Модели языка

Простейшая модель

- **Мешок слов** – появление слова в предложении не зависит от других слов

$$W =< w_1 w_2 w_3 \dots w_n >$$

W = «Мама мыла раму»

- Каждое слово имеет вероятность появления, равную частоте слова во всех текстах

Модель языка

$P(W)$ – вероятность предложения W

$$P(W) = P(w_1 w_2 w_3 \dots w_n) =$$
$$= P(w_1) \cdot P(w_2) \cdot P(w_3) \cdot \dots \cdot P(w_n)$$

$P(w_i)$ – вероятность встретить слово w_i в предложении

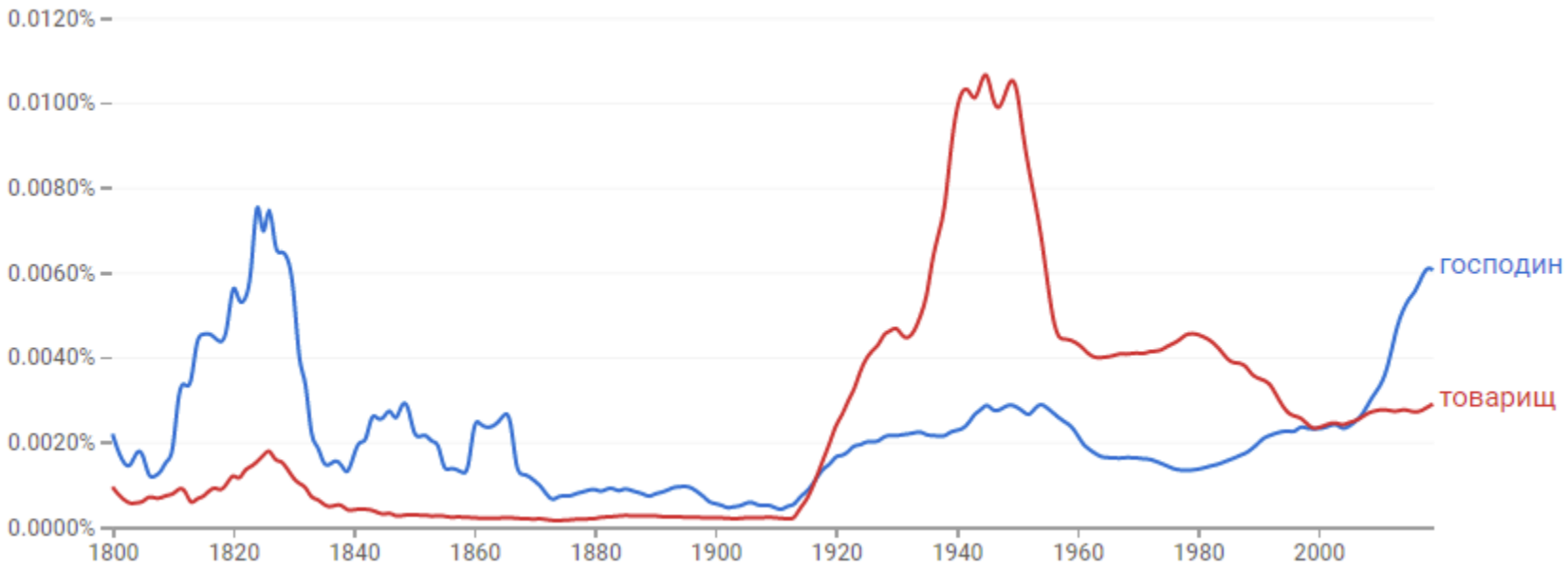
$$P(w_i) = \frac{c(w_i)}{c()}$$

$c(w_i)$ – количество предложений, в которых есть слово w_i

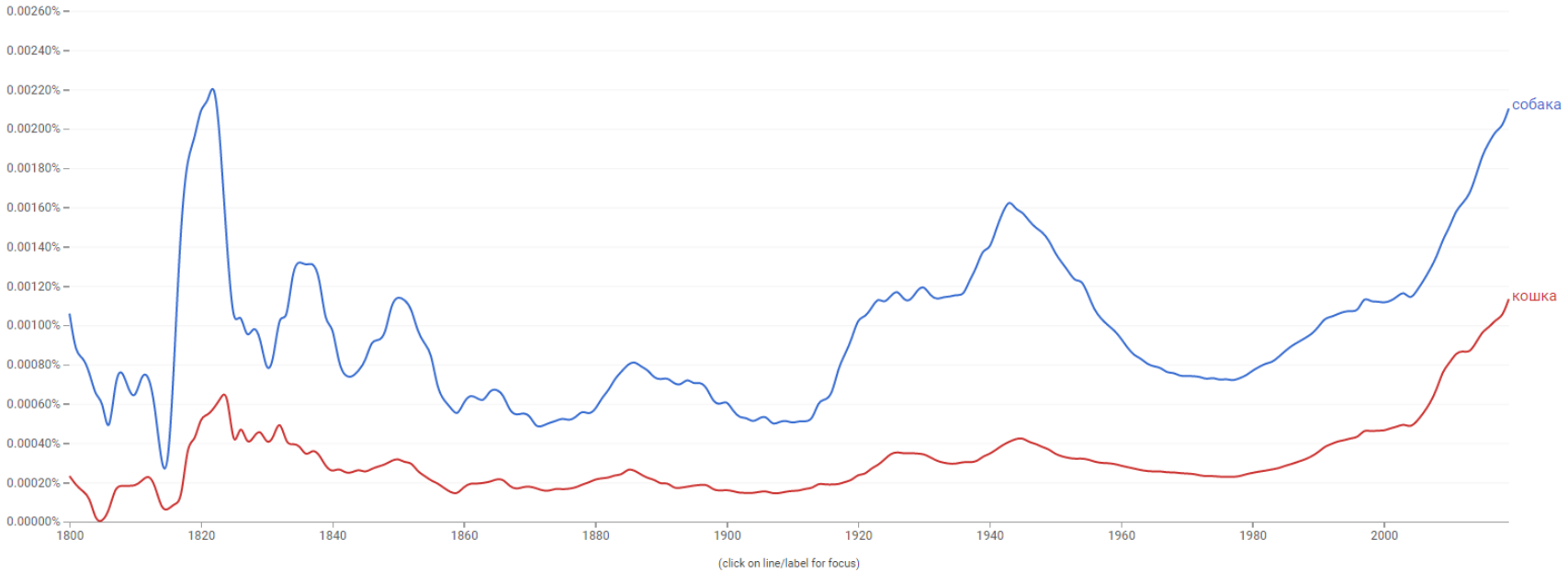
$c()$ – количество предложений в корпусе

