Отчет по практическому заданию по курсу «Тематическое моделирование»

1 Введение

Ставится задача кластеризации текстов новостей. Предполагается, что топики в новостных изданиях излишне размножены и на самом деле для новостного издания было бы достаточно нескольких достаточно общих тем. Приведенное утверждение проверяется в данной рабе. Строится тематическая модель, анализируются выделенные темы и выделяются наиболее явные топики. Кроме того в работе сравниваются модели с различным числом параметров и легуляризацией.

2 Цели и задачи

1. Построить тематическую модель на основе тексов новостного издательства.
2. Оценить достаточное число тем для качественного разделения топиков в новостной ленте.
3. Сравнить качество работы тематических моделей с различным числом тем.

3 Датасет

Для задачи используется датасет «News dataset from Lenta.Ru» с платформы Kaggle.

3.1 Исходные данные

В используемом датасете находятся новостные посты издательства Lenta.ru в период с 1999 по 2019. В датасете более 800000 записей. Имеется как само тело новости, так и ее заголовок, а также название топика, дополнительные теги и дата публикации. В задаче использовалось только тело новостного поста. В качестве априорных тем взяты названия топиков.

Все топики в датасете: ['Первая мировая', 'Все', nan, 'Прибалтика', 'Кино', 'Преступность', 'Общество', 'Происшествия', 'Искусство', 'Бизнес', 'Техника', 'ТВ и радио', 'Политика', 'Пресса', 'Музыка', 'Люди', 'Звери', 'Игры', 'Госэкономика', 'Гаджеты', 'Наука', 'Еда', 'Рынки', 'Деньги', 'Летние виды', 'Интернет', 'Театр', 'Конфликты', 'Реклама', 'Космос', 'Бокс и ММА', 'Футбол', 'Книги', 'Зимние виды', 'Достижения', 'Соцсети', 'Вещи', 'События', 'Средняя Азия', 'Украина', 'Закавказье', 'Белоруссия', 'Молдавия', 'Софт', 'Квартира', 'Город', 'Дача', 'Офис', 'Оружие', 'Мнения', 'Москва', 'Регионы', 'Полиция и спецслужбы', 'Криминал', 'Следствие и суд', 'Движение', 'Производители', 'Мировой бизнес', 'Финансы компаний', 'Деловой климат', 'Мир', 'Россия', 'Часы', 'Явления', 'Стиль', 'Инструменты', 'Вооружение', 'Вкусы', 'Страноведение', 'Госрегулирование', 'История', 'Внешний вид', 'Автобизнес', 'Аналитика рынка', 'Туризм', 'Выборы', 'Экология', 'Мемы', 'Мировой опыт', 'Инновации', 'Хоккей', 'Вирусные ролики', 'Фотография', 'Авто', 'Наследие', 'Преступная Россия', 'Жизнь', 'Киберпреступность', 'Социальная сфера', 'Казахстан', '69-я параллель', 'Экономика', 'Культура', 'Нацпроекты', 'Английский футбол']

3.2 Подготовка данных

Данные обрабатывались следующим образом:

- Приведение в нижний регистр
- Удаление стоп-слов русского языка
- Лемматизация
- Токенизация

Также были удалены слишком редкие темы. А именно те, которые встречались в датасете менее 100 раз.

3.3 Полученные данные: Вариант 1

Кроме того, удалялись новости с темами "Все "69-я параллель"и с неизвестными темами.. Для модели было выбрано 10000 случайных новостных статей.

3.4 Полученные данные: Вариант 2

В ходе экспериментов, было выявлено, что априорные темы слишком сильно пересекаются и даже при ассесорском контроле не всегда удавалось правильно проставить тег. Поэтому было принято решение объединить несколько топиков в один.

Например: Преступность = ['Преступность', 'Полиция и спецслужбы', 'Криминал', 'Следствие и суд', 'Преступная Россия', 'Киберпреступность'].

