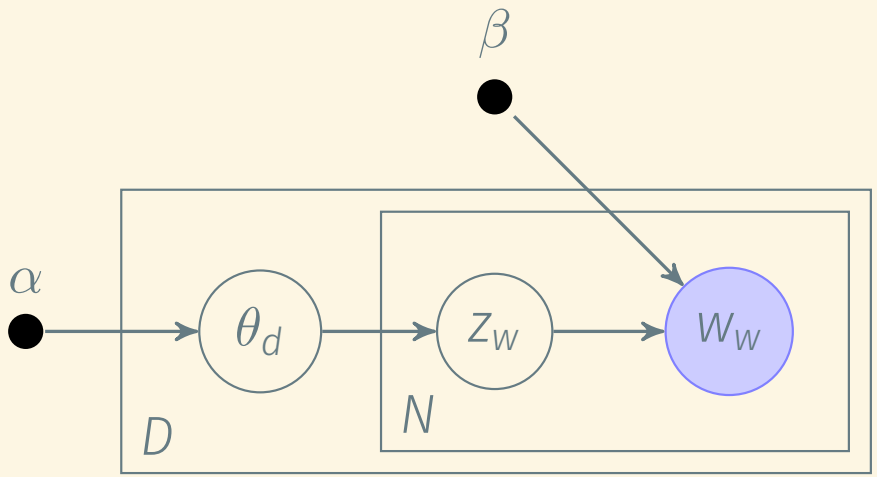
для каждой позиции w в документе d_d

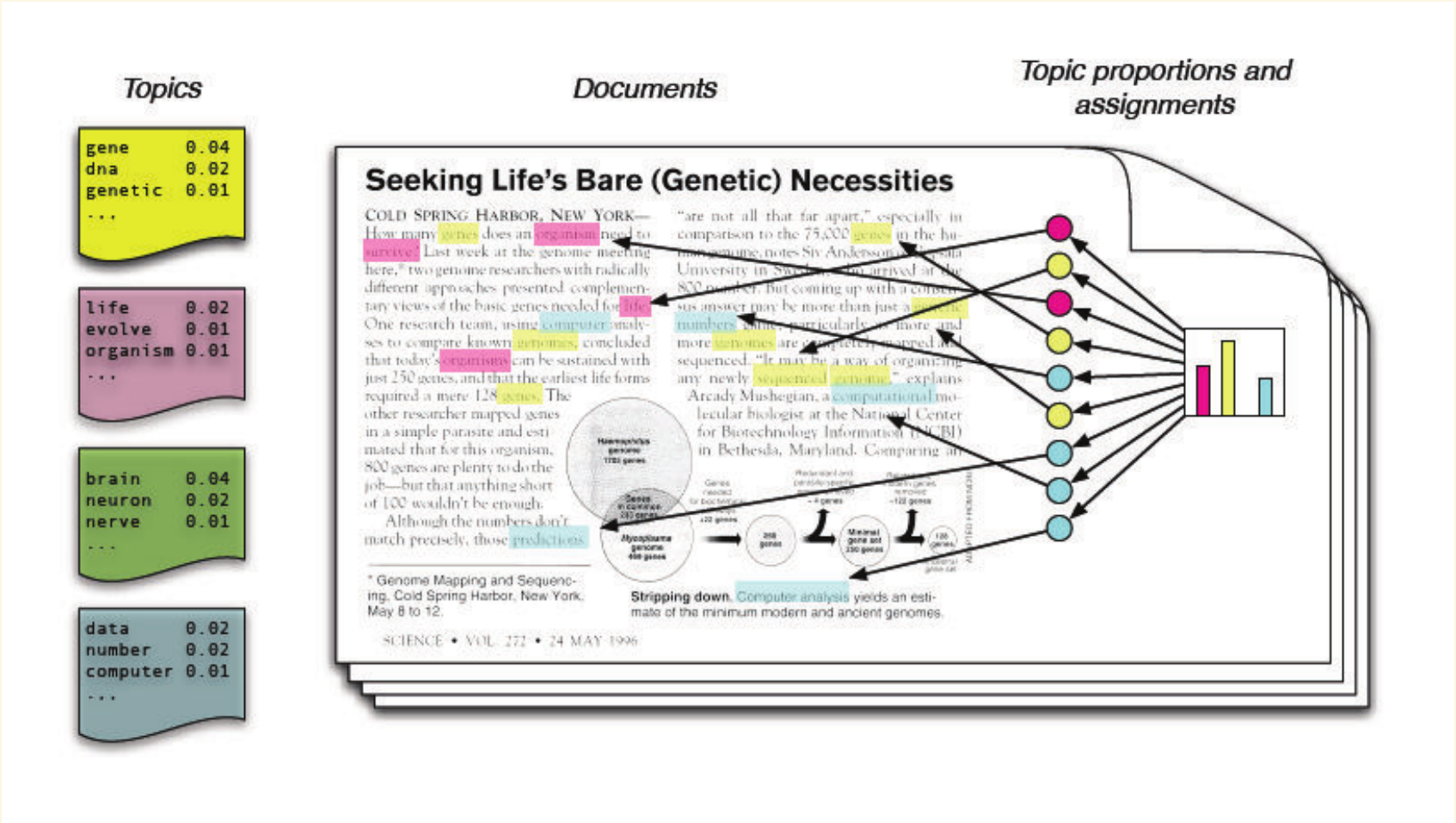
Шаг: Выбрать тему $z_w \sim \text{Multinomial}(\theta_d)$

