Отчёт по индивидуальному заданию
по курсу "Вероятностное тематическое моделирование"

студентки 528 группы ВМК МГУ

Иконниковой Марии

2018 год

1. Введение

Социальные сети так прочно вошли в нашу повседневную жизнь, что многим уже сложно представить себя без своего онлайн-профиля. Информация, получаемая из профилей пользователей, соцсетей может использоваться как в коммерческих целях, так и в социологических исследованиях, поэтому её получение и интерпретация является важной задачей.

Сообщества (группы) социальной сети изначально создавались с целью объединить единомышленников и предоставить им возможность для общения по интересам и обмена тематическими новостями. Поэтому, подписки человека на такие сообщества предоставляют важную информацию о нём и его увлечениях, а тенденции и тренды в этой сфере отражают ситуацию в обществе.

Однако, существуют и такие сообщества, в которых невозможно выделить какую-то конкретную тему: например те, в которых рассказываются жизненные истории. Такие сообщества могут представлять интерес для лиц с совершенно разными увлечениями и менее актуальны для поставленных задач. Исходя из этого, выделение тематических и нетематических сообществ также является важной проблемой.

С учётом всего вышесказанного, в данной работе предполагается сделать некоторые шаги в решении задач исследования актуальных для сообществ социальной сети ВКонтакте тем и выделения тематических сообществ.

2. Цели и задачи

Для выполнения поставленных целей предлагается решить следующие задачи:

- провести эксперименты по тематическому моделированию сообществ социальной сети ВКонтакте с использованием различных регуляризаторов в моделях библиотеки BigARTM;
- оценить интерпретируемость получаемых тем;
- выделить подмножество сообществ, являющихся актуальными для задачи определения интересов пользователя социальной сети.

3. Подготовка данных

3.1. Исходные данные

В качестве исходных данных используются коллекции записей сообществ социальной сети ВКонтакте, полученные кроулингом страниц, активных за предшествующий сбору данных год. В коллекции оставлены только те сообщества, которые являются открытыми

для общего доступа и имеют не менее 100 подписчиков. Итого, имеется около 120 000 сообществ.

3.2. Подготовка данных

В процессе подготовки данных для тематического моделирования были проведены следующие операции над исходными текстами:

- приведение всех букв в нижний регистр
- токенизация с использованием токенизатора библиотеки nltk
- удаление символов, не принадлежащих словам русского языка с помощью регулярных выражений
- определение частей речи с использованием библиотеки nltk и удаление служебных частей речи и местоимений
- лемматизация с использованием лемматизатора библиотеки pymorphy2

Данные тех сообществ, в которых после проведённых операций осталось более 100 различных слов были записаны в виде мешка слов (bag-of-words) в файл в формате vowpal wabbit.

3.3. Построение словаря

Для построения словаря всех используемых в модели слов был использован метод библиотеки BigARTM artm.Dictionary(). Также, с использованием метода dictionary.filter() была построена модифицированная версия исходного словаря со словами, содержащимися не менее чем в 1% документов коллекции и встречающимися не менее 1000 раз.

4. Эксперименты по построению тематических моделей сообществ

Были проведены эксперименты по построению различных тематических моделей сообществ социальной сети ВКонтакте с заданным количеством тем 100, исходным словарём (1 801 829 слов) и:

1) без регуляризаторов, 35 итераций

2) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^7$ после 3 итераций и регуляризатора разреживания матрицы Θ с параметром $\tau = -1{,}5$ после 20 итераций, всего 35 итераций

3) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^7$ после 3 итераций, 35 итераций

4) с добавлением регуляризатора регуляризатора разреживания матрицы Θ с параметром τ = -1,5 после 20 итераций, всего 35 итераций

