НИУ ВШЭ

Школа Филологии

Сильвестров Алексей Сергеевич

Кластеризация больших коллекций документов методами тематического моделирования

Курсовая работа

Научный руководитель:

к.л.н., с.н.с

Бонч-Осмоловская А.А.

Москва — 2015

Оглавление

Аннотация

Данная курсовая работа посвящена исследованию структуры больших коллекций текстовых документов методами тематического моделирования на примере корпуса поэзии stihi.ru.

Тематическое моделирование позволяет построить модель над коллекцией документов, способную определить принадлежность документа к темам, извлеченным из этой коллекции в процессе построения модели.

Введение

Для решения многих лингвистических задач используются текстовые корпуса — специальным образом подобранные и структурированные коллекции текстов. Наиболее информативными являются размеченные корпуса, то есть такие, в которых частям текста приписана лингвистическая информация, например, каждое слово отнесено к той или иной части речи.

Существующие на сегодняшний день корпуса можно разделить на представительные и специализированные. Представительный корпус содержит по возможности все типы письменных и устных текстов, представленные в языке (художественные, учебные, научные, разговорные, и т.п.), пропорционально их доле в языке соответствующего периода. Примером современного представительного корпуса может послужить Национальный корпус русского языка (НКРЯ), на сегодняшний день содержащий свыше 360 млн словоупотреблений‘ [1].

К корпусам второго типа относятся коллекции текстов новостных лент Reuters, газетных статей Wall Street Journal, электронное хранилище художественных текстов "Проект Гутенберг"и другие. Анализу подобного специализированного корпуса - корпуса поэзии stihi.ru - посвящена данная курсовая работа. Корпус поэзии stihi.ru представляет собой коллекцию стихотворных произведений более чем двухсот тысяч авторов.

Задача, представленная в данной курсовой работе - кластеризация и анализ корпуса поэзии - решена методами тематического моделирования. Тематическое моделирование имеет своей целью построение модели коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Тема является результатом одновременной кластеризации и слов, и документов по их семантической близости, причем, документы и слова принадлежат сразу нескольким темам с различными вероятностями [2].

Применительно к анализу художественных текстов, тематическое моделирование позволяет автоматически определить основные темы, затронутые авторами, т.е служит решением самодостаточной задачи в сфере электронного представления гуманитарных знаний.

Постановка задачи

Целью курсовой работы является исследование структуры большой коллекции текстовых документов stihi.ru методами тематического моделирования. Тематическая модель позволит автоматически определить основные темы, характерные для данного корпуса и провести кластеризацию документов корпуса согласно этим темам.

Для достижения данной цели, были поставлены следующие задачи:

1. Исследовать лингвистические особенности корпуса поэзии, существенные для построения тематической модели.
2. Обучить тематическую модель с учетом этих особенностей.
3. Провести анализ и дать интерпретацию кластерам документов, полученных тематической моделью.

Обзор предметной области

В данной главе дан обзор направлений компьютерных наук, с точки зрения исследовательских задач, сформулированных в курсовой работе.

0.1 Кластеризация

Кластеризация документов - задача компьютерной лингвистики, заключающаяся в автоматическом выявлении групп документов, на основе их попарной семантической схожести. Алгоритмы кластеризации включают в себя графовые, статистические и другие методы, общей идеей которых является отсутствие заранее заданных характеристик групп. Одной из целей кластеризации является выявление скрытых закономерностей, если таковые присутствуют, в массиве данных. Например, выявление общих тем в коллекции художественных текстов методами тематического моделирования является результатом процесса кластеризации по их семантической близости.

0.1.1 Векторное представление документа

Для представления документов используется векторная модель (Vector Space Model), ставящая в соответствие каждому слову вес, согласно выбранной функции. Векторное представление позволяет находить близость между документами и таким образом решать задачи кластеризации. Классическим методом построения векторной модели является TF-IDF. Для сравнения векторов документов существует более 70 метрик, одна из них - косинусная мера [3].

Идея работы с векторными пространствами документов получила развитие в виде метода латентно-семантического индексирования [4] . Метод заключается в том, что векторы соответствующие документам из векторного пространства слов проецируются в пространство меньшей размерности, сохраняющее наиболее важную семантическую информацию о документах.

