Статистический вывод: n-граммные модели

глава 6
(Manning,Shutze)
Statistical Natural language Processing

Языковые модели (language models)

- Определение вероятности предложений, последовательностей слов

- Как вероятна каждая последовательность?
  - P (w1, w2, w3,.. wn)
  - P(w5| w1, w2, w3, w4)

- Языковая модель – математическая модель, которая вычисляется вероятность последовательности слов или условную вероятность следования слова в контексте

Возможные применения:

- Распознавание речи
  - P (I saw a van)>P (eyes awe of an)
- Машинный перевод
  - P (high winds tonight)>P(large winds tonight)

- А также:
  - Распознавание сканированного текста
  - Спеллинг
  - Определение авторства
  - Определения языка текста и др.

“Shannon Game”

- Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.

- Предсказание следующего слова на основе (*n-1)* предшествующих слов

- Определение вероятности различных последовательностей на основе «тренировочного» корпуса

Цепи Маркова: предсказание на основе n-грамм

- Предположение Маркова (Марковские цепи) – слово определяется относительно небольшим предшествующим контекстом (несколько слов)
- “n-граммы” = последовательность n слов
  - Униграммы: $P(w_1, w_2, w_3,\ldots)=P(w_1)\ P(w_2)P(w_3)..$
  - биграммы: $P(w_i|w_1, w_2, w_3,\ldots)=P\ (w_i|w_{i-1})$
  - триграммы

Вопрос

- Какая последовательность слов наиболее вероятная по униграммной модели русского языка
  - (в, и, на, с)
  - (ехать, на, автобусе, домой)
  - (на, ехать, домой, автобусе)
  - (улучшить, обеспечение, населения, товарами)

Статистическая надежность vs. Качество предсказания

“большая зеленая _____”

машина? лягушка? лампа? Таблетка?

“Проглотил большую зеленую _____”

таблетку? лягушку?

Статистическая надежность vs. Качество предсказания

- Большая величина n:
  - больше информации о контексте – лучше предсказание продолжения
  - Меньше статистических данных

- Меньшая величина n:
  - Больше примеров в данных, больше статистики
  - Возрастает неопределенность предсказания

Следствие: порождение текстов, похожих на естественные

- Порождение текстов на основе марковских моделей – один из известных видов поискового спама
  - Коллекция текстов нужной тематики
  - Извлечение статистики
- Эксперимент по порождению пьес Шекспира

Порождение Шекспира

- Порождение по униграммам...
  - *Every enter now severally so, let*
  - *Hill he late speaks; or! a more to leg less first you enter*

- На основе биграмм...
  - What means, sir. I confess she? then all sorts, he is trim, captain.
  - Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry.

Порождение Шекспира

- Триграммы
  - Sweet prince, Falstaff shall die.
  - This shall forbid it should be branded, if renown made it empty.

- Тетрограммы
  - What! I will go seek the traitor Gloucester.
  - Will you not tell me who I am?
  - Это выглядит как Шекспир, поскольку это и есть Шекспир

Выбор *n*

Словарь (V) = 20,000 слов

n	**Количество единиц**
2 (биграмм)	**400 000 000**
3 (триграмм)	**8000 000 000 000**
4 (тетраграмм)	**1.6×10^{17}**

Статистическая оценка

- Даны обучающие текстовые данные ...
- Как построить модель (распределение вероятностей), которая может предсказывать продолжение начатого текста

Вероятность появления следующего слова

P (Wn|W1,.., Wn-1)= P(W_1,…..Wn)/P(W_1,…Wn-1)

MLE (максимальное правдоподобие)

Pmle (W_1,…Wn)= C(W1,….,Wn)/N

Pmle (Wn|W_1,.., W_{n-1})= C(W_1,…Wn)/ C(W_1,…Wn-1)

C ()- частота появления подстроки

Для биграмм

Pmle (Wn|Wn-1)=C(Wn-1, Wn)/C(Wn-1)

Пример

- Корпус
  - <s> Он пошел в школу </s>
  - <s> Пошел он в школу</s>
  - <s> Он не любит мясо</s>

- Оценки
  - P(он| <s>)=2/3
  - P (<s> |школу) …
  - P (мясо|он)

Корпус отзывов о ресторанах

- P(english|want)=0.0011
- P(chinese|want)=0.0065
- P(to|want)=0.66
- P(eat|to)=0.28
- P(food|to)=0
- P(want|spend)=0
- P(I, <s>)=0.25

Корпус отзывов о ресторанах

- P(english|want)=0.0011 – знания о мире
- P(chinese|want)=0.0065 – знания о мире
- P(to|want)=0.66 - грамматика
- P(eat|to)=0.28 - грамматика
- P(food|to)=0 – нет значения в корпусе
- P(want|spend)=0 - грамматика
- P(I, <s>)=0.25

Статистическая оценка

Пример:

Корпус: пять романов Джейн Остин

N = 617,091 слов

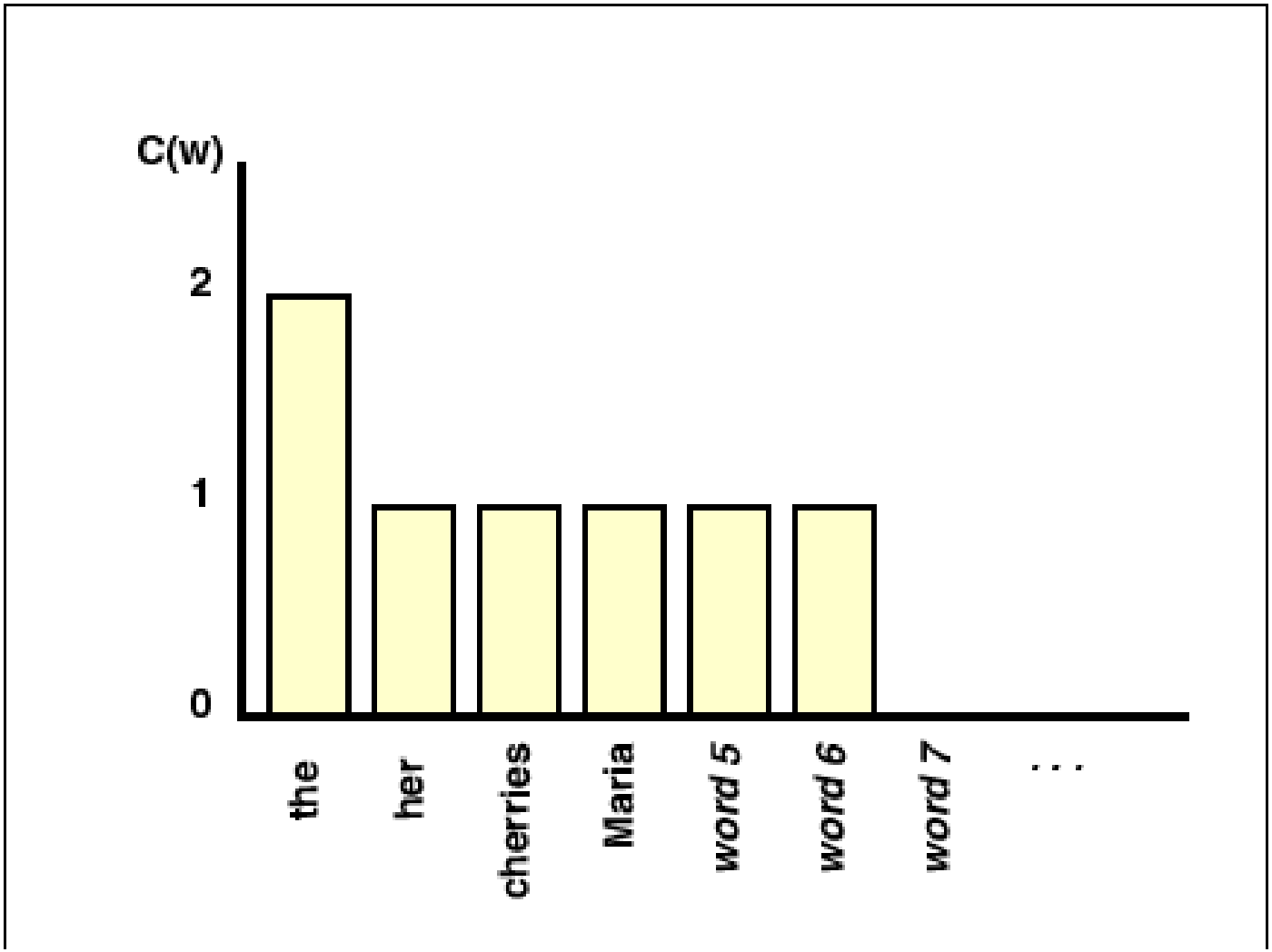
V = 14,585 уникальных слов

Задание: предскажи следующее слово после триграммы “inferior to _____”

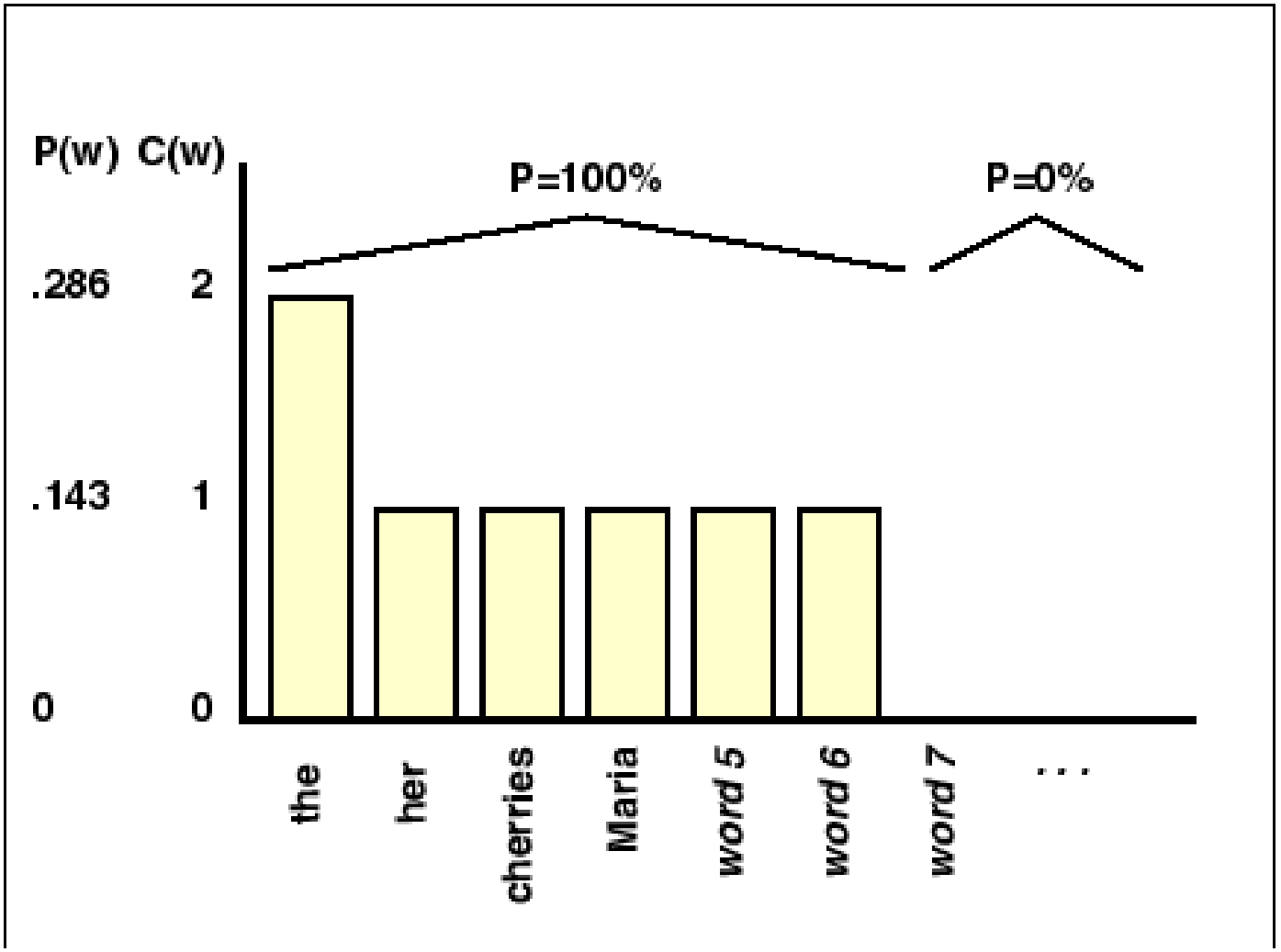
from test data, *Persuasion*: “[In person, she was] inferior to *both* [sisters.]”