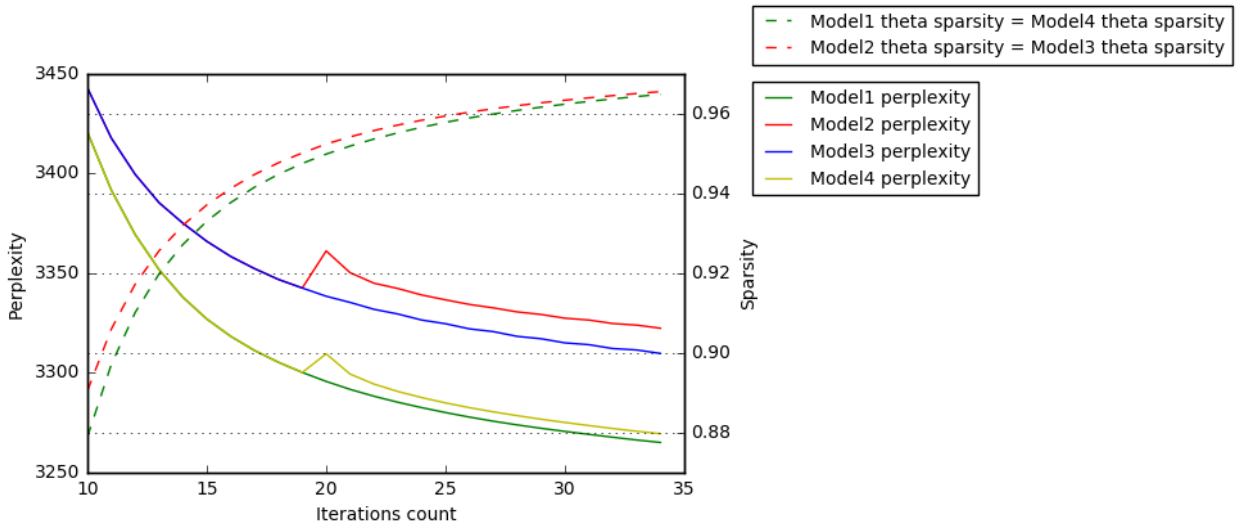

Рис.1 Изменение параметров моделей 1-4 по итерациям

После проведения качественной оценки полученных тем изучением списков 15 топ-слов каждой темы количество хорошо интерпретируемых (экспериментатор понял, что объединяет слова в данной теме), средне интерпретируемых (экспериментатор догадался, что объединяет слова в данной теме, используя некоторые дополнительные знания / несколько слов в построенной теме выбиваются из общей тематики) и плохо интерпретируемых тем распределилось следующим образом:

	модель 1	модель 2	модель 3	модель 4
хорошо	51	61	61	56
средне	15	18	9	14
плохо	34	21	30	30

При этом было отмечено более высокое качество модели во 2 случае (добавление 2 регуляризаторов) по сравнению с другими экспериментами: выбор топ-слов для каждой темы был лучше, модель построила меньше схожих между собой тем.

