Интеллектуальный анализ текстов

Курс «Интеллектуальные информационные системы»

Кафедра управления и интеллектуальных технологий

НИУ «МЭИ»

Весна 2023 г.

Text Mining – интеллектуальный анализ текстов

Интеллектуальный анализ текстов (*ИАТ*, *text mining*) — направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка. (Wikipedia)

- **Категоризация текстов** (*classification*) –
  - отнесении документов из коллекции к одной или нескольким группам (классам, кластерам) схожих между собой текстов
- **Извлечение информации** (*information extraction*) –
  - это задача автоматического извлечения (построения) структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов (распознавание имен людей, названий организаций, поиск ключевых слов для текста, автореферирование)
- **Информационный поиск** (*information retrieval*) –
  - процесс поиска *неструктурированной* документальной информации, удовлетворяющей информационные потребности (процесс выявления в некотором множестве документов всех тех, которые посвящены указанной теме)

Проблемы, возникающие при работе с документами, написанными на ЕЯ

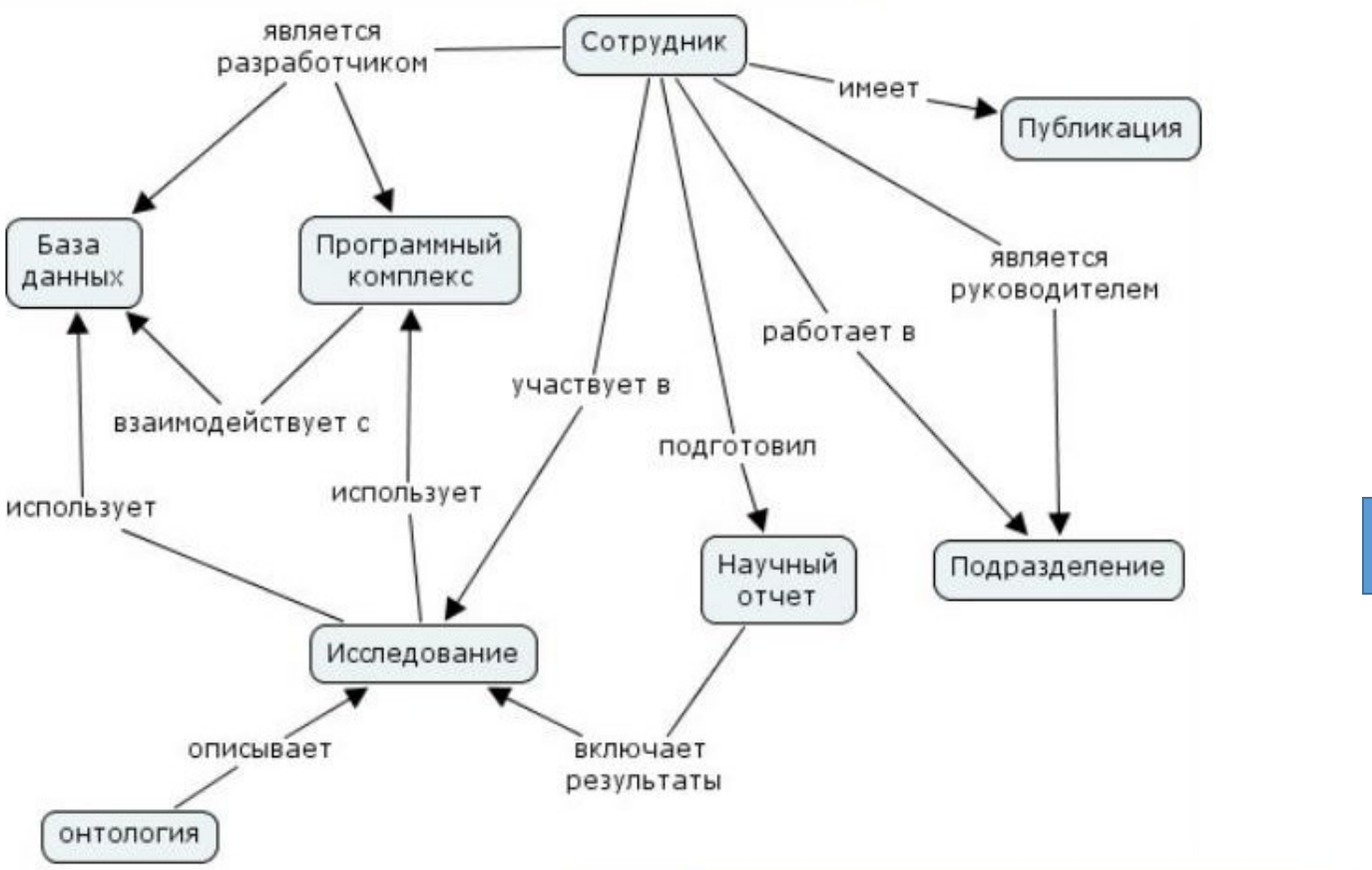
- ***Семантическая неоднозначность:***
  - *Синонимия: экран-дисплей*
  - *Полисемия: команда (судна; футбольная)*
  - *Омонимия: Ключ (родник) – Ключ (от замка)*
  - *Эллипсность: пропуски слов или слова-заменители*
- ***Многообразие средств передачи смысла:***
  - *Лексика ЕЯ*
  - *Контекст*
  - *Ссылки на слова*
- ***Высокая размерность задачи***
- ***Субъективность оценки качества классификации***
- ***Различная длина документов***

