ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА. ЗАДАЧА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Козак А. В.

Теслюк В.Н. - канд. физ.-мат. наук, доцент

Обработка естественного языка (NaturalLanguageProcessing, NLP) – направление такой области прикладной информатики, как искусственный интеллект, и математической лингвистики, изучающее проблемы компьютерного анализа и синтеза естественных языков. На сегодняшний день именно это направление, наряду с компьютерным зрением, набирает все больший интерес у специалистов в области анализа данных (DataScientists). Основная цель NLP – научить компьютер (ВС) понимать “человеческий” язык. NLP решает множество прикладных задач: машинный перевод, стемминг и другие. Мы же остановимся на задаче тематического моделирования.

Тематическое моделирование – одно из современных приложений в машинном обучении к анализу текста. Тематическое моделирование – выявление тематической модели. Тематическая модель корпуса (коллекции) текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. С математической точки зрения, каждую тему представляют как случайную величину. Таким образом, такую модель можно называть вероятностной.

Вероятностная тематическая модель описывает каждую тему дискретным распределением на множестве терминов, а каждый документ дискретным распределением на множестве тем.

Тематическое моделирование имеет ряд приложений: категоризация документов, информационный поиск, выявление трендов в потоках документов и другие. Мы же углубимся в задачу кластеризации текстовых документов.

Задачу кластеризации в машинном обучении можно решить различными методами. Основной и часто используемый – это метод k-means (или метод k средних).

Для задачи тематического моделирования совсем недавно был придуман метод LDA (или латентное размещение Дирихле). Подробнее поговорим про латентное размещение Дирихле и сравним его в задаче разделения документов на темы с методом k средних.

Латентное размещение Дирихле – это модель, которая позволяет нам понять и объяснить наблюдаемые результаты с помощью групп, а также ответить на вопрос: почему в данных существуют схожие части?

В LDA каждый документ может рассматриваться как смесь из различных тем, причем каждый документ разлагается на эти темы только посредством применения LDA. Таким образом, можно сказать (а даже и необходимо), что набор тем – скрытая переменная модели LDA. Отсюда и слово “латентный” в названии метода. Не трудно понять, что латентное размещение Дирихле очень похоже на вероятностный латентный семантический анализ (PLSA). Однако есть одно отличие между этими двумя моделями: в LDA предполагают, что темы имеют распределение Дирихле.

Пусть D – корпус документов, W – множество терминов на всем корпусе документов, а документ d – последовательность терминов. Предположим, что существует конечное множество тем T.

Тогда построить тематическую модель документов D – найти множество тем T, распределения $p(w|t), t \in T$ и распределения $p(t|d), d \in D$.

Примем гипотезу, что появление слов в документе, относящейся к определенной теме, не зависит от самого документа.

Тогда, основываясь на предположение выше, определение условной вероятности и формулу Байеса, получим:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

$$p(w,d) = \sum_{t \in T} p(d)p(w|t)p(t|d).$$

Тогда при дополнительных ограничениях, связанных с латентным размещением Дирихле:

1. векторы документов $\theta_d = (p(t|d), t \in T)$ порождаются одним распределением, а именно распределением Дирихле с параметром α;
2. векторы тем $\varphi_t = (p(w|t), w \in W)$ порождаются одним распределением, а именно распределением Дирихле с параметром β.

Мы можем найти неизвестные параметры LDA с помощью сэмплирования Гиббса, вариационного байесовского вывода или EM-алгоритма по известному корпусу документов.

Проведя исследование в области сравнения работы двух алгоритмов: LDA и k-means. Следует отметить следующее:

1. метод k-means не дает никакого представления о распределениях $p(w|t), t \in T$ и $p(t|d), d \in D$;
2. LDA работает гораздо медленнее метода k-means;
3. LDA лучше понимает природу данных (генерации корпуса документов);

Библиотека БГУИР

4. LDA позволяет определить для документа не одну тему, а несколько, чего лишен метод k-means.

Таким образом, можно сделать вывод, что для задачи вероятностного тематического моделирования лучше использовать вероятностные алгоритмы, а именно для данной задачи имеет место методы LDA (или же его собрат PLSA). При этом в результате работы, обучения на корпусе документов, полученные распределения (скрытые переменные) можно использовать для эффектной визуализации данных: построения облака слов для каждой темы.