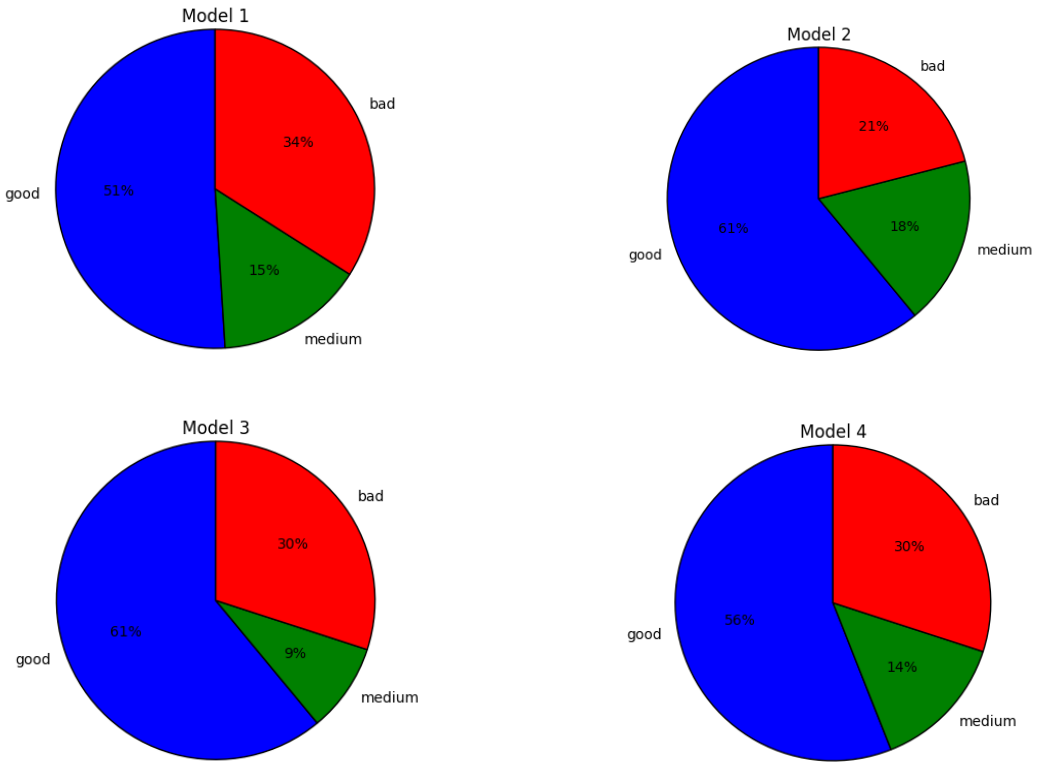

Рис.2 Доли хорошо, средне и плохо интерпретируемых тем в моделях 1-4

Пример:

Хорошо интерпретируемые темы:

['игра', 'устройство', 'компания', 'смартфон', 'приложение', 'пользователь', 'экран', 'также', 'система', 'версия', 'программа', 'планшет', 'мобильный', 'компьютер', 'модель']

['война', 'советский', 'армия', 'ссср', 'русский', 'войско', 'солдат', 'немецкий', 'военный', 'фронт', 'немец', 'сталин', 'офицер', 'орден', 'бой']

Средне интерпретируемые темы:

['ремарк', 'эрих', 'мария', 'одиночество', 'джон', 'сафарли', 'уильям', 'толстой', 'эльчин', 'брэдбрать', 'коэльо', 'чак', 'есенин', 'достоевский', 'рэй']

['забыть', 'скучать', 'хотеться', 'душить', 'боль', 'знаешь', 'нужный', 'уйти', 'уходить', 'ненавидеть', 'ряд', 'улыбаться', 'чувство', 'больно', 'мама']

Плохо интерпретируемые темы:

['донор', 'аниме', 'кровь', 'мочь', 'найти', 'мена', 'база', 'манга', 'поиск', 'наруто', 'проект', 'совершенно', 'волонтёр', 'адресный', 'обширный']

['подробность', 'узнать', 'тег', 'мина', 'бер', 'син', 'белна', 'бар', 'кингисепп', 'татарин', 'татарский', 'башкирский', 'кингисеппский', 'минь', 'синий']

Хорошо интерпретируемые, но бесполезные темы:

['александр', 'сергей', 'андрей', 'владимир', 'алексей', 'евгений', 'россия', 'дмитрий', 'михаил', 'николай', 'иван', 'ольга', 'юрий', 'игорь', 'олег']

['1800', 'октябрь', 'запись', '1900', '2000', 'суббота', 'пятница', '1000', '1600', '1700', '1500', '1200', 'четверг', '1400', 'воскресение']

Затем, после разреживания словаря (см. 3.3), были проведены эксперименты с тем же количеством тем, новым словарём (8171 слово) и:

5) без регуляризаторов, 35 итераций

6) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^3$ после 3 итераций и регуляризатора разреживания матрицы Θ с параметром $\tau = -1{,}5$ после 20 итераций, всего 35 итераций

7) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^3$ после 3 итераций, 35 итераций

8) с добавлением регуляризатора регуляризатора разреживания матрицы Θ с параметром $\tau = -1{,}5$ после 20 итераций, всего 35 итераций