Подходы Text Mining

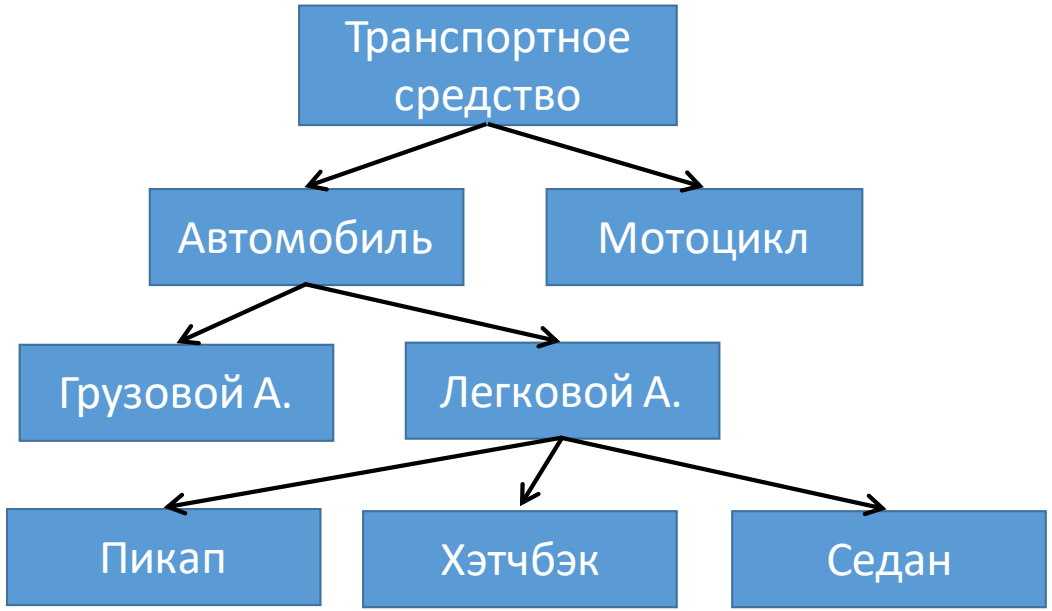
Лингвистический анализ	Статистический анализ
Системы ЛА обычно состоят из модели предметной области, содержащей основные тематические термины и их взаимосвязи, а также специализированной базы данных (БД) грамматических конструкций и семантических правил, свойственных конкретному языку – онтологий и тезаурусов. При этом модель предметной области обычно используется для проведения морфологического анализа, а специализированная БД – для синтаксического и семантического анализа	**Текст – набор ключевых слов. Вес слов зависит от различных факторов, в частности – от частоты встречаемости термина в документе. Предполагается, что появление одних и тех же терминов в различных документах говорит об их подобии**

Онтологии и тезаурусы

Пример онтологии



Пример тезауруса



Институт систем энергетики им Л.А.Мелентьева СО РАН

Что такое текст?

- Текст – конечное множество слов (терминов), объединенных лексическими, грамматическими, смысловыми, частотными отношениями и образующих информативное сообщение.
- Главное в тексте – информация, новая для читателя, которая заключена в авторском изложении, и которую мы хотим извлечь.
- Чем больше информации извлечем – тем лучше.
- Не всегда большой текст = большому количеству информации