Примеры в обучающем корпусе:

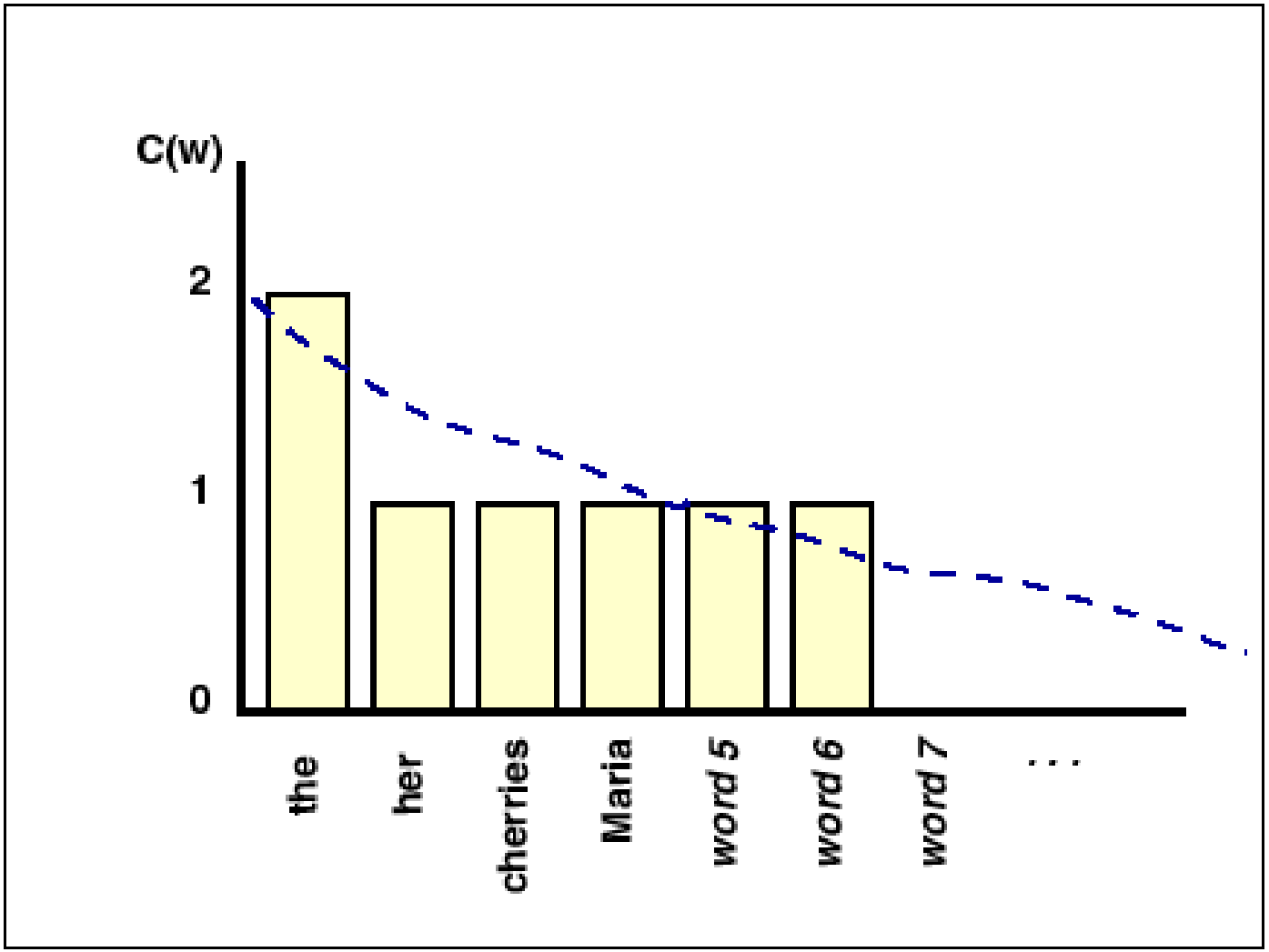
“inferior to _____”