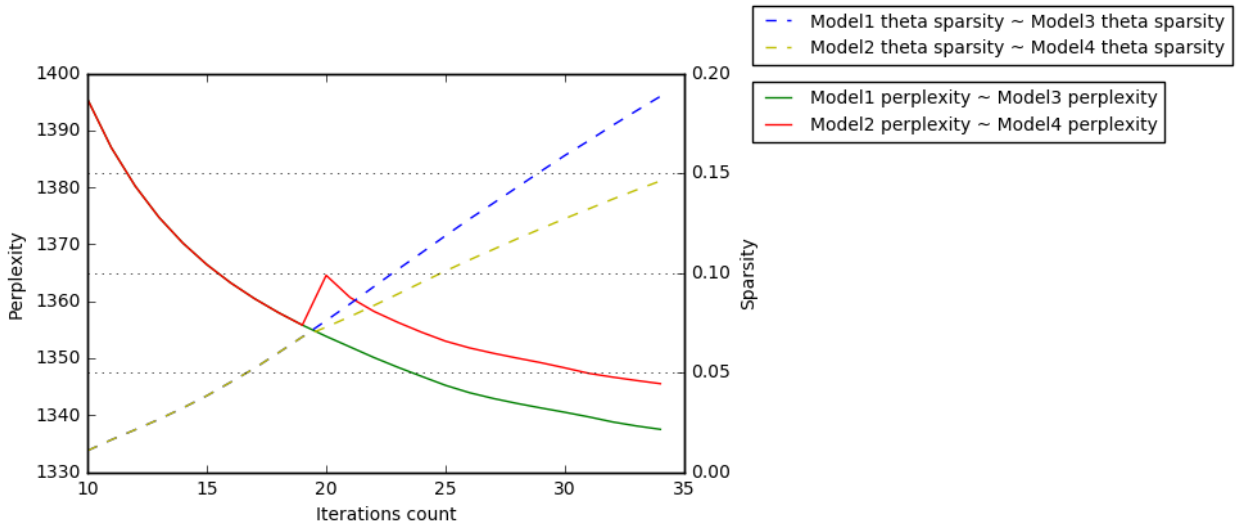

Рис.3 Изменение параметров моделей 5-8 по итерациям

Оценка количества интерпретируемых тем:

	модель 5	модель 6	модель 7	модель 8
хорошо	61	66	72	64
средне	13	15	17	13
плохо	26	19	11	23