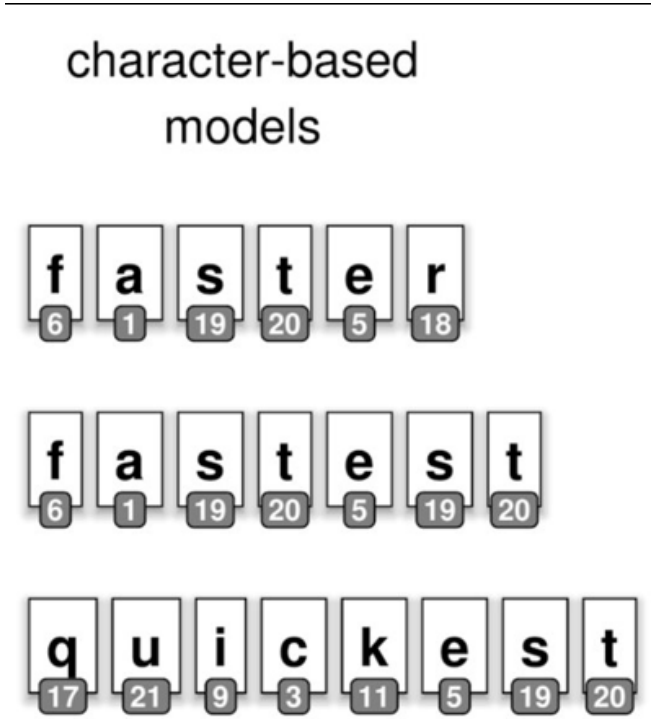
Как текст представляется в математическом виде?

Текст – последовательность токенов.

Токен – единица словаря.
Последовательность из n символов (n варьируется в множестве натуральных чисел {1, 2, 3,…})

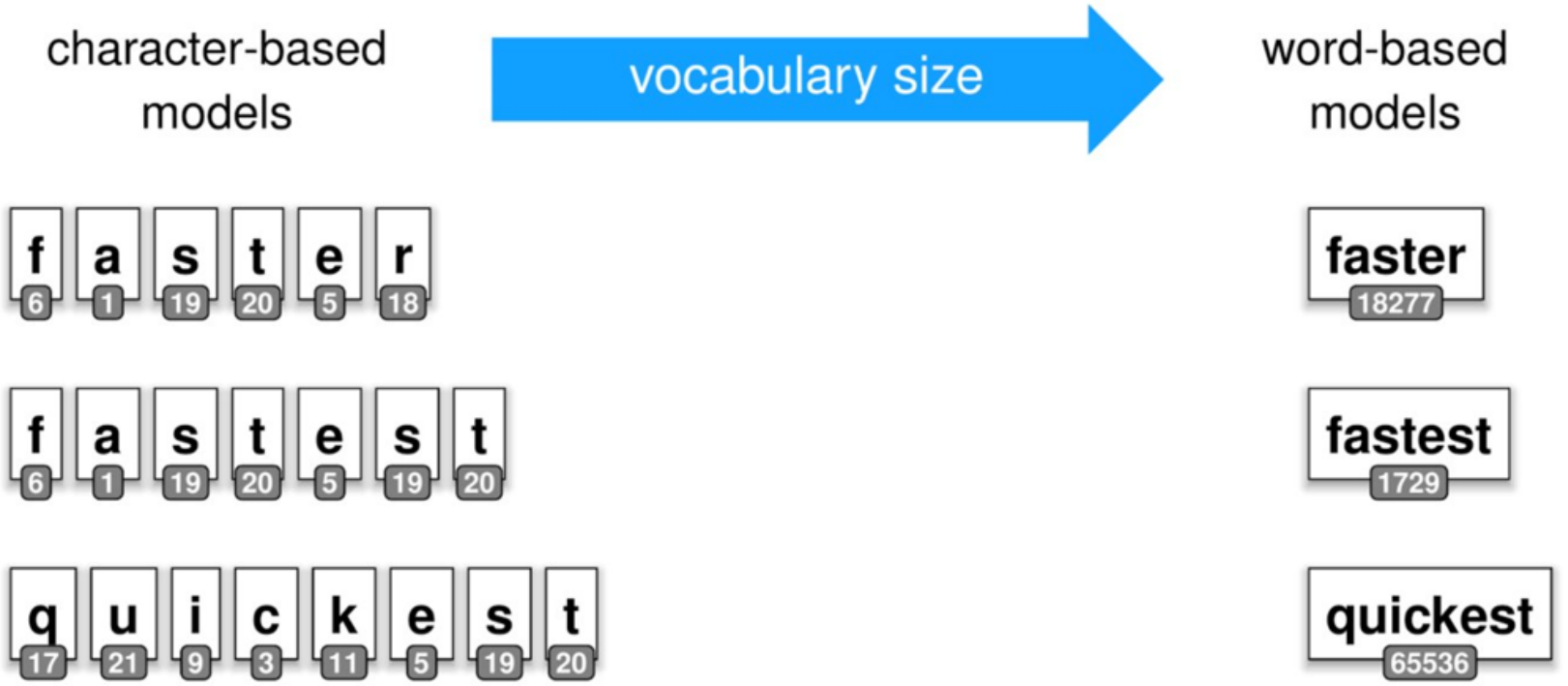
Словарь – всевозможный набор токенов, встречающихся в языке (в генеральной совокупности)

Как можем представить текст?