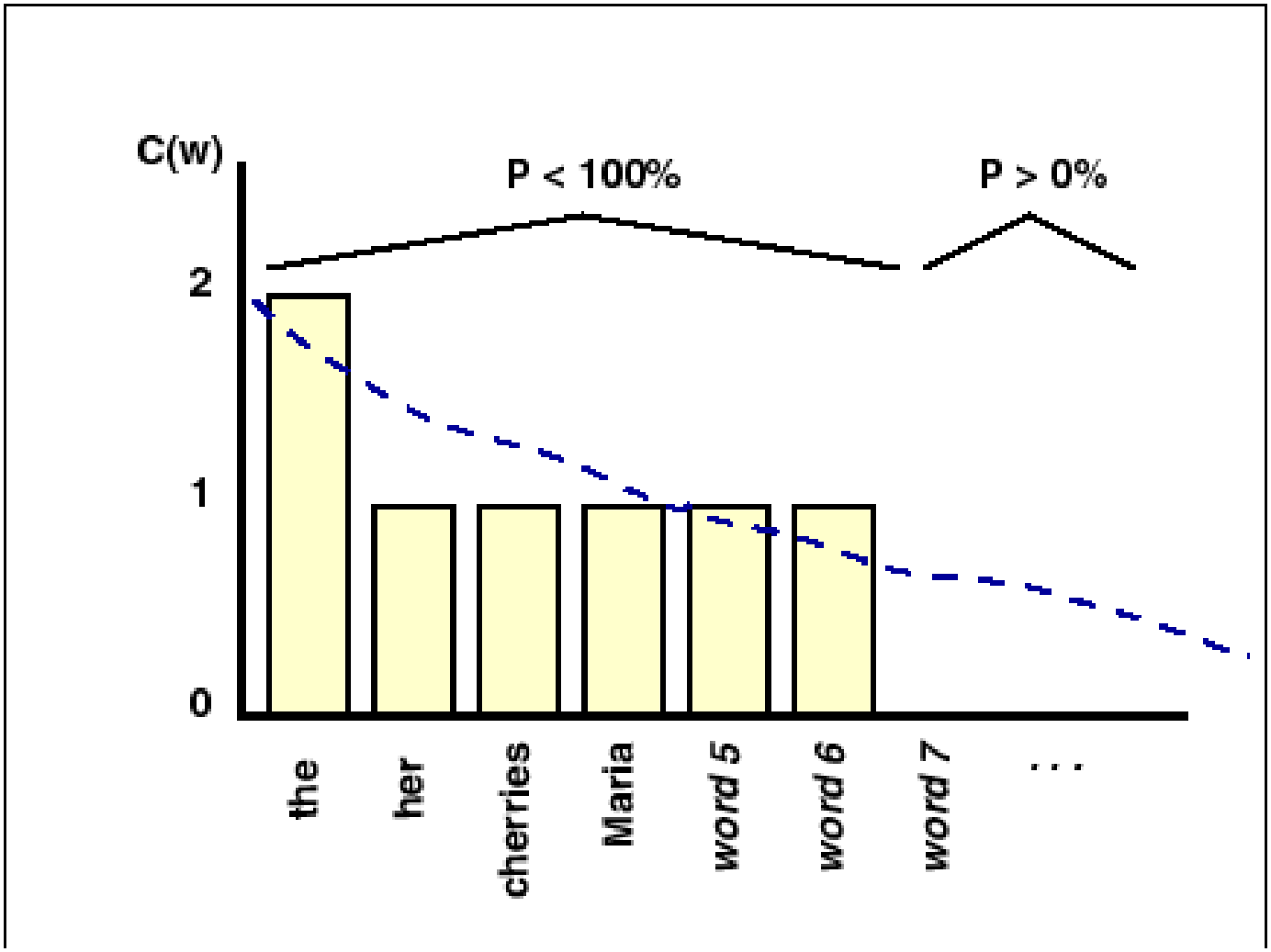
Оценка максимального правдоподобия



Реальное распределение вероятностей:



Реальное распределение вероятностей



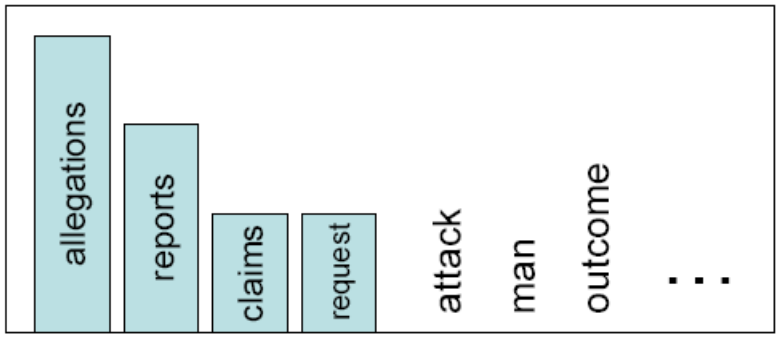
“Smoothing”- сглаживание

- Нужна модель, которая позволяет снизить вероятности уже встреченных событий и повысить вероятность еще не встречавшихся биграмм
- Также называется методы дисконтирования (Discounting methods)

Smoothing is like Robin Hood:
Steal from the rich and give to the poor (in probability mass)

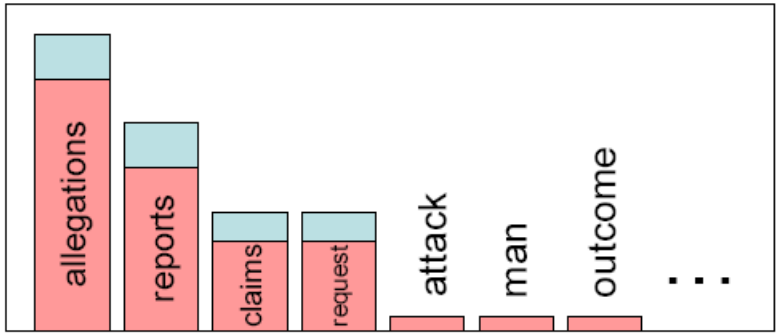
- We often want to make predictions from sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



- Smoothing flattens spiky distributions so they generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



- Very important all over NLP, but easy to do badly!

