Language Modeling

Обобщение и нули

Speech and Language Processing (3rd ed. raft), Dan Jurafsky and James H. Martin. Глава 4.3, стр. 9 - 12
Ведешкин Сергей

Метод визуализации Шеннона

- Выберем случайный биграмм (<s>, w) в зависимости от его вероятности
- Теперь выберите случайный биграмм (w, x) в соответствии с его вероятностью
- И так далее, пока мы не выберем </ s>
- Затем соедините слова вместе

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>
I want to eat Chinese food

Аппроксимация Шекспира

1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have –Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in; –It cannot be but so.

Шекспир как корпус

- N=884,647 ключей, V=29,066
- Шекспир выпустил 300 000 биграмм из V2 = 844 миллиона возможных биграмм.
  - Таким образом, 99,96% возможных биграмм никогда не появятся (имеют нулевые записи в таблице)
- Квадриграммы хуже: то, что генерируется, похоже на Шекспира.

The Wall Street Journal не Шекспир (без обид)

1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Опасности переобучения

- N-граммы работают только для предсказания слов, если тестовый корпус выглядит как учебный корпус
  - В реальной жизни это часто не так
  - Нам нужно обучать прочные модели, которые обобщают!
  - Один вид обобщения: нули!
    - Вещи, которые никогда не встречаются в учебном наборе, Но встречаются в тестовом наборе

Обнуление

- Обучающий набор:
- ... отклонил обвинения
- ... отклонил отчеты
- ... отклонил требования
- ... отклонил запрос

P(“offer” | denied the) = 0

- Набор тестов

... отклонил предложение

... отказал кредит

Нулевая вероятность биграммы

- Биграммы с нулевой вероятностью
  - означает, что мы назначим 0 вероятности тестовому множеству!
- И поэтому мы не можем вычислить perplexity (не можем делить на 0)!

- **Jurafsky D. & Martin J. H.** Speech and Language Processing [электронный ресурс] . — Ch. 4, pp. 1-2 . — 2017. ACM. URL: https://web.stanford.edu/~jurafsky/slp3/5.pdf

- **Shakespeare text statistics** [электронный ресурс] . – 2017.
  URL: https://www.opensourceshakespeare.org/stats/
- **How many words did Shakespeare know?** [электронный ресурс] . – 2017

URL: https://kottke.org/10/04/how-many-words-did-shakespeare-know