В основе латентно-семантического индексирования лежит теорема о сингулярном разложении, согласно которой любая прямоугольная вещественная матрица может быть представлена в виде

$$M = USV$$

Где S - диагональная матрица сингулярных значений матрицы M, U и V - ортогональные матрицы. Если в матрице S выбрать состоящую из ее k первых строк матрицу S_k и соответствующие столбцы матриц U и V

$$M_k = U_k S_k V_k$$

То матрица M_k будет отражать основную структуру различных закономерностей исходной матрицы, например основные темы. Размерность матрицы M_k подбирается эмпирически на основе следующего принципа: матрица размерности, близкой к исходной не дает никаких преимуществ при работе с векторами, матрица небольшой размерности усложнит определение различий между документами.

Метод латентно-семантического индексирования предполагает, что распределение тем по документам равномерно. На смену ему пришли вероятностные модели, позволяющие построить более реалистичные тематические модели.

0.1.2 Вероятностное тематическое моделирование

Вероятностное тематическое моделирование - набор алгоритмов, реализующих вероятностный подход к построению тематической модели. Для каждого документа определяется распределение его слов по темам, последовательность слов не имеет значения.

Первым таким алгоритмом был метод вероятностного латентно-семантического индексирования (Probabilistic Latent Semantic Indexing, PLSI) [5]. Недостатками этой модели были в том числе неприменимость к большим коллекциям документов и склонность к переобучению.

На смену PLSI пришла модель Скрытого размещения Дирихле (Latent Dirichlet Allocation, LDA) — генеративная вероятностная модель, предложенная Блеем и соавторами [6]. Согласно этой модели, процесс генерации документа происходит следующим образом:

1. Случайно выбрать для документа D его распределение по темам θ_D
2. Для каждого слова в документе D:

(a) Случайно выбрать тему t из распределения θ_D

(b) Случайно выбрать слово из распределения слов ϕ_t выбранной темы t

В модели LDA предполагается, что $\theta_D \sim Dir(\alpha)$, $\phi_t \sim Dir(\beta)$, где α и β - гиперпараметры распределения Дирихле. Количество тем - фиксированный параметр, оптимальное значение которого подбирается экспериментально.

0.2 Электронное представление гуманитарных знаний

Электронное представление гуманитарных знаний (Digital Humanities) - область исследований на стыке гуманитарных наук и информационных технологий. В основе этой области лежит использование современных компьютерных методов анализа, обработки и представления данных для работы с культурным наследием человечества.

Применительно к анализу художественных текстов Digital Humanities предлагает решения из области анализа социальных сетей, корпусных технологий, и других методов компьютерной лингвистики [http://litlab.stanford.edu/current-projects/ , http://worldlit.cdh.ucla.edu/]. Применению тематического моделирования в Digital Humanities посвящена статья одного из авторов LDA Дэвида Блея [7]

В статье Topic Modeling and Figurative Language [8] автор сравнивает тематические модели коллекций поэтических текстов и научных статей. Полученная тематическая модель представляет темы, отражающие тип языка (разговорный, литературный) в поэзии. Как лингвист, автор определяет темы как направления дискурса.

0.3 Вывод

Задача анализа тематической структуры большой коллекции документов - корпуса поэзии - решается методами тематического моделирования. Наиболее подходящим считается метод LDA и его современные модификации.

Однако поэтические тексты изобилуют метафорическим языком. Как показывает исследование [8], анализ и интерпретация тем, полученных методом LDA на корпусе поэтических текстов, может отличаться от анализа тематических моделей коллекций документов со слабой художественной составлющей.

Исследование и решение задачи

Основной задачей, поставленной в данной работе, является исследование структуры большой коллекции текстовых документов stihi.ru методами тематического моделирования. Решение данной задачи можно разбить на следующие этапы:

1. Исследование лингвистических особенностей корпуса поэзии, существенных для построения тематической модели.
2. Построение тематической модели с учетом этих особенностей
3. Анализ и интерпретация кластеров документов, полученных тематической моделью.