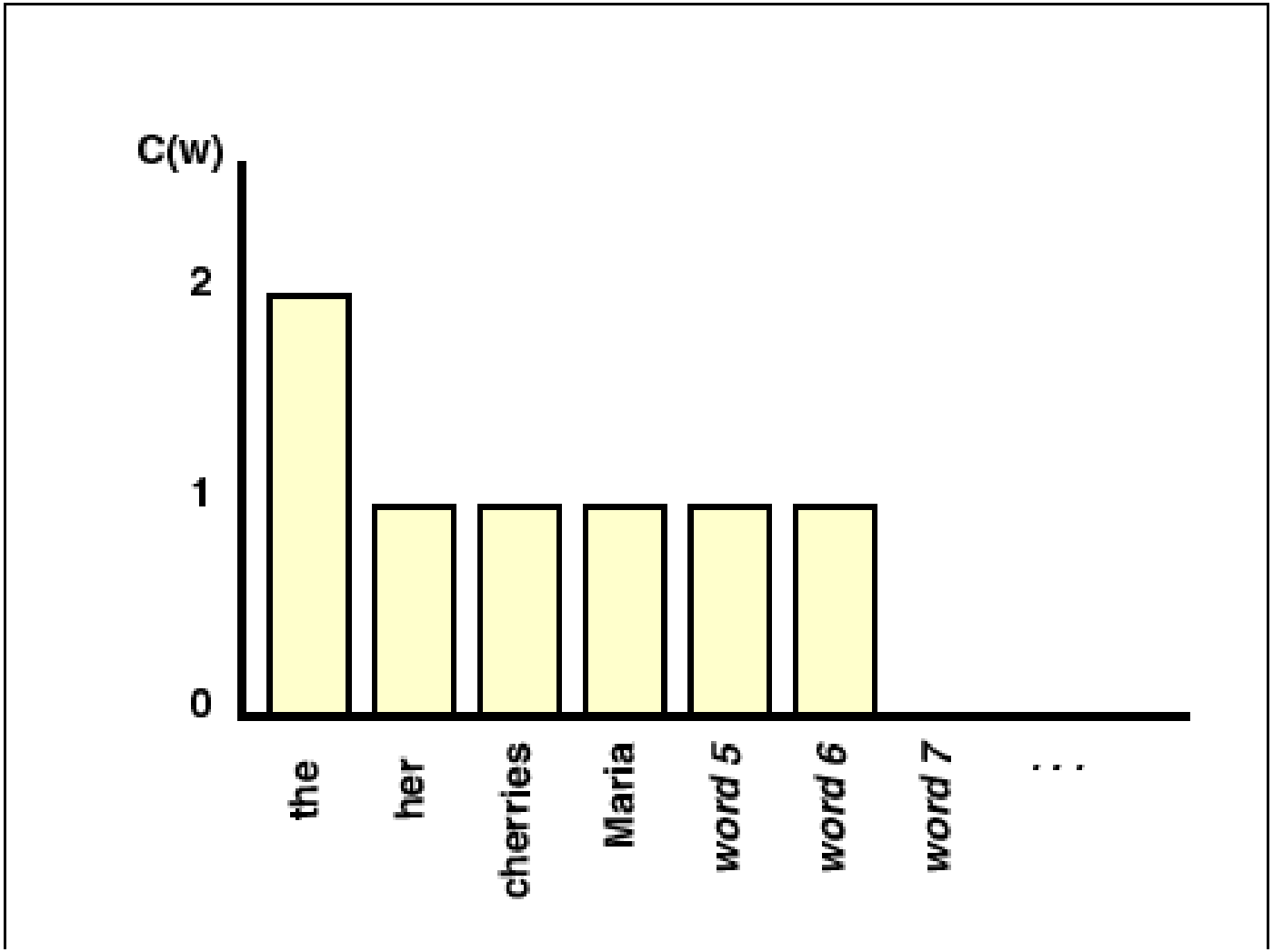
Slide from Dan Klein

Сглаживание ”adding one”

