Текст слайдов

5 мая 2019 г.

1 Титульный слайд

Добрый день. Моя работа о том, как можно использовать статистику совместной встречаемости слов в некотором корпусе текстов для улучшения качества тематических моделей.

2 Тематическое моделирование, модель PLSA

Вначале о задаче тематического моделирования: у нас есть корпус текстов, в каждом мы можем рассчитать частотную оценку условной вероятности слова при условии документа. Эту матрицу частотных оценок можно приблизить моделью методом максимального правдоподобия. В модели введём промежуточную размерность «тема», и цель моделирования — восстановить вероятностные распределения слов над темами и тем над документами.

3 Аддитивная Регуляризация Тематических Моделей

Задача имеет неединственное решение, и к ней можно применять регуляризацию. В регуляризаторах можно формализовать некоторые желаемые свойства тем, и итоговый функционал — оптимизировать EM-алгоритмом.

4 Меры качества тематических моделей

Меры качества — это, стандартно, перплексия, и также есть такая мера качества — когерентность. Она показывает, насколько слова в тема согласована. Берётся m наиболее вероятных слов каждой темы и между парами различных слов считается мера семантической близости, затем значения усредняются по парам и по темам. В качестве функции близости пар слов можно брать PMI — поточечную взаимную информацию.

5 Обоснование PMI

В статье Ньюмэнна было показано, что именно так вычисленная когерентность хорошо коррелирует с человеческими оценками интерпретируемости тем. То есть глобальная цель — получать интерпретируемые темы.

6 Обучение PLSA, рост когерентности

Во время обучения перплексия падает — мы её явно минимизируем, а когерентность растёт, но это уже побочный эффект, хотелось бы иметь возможность её максимизировать напрямую. И у нас есть инструмент аддитивной регуляризации, так что можно ввести регуляризатор когерентности и получать более высокие значения когерентности.

7 Цели и задачи

Какие цели ставились в данной работе. Во-первых, PMI — частотная оценка, её стоит рассчитывать по большим корпусам, таких как англоязычная Википедия. Но вот проблема — она не влазит в оперативную память. Надо придумать и реализовать эффективный алгоритмы подсчёта статистики со-встречаемостей пар слов по большим коллекциям.

Далее, задать вероятность совместно встретить пару слов можно разными способами. Хочется проверить, отличается ли качество при разных способах задания вероятностей.

Также есть разные регуляризаторы когерентности, их тоже хочется сравнить.

И в итоге хотелось бы найти метод улучшения когерентности, который не сильно ухудшает перплексию.

8 Со-встречаемость пар слов

Один из способов, как можно определить вероятность совместно встретить пару слов u и v: можно пройти по коллекции скользящим окном фиксированной ширины и рассчитать счётчики n_{uv} — сколько раз в коллекции слова u и v встретились на таких позициях i, j, что расстояние между i и j не больше заданной константы — ширины окна.

9 Документная со-встречаемость

Другой способ задать вероятность совместно встретить слово u и слово v в коллекции — посчитать долю документов, в которых слова u и v встречались хотя бы раз в некотором окне фиксированной ширины.

10 Регуляризатор когерентности

Перейдём к регуляризаторам. Один был предложен в 2015 году, он минимизирует KL-дивергенцию между столбцами матрицы Ф и нашими оценками вероятностей слов в темах, в которых фигурируют наши оценки вероятностей одного слова при условии другого слова, которые могут быть рассчитаны алгоритмом.

11 Алгоритм сбора статистики со-встречаемостей

Итак, у нас была проблема в том, что коллекция целиком не влезает в оперативную память. Давайте разрежем коллекцию на небольшие наборы документов — батчи, каждый батч параллельно считаем в память, соберём статистику со-встречаемостей на каждом батче, сохраним её во внешнюю память в отсортированном формате. После того, как мы пройдём по всей коллекции, можно будет открыть все получившиеся файлы и отсортировать во внешней памяти. Алгоритм эффективный, с константной сложность по оперативной памяти.

12 Пример работы

На полном тексте англоязычной википедии, которая весит 20 Гб, работает 4 часа.

13 KL регуляризатор, разные типы когерентности

Здесь указаны графики когерентности в результате обучения, по оси x коэффициент регуляризации, по y — когерентность. Верхние 2 графика относятся к регуляризации когерентности, посчитанной на частотах пар слов, нижние 2 — по документным частотам. На левых двух графиках отложены значения когерентности, посчитанной по частотам пар слов, на правых двух — по частотам документов. Видно, что кривые значение скоррелированы и максимизация одной величины неизбежно ведёт к увеличению другой. Однако есть преимущество у документной PMI — если её регуляризовывать, получаются более высокие значения когерентности, также она считается проще, чем PMI на частотах пар слов.

14 Эксперимент с регуляризацией KL

Высокие значения когерентности можно получить ценой большой перплексии, но если не хочется сильно увеличивать перплексию, можно найти небольшой отрезок значений τ, при которых когерентность может возрасти в несколько раз, а перплексия — подняться не выше, чем на 10%. Этого удалось достичь за счёт экспоненциального убывания коэффициента регуляризации и выбора большого числа тем.

15 Результаты, выносимые на защиту

Результаты, выносимые на защиту: придуман и реализован эффективный алгоритм подсчёта статистики со-встречаемостей пар слов в больших текстовых коллекциях; предложен метод улучшения когерентности без большого ущерба перплексии и найдены условия, при которых он работает.