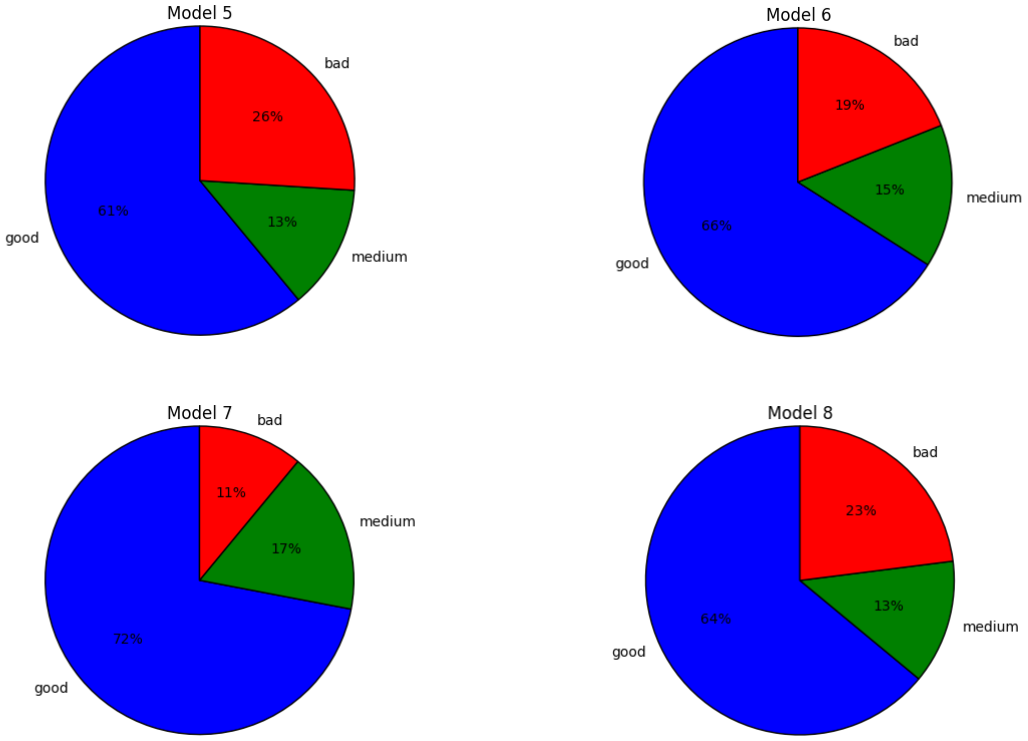

Рис.4 Доли хорошо, средне и плохо интерпретируемых тем в моделях 5-8

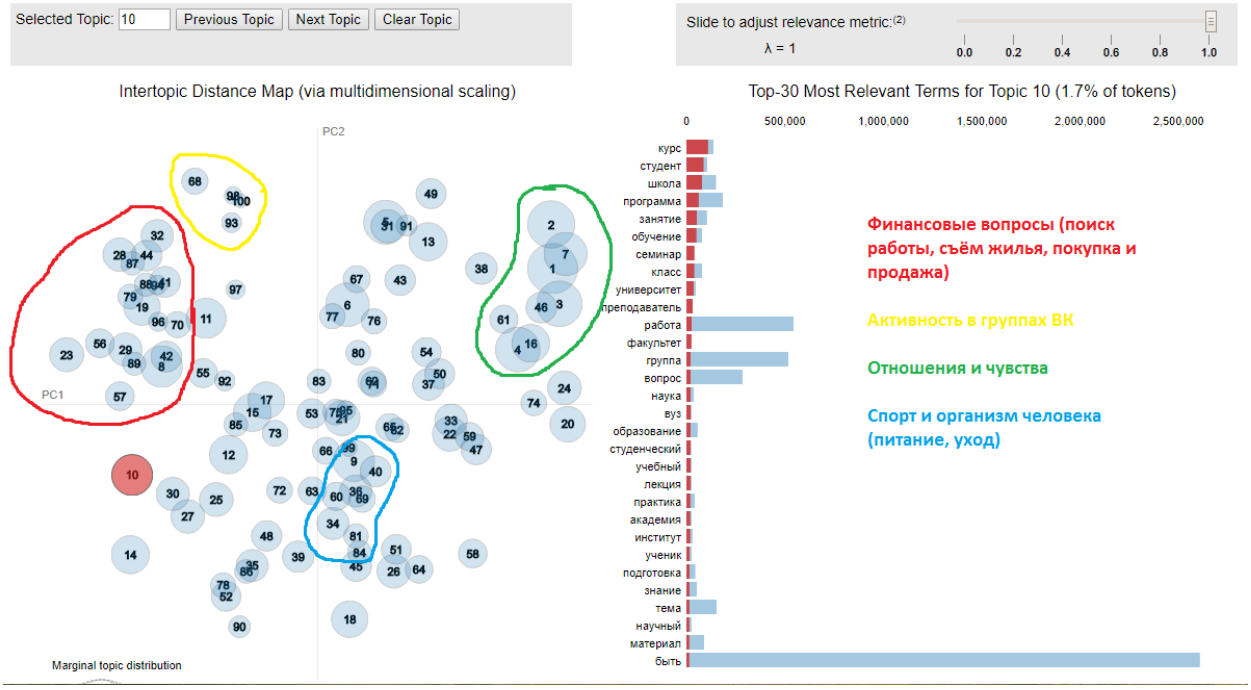

Рис.5 Графическое представление тем и выделение тематических кластеров в модели 7

При построении графического представления тем модели 7 с помощью инструмента pyLDAvis можно отметить, что на полученной диаграмме некоторые сходные между собой темы объединились в тематические кластеры.