Google Books Ngram Viewer

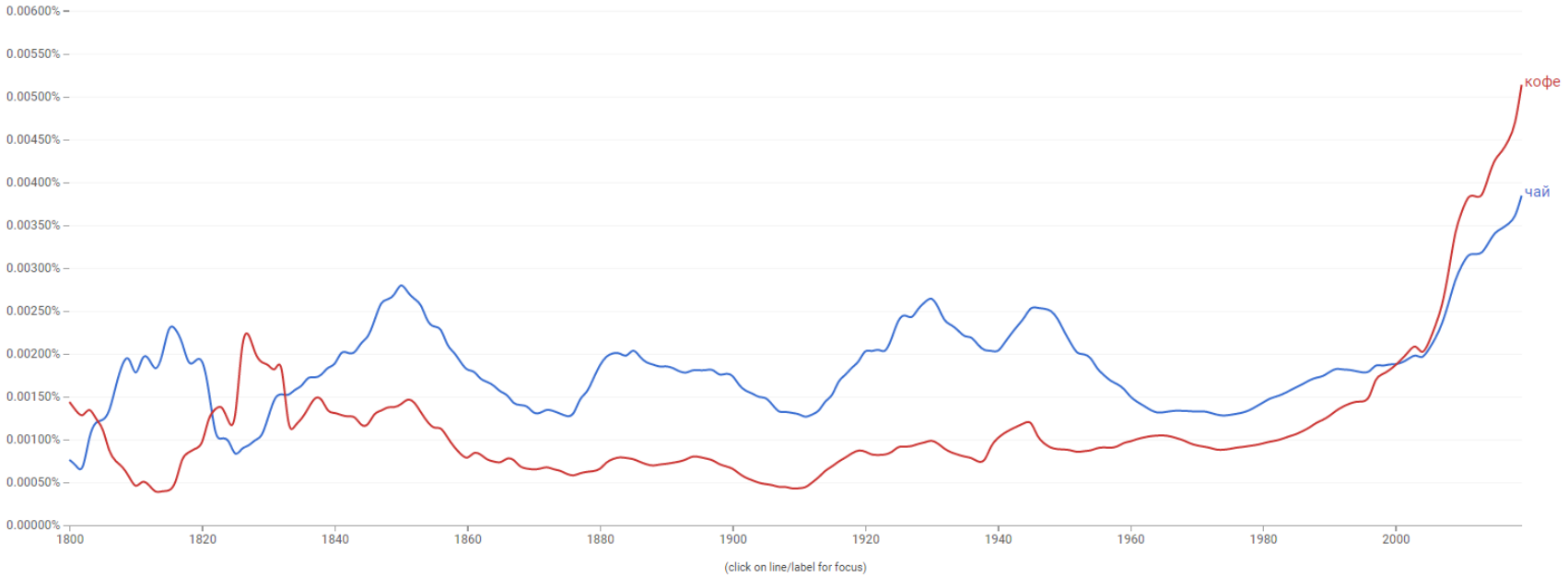
https://books.google.com/ngrams/



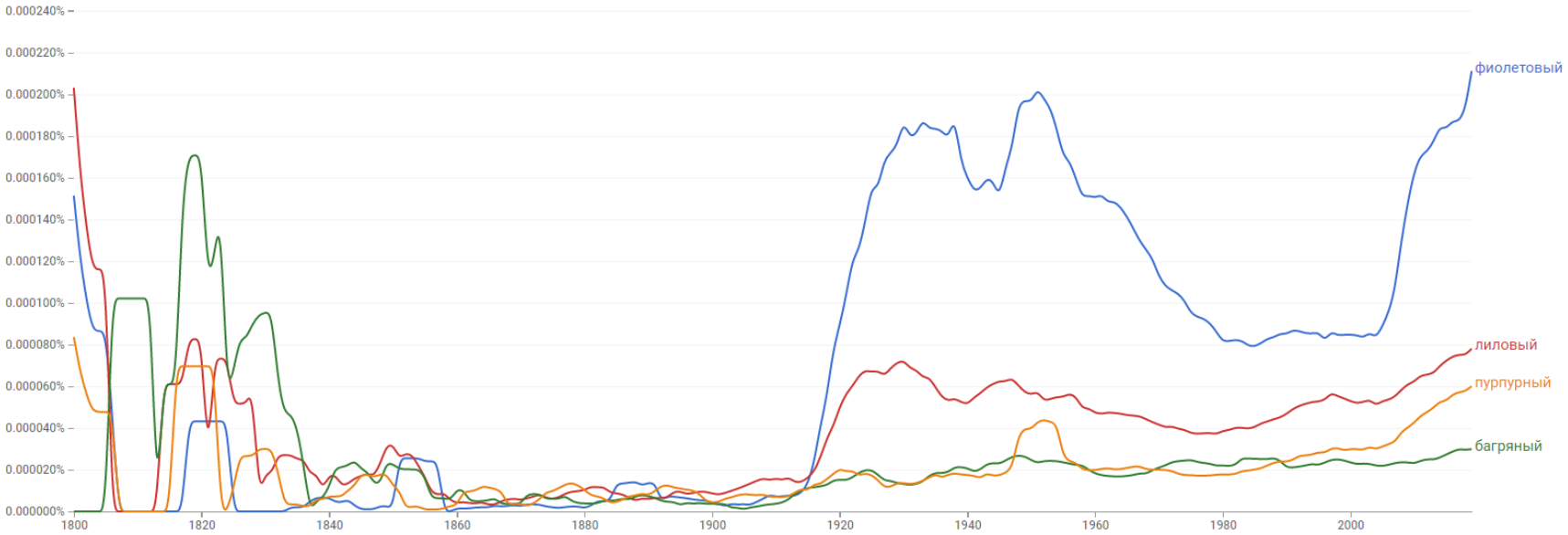
собака и кошка



чай и кофе


0.00600%
0.00550%
0.00500%
0.00450%
0.00400%
0.00350%
0.00300%
0.00250%
0.00200%
0.00150%
0.00100%
0.00050%
0.00000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
кофе
чай
(click on line/label for focus)

