Языковые модели

Квантитативный анализ текста

Кирилл Александрович Маслинский

16.05.2022 / 09

ЕУСПб

Контекст

Контекст

Модель контекста: N-граммы

N-граммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

N-граммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

N-граммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

N-граммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

N-граммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

Вероятность языковых событий

Предсказание слова

искусственный ...?

Предсказание слова

искусственный интеллект?

а в корпусе русской детской литературы XX—XXI вв.?

Вероятность языковых событий

- Классическая вероятность: монетка или кость
- Вероятность, основанная на частотности (урна с шарами, мешок слов)
- В лингвистике считаем события в корпусе
- вероятность = относительная частотность

Пример расчета вероятности слова

Всего слов в корпусе = 46,804,371

искусственный = 121

$$P(\text{искусственный}) = \frac{121}{46804371} \approx 0.000002585 = 2.585 \text{ IPM}$$

Вероятность языковых событий

- Классическая вероятность: монетка или кость
- Вероятность, основанная на частотности (урна с шарами, мешок слов)
- В лингвистике считаем события в корпусе
- вероятность = относительная частотность

Пример расчета вероятности слова

Всего слов в корпусе = 46,804,371

искусственный = 121

$$P(\text{искусственный}) = \frac{121}{46804371} \approx 0.000002585 = 2.585 \text{ IPM}$$

Вероятность пары слов

- $P(\text{искусственный}) = \frac{121}{46804371}$
- $P(\text{интеллект}) = \frac{179}{46804371}$
- $P(\text{искусственный интеллект}) = ?$

Совместная вероятность

$$P(A \wedge B) = P(A) \cdot P(B) \quad | \quad \text{IFF A и B независимы} \qquad (1)$$

Совместная вероятность

$$P(A \wedge B) = P(A) \cdot P(B) \quad | \quad \text{IFF A и B независимы} \qquad (1)$$

$$P(\text{искусственный интеллект}) = \frac{121}{46804371} \cdot \frac{179}{46804371} =$$

$$= 0.00000000000989 = 0.00000989 \text{ IPM}$$

Условная вероятность

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \tag{2}$$

$$P(\text{интеллект}|\text{искусственный}) = \frac{P(\text{искусственный интеллект})}{P(\text{искусственный})} =$$

$$= \frac{\frac{4}{46804371}}{\frac{121}{46804371}} = \frac{4}{121} = 0.033$$

Предсказание слова в контексте

сильный искусственный ...?

Предсказание слова в контексте

сильный искусственный интеллект!

Предсказание слова в контексте

сильный искусственный

$$P(\text{интеллект}|\text{сильный искусственный}) =$$

$$\frac{f(\text{сильный искусственный интеллект})}{f(\text{сильный искусственный})} = \frac{?}{?} = ?$$

Вероятность языковых событий

Коллокации

Модель для коллокаций

Слова, встречающиеся в текстах рядом чаще, чем можно ожидать в результате случайности.

	интеллект	¬ интеллект
искусственный		
¬ искусственный		

Вероятность языковых событий

Языковая модель

Вероятность предложения

- P(Создан сильный искусственный интеллект.) = ?
- P(Создан сильный искусственный крокодил.) = ?

Language model

Языковая модель — приписывает вероятность фрагменту текста (высказыванию, предложению...)

В хорошей модели вероятности языковых фрагментов соответствуют их относительной частотности в текстах.

Иными словами:

- максимизирует вероятность реальных текстов
- минимизирует вероятность нереальных текстов

Униграммная языковая модель

Создан^{31} сильный^{1905} искусственный^{121} интеллект^{71}!

Создан^{31} сильный^{1905} искусственный^{121} крокодил^{442}!

Униграммная языковая модель

Создан[31] сильный[1905] искусственный[121] интеллект[71]!

Создан[31] сильный[1905] искусственный[121] крокодил[442]!

$$\frac{31}{46804371} \cdot \frac{1905}{46804371} \cdot \frac{121}{46804371} \cdot \frac{71}{46804371} = 1.06 \times 10^{-22}$$

$$\frac{31}{46804371} \cdot \frac{1905}{46804371} \cdot \frac{121}{46804371} \cdot \frac{442}{46804371} = 6.58 \times 10^{-22}$$

Крокодил в шесть раз вероятнее!