0.4 Исследование особенностей корпуса

В изначальном виде корпус поэзии преставляет собой коллекцию из более чем 229 тыс. документов. Каждый документ соотвествует автору и содержит все его стихи.

```
<author="Имя Автора" nick="псевдоним">
<div poem id="XXXXXX" url="http://stihi.ru/...">

Стихотворение_1

</div>

...

<div poem id="XXXXXXX" url="http://stihi.ru/...">

Стихотворение_N

</div>
```

Рисунок 1: Разметка корпуса.

В корпусе присутствует разметка, определяющая границы произведений авторов, а так же различная мета-информация.

Алгоритмы тематического моделирования работают с документами по принципу "bag of words- [9]. В данной работе словом считается последовательность символов русского алфавита, с возможным произвольным числом дефисов (для сохранения таких слов, как рок-н-ролл). Так же была проделана первичная работа по нормализации корпуса:

- Удаление знаков препинания.
- Замена ё на е и заглавных букв строчными.
- Удаление мета-информации.

Среди методов нормализации нехудожественных текстов, выделим следующие:

- Удаление междометий и союзов
- Лемматизация
- Введение списков стоп-слов
- Частотная фильтрация, отсеивающая, например, слова с грамматическими ошибками

Из исследования [8] следует, что все эти методы нивелируют дискурсивные особенности художественных текстов. Результаты тематического моделирования на нормализованном корпусе поэзии представлены в следующей секции.

0.5 Тематическое моделирование

Для проведения экспериментов случайным образом было выбрано 12600 нормализованных файлов, затем выборка была разделена на три непересекающиеся части, каждая из которых содержала от 84 до 94 тыс. произведений. В качестве инструмента тематического моделирования был выбран стэнфордский тулкит [10], лемматизация выполнена с помощью программы MyStem [11].

Результаты, полученные на нормализованных выборках без лемматизации представлены в таблицах 1-3. Полученные тематические модели показывают наличие устойчивых тем на различных случайных выборках.

Результаты, полученные на нормализованных выборках с лемматизацией представлены в таблицах 4-6. Полученные тематические модели показывают наличие тех же устойчивых тем на различных случайных выборках. Отметим, что в случае с выборками

лемматизированных документов в топе распределений слов появились глаголы в начальной форме.

Таблица 1: Темы первой нелемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
сердце	любовь	любви	жизни	конкурс	солнце	помню
глаза	сердце	мир	мир	место	ветер	дом
боль	хочу	свет	бог	народ	осень	глаза
свет	любви	море	жить	кот	снег	хотел
мир	люблю	души	людей	дело	ночь	любил
ночь	знаю	душа	путь	дед	небо	знал
мысли	жизни	путь	бога	утра	дождь	друг
сквозь	жить	земли	души	идет	весна	ночь
руки	вновь	небо	мире	стоит	лето	мама
кровь	любить	жизни	друг	друзья	зима	видел
слова	глаза	сердце	любовь	рот	небе	стало
тело	счастье	земле	народ	вроде	солнца	вчера
небо	слова	волны	земле	новый	звезды	тихо
боли	прости	сердца	вновь	дома	свет	домой
души	душе	любовь	коль	мол	вновь	вместе
душу	знаешь	земля	смерть	нос	вечер	шел
смерть	моей	небес	дети	юрий	окном	одна
страх	друг	моря	стихи	денег	сердце	ушел

Таблица 2: Темы второй нелемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
хочу	народ	ветер	мама	жизни	любовь	ночь
знаю	жизни	осень	дом	мир	любви	глаза
сердце	людей	солнце	домой	путь	сердце	сердце
жить	дети	небо	дома	души	глаза	свет
любовь	россии	снег	утра	душа	люблю	сквозь
жизни	стихи	лето	утром	бог	вновь	кровь
люблю	поэт	дождь	ноги	свет	счастье	руки
любить	друзья	весна	спать	любовь	моей	боль
друг	детей	небе	кот	любви	свет	небо
прости	война	зима	друг	пути	душа	тело
знаешь	мир	город	мать	жить	ночь	мысли
хочется	друг	море	жена	бога	взгляд	мир
вновь	войны	ночь	дед	земле	сон	слова
слова	жить	звезды	дело	душе	душе	души
боль	коль	неба	старый	судьбы	счастья	душу
глаза	мать	солнца	нос	душу	глаз	слов
любви	честь	облака	новый	мире	руки	ночи
забыть	россия	птицы	руки	людей	сердца	тени