Список использованных источников:

1. Луис Педро Коэльо, Вилли Ричард - Построение систем машинного обучения на языке Python,
2. Christopher D. Manning, Hinrich Schütze - Foundations of Statistical Natural Language Processing,
3. Diane J. Hu - Latent Dirichlet Allocation for Text, Images, and Music.

ДЕМОНСТРАЦИЯ ПРИНЦИПА РАБОТЫ ШИФРОВАЛЬНОЙ МАШИНЫ «ЭНИГМА»

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Коршунов А.А.

Стройникова Е.Д. – ассистент кафедры информатики

Для демонстрации принципа работы шифровальной машины «Энигма» разработана учебная программа с использованием языка программирования C# на основе алгоритма шифрования.

Со времён изобретения письменности людьми движет желание скрыть написанное от посторонних глаз. Было изобретено много простых способов защитить текст, адресованный конкретному человеку, таких, как стеганография и простейшие шифры перестановки («атбаш»).

Как показала практика, стеганографию не всегда возможно применить, а простейшие шифры очень быстро поддаются криптоанализу. Вследствие этого криптографы стали придумывать различные занимательные способы запутать своих «противников». Появились такие шифры, как аффинный, полиалфавитный (простейший – диск Альберти). Данные способы сокрытия информации были намного более эффективными, т.к. давали большее количество комбинаций, а соответственно усложняли расшифровку. В механизме шифровальной машины «Энигма» используется некое подобие диска Альберти, объединённого с шифратором Джефферсона.

В стандартной механической версии данной машины использовано 3 ротора и 1 рефлектор. На каждом из роторов имеется 2 алфавита: «принимающий» букву и «отдающий» букву. Соответственно на принимающем роторе имеем обычный алфавит, на отдающем – шифроалфавит. Буква, попадающая в данный механизм, проходит каждый из роторов в одном направлении, доходит до рефлектора и возвращается обратно, после чего происходит поворот заданного (не обязательно первого) ротора на 1 позицию. Таким образом, с помощью несложных преобразований, ускоренных электрической схемой, получаем несколько раз зашифрованную букву. Следует заметить, что после каждого поворота ротора шифроалфавит будет изменён. Следовательно, можно не бояться, что одна и та же буква открытого текста будет зашифрована одинаково более 2-х раз подряд. Также данная машина даёт возможность изменить входящие в неё буквы ещё до преобразования. Это как бы добавляет ещё один ротор, но уже статический. Конечно же, стоит отметить, что все роторы «Энигмы» можно ставить на любую позицию, что позволяет варьировать начало шифрования. На рисунке 1 приведена иллюстрация зашифровывания буквы в более ранних версиях «Энигмы»:

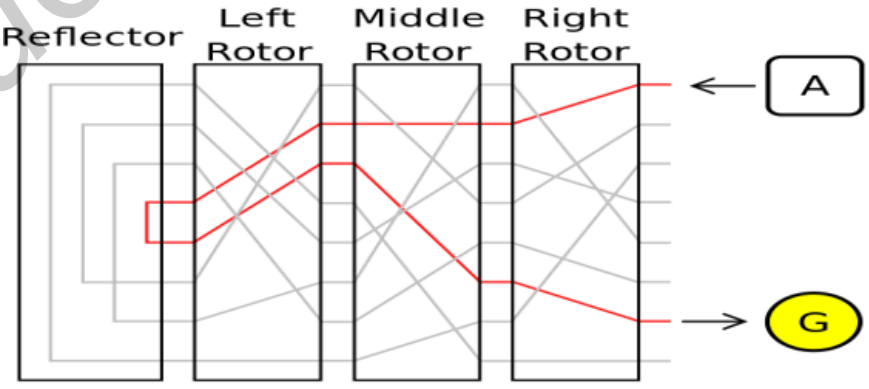

Рис. 1 – Процесс шифрования

Библиотека БГУИР