Лингвистическая модель

Автор
Ребрик Юрий

Руководитель
Игорь Куралёнок

весна 2017

Мотивация

Предсказывать запросы пользователей в поисковую систему

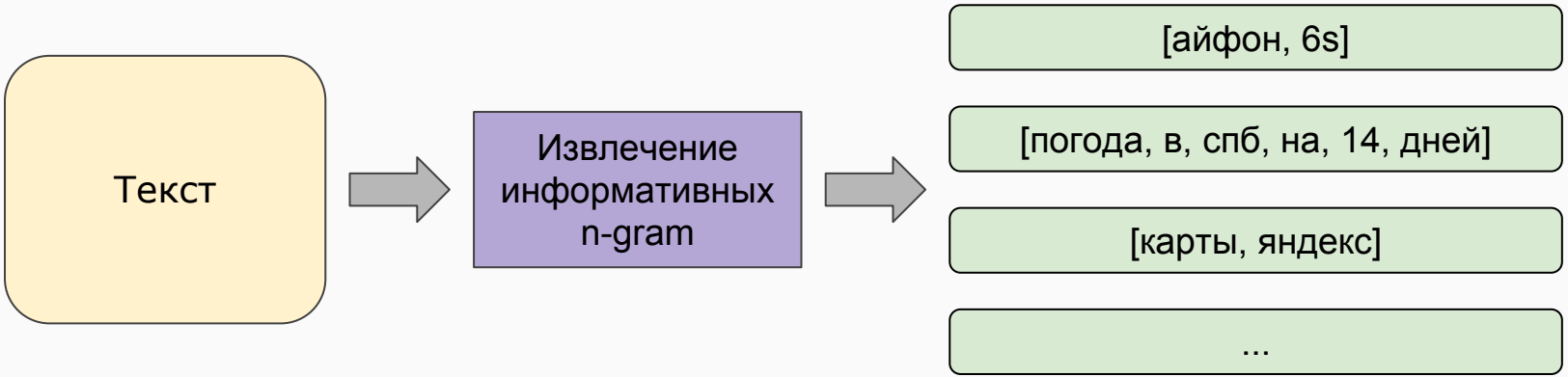
Цель

Придумать и реализовать лингвистическую модель для предсказания запросов пользователей

Задачи

1) Описание лингвистической модели
2) Реализация модели
3) Проведение бенчмарков
4) Выводы

N-grams


Текст
Извлечение информативных n-gram
[айфон, 6s]
[погода, в, спб, на, 14, дней]
[карты, яндекс]
...

Модель

$$P(q_2|q_1) = \sum_{\pi} P(q_2|q_1, \pi) P(\pi|q_1)$$

$$P(q_2|q_1, \pi) = \prod_{s_i \in q_1} P(\pi_i|s_i) = \prod_{s_i \in q_1} \prod_{t_j \in \pi_i} \frac{e^{\gamma_{ij}}}{\sum_k e^{\gamma_{ik}} + 1}$$

$$P(\pi|q_1) = \frac{\prod_{s_i \in q_1} Poiss_{mean_{s_i}}(|\pi_i|)}{\sum_{\pi'} \prod_{s_i \in q_1} Poiss_{mean_{s_i}}(|\pi'_i|)}$$

Урна Блэквела-МакКвина

$$t_{n+1}|t_1, t_2, \ldots, t_n \sim \begin{cases} new, & \text{with probability } \frac{\alpha}{\alpha+n} \\ \delta_c, & \text{with probability } \frac{\sum_{j=1}^{n}[t_j=c]}{\alpha+n} \end{cases}$$

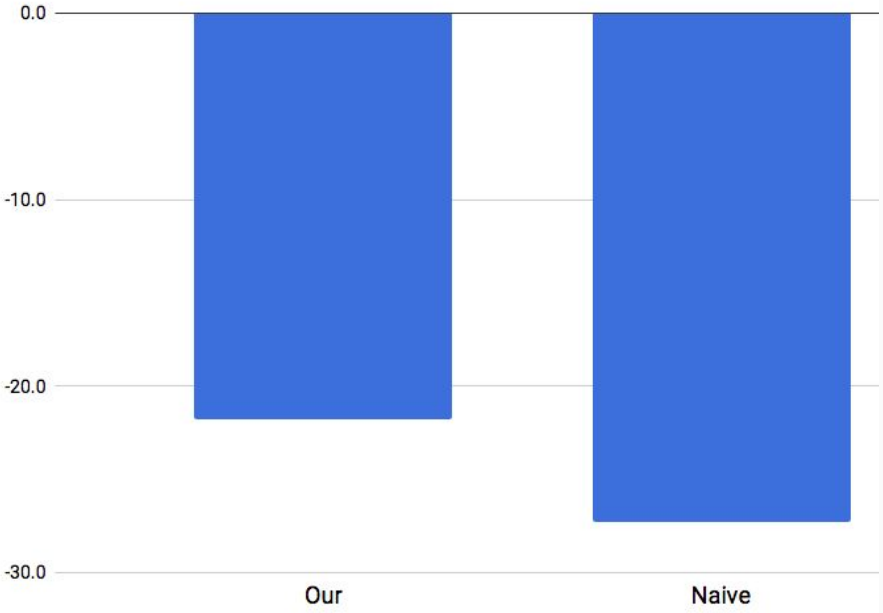
Наивная модель

$$P(q_2|q_1) = \max_{\pi} P(q_2|q_1, \pi) = \max_{\pi} \prod_{s_i \in q_1} P(\pi_i|s_i)$$