Слова состоят из букв. Представляем слова как набор букв.
Размер словаря небольшой (~100 для одного языка).
Но размер текста очень большой

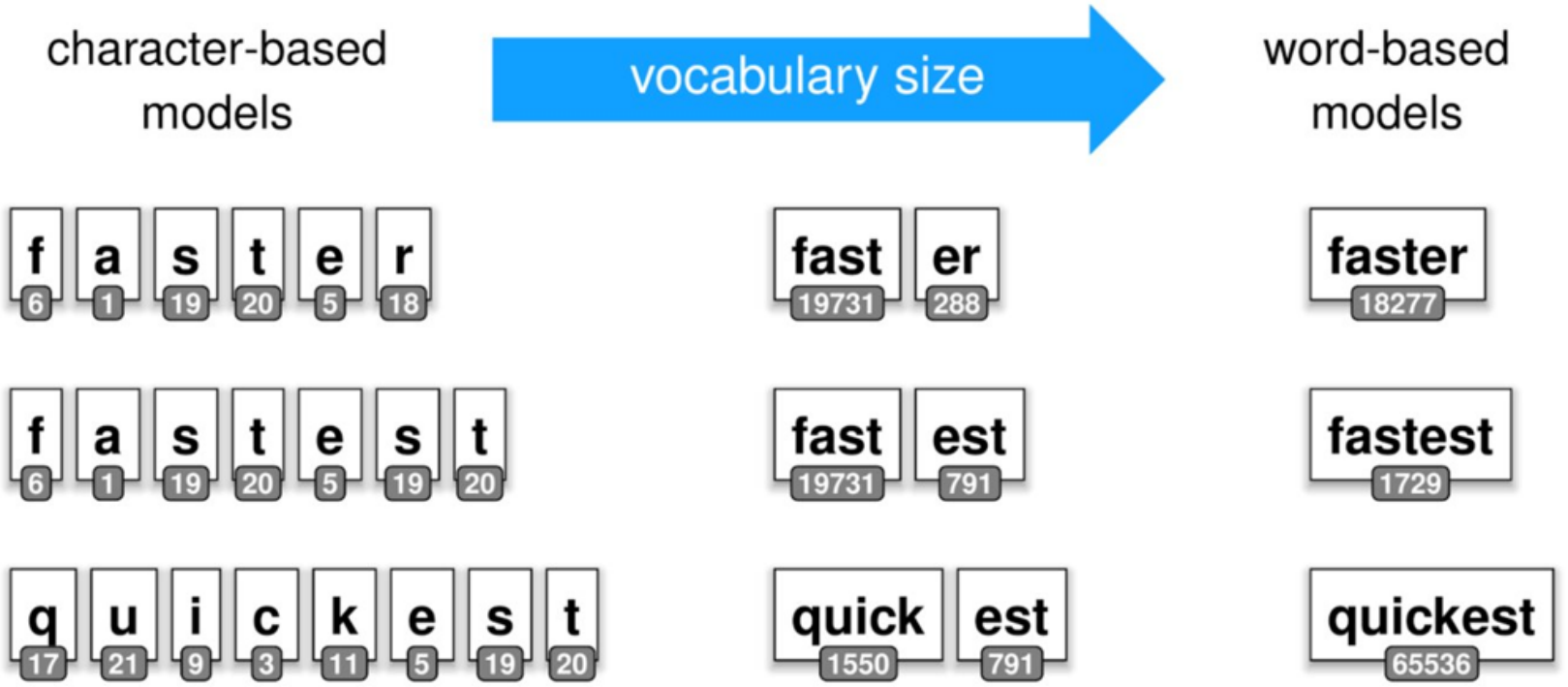
Как можем представить текст?



Слова состоят из букв. Представляем слова как набор букв.
Размер словаря небольшой (~100 для одного языка).
Но размер текста очень большой

Длина последовательности меньше, но на порядки возрастает длина словаря

Как можем представить текст?