Для дальнейших экспериментов была выбрана лучшая с точки зрения экспериментатора модель из полученных ранее - модель 7 (сокращённый словарь, регуляризатор разреживания матрицы Θ). Из построенных тем были выбраны интересующие нас для дальнейшего исследования, была проведена фильтрация документов коллекции: в коллекции остались только те документы, сумма вероятностей оставшихся тем для которых не менее 0,6. Словарь новой коллекции был ещё раз разрежен: слова встречаются не менее чем в 1% документов коллекции и не менее 100 раз. Были проведены эксперименты с тем же количеством тем (100), новым словарём (9068 слов) и:

9) без регуляризаторов, 35 итераций

10) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^5$ после 3 итераций и регуляризатора разреживания матрицы Θ с параметром $\tau = -1{,}5$ после 20 итераций, всего 35 итераций

11) с добавлением регуляризатора декорреляции тем с параметром $\tau = 10^2$ после 3 итераций, 35 итераций

12) с добавлением регуляризатора разреживания матрицы Θ с параметром $\tau = -1{,}5$ после 20 итераций, всего 35 итераций

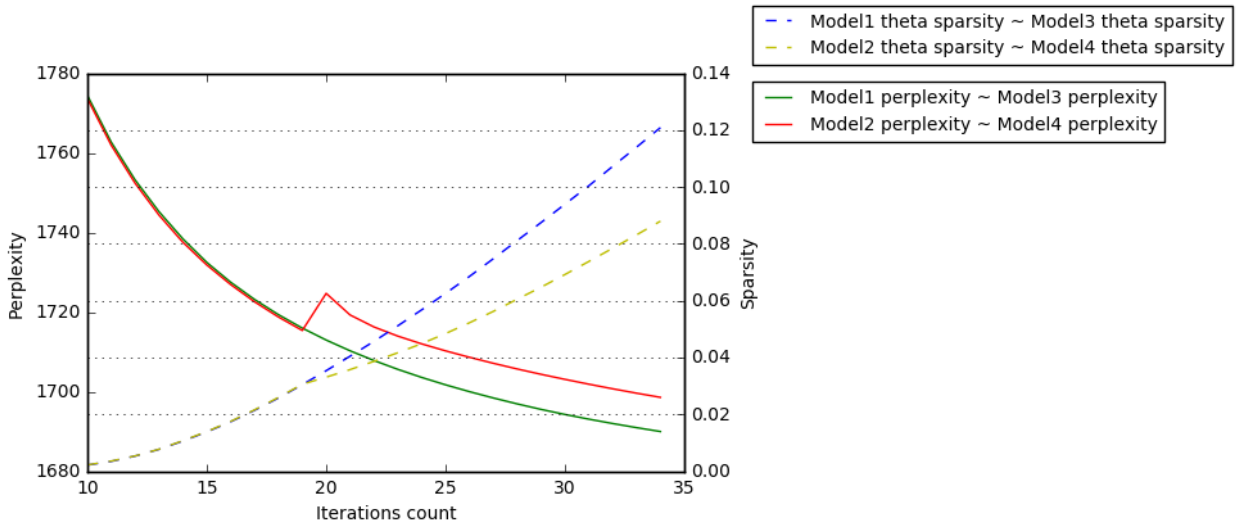

Рис.6 Изменение параметров моделей 9-12 по итерациям

Оценка количества интерпретируемых тем:

	модель 9	модель 10	модель 11	модель 12
хорошо	58	67	60	64
средне	20	16	21	14
плохо	22	17	19	22

