Задание по курсу «Численные методы обучения по прецедентам»

Куратор: Потапенко Анна Александровна, anya_potapenko@mail.ru

Название: Визуализация и частичная разметка тематической структуры текстов.

Задача: Вероятностная тематическая модель описывает вероятности появления слов $w \in W$ в документах $d \in D$ через латентные темы $t \in T$:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}.$$

Для анализа качества тематическоой модели и постепенного отказа от гипотезы «мешка слов» предлагается:

- Разработать среду визуализации, отображающую вероятности $p(t|d,w) \propto \phi_{wt}\theta_{td}$ принадлежности слов к темам, полученные с помощью тематической модели.
- Предусмотреть возможность ручной разметки принадлежности слов к темам для дальнейшего использования в оценке качества модели или в частичном обучении.
- С помощью разработанной среды сравнить стадартную модель PLSA (Probabilistic Latent Semantic Analysis) и регуляризованную модель с разреженными различными предметными и сглаженными фоновыми темами [1].

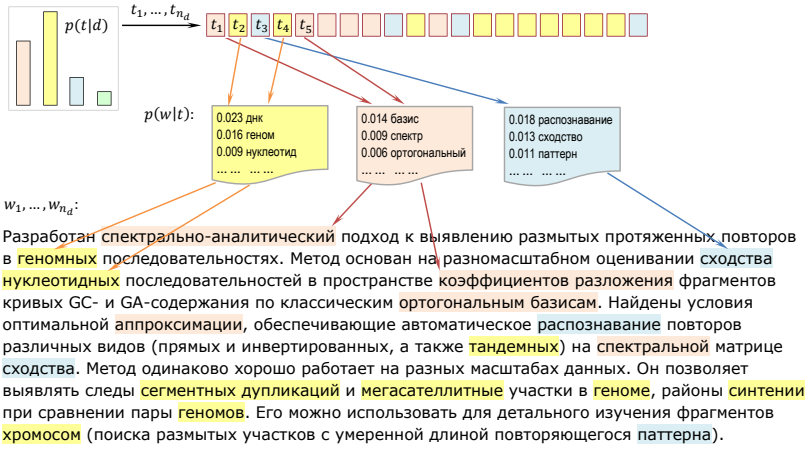
Данные для вычислительного эксперимента — коллекция статей конференций ММРО-ИОИ за несколько лет (коллекция доступна в формате исходных текстов, а также в векторизованном формате «мешка слов»).

Среда визуализации на вход принимает:

- исходные текстовые документы;
- выход тематической модели (напр. матрицы Φ и Θ), позволяющий рассчитать вероятности $p(t|d,w)$ принадлежности слов к темам (тематическая модель может быть построена отдельно).

Базовый алгоритм.

Слова в текстах последовательно сканируются, если слово найдено в словаре тематической модели, то для него находится тема с максимальной вероятностью, и слово окрашивается в соответсвующий цвет. Отдельно окрашиваются фоновые (нетематические) слова.



Литература:

[1] *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization Analysis of Images, Social Networks, and Texts AIST-2014.— CCIS 436, Springer.