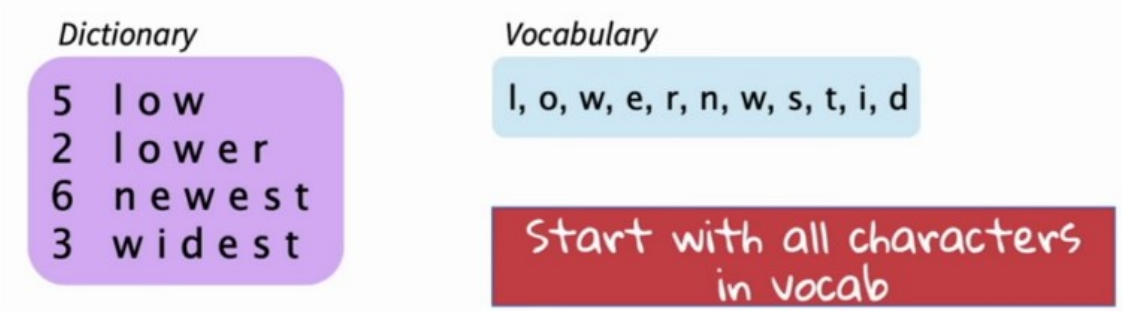


Слова состоят из букв. Представляем слова как набор букв.
Размер словаря небольшой (~100 для одного языка).
Но размер текста очень большой

Золотая середина – самые частовстречаемые буквосочетания.
Смысл и в корне, и в суффиксе

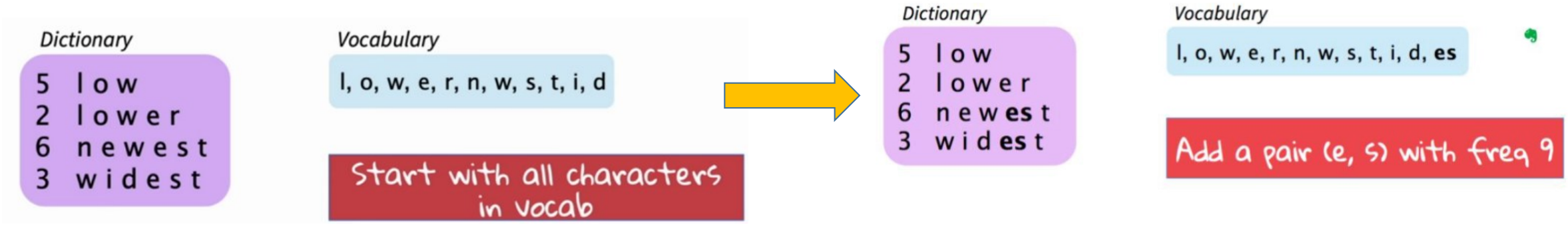
Длина последовательности меньше, но на порядки возрастает длина словаря

Как получить набор таких сочетаний (токенов)?



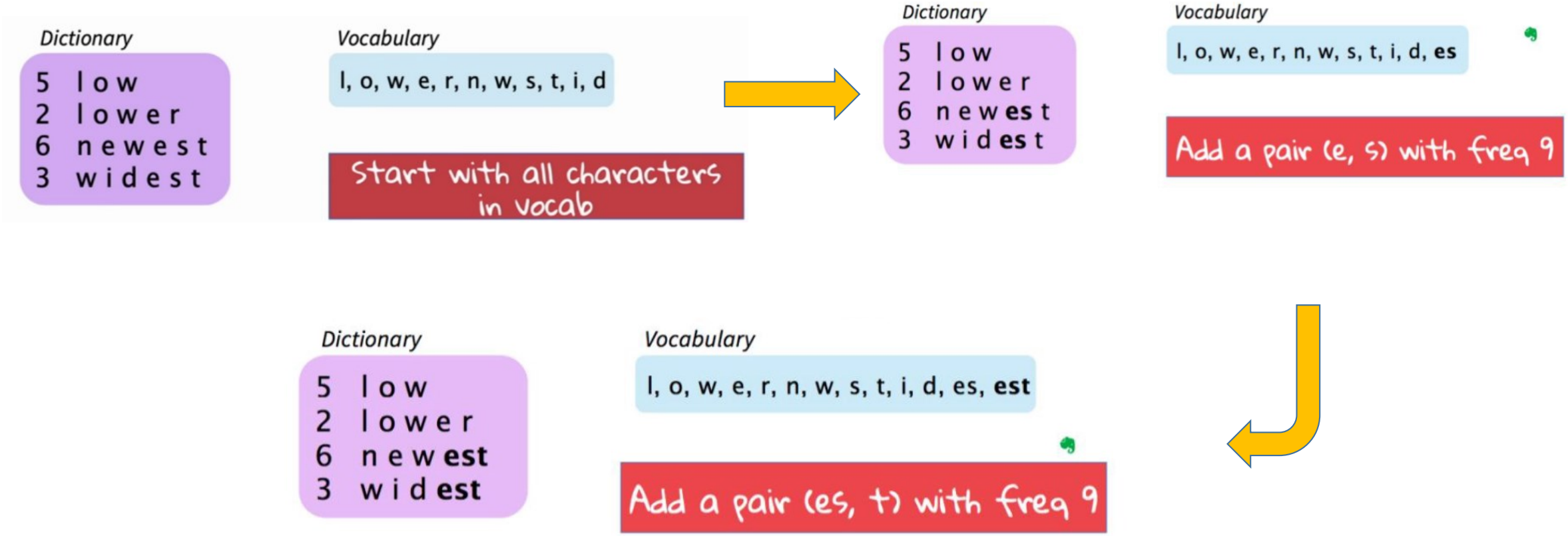
Строим словарь токенизатора:
Инициализируем словарь набором уникальных символов

Как получить набор таких сочетаний (токенов)?