Chain rule of probability

$$P(A, B) = P(B|A)P(A) \tag{3}$$

$$P(W_1, W_2, W_3, W_4) =$$

$$P(W_4|W_1, W_2, W_3) \cdot P(W_3|W_1, W_2) \cdot P(W_2|W_1) \cdot P(W_1)$$

P(Создан сильный искусственный крокодил) = P(Создан)
P(сильный|Создан) P(искусственный|Создан сильный)
P(крокодил|Создан сильный искусственный)

Цепь Маркова

- система с конечным числом состояний
- следующее состояние зависит только от текущего

Применительно к тексту:

Следующее слово зависит только от предыдущего (N предыдущих)

Биграммная языковая модель

Markov assumption:

$$P(\text{Создан сильный искусственный интеллект}) \approx$$

$$P(\text{интеллект}|\text{искусственный}) \cdot P(\text{искусственный}|\text{сильный})$$

$$\cdot P(\text{сильный}|\text{создан}) =$$

$$\frac{2.21}{45.65} \cdot \frac{0.01}{189.7} \cdot \frac{0.003}{16.28} \approx 4.7 \times 10^{-10}$$

Биграммная языковая модель

Markov assumption:

$$P(\text{Создан сильный искусственный крокодил}) \approx$$

$$P(\text{крокодил}|\text{искусственный}) \cdot P(\text{искусственный}|\text{сильный})$$

$$\cdot P(\text{сильный}|\text{создан}) =$$

$$\frac{0}{45.65} \cdot \frac{0.01}{189.7} \cdot \frac{0.003}{16.28} \approx 0$$

Биграммная языковая модель

Markov assumption:

$$P(\text{Создан сильный искусственный крокодил}) \approx$$

$$P(\text{крокодил}|\text{искусственный}) \cdot P(\text{искусственный}|\text{сильный})$$

$$\cdot P(\text{сильный}|\text{создан}) =$$

$$\frac{0.001}{45.65} \cdot \frac{0.01}{189.7} \cdot \frac{0.003}{16.28} \approx 2.13 \times 10^{-13}$$

Проблема редких слов

Проблема недостаточных данных (sparse data)

- В результате обучения на конечном корпусе очень многие N-граммы получат нулевую вероятность
- Хотя в действительности должны иметь ненулевую (встречаются в текстах)

Если в предложении встречается слово, отсутствовавшее в корпусе, вероятность такого предложения в модели равна нулю!

Задача

Добиться, чтобы модель приписывала ненулевую вероятность любому тексту.

Сглаживание Лапласа

Добавить 1 ко всем частотам в корпусе

Для униграммной модели (сглаженная относительная частотность слова):

$$P_{Laplace} = \frac{f+1}{N+V} \tag{4}$$

Другие методы сглаживания

Идея

Оценить вероятность никогда не встречавшихся N-грамм на основании вероятности N-грамм, встречавшихся один раз

- Good-Turing discounting
- Kneser-Ney smoothing

Биграммная языковая модель

Markov assumption: P(Создан сильный искусственный крокодил)
$\approx$ P(крокодил|искусственный) P(искусственный|сильный)
P(сильный|создан)

$$\frac{0.001}{45.65} \cdot \frac{0.01}{189.7} \cdot \frac{0.003}{16.28} \approx 2.13 \times 10^{-13}$$

$$\frac{0.001}{45.2} \cdot \frac{0.01}{187.8} \cdot \frac{0.004}{16.12} \approx 2.92 \times 10^{-13}$$

Интеллект:

4.7×10^{-10}

6.4×10^{-10}

Триграммная языковая модель

$$P(\text{Создан сильный искусственный интеллект}) \approx$$

$$P(\text{интеллект}|\text{сильный искусственный}) \cdot$$

$$P(\text{искусственный}|\text{создан сильный})$$

Конечный автомат