Шаг: Выбрать слово w_w из $p(w_w|z_w, \beta)$,

мультиномиального распределения над словами, зависящего от темы и β.

Графическая модель LDA





Slide by David Blei

Тематическое моделирование: процедура

Стоп-слова

- Статический список: без более бы был была были было быть в вам вас весь во вот все всего всех вы где да даже для ...
- Динамический список:
  - Слишком частотные (N самых частотных)
  - Слишком редкие (порог: не менее чем в F документов)
  - Слишком короткие (меньше M букв)
  - Не существительные
  - Имена собственные

Размер документа имеет значение

- Разбиение длинных текстов — романы
- Объединение коротких текстов — сообщения в чате
- Оптимальный(?) масштаб — 100—1000 слов (от абстракта до статьи)

Главный принцип — единство контекста.

Черная магия с разоблачением

- Мы не можем наблюдать «темы».
- Мы наблюдаем только документы.
- LDA должен «восстановить» дискурсы, которые породили документы.
- Невозможно точно восстановить темы: слишком много неизвестных.

Gibbs sampling

Представим, что мы почти решили проблему:

Мы знаем, к какой теме принадлежит каждое слово коллекции, кроме одного слова.

Как решить, к какой теме принадлежит это слово?

Gibbs sampling

Рассмотрим два вопроса (для каждой темы Z):

1. Как часто слово принадлежит к теме Z?
   Если данное слово часто возникает при обсуждении Z, то и рассматриваемое словоупотребление может относиться к Z.
2. Насколько часто тема Z встречается в этом документе?
   Сколько слов документа отнесено к теме Z. Если Z уже обсуждается в данном документе, вероятно, что и данное слово к ней относится.

[Collapsed] Gibbs sampling

$$P(Z|W,D) = \frac{\text{частотность W в Z} + \beta_W}{\text{всего словоупотреблений в Z} + \beta} \cdot (\text{слов из Z в D} + \alpha)$$

Z — тема;

W — слово;

D — документ;

α, β — гиперпараметры. Позволяют учесть вероятность, что W может относиться к Z, даже если оно там раньше не встречалось.

Collapsed Gibbs sampling

1. Назначаем всем словам в коллекции произвольные темы.
2. Поочередно для каждого слова заново <случайным образом> выбираем тему в соответствии с вероятностями тем для этого слова, вычисленными по приведенной выше формуле.