Строим словарь токенизатора:
Инициализируем словарь набором уникальных символов
Добавляем наиболее частые пары символов

Как получить набор таких сочетаний (токенов)?



Строим словарь токенизатора:
Инициализируем словарь набором уникальных символов
Добавляем наиболее частые пары символов.
Добавляем наиболее частые тройки символов….
Повторяем, пока не достигнем заданного размера словаря.

Модели представления текстовых документов

- *Неструктурированная модель* – «мешок слов» (“bag of words”) – каждый термин рассматривается в качестве независимой случайной величины. Не учитываются возможные связи с другими словами в тексте.
- *Векторное представление слова (word embedding)*

Мешок слов

Векторная модель документа:

Документ:

$$\vec{X}_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_j^{(M)} \end{bmatrix}$$

Матрица «Документ-термин»:

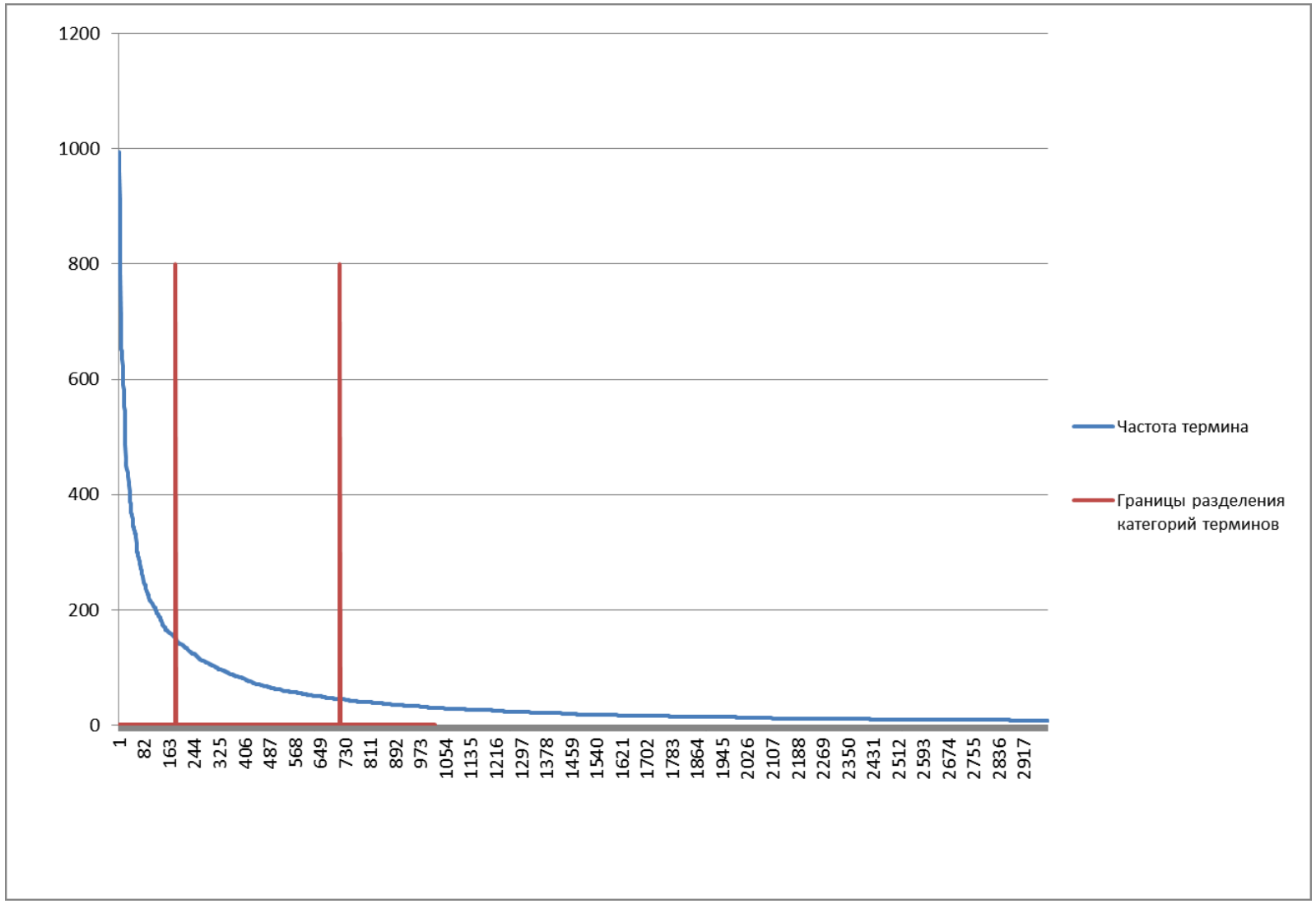
$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \dots & x_N^{(M)} \end{pmatrix}$$

