Тематическое моделирование в Digital Humanities

Даня Скоринкин, DH HSE

22 апреля 2020

Автоматический анализ тематики текста

от частотных слов — к вероятностному тематическому моделированию

Часть II: Тематическое моделирование

Даня Скоринкин, DH HSE

22 апреля 2020

Зачем вообще нужно тематическое моделирование?

- Понять тематический состав корпуса
- Разложить (кластеризовать) документы по темам
- Организовать «тематический поисковик»

Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Зачем вообще нужно тематическое моделирование?

- Понять тематический состав корпуса
- Разложить (кластеризовать) документы по темам
- Организовать «тематический поисковик»

Что здесь «тема»?

- Тема — это бессмысленный с точки зрения человека список слов
- Слова «сбиваются» в тему, если они встречаются в группе документов с похожей частотой
- Интерпретация темы остается за исследователем

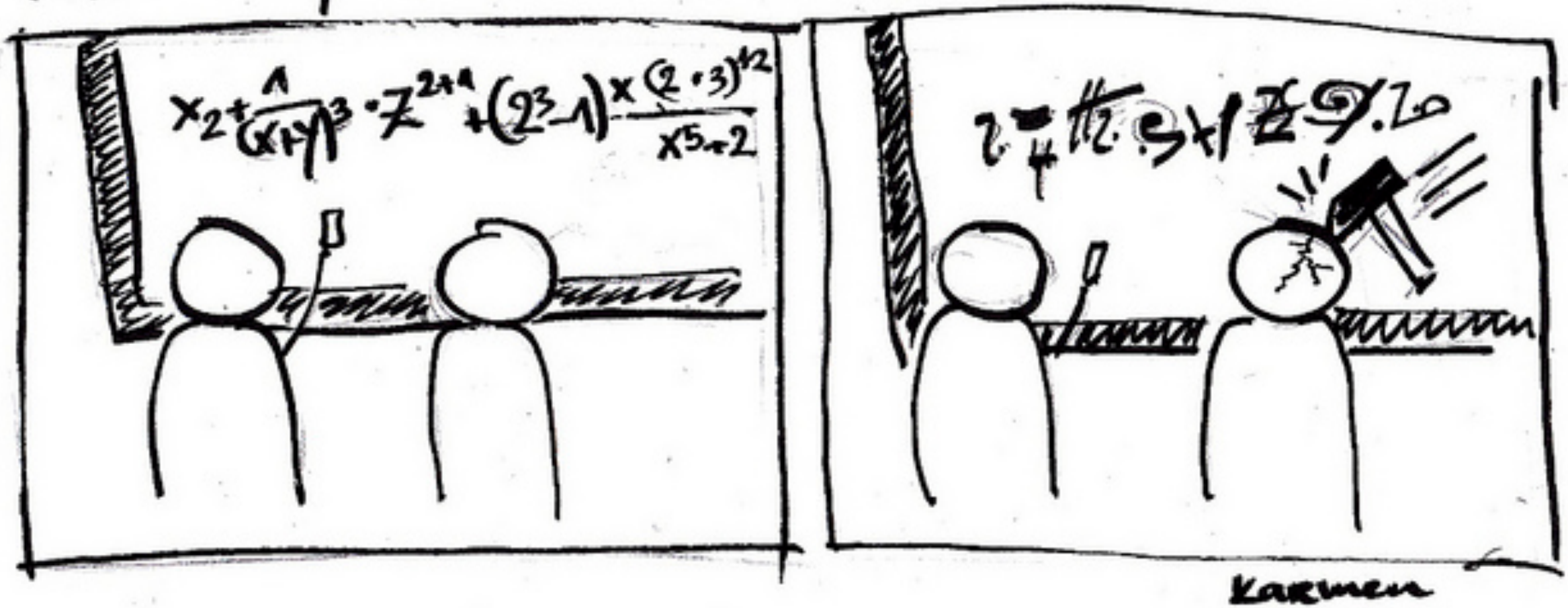
Как работает такое собираение слов в темы?

- LSA (латентно-семантический анализ)
- pLSA (вероятностный латентно-семантический анализ)
- LDA (латентное размещение Дирихле)

Каноничные статьи:

- Thomas Hoffman. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. — 1999.
  (этот человек придумал PLSA)
- David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation // Journal of Machine Learning Research. — 2003
  (эти люди придумали LDA)
- Воронцов К.В. Вероятностное тематическое моделирование.
  (это один из самых знающих русских специалистов)

THE WAY I SEE MATH...

