Выделение тем текстовой коллекции

1 Данные

Для моделирования была использована англоязычная коллекция токсичных комментариев:

2 Предобработка

С коллекцией были проведены следующие операции:

- приведение символов к нижнему регистру с помощью регулярных выражений удалены ссылки
- удаление слов с длиной<3 символов
- удаление стоп слов и лемматизация с помощью библиотеки ntlk
- построен словарь с помощью BigARTM

3 Эксперименты

Попытаемся выделить из полученного словаря 20 тем с помощью различных моделей. Для этого будем делать 100 проходов по коллекции, так как этого достаточно для сходимости перплексии. Использованные модели:

- LDA
- ARTM
  - без разреженностью модальности меток классов и регуляризаторами декорреляции
  - с разреженностью модальности меток классов и регуляризаторами декорреляции

3.1 LDA

Хорошо интерпретируемые темы:
Topic 3: ['wikipedia', 'help', 'page', 'links', 'pages', 'welcome', 'ask', 'talk', 'style', 'question']
Topic 5: ['http', 'com', 'www', 'org', 'school', 'city', 'new', 'states', 'united', 'university']
Topic 6: ['sources', 'source', 'information', 'references', 'reliable', 'reference', 'book', 'material', 'original', 'facts']
Topic 9: ['term', 'definition', 'theory', 'science', 'scientific', 'human', 'one', 'knowledge', 'study', 'many']
Topic 19: ['would', 'case', 'comment', 'best', 'point', 'opinion', 'matter', 'agree', 'whether', 'support']
Средне интерпретируемые темы:
Topic 0: ['article', 'deletion', 'page', 'please', 'wikipedia', 'deleted', 'articles', 'may', 'add', 'speedy']
Topic 1: ['category', 'people', 'live', 'country', 'history', 'american', 'white', 'british', 'also', 'greek']
Topic 2: ['edit', 'please', 'user', 'wikipedia', 'page', 'edits', 'wiki', 'editing', 'stop', 'talk']
Topic 8: ['war', 'anti', 'political', 'jews', 'god', 'jewish', 'government', 'pig', 'church', 'freedom']
Topic 12: ['use', 'image', 'link', 'copyright', 'wikipedia', 'fair', 'images', 'page', 'media', 'free']
Topic 15: ['discussion', 'consensus', 'issue', 'involved', 'ago', 'dispute', 'editors', 'project', 'comments', 'months']
Плохо интерпретируемые темы:
Topic 4: ['f*ck', 'shit', 'f*cking', 'hey', 'ass', 'gay', 'dont', 'love', 'life', 'lol']
Topic 7: ['name', 'utc', 'english', 'suck', 'language', 'names', 'word', 'used', 'german', 'july']
Topic 10: ['like', 'know', 'think', 'people', 'one', 'really', 'would', 'get', 'say', 'something']
Topic 13: ['list', 'n*g*er', 'film', 'music', 'series', 'show', 'band', 'also', 'song', 'video']
Topic 14: ['article', 'section', 'think', 'articles', 'would', 'also', 'needs', 'one', 'could', 'version']
Topic 16: ['talk', 'page', 'thanks', 'work', 'see', 'sorry', 'time', 'good', 'hate', 'thank']
Topic 11: ['die', 'law', 'dead', 'fag', 'court', 'cool', 'money', 'death', 'day', 'must']
Topic 17: ['game', 'number', 'system', 'penis', 'team', 'fish', 'small', 'games', 'numbers', 'space']
Topic 18: ['block', 'personal', 'admin', 'attack', 'wikipedia', 'attacks', 'bad', 'faith', 'editor', 'moron']