$x_j^{(i)}$ - Вес *термина* i в документе j
i = 1..M, j = 1..N

Размерность матрицы крайне высокая, $M \to 10^4 - 10^5$

Вместо термина (или слова) могут использоваться токены или n-граммы – частный случай токенов при заданных заданных n:
Триграммы Hello world: Hel, ell, llo, lo ,o w, wo, orl, rld.

Выявление информативных признаков



Закон Ципфа:
$w_n = w_1/n$

Предварительная обработка документов

Стемминг и отсечение стоп-слов

Стемминг — это процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова. Наиболее известный алгоритм – Алгоритм Портера

Стоп-слова (шумовые слова) – слова, не несущие смысловой нагрузки – частицы, предлоги, местоимения,..

Семантическая интерпретация в системах компьютерного анализа текста

Описывается подход к построению семантического компонента в системах компьютерного анализа текста на естественном языке. Подход основан на применении специальных шаблонов к сети синтактико-семантических отношений между словами текста, которая строится синтаксическим анализатором. Шаблоны определяют способ интерпретации фрагментов сети в заданные фреймы с идентификацией участников ситуаций и их ролей.
Ключевые слова: компьютерный анализ текста, семантическая интерпретация, семантическая сеть, синтаксический анализ, фреймы.

Semantic Interpretation in Computer Text Analysis Systems

The article describes an approach to semantic component building in computer text analysis systems for a natural language text. The approach is based on applying special patterns to a net of syntactic and semantic relations between words in a text, which is formed by a syntactic parser. The patterns define the way to interpret parts of the net according to given frames with identification of participants of the situation and their roles.
Keywords: text mining, semantic interpretation, semantic network, syntactic parser, frames.

Определение весов терминов

Название	Формула
Логическое взвешивание	$x_j^{(i)} = \begin{cases} 1, & f_{ij} > 0 \\ 0, & f_{ij} = 0 \end{cases}$
Взвешивание частотой слова (term frequencies, tf)	$x_j^{(i)} = f_{ij}$
tf–idf - взвешивание (term frequencies – inverse document frequencies)	$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right)$
tfc - взвешивание	$x_j^{(i)} = \frac{f_{ij} \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[f_{ij} \log\left(\frac{N}{N_i}\right)\right]^2}}$

Определение весов терминов (2)

Название	Формула
ltc – взвешивание. Данный подход заключается в использовании логарифма частоты слова вместо f_{ij}. Это позволяет сократить характерный для большинства текстовых документов существенный разброс в частотах различных терминов	$x_j^{(i)} = \frac{\log(f_{ij}+1)\log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[\log(f_{ij}+1)\log\left(\frac{N}{N_i}\right)\right]^2}}$
atc – взвешивание. При таком взвешивании веса будут изменяться от 0,5 до 1, что в ряде случаев приводит к улучшению качества классификации, позволяя учесть значимые термины, имеющие редкую встречаемость в конкретной выборке	$x_j^{(i)} = \frac{(0{,}5+0{,}5\frac{f_{ij}}{\max f})\log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{k=1}^{Mk}\left[(0{,}5+0{,}5\frac{f_{ij}}{\max f})\log\left(\frac{N}{N_k}\right)\right]^2}}$

Кроме взвешивания применяются и другие методы выявления информативных терминов:

- Факторный и компонентный анализ (переход к новой системе признаков)
- Статистический подход (Хи-квадрат критерий)
- Теоретико-информационный подход

Факторный анализ (ФА) и Метод Главных Компонент (МГК)

- ФА: различные признаки являются одним и тем же явлением, следовательно можно создать новые переменные – «факторы», позволяющие «вскрыть» логическую структуру выборки.
- МГК: переход к новым переменным, которые являются линейной комбинацией исходных.

Проведение снижения размерности с помощью ФА и МГК особенно эффективно для отображения объектов в трехмерное пространство и на плоскость.