Модель будет постепенно улучшаться, так что:

- Слова будут все чаще относиться к темам, где они уже распространены.
- Темы, которые распространены в документе, будут распространяться в нем еще.

Принцип: Rich get richer

Фрида Вигдорова, Это мой дом (1961)

Меня вызвали в <>. Я шел по длинному полутемному коридору, и вдруг откуда-то выскочила девчонка лет одиннадцати.

<...>

Вызвала меня <>. Когда я вошел, она бегло посмотрела в мою сторону, уронила: <...> – Привычку командовать надо оставить, то-ва-рищ <>! Я вижу, что слухи о вашем самомнении не преувеличены. Я вызвала вас для того, чтобы сказать: неприлично директору детдома заниматься саморекламой!

И вот тут я делаю непозволительную глупость. При этом имени я встаю и, не прощаясь, покидаю кабинет <>. <...> Иду по коридору, стиснув зубы, и злюсь на себя.

Оборачиваюсь. Девчонка со всех ног удирает от меня и, еще два раза выкрикнув ехидным голосом: «Цыган! Цыган!» – скрывается за дверью в конце коридора.

– Ну нет! Не уйдешь!

Иду за ней, дергаю дверь – она заперта изнутри. Стою тихо, жду. Дверь чуть приоткрывается, в щелку виден кончик вздернутого носа, и дверь снова захлопывается.

Гадание на темах

Jonathan Chang и др. “Reading tea leaves: How humans interpret topic models”. В: *Advances in neural information processing systems*. 2009, с. 288—296

Гадание на темах

Jonathan Chang и др. “Reading tea leaves: How humans interpret topic models”. В: *Advances in neural information processing systems*. 2009, с. 288—296

глаз лицо губа рука взгляд бровь волос голос улыбка нос лоб взглядывать щека подымать темный плечо строгий широкий рот повертываться черный ухо палец открытый словно выражение высокий бледный густой весить прямой подбородок звать угол чувствовать круглый вспыхивать похожий покраснеть сводить слегка несколько спокойно дело ресница левый живой поглядеть успевать

Интерпретация:

глаз лицо губа рука взгляд бровь волос голос улыбка нос лоб взглядывать щека подымать темный плечо строгий широкий рот повертываться черный ухо палец открытый словно выражение высокий бледный густой весить прямой подбородок звать угол чувствовать круглый вспыхивать похожий покраснеть сводить слегка несколько спокойно дело ресница левый живой поглядеть успевать

Интерпретация: портретные описания

глаз лицо губа рука взгляд бровь волос голос улыбка нос лоб взглядывать щека подымать темный плечо строгий широкий рот повертываться черный ухо палец открытый словно выражение высокий бледный густой весить прямой подбородок звать угол чувствовать круглый вспыхивать похожий покраснеть сводить слегка несколько спокойно дело ресница левый живой поглядеть успевать

Оценка качества и подбор параметров модели

Разновидности неудачных тем

Chained every word is connected to every other word through some pairwise word chain, but not all word pairs make sense.
fatty ← acids → nucleic