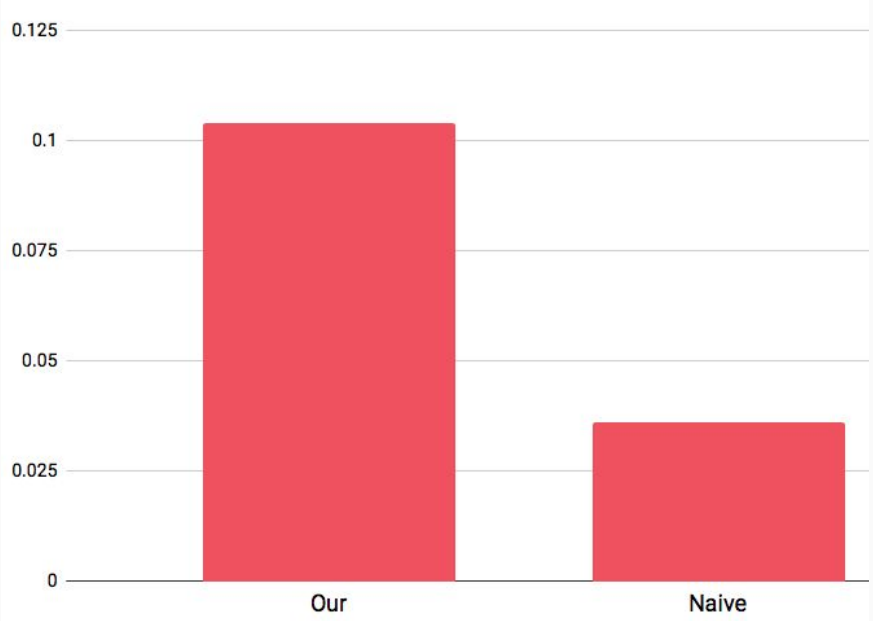
$$P(\pi_i|s_i) = \prod_{t \in \pi_i} \frac{count_{s_i t} + 1}{\sum_u (count_{s_i u} + 1)}$$

Бенчмарк

Log probability



std



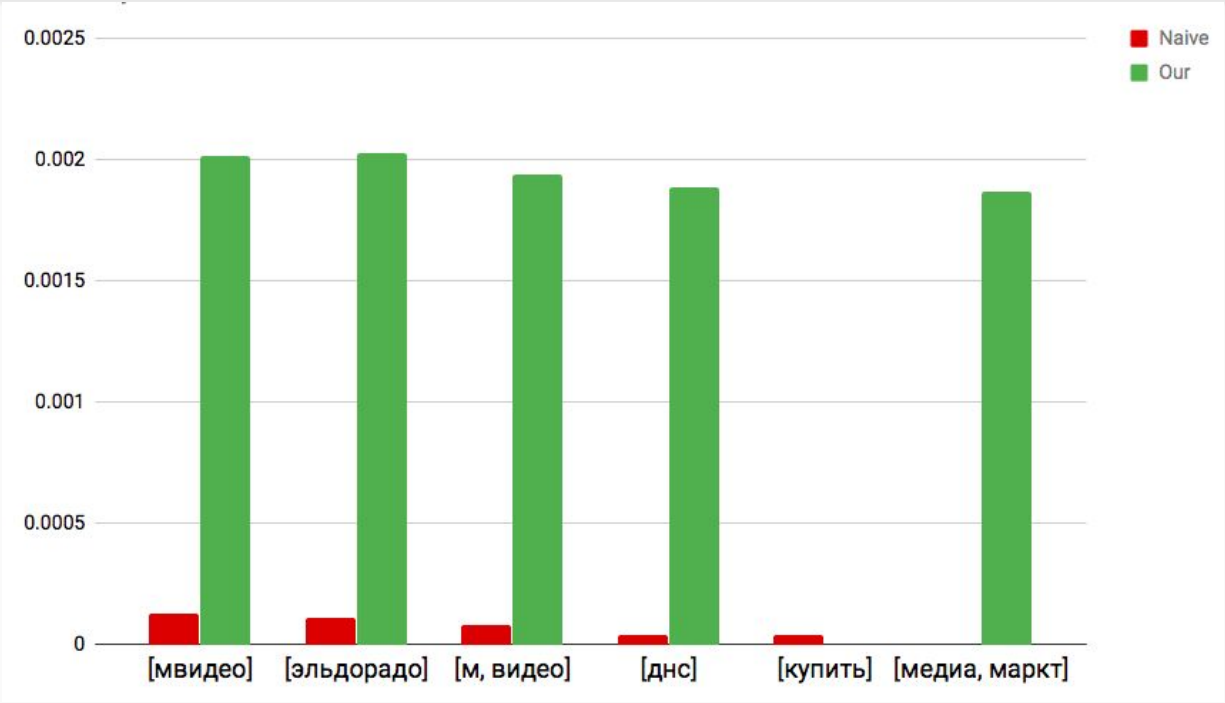
Выводы

- Сырые реальные данные - это боль!
- Тестирование на готовых датасетах

Перспективы

- сравнение модели с алгоритмом из статьи

Пример


0.0025
0.002
0.0015
0.001
0.0005
0
Naive
Our
[мвидео]
[эльдорадо]
[м, видео]
[днс]
[купить]
[медиа, маркт]

Project

https://github.com/spbsu-ml-community/jmll/tree/rebryk

Contacts

https://github.com/Rebryk