Отчет по индивидуальному заданию по курсу "Вероятностные тематические модели"студента 6 курса группы М05-104а Панкратова Виктора Владимировича

1 Введение

Рассматривается задача категоризации небольшой текстовой коллекции. Предполагается, что применение тематической модели не позволит задать число тем равное числу категорий и получить искомое соответствие категории и документа. В работе проверяется предыдущее утверждения: приводится процесс построения модели, описывается подбор регуляризаторов и количества тем, необходимых для описания документов коллекции. Также проводится сравнение результатов построения тематической модели и других стандартных методов для выбранной задачи категоризации.

2 Цели и задачи

- Построить тематическую модель для коллекции новостей BBC 2004-2005 годов небольшого размера, используя библиотеку BigARTM
- Путем оценки полученных тем на интерпретируемость оценить сверху и снизу необходимое число тем модели
- Сравнить стандартные решения задачи категоризации и построение тематической модели с тем же числом тем

Для оценки качества решения задачи категоризации используются f1-score и accuracy.

3 baseline решения

Для сравнения использованы три способа решения задачи категоризации. Первый - naive bayes и accuracy 0.959. Второй - логистическая регрессия и accuracy 0.980. Третий - SVC и accuracy 0.979. С данными значениями будут сравниваться полученные результаты.

4 Подготовка данных

4.1 Исходные данные

Данные - коллекция новостей BBC за 2004-2005 год, доступная по (ссылке на kaggle). Она состоит из 2225 документов, принадлежащих одной из пяти категорий: бизнес, спорт, развлечения, политика, технологии.

4.2 Подготовка данных

Подготовка данных включает в себя несколько этапов:

- Приведение букв в нижний регистр
- Удаление стоп слов
- Лемматизация
- Токенизация

Все указанные действия проводились с помощью функции gensim.utils.lemmatize. Так как рассматривается коллекция новостей, имена собственные и прочие не словарные слова не удалялись. Дополнительно исключались слова, общее число которых во всех документах не превосходило трех.

4.3 Полученные данные

На основании предыдущих результатов была создана матрица n_{wd}, в строке i и столбце j которой описывается число встреченных слов i в документе j и словарь, сопоставляющий номер слова его строке в матрице n_{wd}.Итоговое среднее число слов в документе - 131, а максимальное - 898.

5 Эксперименты,простейшая модель, 5 тем

5.1 Построение и обучение

Была построена тематическая модель для получения пяти тем по всей коллекции. Обучение велось до тех пор, пока перплексия не закончит изменения. В данном эксперименте это означает относительную разницу

значений между последовательными итерациями 0.0001 процент или 100 итераций. В последующих обучение останавливалось, когда переплексия увеличивалась по сравнению с предыдущей итерацией и отклонялась не более чем на 0.001 процент. Предыдущие критерии проверялись по добавленной и выводимой в BigARTM метрике перплексии.

5.2 Результаты и их анализ

Выведем наиболее вероятные 20 слов каждой темы:

1. 'go/VB', 'mr/JJ', 'spokesman/NN', 'person/NN', 'law/NN', 'leader/NN', 'issue/NN', 'election/NN', 'year/NN', 'labour/NN', 'claim/VB', 'take/VB', 'also/RB', 'party/NN', 'make/VB', 'minister/NN', 'tell/VB', 'government/NN', 'mr/NN', 'say/VB'

2. ['make/VB', 'new/JJ', 'share/NN', 'economy/NN', 'price/NN', 'business/NN', 'analyst/NN', 'growth/NN', 'expect/VB', 'however/RB', 'last/JJ', 'firm/NN', 'rise/VB', 'month/NN', 'bn/NN', 'also/RB', 'company/NN', 'market/NN', 'year/NN', 'say/VB']

3. ['work/VB', 'use/VB', 'see/VB', 'number/NN', 'go/VB', 'year/NN', 'want/VB', 'time/NN', 'take/VB', 'service/NN', 'technology/NN', 'many/JJ', 'new/JJ', 'get/VB', 'way/NN', 'make/VB', 'also/RB', 'used/VB', 'say/VB', 'person/NN']

