

МАРКОВСКИЕ ЦЕПИ И СТАТИСТИЧЕСКИЕ ЯЗЫКОВЫЕ МОДЕЛИ

Катя Герасименко

10.01.2019

По материалам Маши Шеяновой

Задача: рекомендательная система по новостям — по новости порекомендовать другие новости из нашего корпуса

Корпус: `lenta_articles_full.zip`

Как можно делать:

- поиск по словам (обратный индекс, BM-25)
- сравнение векторов (word2vec, тематическое моделирование)
- дополнительная фильтрация по темам (классифицировать по теме, предлагать статьи только из самых вероятных тем)
- можно попробовать машинное обучение с учителем, потому что у нас есть данные о том, какие статьи похожи (на одну тему), а какие - нет.
- любые другие ваши идеи

Проект — это один или несколько файлов с кодом (относящимся только к проекту) и все необходимые сохраненные файлы с данными.

Оцениваться будут:

- не падает ли программа
- корректность поиска релевантных новостей
- подход
- чистота и красота кода

На зачете можно будет делать проект и мы будем помогать.

Дедлайн сдачи — **12 января в 13.00**, но чем раньше — тем лучше.

INTRO

ЧТО МЫ НАЗЫВАЕМ СТАТИСТИЧЕСКОЙ ЯЗЫКОВОЙ МОДЕЛЬЮ?

Определение из Википедии:

A **statistical language model** is a probability distribution over sequences of words.

По-русски:

Статистическая языковая модель — это распределение вероятностей по последовательностям слов.

ЧТО УМЕЕТ ЯЗЫКОВАЯ МОДЕЛЬ?

Какое слово в последовательности вероятнее:

Поезд прибыл на

- вокзал
- север

Какая последовательность вероятнее:

- Вокзал прибыл поезд на
- Поезд прибыл на вокзал

ЗАЧЕМ ЭТО МОЖЕТ БЫТЬ ПОЛЕЗНО?

Языковые модели принождаются в огромном количестве задач:

- спеллчекинг
- автодополнение
- распознавание речи
- распознавание символов (Optical Character Recognition, OCR)
- машинный перевод
- реферирование текста
- порождение текста

С распознавания речи всё началось.

Что нужно для распознавания речи?

- акустическая модель (представление о фонетике языка)
- лексическая модель (о том, какими могут быть слова)
- языковая модель (вероятности последовательностей слов)

Яндекс.Рефераты (<https://yandex.ru/referats/>):

Точка перегиба оправдывает экзистенциальный принцип восприятия, открывая новые горизонты.

Плазменное образование восстанавливает элементарный платежный документ, даже с учетом публичного характера данных правоотношений.

Закон, основываясь на парадоксальном совмещении исключающих друг друга принципов характерности и поэтичности, предоставляет абстрактный голос персонажа.

ПОРОЖДЕНИЕ ТЕКСТА. ВО-ПЕРВЫХ, ЭТО ВЕСЕЛО!

Ветхий Алгоритм (https://twitter.com/alg_testament):



Ветхий Алгоритм @alg_testament · Jul 24

эта процедура должна работать не так, отец мой, отец мой, не бесчести имени Бога твоего.

Translate Tweet



1



33



94



Ветхий Алгоритм @alg_testament · Jul 22

Сигналы, генерируемые терминалом, возникают, когда пользователь входит в дом Господень

Translate Tweet



39



103



Ветхий Алгоритм @alg_testament · Jul 22

И поставлю Себе священника верного; он будет обладать правами суперпользователя.

Translate Tweet



1



82



217



ПОРОЖДЕНИЕ ТЕКСТА. ЧТО ЕЩЁ?

Чатботы!

Например, когда Алиса понимает, что пользователь хочет поговорить, она включает болталку — порождение текста.

Это может быть не только развлечение, но и психологическая помощь человеку.

Чтобы перевести последовательность слов правильно, нам мало знать самый вероятный перевод.

Важно знать, насколько вообще вероятна подобранная последовательность **в целевом языке**.

НЕМНОГО ТЕОРИИ ВЕРОЯТНОСТЕЙ

Вероятностью события A называют отношение числа m **благоприятствующих** этому событию исходов к общему числу n **всех элементарных исходов** (равновозможных несовместных).

$$P(A) = \frac{m}{n}$$

Пример: бросок монетки. O — выпал орёл, P — выпала решка.
Элементарные исходы: $\{O, P\}$.

Равновозможность — монетка честная, $P(O) = P(P)$.

$$P(O) = \frac{1}{2}$$

Пусть есть некоторый корпус (26 слов):

I wanna sleep. I wanna play with you. You wanna skate with me. You want to sleep in your bed. You want to play with me.

Событие X — мы случайным образом выбрали из текста слово **wanna**. Оцените $P(X)$?

(26 слов)

I *wanna* sleep. I *wanna* play with you. You *wanna* skate with me. You want to sleep in your bed. You want to play with me.

Событие X — мы случайным образом выбрали из текста слово *wanna*. Оцените $P(X)$?

$$P(X) = \frac{3}{26}$$

Событие Y — мы случайным образом выбрали из текста слово *sleep*. Оцените $P(Y)$?