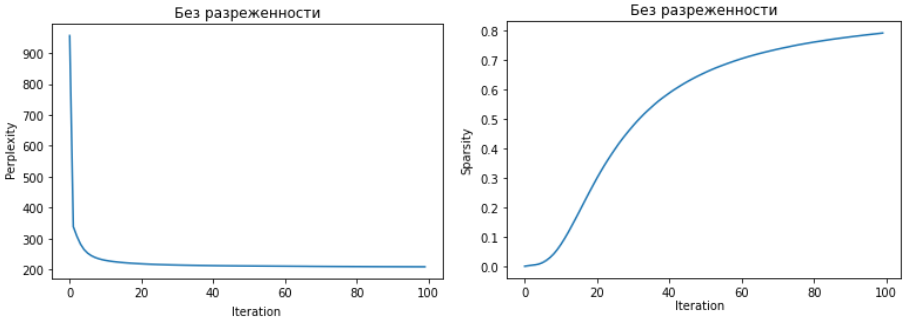

Рис. 1: ARTM без разреженности модальности

3.2 ARTM

Хорошо интерпретируемые темы:
topic_0 : $['god','anti','fat','party','jew','jews']$
$topic_2$: $['school','redirect','die','year','city','university']$
$topic_3$: $['site','www','gay','info','news','show']$
$topic_6$: $['number','system','theory','definition','science','level']$
$topic_7$: $['english','word','language','correct','called','term']$
$topic_8$: $['agree','needs','version','mention','topic','perhaps']$
$topic_9$: $['references','reference','reliable','book','material','original']$
Средне интерпретируемые темы:
$topic_1$: $['sorry','suck','care','dont','call','live']$
$topic_10$: $['trying','give','try','delete','review','template']$
$topic_12$: $['opinion','clearly','matter','others','simply','evidence']$
$topic_15$: $['editor','post','revert','consensus','issue','category']$
$topic_17$: $['fair','images','media','created','listed','jpg']$
$topic_19$: $['n*g*er','war','states','state','country','government']$
Плохо интерпретируемые темы:
$topic_4$: $['shit','f*cking','hey','bad','love','lol']$
$topic_5$: $['american','known','white','small','black','human']$
$topic_11$: $['got','else','yes','come','last','thought']$
$topic_13$: $['style','hello','write','check','great','contributions']$
$topic_14$: $['life','day','ass','away','nice','game']$
$topic_16$: $['speedy','leave','adding','notable','hate','guidelines']$
$topic_18$: $['admin','users','account','request','attack','address']$

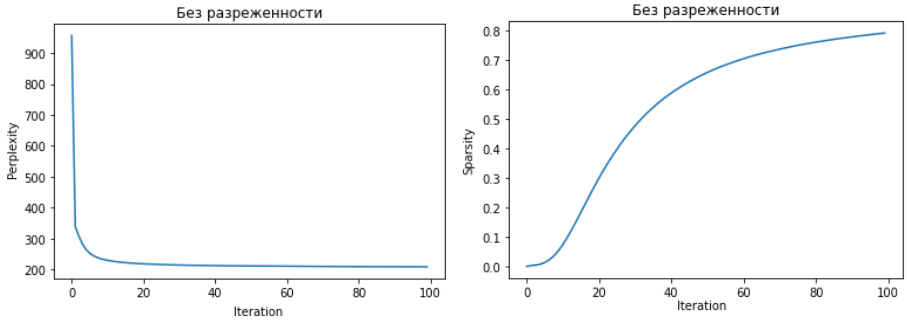

Рис. 2: ARTM с разреженностью

Хорошо интерпретируемые темы:

topic_0 : $['god','anti','fat','party','jew','jews']$
$topic_2$: $['school','redirect','die','year','city','university']$
$topic_3$: $['site','www','gay','info','news','show']$
$topic_6$: $['number','system','theory','science','level','set']$
$topic_7$: $['english','word','language','correct','term','called']$
$topic_8$: $['agree','needs','perhaps','version','mention','important']$
$topic_9$: $['references','reference','reliable','says','book','material']$

Средне интерпретируемые темы:

$topic_1$: $['sorry','suck','care','dont','call','live']$
$topic_15$: $['editor','post','consensus','revert','issue','category']$
$topic_17$: $['fair','created','images','media','listed','jpg']$
$topic_19$: $['n*g*er','war','states','country','state','government']$
$topic_12$: $['opinion','clearly','pov','view','matter','simply']$

Плохо интерпретируемые темы:

$topic_4$: $['shit','f*cking','hey','bad','love','lol']$
$topic_5$: $['american','known','white','family','black','small']$
$topic_10$: $['trying','give','try','delete','review','template']$
$topic_11$: $['last','got','else','yes','come','long']$
$topic_13$: $['style','hello','check','great','write','contributions']$
$topic_14$: $['life','day','away','nice','ass','game']$
$topic_16$: $['speedy','leave','adding','notable','hate','guidelines']$
$topic_18$: $['continue','admin','users','account','request','attack']$

4 Выводы

- Все модели сошлись по перплексии.
- Модель с использованием регуляризаторов и разреживанием матрицы показывает наибольшее число интерпретируемых тем
- Модель без разреживания идет за ней
- LDA показал наибольшее число плохо интерпретируемых тем.

В целом всем моделям удалось выделить наиболее различные темы.