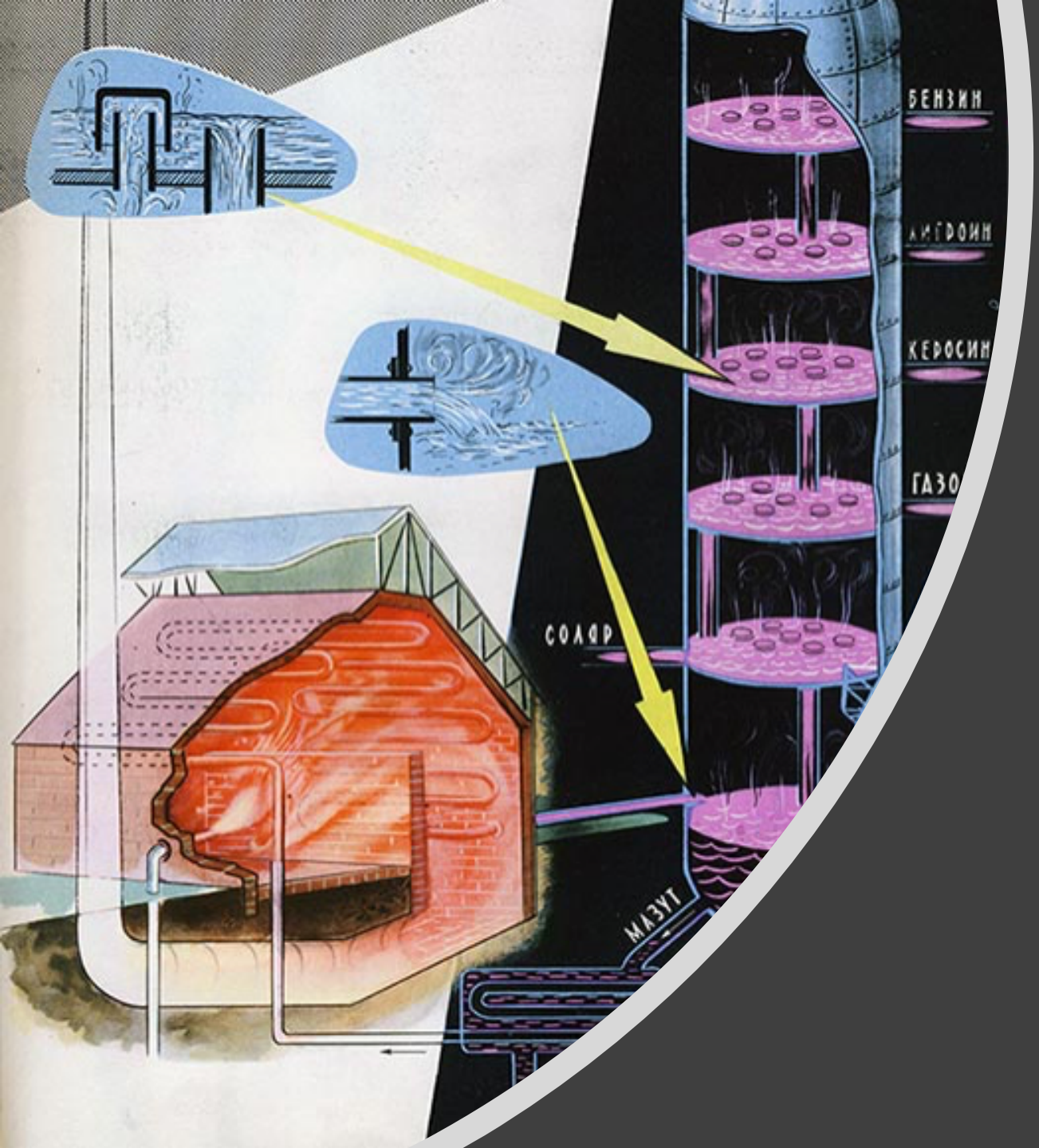


Попытки простых объяснений

- Владимир Селеверстов. Как понять, о чем текст, не читая его?// https://sysblok.ru/knowhow/kak-ponjat-o-chem-tekst-ne-chitaja-ego/
- Text Mining 101: Topic Modeling // https://kdnuggets.com/2016/07/text-mining-101-topic-modeling.html
- Introduction to Topic Modeling (Analytics Vidhya) // https://youtu.be/p1I9Sa1lRvk
- Ted Underwood. Topic modeling made just simple enough https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Общая идея вероятностных тем. моделей:

- Берем коллекцию документов
- *Случайно* присваиваем каждому слову тему
- Смотрим вероятность появления слов одной темы рядом
- Много-много раз переназначаем темы, пока вероятности совместного появления слов одной темы не станут высокими
- Для каждой получившейся темы подсчитываем наиболее характерные слова
- Для каждого документа подсчитываем распределение тем



Мне нравится метафора разгонки (фракционирования) нефти

But the practice of topic modeling makes good sense on its own, without proof, and does not require you to spend even a second thinking about “Dirichlet distributions.”

Ted Underwood. Topic modeling made just simple enough // https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Спасибо за внимание

Продолжение следует