4. ['see/VB', 'director/NN', 'play/VB', 'new/JJ', 'time/NN', 'show/NN', 'win/VB', 'last/JJ', 'best/JJ', 'go/VB', 'take/VB', 'award/NN', 'first/JJ', 'make/VB', 'star/NN', 'include/VB', 'film/NN', 'also/RB', 'say/VB', 'year/NN']

5. ['back/RB', 'side/NN', 'also/RB', 'old/JJ', 'world/NN', 'come/VB', 'team/NN', 'take/VB', 'get/VB', 'last/JJ', 'player/NN', 'make/VB', 'first/JJ', 'play/VB', 'go/VB', 'time/NN', 'year/NN', 'game/NN', 'win/VB', 'say/VB']

Заметим, что темы получились интерпретируемыми. По первой теме понятно, что она описывает политику, по второй - бизнес, по пятой - спорт. Менее доступно описание остальных двух, но зная исходные категории, можно понять, что они восстановились

Полученные темы недостаточно хорошо описывают коллекцию. Во-первых, они смешались: например, слово "год"есть в выборке для всех

пяти тем. Во-вторых, как уже было указано, некоторые темы сложно интерпретировать.

Рассмотрим исходную задачу категоризации. Результаты приведены как f1-score для каждой темы:

$$'business' : 0.855,' politics' : 0.354,' entertainment' : 0.919,' sport' : 0.917,' tech' : 0.516$$

Accuracy для данной задачи - 0.599, что является слишком низким значением для такой задачи.

Итог: темы выделены неудачно, необходимо улучшать модель.

6 Эксперименты, добавление декоррелятора, модель для пяти тем

В модель, описанную в предыдущем пункте был добавлен регуляризатор декоррелирования для матрицы Ф. Указанные выше метрики практически не изменялись при любом из рассматриваемых коэффициентов регуляризации. Для сравнения ниже приведена зависимость Accuracy от коэффициента регуляризации:

1. $\tau = 0.1, Accuracy = 0.599$
2. $\tau = 1, Accuracy = 0.599$
3. $\tau = 10, Accuracy = 0.599$
4. $\tau = 100, Accuracy = 0.599$
5. $\tau = 1000, Accuracy = 0.599$
6. $\tau = 10000, Accuracy = 0.599$
7. $\tau = 100000, Accuracy = 0.598$
8. $\tau = 1000000, Accuracy = 0.595$
9. $\tau = 3000000, Accuracy = 0.546$
10. $\tau = 5000000, Accuracy = 0.555$
11. $\tau = 10000000, Accuracy = 0.518$

Для $\tau = 1000000$ рассмотрим полученные темы:

- ['club/NN', 'world/NN', 'injury/NN', 'season/NN', 'back/RB', 'final/JJ', 'coach/NN', 'champion/NN', 'match/NN', 'england/NN', 'old/JJ', 'side/NN', 'get/VB', 'first/JJ', 'team/NN', 'go/VB', 'play/VB', 'player/NN', 'win/VB', 'game/NN']
- ['oscar/NN', 'prize/NN', 'go/VB', 'song/NN', 'play/VB', 'win/VB', 'singer/NN', 'first/JJ', 'star/VB', 'director/NN', 'movie/NN', 'tv/NN', 'actor/NN', 'music/NN', 'best/JJ', 'show/NN', 'include/VB', 'award/NN', 'star/NN', 'film/NN']
- ['help/VB', 'number/NN', 'election/NN', 'come/VB', 'make/VB', 'use/VB', 'labour/NN', 'see/VB', 'system/NN', 'technology/NN', 'work/VB', 'want/VB', 'go/VB', 'many/JJ', 'service/NN', 'new/JJ', 'get/VB', 'way/NN', 'used/VB', 'person/NN']
- ['continue/VB', 'biggest/JJ', 'bank/NN', 'cost/NN', 'expect/VB', 'sale/NN', 'month/NN', 'economic/JJ', 'however/RB', 'business/NN', 'economy/NN', 'firm/NN', 'share/NN', 'price/NN', 'analyst/NN', 'growth/NN', 'company/NN', 'rise/VB', 'bn/NN', 'market/NN']
- ['issue/NN', 'agree/VB', 'statement/NN', 'public/JJ', 'court/NN', 'member/NN', 'former/JJ', 'deal/NN', 'claim/VB', 'law/NN', 'case/NN', 'spokesman/NN', 'decision/NN', 'minister/NN', 'government/NN', 'tell/VB', 'mr/NN', 'make/VB', 'also/RB', 'year/NN']

Итог:применение регуляризатора улучшило интерпретируемость тем в тематической модели, но не качество категоризации.