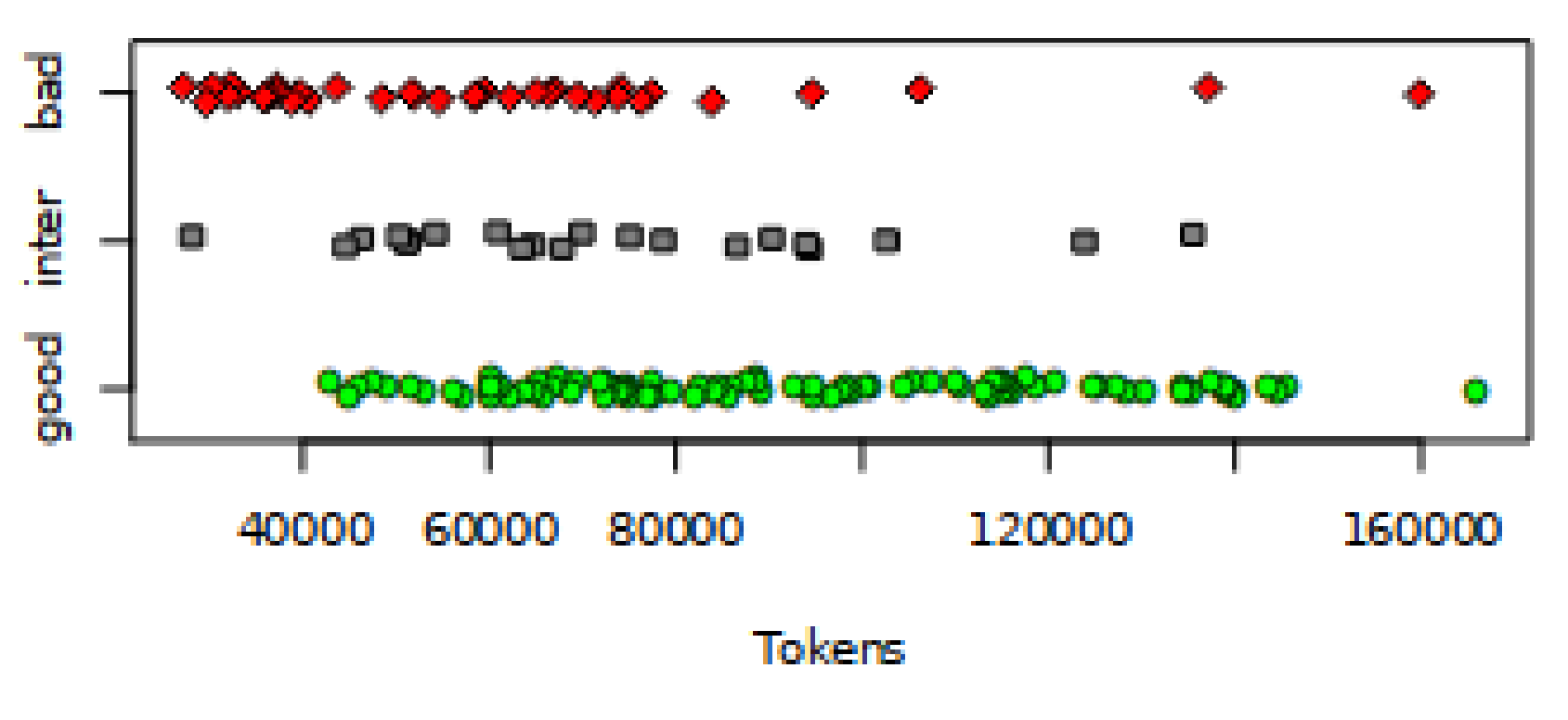
Intruded either two or more unrelated sets of related words, joined arbitrarily, or an otherwise good topic with a few "intruder" words.

Random no clear, sensical connections between more than a few pairs of words

Unbalanced the top words are all logically connected to each other, but the topic combines very general and specific terms

[0] David Mimno и др. "Optimizing semantic coherence in topic models". В: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, с. 262—272