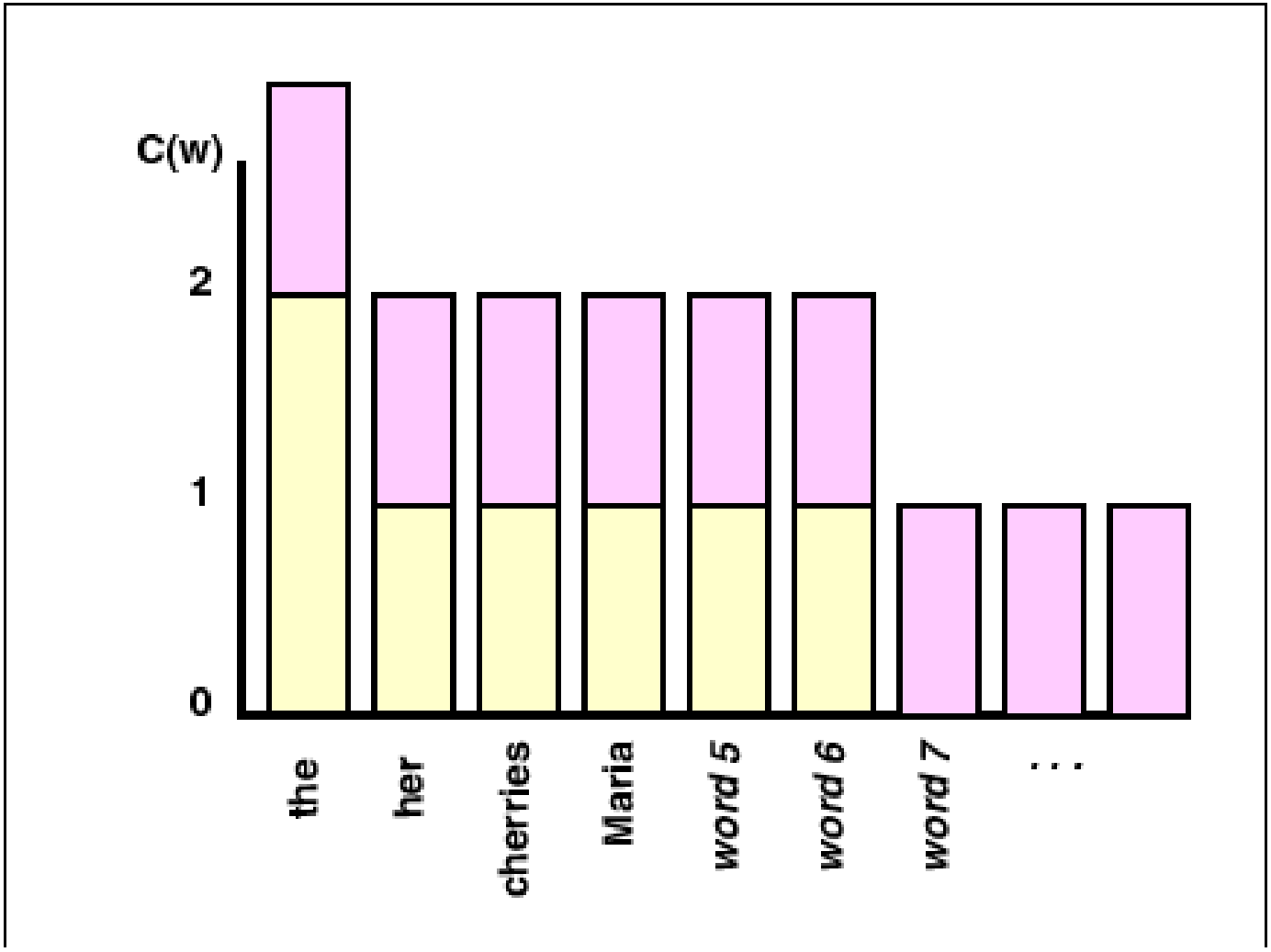
- Правило Лапласа оценки вероятности следующего результата S,F
  - P(s)=(s+1)/(s+f+2)
  - Предположение о равновероятном распределении: Uniform Prior

- Сглаживание
  - Plap=(C(W1,…Wn)+1)/(N+B)
  - C(W1,…Wn) – частотность n-граммы
  - N - число слов в корпусе
  - B – число возможных значений предсказываемой величины

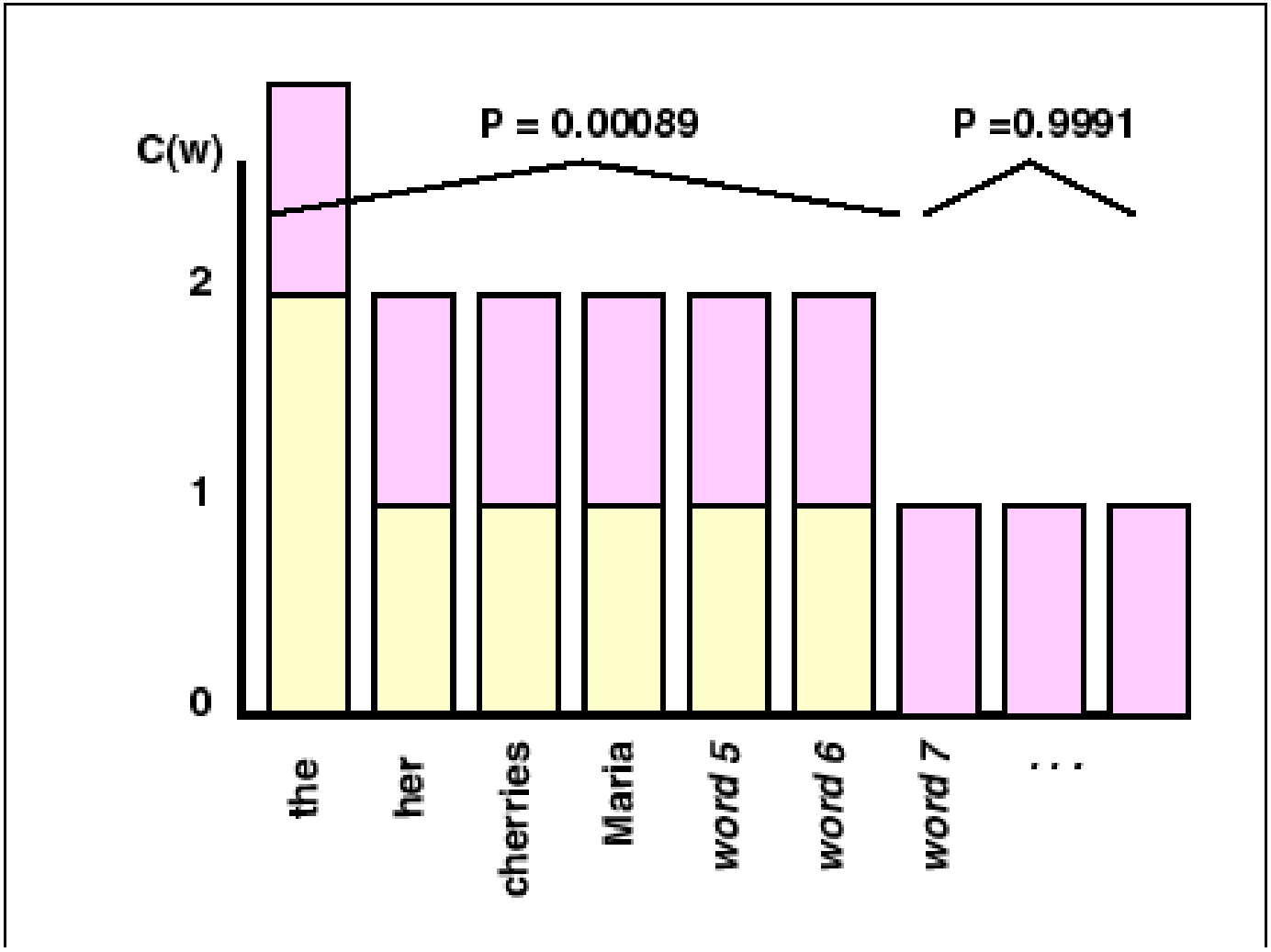
LaPlace's Law
(adding one)