7 Эксперименты,увеличение числа тем до 8

Построена аналогичная прошлому пункту модель. Коэффициент регуляризатора декоррелирования был установлен значением 100000. Для интерпретируемости тем стало необходимо, чтобы темы пересеклись, поэтому выбор не согласуется с результатами предыдущего пункта. Число тем увеличено до 8. Итоговые темы(20 наиболее вероятных слов приведены только для не описанных ранее):
судебная(['law/NN', 'face/VB', 'tell/VB', 'firm/NN', 'trial/NN', 'action/NN',

'lawyer/NN', 'month/NN', 'former/JJ', 'charge/NN', 'take/VB', 'also/RB', 'legal/JJ', 'claim/VB', 'company/NN', 'last/JJ', 'case/NN', 'year/NN', 'court/NN', 'say/VB']),
бизнес, **политическая**,
выборы(['michael/NN', 'issue/NN', 'also/RB', 'mr/VB', 'mr/JJ', 'go/VB', 'campaign/NN', 'make/VB', 'person/NN', 'tell/VB', 'prime/JJ', 'labour/NN', 'blair/NN', 'minister/NN', 'leader/NN', 'party/NN', 'tony/JJ', 'election/NN', 'say/VB', 'mr/NN']),
технологии,**спорт**,**фильмы**,
музыка(['rock/NN', 'play/VB', 'last/JJ', 'album/NN', 'time/NN', 'take/VB', 'singer/NN', 'new/JJ', 'band/NN', 'first/JJ', 'song/NN', 'make/VB', 'star/NN', 'show/NN', 'go/VB', 'also/RB', 'include/VB', 'music/NN', 'say/VB', 'year/NN']).

Итог: некоторые из описанных тем уже слабо интерпретируемы(выборы), дальше увеличивать число тем нет смысла. Пример приведен в секции "Описание дополнительных экспериментов"далее. Среди тем есть пересечения, например, слова law,tell,say принадлежат и судебной и политической теме, но это желаемый результат, поэтому кэффициент регуляризатора декоррелирования считается подобранным верно.

8 Эксперименты, конечная модель

В окончательной модели был установлен коэффициент регуляризатор декоррелирования для матрицы Φ до 100000 и введен регуляризатор декоррелирования для матрицы Θ с коэффициентом 1000000. Добавление других регуляризаторов лишь ухудшало результат и поэтому оставлены только описанные два. Изменение коэффициентов также не увеличило описанные метрики. Получившиеся темы:

- ['also/RB', 'decision/NN', 'secretary/NN', 'case/NN', 'take/VB', 'mr/JJ', 'mr/VB', 'spokesman/NN', 'make/VB', 'law/NN', 'leader/NN', 'issue/NN', 'election/NN', 'claim/VB', 'party/NN', 'tell/VB', 'minister/NN', 'government/NN', 'say/VB', 'mr/NN']
- ['sale/NN', 'economic/JJ', 'last/JJ', 'share/NN', 'economy/NN', 'price/NN', 'expect/VB', 'business/NN', 'analyst/NN', 'however/RB', 'growth/NN', 'month/NN', 'also/RB', 'firm/NN', 'rise/VB', 'bn/NN', 'company/NN', 'market/NN', 'year/NN', 'say/VB']