фиолетовый, лиловый, пурпурный



Биграммная языковая модель

Панграмма

Съешь ещё этих мягких французских булок, да выпей же чаю

Панграмма

Съешь ещё этих мягких французских булок, да выпей же чаю

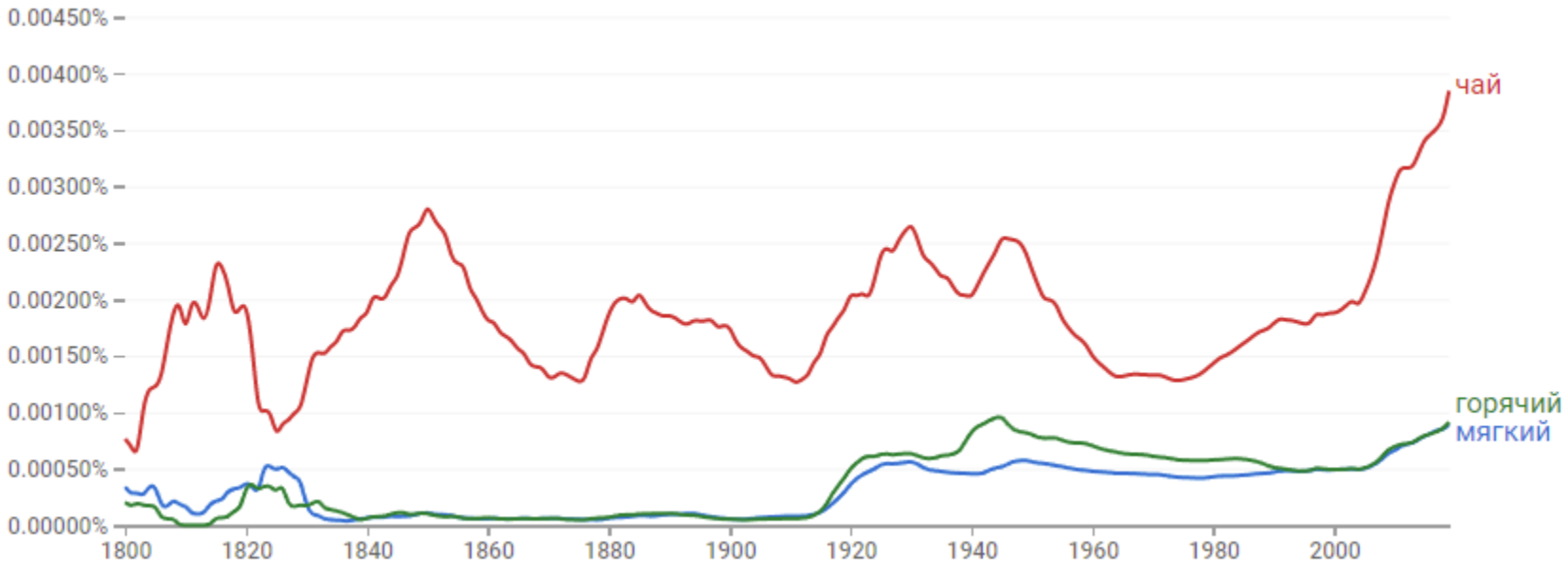
Съешь чаю, да выпей булок

Выпей мягкого чаю

Съешь драчливых булок

Горячий чай, мягкий чай

Чай, мягкий, горячий



Предсказание следующего слова

- круглый …
- выпил …
- яблоко от …

Биграммы

НКРЯ (ruscorpora.ru):

- Горячего …
- 67 сочетаний
- 1054 вхождения

№	Вхождения	Документы	Фрагмент
1	118	109	**горячего** чая
2	85	82	**горячего** чаю
3	66	57	**горячего** воздуха
4	54	53	**горячего** и
5	50	48	**горячего** кофе
6	50	41	**горячего** водоснабжения
7	41	36	**горячего** сердца
8	37	36	**горячего** молока
9	33	28	**горячего** пара
10	21	20	**горячего** желания
11	21	19	**горячего** ветра
12	20	20	**горячего** солнца
13	19	18	**горячего** сочувствия
14	18	18	**горячего** копчения
15	18	18	**горячего** супа
16	18	17	**горячего** до
17	17	16	**горячего** вина
18	16	10	**горячего** боя
19	15	15	**горячего** спора
20	13	13	**горячего** дыхания

Вероятность предложения

$$P(W) = P(w_1 w_2 w_3 \ldots w_N) =$$

$$= P(w_1|\emptyset) \cdot P(w_2|w_1) \cdot P(w_3|w_2) \cdot \ldots \cdot P(w_N|w_{N-1}) \cdot P(\emptyset|w_N)$$

Биграммная модель

$P(w_i|w_{i-1})$ – вероятность того, что слово w_i стоит в предложении после слова w_{i-1}

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

$c(w_{i-1}w_i)$ – количество предложений, в которых есть последовательность слов $w_{i-1}w_i$

$c(w_{i-1})$ – количество предложений, в которых есть слово w_{i-1}

Сравнение моделей по качеству

- PP – перплексия (коэффициент неопределенности) модели на предложении

$$PP(W) = P(w_1 w_2 w_3 \dots w_N)^{\frac{1}{N}}$$

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

Пример

Корпус:

Стол стул шкаф.

Диван стул шкаф.

Стул табурет диван.

Табурет диван люстра.

Стул диван люстра.

1. Построить языковую модель коллекции
2. Найти вероятность предложения:
   - Стол стул диван.
   - Диван стул люстра.
   - Стул шкаф.

