**Латентно-семантический анализ текстов портала SmartLab**

*Воробьева Марья*

*28.02.19*

Доклад является частью прикладных семинаров, посвященных изучению тематического моделирования методами BigARTM и новым методом тематического моделирования, предложенным исследовательской группой, в основу которого лёг категориальный метод главных компонент (CatPCA). В работе была построена модель тематического моделирования методом CatPCA на данных Смартлаба за январь 2019 года. Всего было собрано 35 000 постов.

Было выделено 68 содержательных тем. Для определения количества тем был выбран порог в 50% общей объяснённой дисперсии переменных. При построении тематической модели методом CatPCA в качестве исходных переменных выступают слова-токены. При построении модели был выделен ряд проблем рассматриваемого метода тематического моделирования CatPCA, предложенного группой.

Первая проблема связана с тем, что в одну тему вошли практически только имена. При этом данная тема объясняет больше всего процентов общей дисперсии исходных переменных. Ещё одной особенностью тематического моделирования CatPCA выступает то, что количество слов в темах уменьшается с каждой новой добавленной темой. Это может быть связано с одним из свойств факторного анализа методом главных компонент: первый фактор объясняет больше всего дисперсии переменных. Другой выявленной проблемой предложенного алгоритма выступает то, что, судя по полученным результатам, он объединяет в темы слова-синонимы. Также данный метод автоматически сортирует токены внутри темы по алфавиту. Скорее всего существует проблемы с сортировкой исходных данных. Ещё одна проблема работы механизма – временные затраты. Модель на 35000 постов и 1139 уникальных слов – переменных сводилась в тематическую модель около 4 часов.

В результате построения модели удалось, однако, выявить пул содержательно интерпретируемых тем постов портала Смартлаб за январь 2019 года. Так, одна из выделенных тем относится к нефти и нефтегазовому делу (слова: бензин, нефть, баррель, Саудовская, Аравия). В другую тему вошли слова поздравления с новым годом и рождеством пользователей Смартлаба (слова: доброго, Рождества, Смартлаб, успешный, уверенный, пожелание). Ещё одна из тем описывает доступ к глобальной сети интернет (слова: глобальный, мобильный, цифровой, доступ). Ещё одна тема посвящена валюте и казначейству (слова: золотой, казначейский, национальный, Пекин, юань, золотой, запас). Также удалось обозначить тему строительства (слова: город, дорога, материал, предприятие, парк).