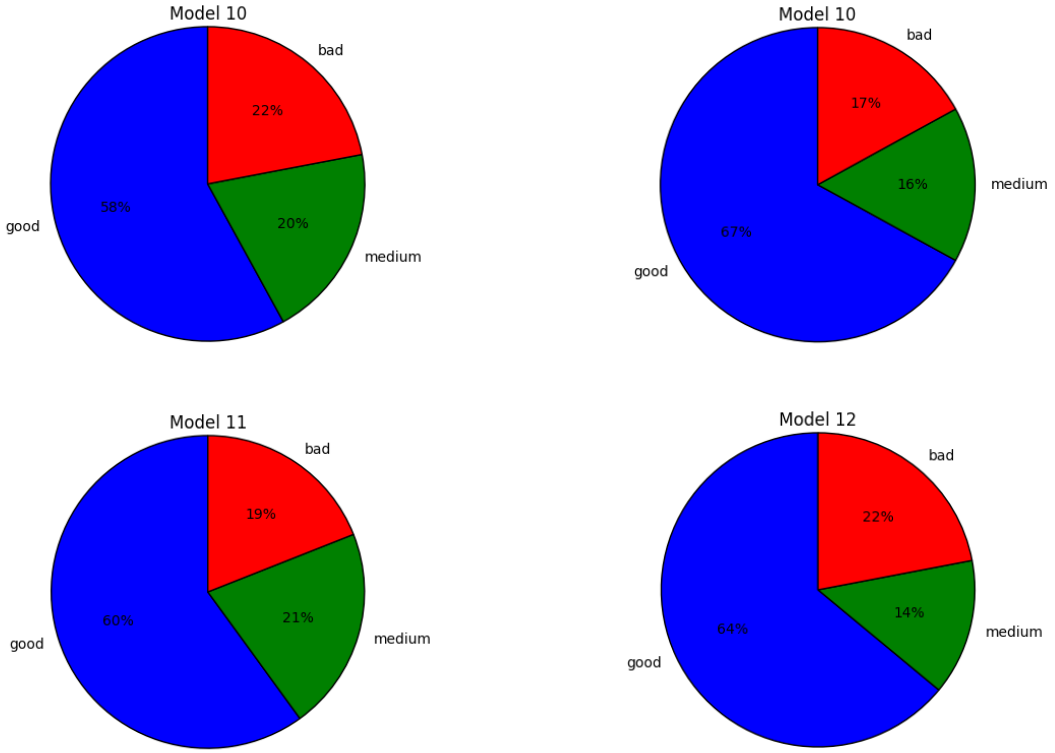

Рис.7 Доли хорошо, средне и плохо интерпретируемых тем в моделях 9-12

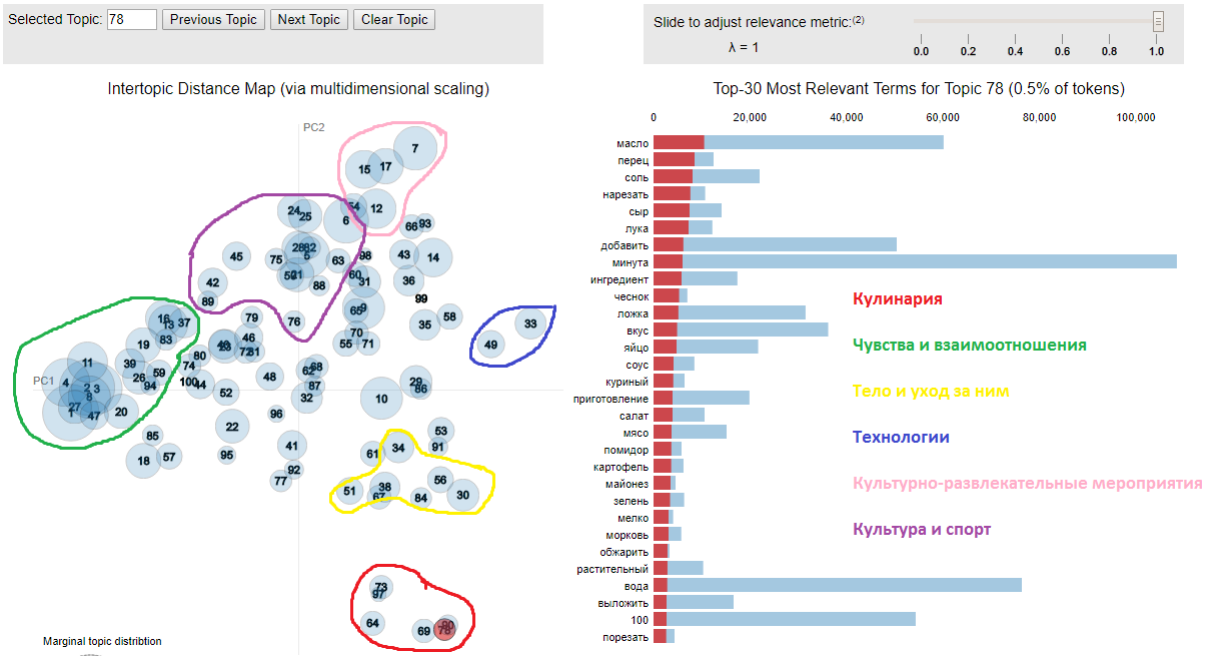

Рис.8 Графическое представление тем и выделение тематических кластеров в модели 10

В данных моделях выделяются уже более узконаправленные темы. Несмотря на то, что по-прежнему присутствуют неинтерпретируемые темы, понятность интерпретируемых тем и выбор топ-слов в них улучшились.