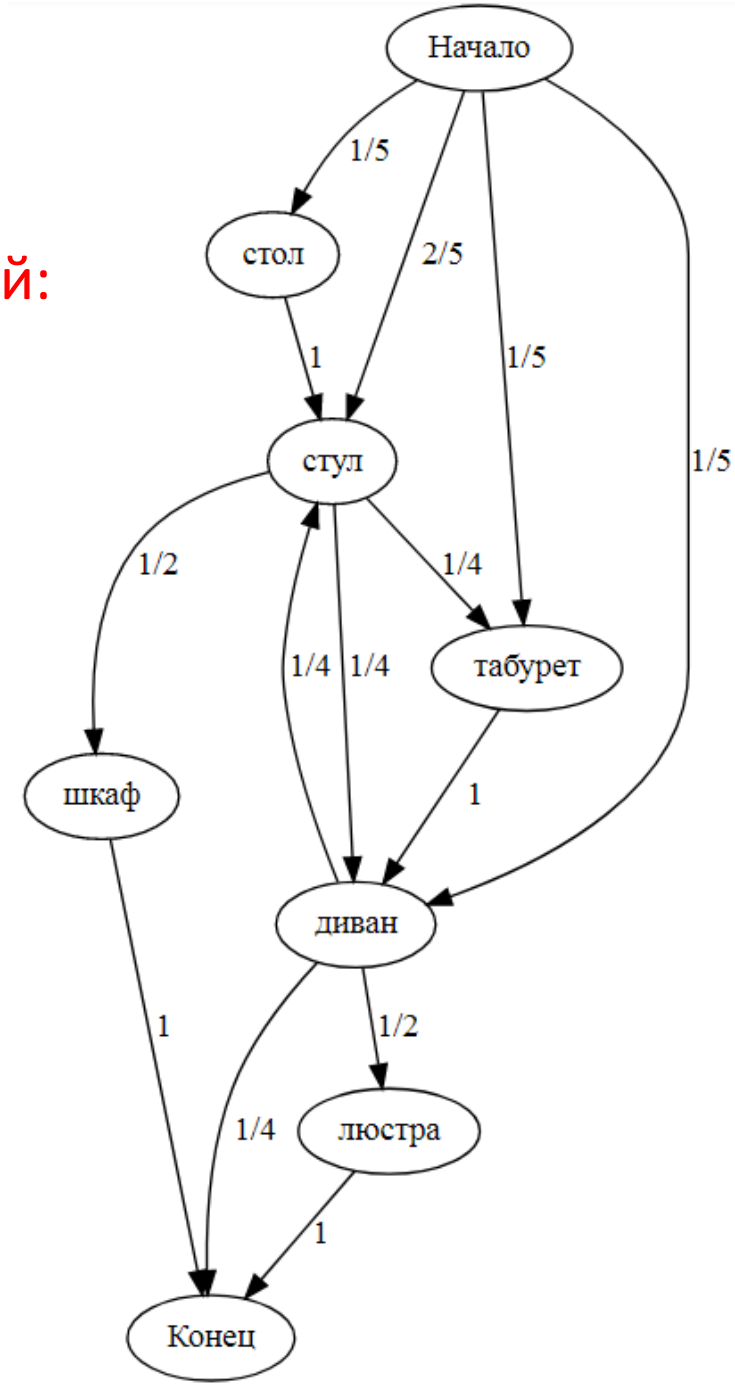
Пример

Найти вероятность предложений:

Стол стул диван.

Диван стул люстра.

Стул шкаф.