Таблица 3: Темы третей нелемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic6
б	ветер	жизни	любовь	бог	сквозь	стихи
дом	солнце	хочу	сердце	мир	свет	жизни
домой	осень	жить	люблю	жизни	мир	друг
дома	снег	сердце	любви	бога	небо	пишу
мама	небо	любовь	глаза	души	ночь	поэт
мать	весна	хочется	знаю	путь	души	слова
утра	дождь	любить	вновь	любовь	ночи	стих
боб	лето	счастье	слова	любви	мысли	вопрос
ноги	зима	любви	ночь	свет	море	дело
утром	небе	знаю	моей	душа	ветер	коль
кот	ночь	мир	друг	людей	глаза	ответ
глаза	солнца	мысли	прости	земле	сердце	слово
стало	вновь	понять	одна	смерть	тени	дела
стоит	птицы	душе	взгляд	господь	руки	людей
ночью	весны	боль	свет	сердце	город	новый
старый	листья	душу	хочу	богу	слов	дети
дело	окном	душа	душе	дух	звезды	друзья
дверь	свет	вновь	вместе	вновь	глаз	писать
хотел	город	забыть	ночи	пред	жизни	жить

Таблица 4: Темы первой лемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
небо	любить	бог	слово	давать	сердце	море
ветер	любовь	мир	стих	дело	душа	волна
солнце	знать	душа	мысль	жить	глаз	любовь
ночь	друг	земля	рука	друг	рука	цветок
дождь	хотеть	путь	окно	идти	любовь	нежный
снег	жить	свет	стена	ребенок	боль	прекрасный
свет	понимать	жизнь	лицо	пойти	ночь	глаз
окно	душа	давать	строка	б	слеза	вода
весна	жизнь	жить	писать	мама	сон	берег
звезда	забывать	человек	черный	дом	кровь	роза
сон	слово	небо	город	народ	уходить	красота
земля	счастие	век	старый	страна	мой	рука
осень	ждать	сила	новый	знать	слово	сад
зима	сердце	судьба	чай	сидеть	огонь	белый
белый	прощать	смерть	пустой	пить	тело	цвет
лето	один	идти	палец	хотеть	губа	душа
птица	верить	святой	строчка	взять	оставаться	песня
утро	уходить	дорога	голова	стоять	взгляд	петь
день	судьба	дух	утро	деньги	забывать	река
дом	сказать	война	точка	рука	вновь	взгляд

Таблица 5: Темы второй лемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
любовь	давать	знать	небо	весна	бог	глаз
любить	дом	слово	ветер	любовь	мир	душа
душа	рука	стих	земля	свет	душа	сердце
сердце	пить	писать	вода	душа	жизнь	рука
знать	пойти	друг	солнце	зима	земля	ночь
жить	б	понимать	дождь	снег	человек	сон
хотеть	нога	хотеть	море	солнце	давать	боль
уходить	мама	мысль	птица	лето	жить	свет
прощать	ребенок	думать	река	осень	век	тело
забывать	идти	смотреть	лес	сон	свет	слово
жизнь	дело	ждать	волна	цветок	сила	губа
друг	жена	один	трава	ночь	смерть	кровь
счастие	друг	хотеться	облако	глаз	путь	огонь
понимать	сидеть	жить	белый	нежный	война	взгляд
судьба	взять	становиться	дорога	тепло	слово	тень
давать	жить	сказать	город	сердце	судьба	небо
ждать	выходить	оставаться	звезда	счастие	поэт	слеза
боль	голова	видеть	поле	белый	народ	тишина
один	деньги	ночь	ночь	милый	святой	мой
верить	стол	любить	лететь	песня	страна	мысль