LaPlace's Law
(adding one)



LaPlace's Law



Сглаживание Лапласа

- Для униграмм:
  - Добавляем 1 к частоте каждого слова
  - Нормализуем N (#tokens) + V (#types)
  - Исходная вероятность униграммы $$P(w_i) = \frac{c_i}{N}$$
  - Новая вероятность униграммы $$P_{LP}(w_i) = \frac{c_i + 1}{N + V}$$
  - Для биграмм
  - исходная $$P(w_n | w_{n-1}) = \frac{c(w_n | w_{n-1})}{c(w_{n-1})}$$
  - новая $$P(w_n | w_{n-1}) = \frac{c(w_n | w_{n-1}) + 1}{c(w_{n-1}) + V}$$

Однако

- Для больших словарей – биграмм всегда много – закон Лапласа дает слишком много вероятности не встречавшимся событиям

- (Church, Gale, 1991)
- Corpus Associated Press – 44 млн. слов – разделили на две части и пытались предсказать поведение на второй части
- - 400653 разных слов
- - $1.6*10^{11}$ число возможных биграмм

Предсказание числа биграмм

- F_{lap}= ((r+1)/(N+B))*N
- Биграммы встречались r – раз в одной половине корпуса
- Нужно предсказать, сколько раз такие биграммы встретятся во второй половине корпуса

- N=22 млн. слов
- V= 273266 разных слов
- B=V*V

Закон Лапласа и реальные частоты

R=fmle	flap	femp
0	0.000137	0.000027
1	0.000274	0.448
2	0.000411	1.25
3	0.000548	2.24
4	0.000685	3.23
5	0.000822	4.21
6	0.000959	5.23

Flap – предсказание средней частоты во второй части по Лапласу

Femp – реальная средняя частота во второй части

Lidstone's Law

$$P_{Lid}(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n) + \lambda}{N + B\lambda}$$

P = вероятность n-граммы

C = частота n-gram в обучающей колллекции

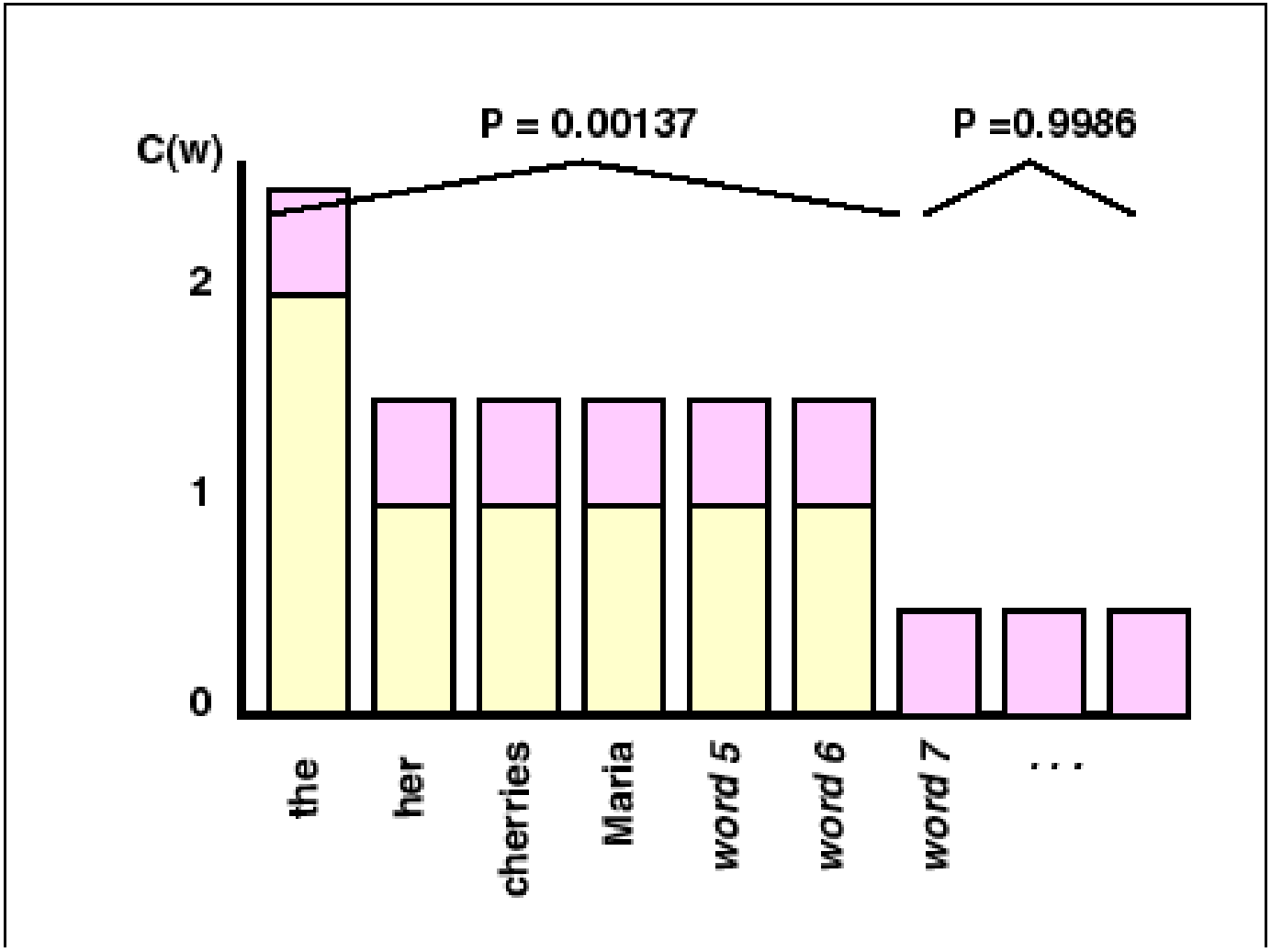
N = количество n-грамм в обучающих данных