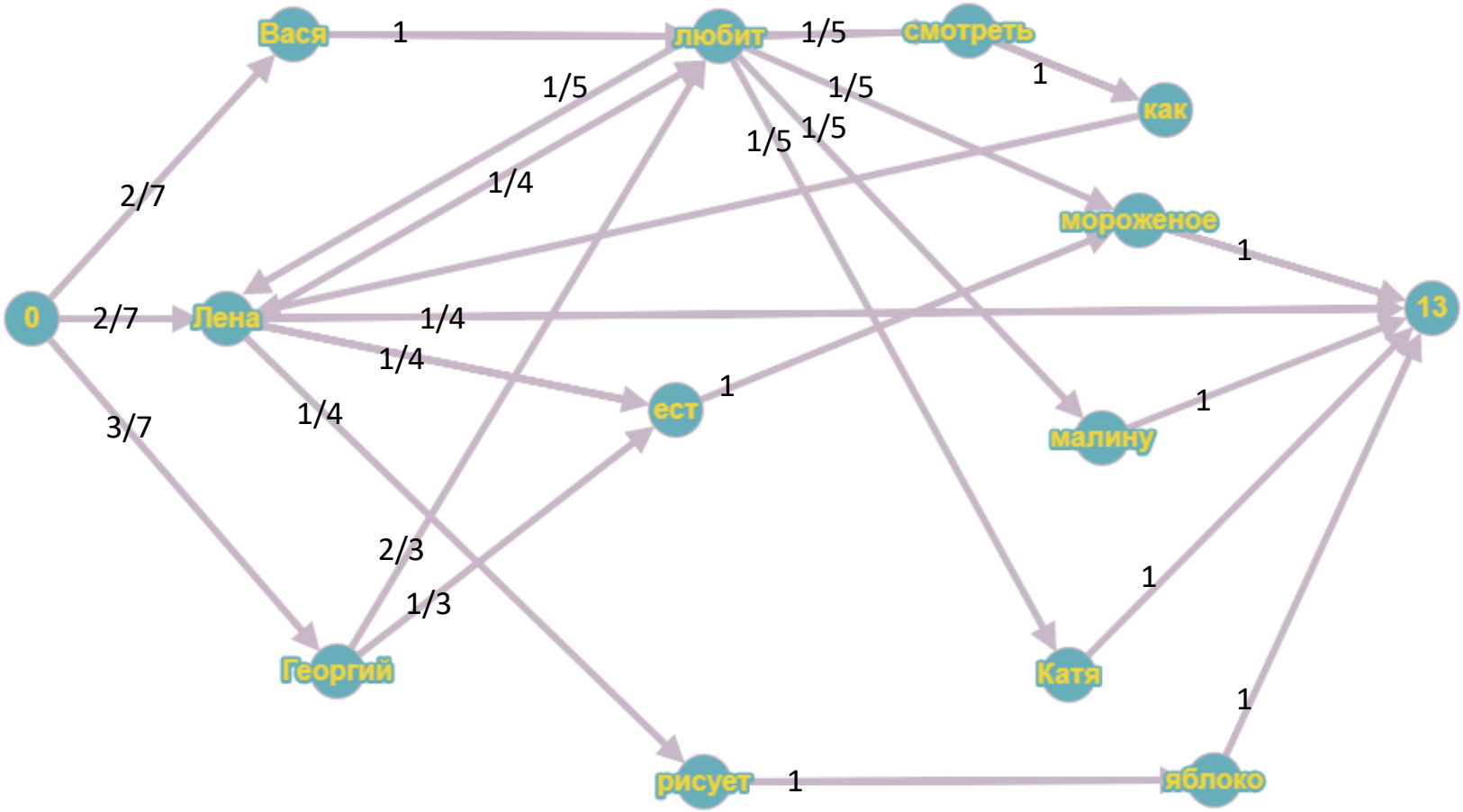


Задание

Корпус:

- *Вася любит мороженое.*
- *Лена любит малину.*
- *Вася любит Лену.*
- *Георгий ест мороженое.*
- *Лена рисует яблоко.*
- *Георгий любит Катю.*
- *Георгий любит смотреть, как Лена ест мороженое.*

1. Рассчитайте вероятности для предложений:
   - *Вася любит Катю.*
   - *Лена любит мороженое.*
   - *Лена рисует малину.*
2. Вычислите перплексию модели на предложении:
   - *Георгий любит малину.*


0
Вася
Лена
Георгий
любит
смотреть
как
мороженое
ест
малину
рисует
Катя
яблоко
13
2/7
3/7
1
1/5
1/4
1/3
2/3

Сглаживание

- языковая модель должна обобщать (а не повторять) данные, на которых она обучалась
- всегда будут новые последовательности слов, которые не встречались в корпусе для обучения

Сглаживание

- Лапласа
- Откат
- Интерполяция
- Кнесера-Нея (Kneser-Ney)
- Гуда-Тьюринга (Good-Turing)
- Виттена-Белла (Witten-Bell)
- …

Сглаживание Лапласа

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i) + 1}{c(w_{i-1}) + V}$$

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i) + \alpha}{c(w_{i-1}) + \alpha \cdot V}$$