Таблица 6: Темы второй лемматизированной подвыборки

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
бог	знать	слово	весна	вода	любовь	пить
мир	друг	стих	снег	небо	душа	давать
давать	любить	любить	ветер	море	сердце	сидеть
душа	жить	любовь	небо	черный	свет	б
человек	уходить	писать	дождь	ветер	глаз	пойти
жить	хотеть	душа	солнце	волна	небо	дело
земля	ждать	рука	окно	город	мир	взять
жизнь	забывать	мысль	осень	стена	ночь	рука
сила	понимать	сердце	зима	рука	сон	дом
страна	жизнь	глаз	ночь	земля	боль	утро
война	прощать	чувство	лето	река	рука	мама
народ	оставаться	поэт	белый	сквозь	мечта	нога
святой	судьба	женщина	лист	камень	звезда	идти
век	идти	взгляд	утро	берег	мой	спать
дух	один	знать	сон	кровь	счастие	голова
грех	верить	мой	свет	тень	огонь	работа
путь	день	губа	земля	белый	вновь	жить
дело	проходить	строка	лес	дорога	слеза	глаз
россия	путь	хотеть	цветок	гора	жить	друг
божий	находить	друг	птица	лицо	любить	хотеть

0.6 Анализ и интерпретация кластеров полученных тематической моделью.

Модель LDA, выбранная для выполнения данного исследования, предполагает, что количество тем в анализируемой выборке подбирается эмпирически. Одним из известных недостатков классической модели LDA является то, что при большом количестве тем некоторые темы являются подтемами других [2].

Анализ документов, имеющих высокую вероятность принадлежности к теме $Topic_0$, первой лемматизированной подвыборки показал, что наиболее близкие к этой теме стихи (приблизительно первые 1500 стихов, вероятность $0.99 - 0.7$) действительно имееют своей основной темой описание природных явлений. Стихи с вероятностью, стремящейся к 0.6, активно используют слова из природной темы и не сильно отличаются от верхней трети. Стихи, вероятность принадлежности которых к природной теме $0.4 - 0.3$ испытывают сильное влияние других тем. Остальные темы имеют схожую структуру.

Таблица 7: Природа и природные явления

$Topic_0$
небо
ветер
солнце
ночь
дождь
снег
свет
окно
весна
звезда
сон
земля
осень
зима
белый
лето
птица
утро
день
дом

Таким образом, эксперименты показали, что корпус поэзии содержит несколько наиболее общих устойчивых тем. Причем, в отличие от исследования Лизы Роди [8], в данной работе тематическое моделирование позволило получить темы, которым можно дать классическую интерпретацию. Эти результаты были получены за счет применения методов нормализации, характерных для задач анализа нехудожественных текстов.

Отдельно отметим, что тематическая модель, полученная на описанных выше подмножествах корпуса поэзии мало отличается от моделей, обученных на выборках большего размера, по структуре тем.

Заключение

В рамках курсовой работы были решены следующие задачи:

1. Исследованы лингвистические особенности корпуса поэзии, существенные для построения тематической модели.
2. С учетом этих особенностей обучена тематическая модель.
3. Проведен анализ и дана интерпретация кластерам документов, полученных тематической моделью.

Список литературы

1. Национальный Корпус Русского Языка, http://www.ruscorpora.ru/.
2. *Коршунов Антон Гомзин Андрей.* Тематическое моделирование текстов на естественном языке // *Труды Института системного программирования РАН: журнал.* — 2012.
3. *S. Choi S. Cha C. C. Tappert.A.* A Survey of Binary Similarity and Distance Measures // *Journal of Systemics, Cybernetics and Informatics.* — 2010. — 7. — Vol. 8, no. 1. — Pp. 43–48.
4. *Dumais S. T. Furnas G. W. Landauer T. K., Deerwester S.* Using latent semantic analysis to improve information retrieval // *In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285.* — 1988.
5. *Hoffman Thomas.* Probabilistic Latent Semantic Indexing // *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.* — 1999.
6. *D. Blei A. Ng, Jordan M.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — January 2003. — no. 3.
7. *Blei. D.* Topic modeling and digital humanities. // *Journal of Digital Humanities, Vol. 2, No. 1.* — Winter 2012.
8. *Rhody Lisa.* Topic modeling and figurative language. // *Journal of Digital Humanities, Vol. 2, No. 1.* — Winter 2012.
9. *Jurafsky D., Martin J. H.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. — Upper Saddle River, NJ: Prentice-Hall,, 2000.
10. Stanford tmt, http://nlp.stanford.edu/software/tmt/tmt-0.4/.
11. MyStem, https://tech.yandex.ru/mystem/.