Вероятность наступления события A , при условии наступления события B , называется условной вероятностью A (при данном условии) и обозначается $P(A|B)$.

Пусть AB — события A и B , произошедшие одновременно, или одно за другим. Тогда $P(A|B)$ будет равна $P(A|B) = n_{AB}/n_B$.

Подставляем числитель и знаменатель в формулу вероятности:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Теорема умножения вероятностей:

$$P(AB) = P(A|B) \times P(B)$$

I **wanna** *sleep*. I **wanna** play with you. You **wanna** skate with me. You want to *sleep* in your bed. You want to play with me.

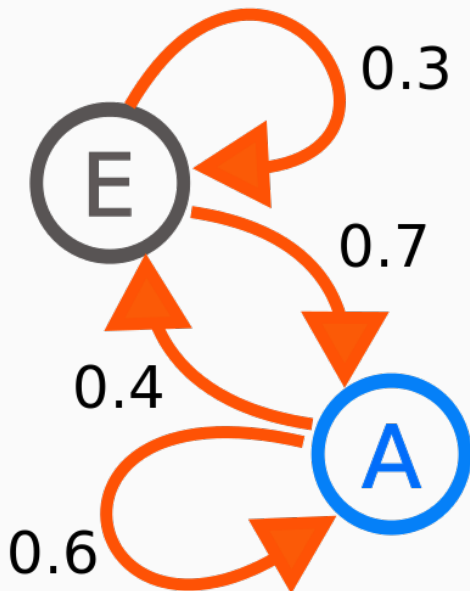
Дано: мы уже сказали слово "wanna" (событие X произошло).
Теперь мы хотим случайным образом выбрать слово после wanna.

Событие Y — мы случайным образом выбрали из текста слово *sleep*. Оцените $P(Y|X)$, то есть вероятность, что следующим после wanna словом мы выбрали из текста слово *sleep*.

МАРКОВСКИЕ ЦЕПИ

Последовательность случайных событий с конечным / счётным числом исходов, такая, что при фиксированном настоящем будущее независимо от прошлого. Иными словами — следующее событие зависит только от настоящего.

Можно смотреть на марковскую цепь как на частный случай **взвешенного конечного автомата**.



ЧТО УМЕЕТ ЯЗЫКОВАЯ МОДЕЛЬ — ФОРМАЛЬНО

- Оценивать вероятность той или иной последовательности слов в языке

$$P(W) = P(w_1, \dots, w_n)$$

(дискриминативная модель)

- Предсказывать наиболее вероятное следующее слово при условии уже известного ряда слов

$$P(w_n | w_1, \dots, w_{n-1})$$

(генеративная модель)

Статистической языковой моделью называется такая модель, которая умеет делать хотя бы один из пунктов.

Пусть $w_{1:n} = w_1, \dots, w_m$ – последовательность слов.

Точная оценка вероятности этой последовательности — **цепное правило**:

$$P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)\dots P(X_n|X_1, \dots, X_{n-1}))$$

Вероятность следующего слова:

$$P(X_n|X_1, \dots, X_{n-1})) = \frac{P(X_1, \dots, X_n)}{P(X_1, \dots, X_{n-1}))}$$

Но оценить $P(w_k|w_{1:k-1})$ не легче!

Мы пользуемся **марковским предположением**:

текущее состояние зависит лишь от конечного числа
предыдущих состояний

Иными словами:

$$P(w_i | w_1 \dots w_{i-1}) \approx P(w_i | w_{i-n} + 1 \dots w_{i-1})$$

Переходим к n-граммам: $P(w_{i+1}|w_{1:i}) \approx P(w_{i+1}|w_{i-n:i})$, то есть, учитываем $n - 1$ предыдущее слово. Т.е. используем Марковские допущения о длине запоминаемой цепочки.

Модель

- униграмм: $P(w_k)$
- биграмм: $P(w_k|w_{k-1})$
- триграмм: $P(w_k|w_{k-1}w_{k-2})$

- Незнакомые слова
- Нули в матрице переходов -> когда считаем вероятность последовательности, ноль в произведении обнуляет всю последовательность
- Концептуальная проблема: помнит только несколько предыдущих слов

Решения:

- От незнакомых слов:
 - иметь фиксированный словарь и в корпусе заменять на '`<UNK>`' все, что не в этом словаре
 - заменять в корпусе редкие слова на '`<UNK>`'
- От нулей в переходах - smoothing
 - Discounting - давать нулям ненулевые вероятности, «отнимая» их от вероятностей существующих переходов
 - Backoff, interpolation - использовать данные о вероятностях меньшего контекста (триграммы -> биграммы -> униграммы)

Нейронные модели показывают лучшие результаты, потому что лучше обобщают поведение похожих слов и запоминают всю историю.

Бывают пословные и посимвольные.

Пословные используют векторные представления слов (эмбединги), посимвольные — one-hot encoding. Ключевые слова:

- RNN — рекуррентные нейронные сети
- LSTM — long short-term memory — подвид RNN с долговременной памятью

СПАСИБО ЗА ВНИМАНИЕ!
Вопросы?