Итого в датасете остались следующие топики: ['ТВ и радио', 'Пресса', 'Звери', 'Игры', 'Еда', 'Реклама', 'Бокс и ММА', 'Футбол', 'Оружие', 'Движение', 'Часы', 'Вооружение', 'Вкусы', 'История', 'Туризм', 'Экология', 'Хоккей', 'Фотография', 'Авто', 'Наследие', 'Политика', 'Преступность', 'Интернет', 'События', 'Технологии', 'Общество', Бизнес', 'Культура', 'Стиль']

4 Эксперименты

В качестве тематической модели использовалась модель BigARTM.

4.1 Эксперимент с изначальными топиками

При запуске модели с 6 темами, были получены следующие топики:

- topic0 ['данные', 'новый', 'составлять', 'первый', 'россия', 'доллар', 'составить', 'миллиард', 'тысяча', 'миллион', 'рубль', 'процент', 'компания']
- topic1 ['состояться', 'игра', 'место', 'победа', 'время', 'клуб', 'российский', 'чемпионат', 'пройти', 'матч', 'мир', 'команда', 'первый']
- topic2 ['несколько', 'день', 'появиться', 'женщина', 'издание', 'рассказать', 'однако', 'время', 'пользователь', 'слово', 'the', 'опубликовать', 'человек']
- topic3 ['заявить', 'владимир', 'слово', 'пресс', 'новый', 'военный', 'международный', 'москва', 'территория', 'сторона', 'президент', 'россия', 'российский']

- topic4 ['время', 'сказать', 'министр', 'власть', 'отметить', 'решение', 'российский', 'государство', 'слово', 'президент', 'глава', 'россия', 'заявить']

- topic5 ['уголовный', 'погибнуть', 'служба', 'пострадать', 'результат', 'задержать', 'время', 'сотрудник', 'полиция', 'данные', 'дело', 'произойти', 'человек']

Сопоставить реальные темы таким топикам очень сложно и даже при выборе ['Деньги', 'Футбол', 'Соцсети', 'Украина', 'Политика', 'Преступная Россия'], точность классификации составила 3.92%.

4.2 Эксперимент с измененными топиками

В случае использования объединённых тем, получилось следующее сопоставление: ['Бизнес', 'Футбол', 'Культура', 'Общество', 'Политика', 'Преступность'], которые дало точность 27.82.

4.3 Эксперимент с измененными топиками и большим числом тем

При попытке добавить два дополнительных топика:

- topic6 ['решение', 'пост', 'человек', 'власть', 'киев', 'бывший', 'владимир', 'суд', 'депутат', 'глава', 'президент', 'украинский', 'украина']

- topic7 ['сообщаться', 'мужчина', 'местный', 'место', 'инцидент', 'пострадать', 'данные', 'результат', 'время', 'погибнуть', 'город', 'человек', 'произойти']

появлялась неоднозначность в теме 'Политика', поскольку новый топик (tipic6) подходит под нее, и в итоге качество падало до 21.13. Несмотря на то, что новый topic7 подходил под новую тему 'События'

4.4 Эксперимент с регуляризацией

При попытке добавить регуляризацию на декорреляцию матрицы ϕ качество алгоритма лишь падало, а топ слов в темах практически не менялся:

- $\tau = 100$: точность 12.12%
- $\tau = 10000$: точность 16.9%
- $\tau = 1000000$: точность 15.58%

5 Вывод

Были проведены эксперименты по тематическому моделированию новостей издательства Lenta.ru с использованием тематических моделей библиотеки BigARTM. Из полученнных результатов можно сделать выводы, что:

1. В Lenta.ru действительно слишком много тегов новостей, которые пересекаются между собой, что может приводить к заблуждению относительно статьи
2. Предложенное простое объединение тем улучшает качество тематической модели.
3. Тематическая модель работает с 6 темами лучше, чем с 8, что может говорить либо о том, что для новостей можно выделить 6 основных тем, либо о том, что в данном корпусе есть акцент в сторону тех или иных новостей.
4. На предложенном корпусе регуляризация декорреляцией ϕ не дает улучшения тематической модели.