- ['user/NN', 'work/VB', 'labour/NN', 'use/VB', 'time/NN', 'number/NN', 'want/VB', 'system/NN', 'take/VB', 'many/JJ', 'technology/NN', 'service/NN', 'new/JJ', 'make/VB', 'get/VB', 'also/RB', 'way/NN', 'say/VB', 'used/VB', 'person/NN']
- ['tv/NN', 'last/JJ', 'play/VB', 'actor/NN', 'music/NN', 'director/NN', 'go/VB', 'win/VB', 'take/VB', 'make/VB', 'best/JJ', 'first/JJ', 'show/NN', 'say/VB', 'also/RB', 'award/NN', 'star/NN', 'include/VB', 'year/NN', 'film/NN']
- ['match/NN', 'back/RB', 'come/VB', 'world/NN', 'side/NN', 'old/JJ', 'take/VB', 'make/VB', 'team/NN', 'last/JJ', 'get/VB', 'first/JJ', 'year/NN', 'go/VB', 'time/NN', 'player/NN', 'play/VB', 'say/VB', 'win/VB', 'game/NN']

Ниже приведен f1-score для задачи категоризации:

$'business' : 0.847,' politics' : 0.723,' entertainment' : 0.918,' sport' : 0.965,' tech' : 0.870$

Значение Accuracy: 0.865.

Итог: результаты, полученные описанными в начале решениями baseline достигнуты не были. Регуляризаторы, кроме регуляризатора декоррелирования, для данной задачи не дают лучшего решения.

9 Описание дополнительных экспериментов

1)Был повторен первый эксперимент с другим порядком строк матрицы. Результаты - темы чуть хуже интерпретируются, но f1-score для них:

$'business' : 0.855,' politics' : 0.753,' entertainment' : 0.919,' sport' : 0.965,' tech' : 0.892$

Accuracy возросла до 0.878. Итог:результаты не возобновляемые и поэтому все эксперименты в отчете повторялись несколько раз. Выбирался худший результат. Лучший результат для Accuracy из всех моделей был 0.958 и достигнут на конечной модели.

2)Коллекция была ограничена: были взяты только документы с изначальным числом слов больше 300. Все метрики повысились, Accuracy равна 0.945 для первой описанной модели. Итог: многие плохие результаты модели связаны с малой длиной документа.
3)Построена аналогичная эксперименту с добавлением декоррелятора модель с единственным изменением: число тем увеличено до 10. Некоторые темы получились повторяющимися:

- ['saturday/NN', 'chance/NN', 'goal/NN', 'ireland/NN', 'play/VB', 'nation/NN', 'match/NN', 'half/NN', 'come/VB', 'injury/NN', 'second/JJ', 'england/NN', 'win/VB', 'coach/NN', 'cup/NN', 'team/NN', 'season/NN', 'side/NN', 'player/NN', 'game/NN']
- ['french/JJ', 'wimbledon/NN', 'third/JJ', 'title/NN', 'event/NN', 'champion/NN', 'final/JJ', 'roger/NN', 'israel/NN', 'match/NN', 'grand/JJ', 'final/NN', 'tournament/NN', 'set/NN', 'american/JJ', 'tennis/NN', 'seed/NN', 'round/NN', 'open/JJ', 'australian/JJ']

Некоторые темы получились не интерпретируемыми, например ['back/RB', 'something/NN', 'never/RB', 'give/VB', 'much/JJ', 'feel/VB', 'way/NN', 'lot/NN', 'happen/VB', 'really/RB', 'world/NN', 'add/VB', 'old/JJ', 'tell/VB', 'thing/NN', 'think/VB', 'want/VB', 'know/VB', 'see/VB', 'get/VB']

Итог: количество тем модели должно быть меньше 10, из чего мы получаем оценку сверху.

10 Выводы

Были проведены эксперименты для коллекции новостей BBC. Было получено, что несмотря на то, что простые решения, такие как наивный байес или логистическая регрессия могут решить задачу категоризации для пяти категорий с высокой точностью, тематическая модели не может выделить ровно 5 тем и необходимо 7-8. Также было обнаружено, что тематическая модель может использоваться как классификатор, однако необходима точная настройка регуляризаторов и их коэффициентов. Даже при такой настройке результаты применения тематической модели как решения задачи категоризации оказываются не лучше результатов, полученных стандартными решениями той же задачи, но могут быть к ним близки.