Вероятностные алгоритмы тематического моделирования документов: Probabilistic Latent Semantic Analysis и Latent Dirichlet Allocation

УЧЕБНЫЙ КУРС “ТЕОРИЯ БАЙЕСОВСКИХ СЕТЕЙ”

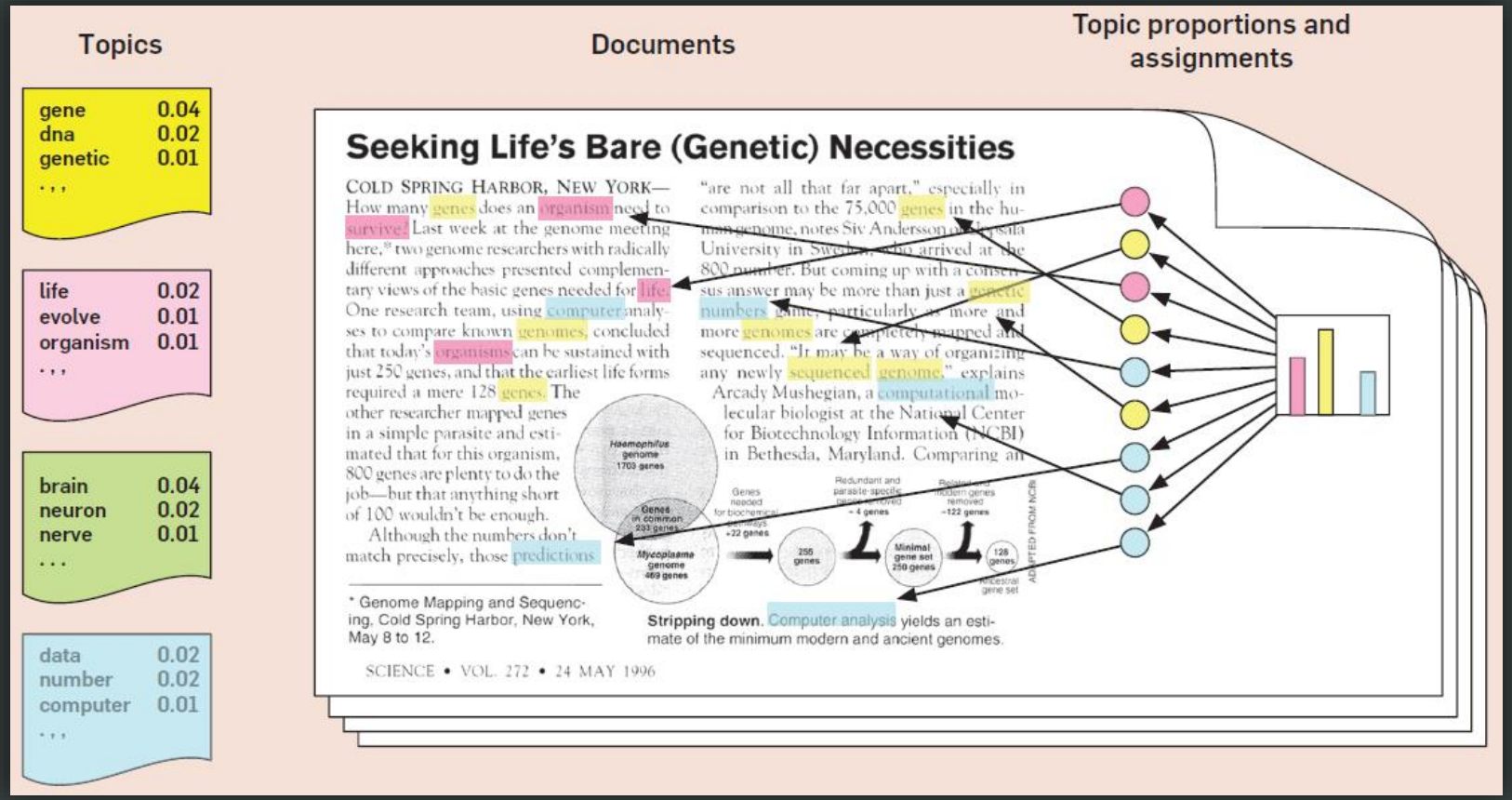
Презентацию подготовил:
Погорелов Петр Глебович,
магистр математико-механического факультета СПбГУ