Статистический подход выявления информативных признаков

X \ Q	Q_1	Q_2	...	Q_K	$\sum_{k=1}^{K} n_{ik} = n_{i*}$
$x^{(1)}$	n_{11}	n_{12}	...	n_{1K}	n_{1*}
$x^{(2)}$	n_{21}	n_{22}	...	n_{2K}	n_{2*}
...	...	...	...	...	...
$x^{(M)}$	n_{M1}	n_{M2}	...	n_{MK}	n_{M*}
$\sum_{i=1}^{M} n_{ik} = n_{*k}$	n_{*1}	n_{*2}	...	n_{*K}	$n_{**} = N$

n_{ik} – клеточная частота – число объектов в выборке, обладающих данным сочетанием переменных $\{x^{(i)}, Q_k\}$

Проверяется гипотеза H_0: $n_{ik} - \hat{n}_{ik} = 0$

$$\hat{n}_{ik} = P(x^{(i)}, Q_k) = P(x^{(i)})P(Q_k) = N\frac{n_{i*}}{N}\cdot\frac{n_{*k}}{N} = \frac{n_{i*}n_{*k}}{N}$$

$$\chi^2 = \sum_{i=1}^{M}\sum_{k=1}^{K}\frac{(n_{ik} - \hat{n}_{ik})^2}{\hat{n}_{ik}}$$

Гипотеза о независимости отвергается с уровнем значимости α, если рассчитанная величина χ^2 превышает критическое значение $\chi^2_{\alpha,S}$

где S=(M-1)(K-1)

Частный случай Хи-квадрат критерия

Q_k / X	Принадлежность классу Q_k	Непринадлежность классу Q_k	$\sum$
Наличие признака $x^{(i)}$	A	B	$A+B$
Отсутствие признака $x^{(i)}$	C	D	$C+D$
$\sum$	$A+C$	$B+D$	N

$$\chi^2(x^{(i)},Q_k) = N\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1*}n_{2*}n_{*1}n_{2*}} = N\frac{(AD-CB)^2}{(A+B)(C+D)(A+C)(B+D)}$$

$$\chi^2_{средний}(x^{(i)}) = \sum_{k=1}^{K} P(Q_k)\chi^2(x^{(i)},Q_k)$$

$$\chi^2_{\max 1}(x^{(i)}) = \max_{k=1}^{K} P(Q_k)\chi^2(x^{(i)},Q_k)$$

$$\chi^2_{\max 2}(x^{(i)}) = \max_{k=1}^{K} \chi^2(x^{(i)},Q_k)$$

$$P(Q_k) = \frac{N_k}{N}$$

Недостатки χ^2 - критерия:

- Вычислительная сложность
- Невысокая точность для редких терминов

Критерий взаимной информации (Mutual information)

Взаимная информация, как среднее количество информации, содержащееся в X относительно Q:

$I(X,Q) = H(X) - H(X \mid Q)$, где $H(X), H(X \mid Q) -$ соответственно энтропия и условная энтропия.

$$I(X,Q) = -\sum_{i=1}^{M} P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^{M} \sum_{k=1}^{K} P(x^{(i)}, Q_k) \log \frac{P(x^{(i)}, Q_k)}{P(Q_k)}$$

$$P(x_i) = P(Q_1)P(x^{(i)} \mid Q_1) + P(Q_2)P(x^{(i)} \mid Q_2) + \ldots + P(Q_K)P(x^{(i)} \mid Q_K) \quad = \quad P(x^{(i)}) = \sum_{k=1}^{K} P(x^{(i)}, Q_k)$$

Критерий взаимной информации (Mutual information) (2)

$$I(X,Q) = -\sum_{i=1}^{M} P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^{M}\sum_{k=1}^{K} P(x^{(i)},Q_k) \log \frac{P(x^{(i)},Q_k)}{P(Q_k)} =$$

$$= \sum_{i=1}^{M}\sum_{k=1}^{K} P(x^{(i)},Q_k) \log \frac{P(x^{(i)},Q_k)}{P(Q_k)P(x^{(i)})}.$$

$$P(x^{(i)}) = \frac{A+B}{N}$$

$$P(Q_k) = \frac{A+C}{N}$$

$$P(x^{(i)},Q_k) = \frac{A}{N}$$

$$MI(x^{(i)},Q_k) = \log_2 \frac{AN}{(A+B)(A+C)}$$

$$I_{сред}(X,Q) = \sum_{k=1}^{K} P(Q_k) MI(x^{(i)},Q_k)$$

$$I_{\max}(X,Q) = \max_{k=1}^{K} \{MI(x^{(i)},Q_k)\}$$

Данный критерий, в отличие от Хи-квадрат, большие веса дает редким признакам