- V – размер словаря
- α – коэффициент

Интерполяция

$$\hat{P}(w_i|w_{i-2}w_{i-1}) =$$
$$= \lambda_1 P(w_i|w_{i-2}w_{i-1}) +$$
$$+ \lambda_2 P(w_i|w_{i-1}) +$$
$$+ \lambda_3 P(w_i)$$

$$\sum_i \lambda_i = 1$$

Задание

Пусть есть корпус, содержащий 10000 предложений, размер словаря - 1500 уникальных слов (включая специальные "слова" – маркеры начала *<s>* и конца *</s>* предложений). Некоторые частоты униграмм:

ем – 100
дуриан – 1
и – 5000
не – 3000
морщусь – 50

и биграмм:

<s> ем – 20
ем дуриан – 0
дуриан и – 0
и не – 300
не морщусь – 15
морщусь </s> – 5

Примените сглаживание Лапласа для оценки вероятностей биграмм и оцените на их основе вероятность предложения: *<s>* Е*м дуриан и не морщусь </s>*

В качестве ответа запишите **натуральный логарифм** оценки вероятности предложения.

Триграммная скрытая Марковская модель для определения части речи

Триграммная скрытая Марковская модель

Ψ – множество словоформ {слон, слону,…}

Ω – множество тегов {N, V, P,…} (части речи)

Предложение – цепочка словоформ:

$w_1 w_2 w_3 \ldots w_n$, $w_i \in \Psi$,

которой соответствует цепочка тегов:

$t_1 t_2 t_3 \ldots t_n$, $t_i \in \Omega$

P(w|t) – вероятность того, что тегу t соответствует слово w

P(стекло|N)

$$p(w \mid t) = \frac{c(t \to w)}{c(t)}$$

c(t) – сколько раз встретился тег t в корпусе

c(t$\to$w) – сколько раз тегу t соответствовала словоформа w

P(t|u, v) – вероятность появления тега t при условии, что перед ним находятся теги u и v

P(N|VP)

$$p(t \mid u,v) = \frac{c(uvt)}{c(uv)}$$

c(uvt) – сколько раз в корпусе встретилась цепочка uvt

c(uv) – сколько раз в корпусе встретилась цепочка uv

Учитываются начало и конец предложения:

$P(t_1|**)$, $P(t_2|*t_1)$, $P(STOP|t_{n-1}t_n)$

Тогда

$$P(w_1w_2 \dots w_n, t_1t_2 \dots t_n) = \prod_{i=1}^{n+1} P(t_i|t_{i-2}t_{i-1}) \times \prod_{i=1}^{n} P(w_i|t_i)$$

$t_0 \equiv t_{-1} \equiv *$, $t_{n+1} \equiv STOP$

Дождь стучит в стекло

Существительное глагол предлог существительное (NVPN)

$$P(\text{стучит в стекло, VPN STOP}) = P(V|*,*) \times$$
$$\times P(P|*, V) \times P(N|V, P) \times P(STOP|P, N) \times$$
$$\times P(\text{стучит}|V) \times P(\text{в}|P) \times P(\text{стекло}|N)$$

$$P(N \mid *,*) = \frac{c(\text{N в начале предложения})}{c(\text{предложений})}$$

$$P(N \mid *,P) = \frac{c(\text{PN в начале предложения})}{c(\text{P в начале предложения})}$$

$$P(STOP \mid V,P) = \frac{c(\text{VP в конце предложения})}{c(VP)}$$

Алгоритм Витéрби

P(стучит в стекло, VPN STOP)
P(стучит в стекло, VPV STOP)
P(стучит в стекло, VPP STOP)
P(стучит в стекло, VNV STOP)
P(стучит в стекло, VNP STOP)
P(стучит в стекло, VNN STOP)
P(стучит в стекло, VVV STOP)
P(стучит в стекло, VVP STOP)
P(стучит в стекло, VVN STOP)
…