Topic size



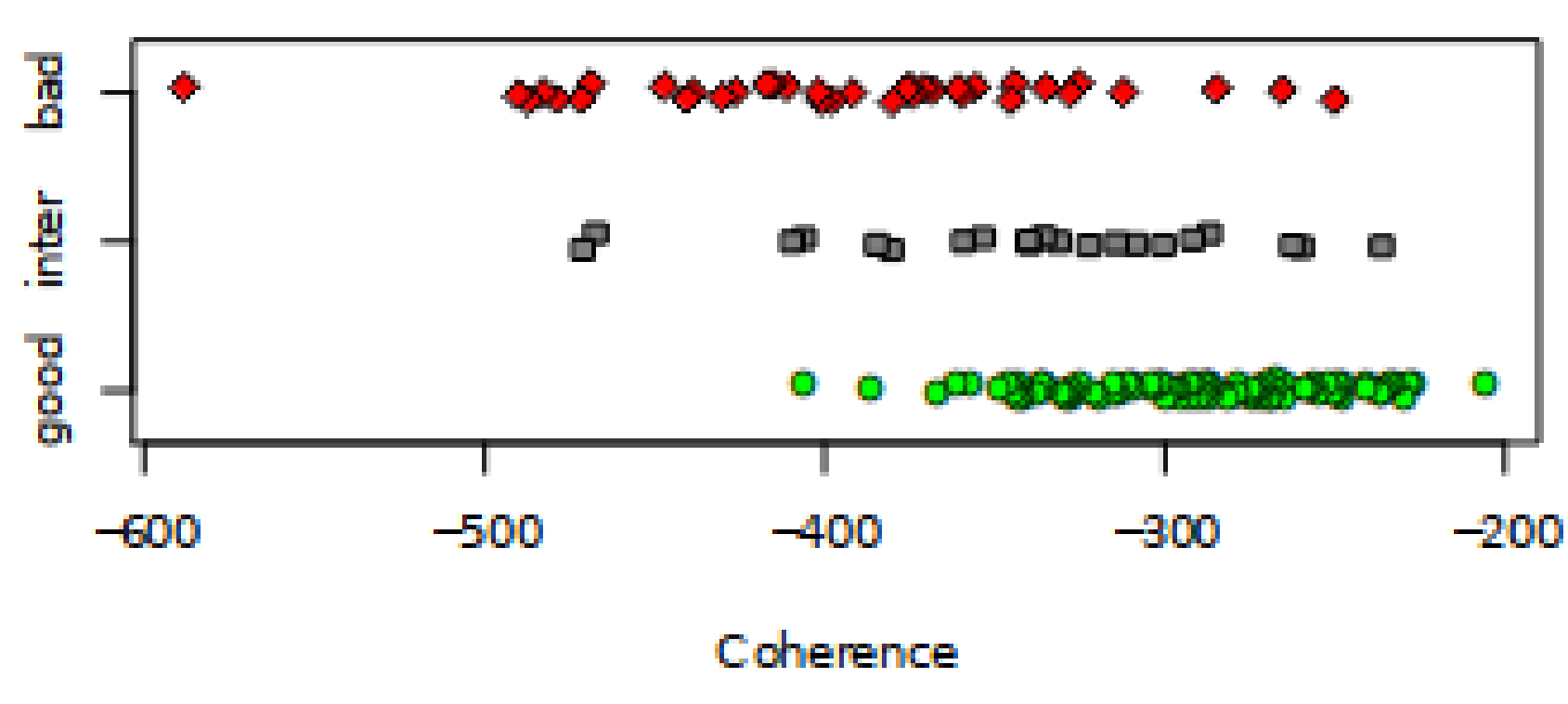
Topic coherence

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(w_m^{(t)}, w_l^{(t)}) + 1}{D(w_l^{(t)})}$$

, где

- $D(w)$ — документная частота слова w
- $D(w, w')$ — совместная документная частота слов w и w'
- $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ — список M самых вероятных слов в теме t

Topic coherence



Чему равно K (количество тем)?

David Mimno
@dmimno

“Topic” models are just machines for finding groups of words that occur together. Themes are one of many ways to produce those groups, but they are not defined by them. To say that there is one “optimal” “topic” model is insulting to the complexity of human communication.

3:59 AM · Oct 28, 2021 · Twitter for iPhone

Summary

LDA: Итоговый рецепт

- Дистрибутивная гипотеза
- Корпус текстов
- Границы документов
- Отбор и нормализация слов
- Матрица термов-документов
- Тема как мультиномиальное распределение над словами
- Модель распределения тем в документах по Дирихле
- Collapsed Gibbs Sampling
- Интерпретация