План

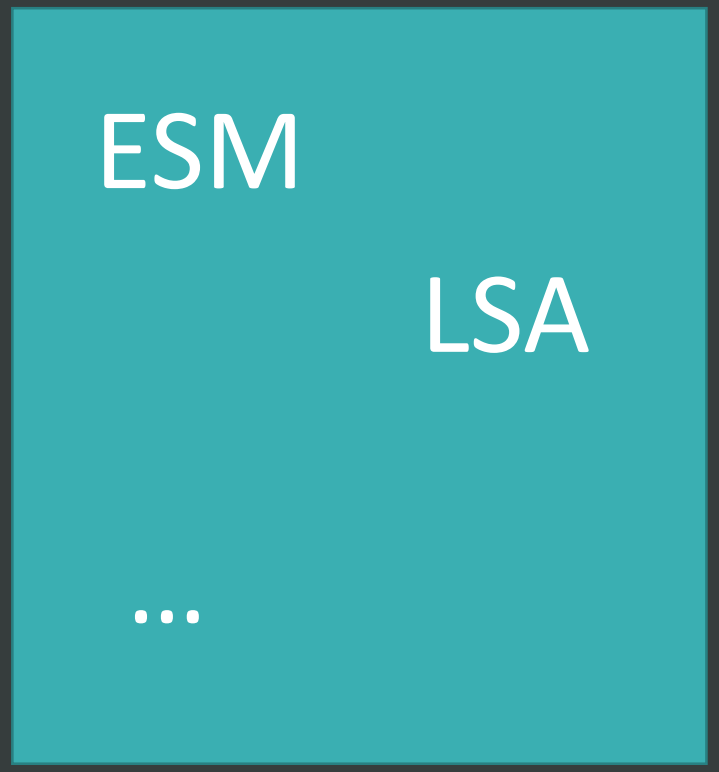
- Сформулировать постановку задачи тематического моделирования,
- сформулировать описание модели **Probabilistic Latent Semantic Analysis**, привести пример использования,
- сформулировать описание модели **Latent Dirichlet Allocation**, привести пример использования.

Задача тематического моделирования

Тематическое моделирование – одна из задач обработки естественных языков (NLP) без учителя, очень близкая к идее мягкой кластеризации.



Модели тематического моделирования



Не вероятностные модели (факторизация, information retrieval..)

Вероятностные модели

Модель PLSA

Модель PLSA (общий вид)

В модели фигурируют 3 величины:

- $d \in D$ – документ из корпуса D (*величины*)
- $w \in W$ – слово из словаря W (*величины*)
- $z \in Z$ – тематика из набора тематик Z (*величины*)

$$P\{d,w\} = \sum_{z}^{Z} P\{z\}P\{d \mid z\}P\{w \mid z\} = P\{d\} \sum_{z}^{Z} P\{z \mid d\}P\{w \mid z\}.$$

Модель PLSA (процесс порождения документов)

$$P\{d, w, z\} = P\{d\}P\{z \mid d\}P\{w \mid z\}$$

- Выбираем документ d_k с вероятностью $P\{d\}$,
- генерируем слово w_i для документа d, $i = 1..|W|$:
  - выбираем тематику $z_i \sim p(z \mid d_k)$ – мульт-е р-е.
  - выбираем слово $w_i \sim p(w \mid z_i)$ – мульт-е р-е.

Модель PLSA (пример использования)

Исходная информация об эксперименте:

- набор данных: 20 Nesgroups Dataset,
- документы взяты из категорий: религия, разработка под Windows,
- количество тем: 2
- количество документов: 1192

topic_0	god	edu	can	one	will	re	people	subject	lines	jesus
topic_1	com	edu	window	can	subject	lines	file	organization	use	mit

Модель PLSA (недостатки)

Оцениваемые параметры в модели LDA – это матрицы вероятностей $P\{z \mid d\}$ и $P\{w \mid z\}$, общее число параметров: $|\mathrm{Z}| * (|\mathrm{W}| + |\mathrm{D}|)$. Их число линейно увеличивается при росте выборки. Более того, модель становится непригодной для новых данных.