Возникли ситуации, когда темы из моделей 5-8 разделились на несколько подтем в моделях 9-12, также выделились некоторые из тем, которые были слишком узки для изначальных моделей.

Пример (набор тем о кулинарии и питании, названия тем условные)**:**

1) ['вода', 'полезный', 'чаять', 'растение', 'витамин', 'масло', 'также', 'кофе', 'день', 'напиток', 'вещество', 'продукт', 'чай', 'свойство', 'мёд'] – напитки

2) ['организм', 'жир', 'продукт', 'день', 'питание', 'пища', 'диета', 'вес', 'есть', 'белок', 'витамин', 'количество', 'быть', 'вещество', 'большой'] – диета и полезное питание

3) ['ложка', 'вода', 'минута', 'рецепт', 'стакан', 'сок', 'молоко', 'столовый', 'добавить', 'шоколад', 'яблоко', 'смесь', 'чайный', 'приготовить', 'мёд'], ['масло', 'яйцо', 'соль', 'добавить', 'мука', 'ингредиент', 'перец', 'сыр', 'нарезать', 'приготовление', 'лука', 'минута', 'тесто', 'сливочный', 'ложка'] – кулинария

→

1) ['мёд', 'вода', 'сок', 'чаять', 'чайный', 'ложка', 'лист', 'добавить', 'вкус', 'чай', 'стакан', 'кофе', 'лимон', 'корица', 'орех'] – напитки

2) ['организм', 'жир', 'мышца', 'углевод', 'питание', 'количество', 'белок', 'мышечный', 'также', 'вес', 'продукт', 'быть', 'витамин', 'являться', 'уровень'] – полезное питание

['продукт', 'день', 'диета', 'овощ', 'есть', 'пища', 'полезный', 'фрукт', 'витамин', 'организм', 'завтрак', 'сок', 'яблоко', 'вода', 'салат'] – диета

3) ['мука', 'масло', 'тесто', 'яйцо', 'минута', 'сахар', 'добавить', 'ложка', 'приготовление', 'форма', 'ингредиент', 'молоко', 'стакан', 'сахара', 'сливочный'] – кондитерское искусство

['масло', 'перец', 'соль', 'нарезать', 'сыр', 'лука', 'добавить', 'минута', 'ингредиент', 'чеснок', 'ложка', 'вкус', 'яйцо', 'соус', 'куриный'] – приготовление «несладких» блюд

['рецепт', 'банка', 'вкусный', 'есть', 'посуда', 'шоколад', 'готовить', 'приготовить', 'продукт', 'блюдо', 'вода', 'пирог', 'хлеб', 'кухня', 'добавлять'] – общая кулинария

4) ['блюдо', 'ресторан', 'напиток', 'вино', 'пиво', 'вкус', 'кухня', 'вкусный', 'меню', 'кофе', 'пицца', 'гость', 'коктейль', 'еда', 'кафе'] – ресторанный бизнес

5. Выводы

Были проведены эксперименты по тематическому моделированию сообществ социальной сети ВКонтакте с использованием различных регуляризаторов в моделях библиотеки BigARTM и различных размеров словарей. Из полученнных данных можно видеть, что:

1) Для исходного словаря коллекции без сокращений использование моделей с регуляризаторами декорреляции тем и разреживания матрицы вероятностей тем в документах может увеличить количество интерпретируемых тем (до 20%) и улучшить их качество.

2) Доля интерпретируемых с точки зрения эксперта тем для модели с сокращённым словарём выше, чем для полного словаря; использование регуляризаторов опять же даёт лучшие результаты.

3) Доля интерпретируемых тем при выборе для изучения подмножества исходных тем существенно не меняется (отдельные результаты могут быть даже ниже), но модель получает возможность выделить более узконаправленные темы и определить подтемы в исходных темах.