B = количество типов (разных n-грамм)

λ : ($0 << \lambda << 1$)

M.L.E: $\lambda = 0$
LaPlace's Law: $\lambda = 1$
Jeffreys-Perks Law: $\lambda = ½$

Jeffreys-Perks Law



Тестирование моделей

- Hold out ~ 5 – 10% для тестирования
- Hold out ~ 10% для подбора параметров (smoothing)
- Для тестирования: полезно тестировать на разных коллекциях, и исследовать поведение моделей

Кросс-валидация на двух частях (a.k.a. deleted estimation)

- **Use data for both training and validation**

Divide test data into 2 parts

(1) Train on A, validate on B

(2) Train on B, validate on A

Combine two models

Кросс-валидация

Two estimates:

$$P_{ho} = \frac{T_r^{01}}{N_r^0 N} \qquad P_{ho} = \frac{T_r^{10}}{N_r^1 N}$$

N_r^a = number of n-grams occurring r times in a-th part of training set

T_r^{ab} = total number of those found in b-th part

Combined estimate:

$$P_{ho} = \frac{T_r^{01} + T_r^{10}}{N(N_r^0 + N_r^1)}$$

(arithmetic mean)

Комбинирование оценок

- Иногда триграммная модель – лучшая, иногда – биграммная, иногда - униграммная

- Как сделать модель, которая использует несколько видов биграмм?

Простая линейная интерполяция
(a.k.a., finite mixture models;
a.k.a., deleted interpolation)

$$\boldsymbol{P_{li}(w_n \mid w_{n-2}, w_{n-1}) =}$$

$$\lambda_1 P_1(w_n) + \lambda_2 P_2(w_n \mid w_{n-1}) + \lambda_3 P_3(w_n \mid w_{n-2}, w_{n-1})$$

- Взвешенное среднее униграмм, биграмм и триграмм

Katz's Backing-Off

- Используем *n-gram* вероятность, когда достаточно данных
  - (когда частота > k; k usu. = 0 or 1)
- Если нет, то переходим ("back-off") на *(n-1)-gram* вероятность
- (Повторяем при необходимости)

Как оценить качество языковой модели?

- Имеется две языковые модели
  - Какая из них лучше подходит для корпуса?
  - = Какая из них лучше предсказывает следующее слово.
  - = предсказывает более правильные предложения

Лучшее тестирование – внешнее (extrinsic)

- Поместить языковую модель в задачу
  - Спеллер, машинный перевод
- Выполнить задачу и проверить качество
  - Сколько слов обработано правильно
  - Сколько слов переведено правильно
  - Сравнить качество моделей A и B
  - Это правильно, но дорого

Внутреннее (intrinsic) тестирование

- Разделение корпуса на две части
  - Обучение
  - Тестирование
- Вычисление перплексии (perplexity)

$$\begin{aligned} \mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

- Чем меньше перплексия – тем лучше,
- Модель меньше удивляется продолжению предложения

Как вычислять перплексию

- Перплексия для униграмм
- Перплексия для биграмм
- Перплексия для предложения, состоящего из случайно последовательности цифр=10
- Перплексия – это среднее число вариантов, из которых происходит выбор на каждом шаге.
- Перплексия: расчет

$$PP = 2^{-\frac{1}{N}\sum_{1}^{N} \log_2 p(x)}$$

Перплексия для корпуса Wall Street Journal

- Обучающая коллекция – 38 млн. слов
- Тестовая коллекция – 1.5 млн. слов
- Перплексия.Униграммы – 962
- Перплексия. Биграммы – 170
- Перплексия. Триграммы - 109

Проблемы языковых моделей

- «Длинные зависимости» (long dependences)
  - Дом, в котором я был вчера вечером, состоял из пяти комнат
  - Биграмма (вечером состоял) – не очень вероятна

Заключение

- Статистические методы обработки текстов применяют различные методы сглаживания для оценки вероятности еще не встречавшихся событий
- Применяются специальные методы тестирования и настройки моделей
  - Held-out data – данные для настройки параметров
  - Testing data – тестовая коллекция
  - Cross-validation – кросс-проверка

Задание к 2 октября

- Выделить десятую часть из Вашего корпуса
- Вычислить вероятности
- Посчитать перплексию на другой части текста (15% общей длины)
  - По униграммам
  - По биграммам
  - В обоих случаях используем закон Jeffreys-Perks Law: $\lambda = ½$
- Отчет
  - Название текста
  - Формулы
  - Необходимые данные
  - Результат вычислений