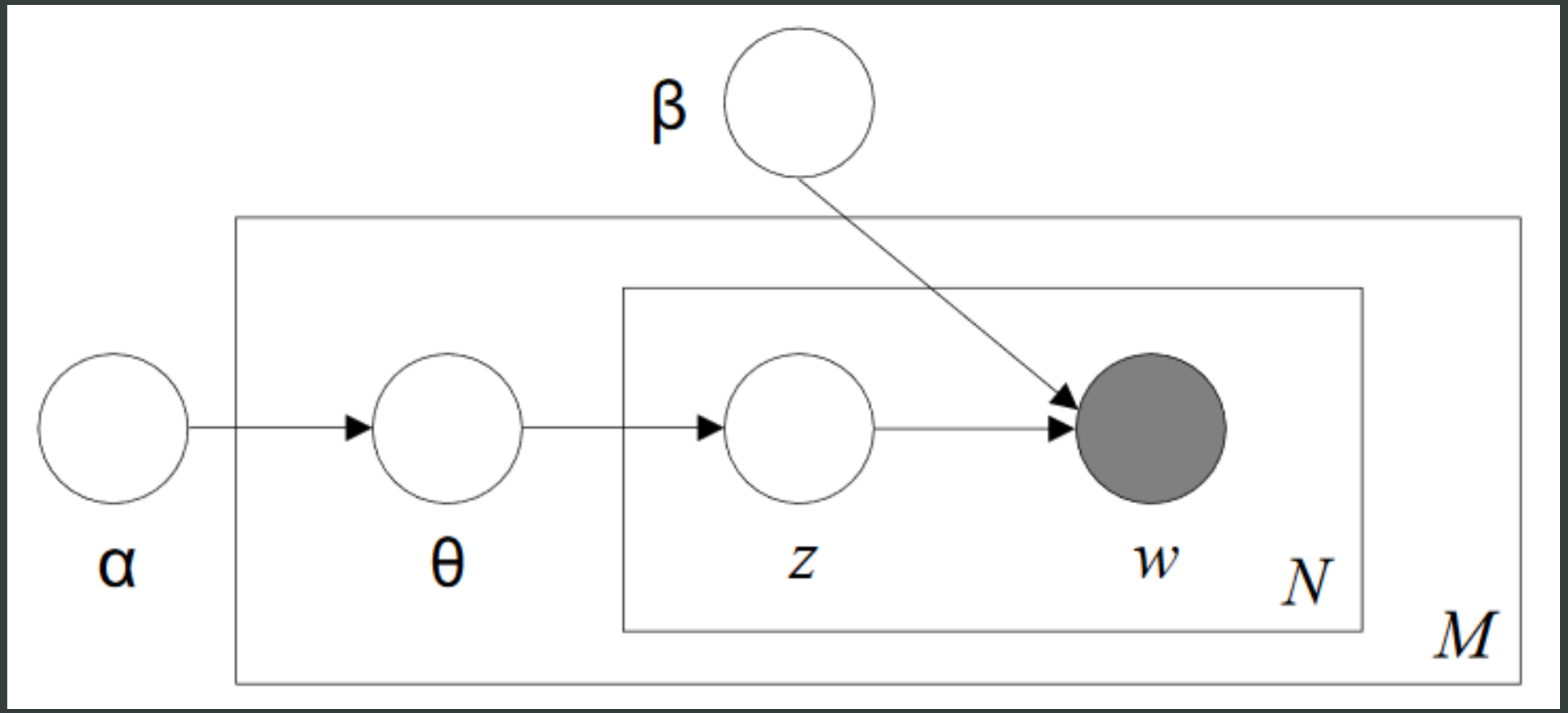
Модель LDA

Модель LDA

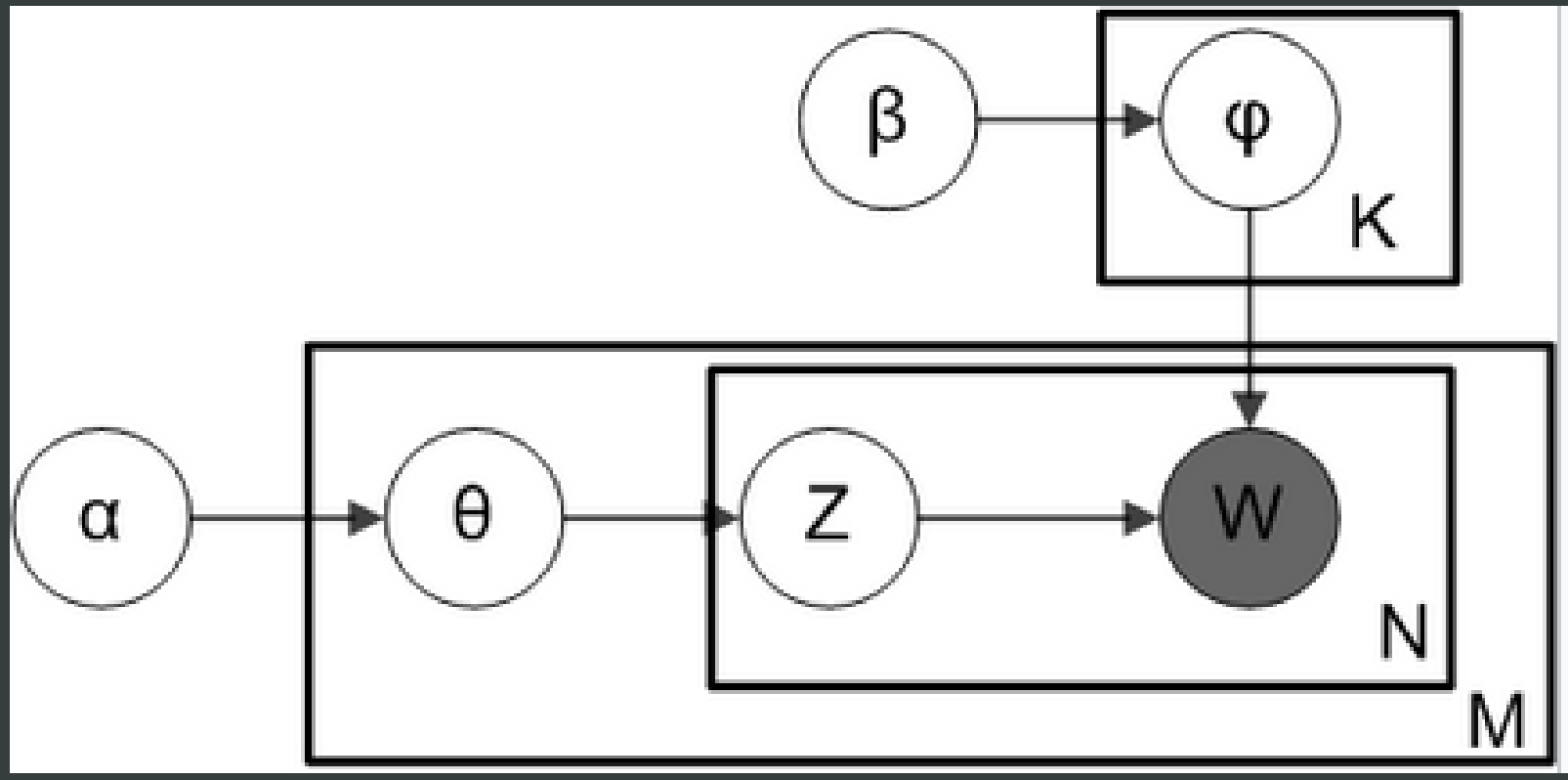
Недостаток PLSA модели – большое количество параметров для оценки, что может привести к оверфиттингу. Эта проблема решается в подходе LDA, где число параметров = $|Z| * (|W| + 1)$.

Так же, модель отвязывается от понятия “документ” как некоторой самостоятельной единицы данных. Под документом теперь подразумевается некоторый набор слов.

Модель LDA (тематики слов распределы мультиномиально)



Модель LDA (тематики слов имеют сопряженное априорное распределение Дирихле)



Модель LDA (генеративный процесс)

$$P\{w, z, \theta\} = P\{\theta|\alpha\} \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

- выбираем длину документа N
- генерируем вектор вероятностей $\theta \sim Dir(\alpha)$
- генерируем слова для документа длины N:
  - выбираем тематику $z_i \sim Multinomial(\theta)$
  - выбираем слово $w_i \sim Multinomial(\beta_{z_i})$

Модель LDA (пример использования)

Исходная информация об эксперименте:

- набор данных: 20 Nesgroups Dataset,
- документы взяты из категорий: религия, разработка под Windows,
- количество тем: 2
- количество документов: 1192

topic_0	god	edu	can	one	will	people	subject	re	jesus	lines
topic_1	edu	com	can	window	subject	lines	file	organization	use	mit

Спасибо за внимание!