Лекция 5: Векторизация текста и модель «мешка слов»

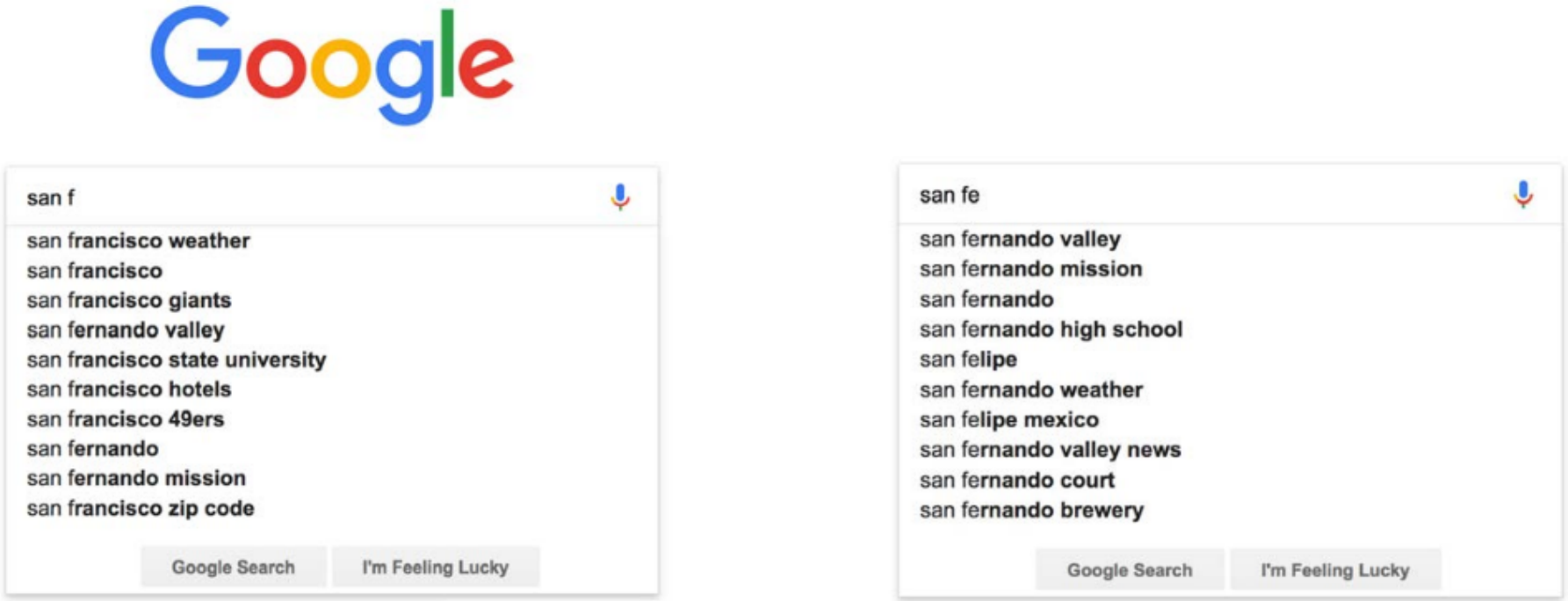
14 апреля 2026 г.

Почему именно обработка естественного языка (NLP)?

- Человеческие знания (по большей части) представляют собой текст на естественном языке.
- Интернет — это (по большей части) текст на естественном языке.
- Человеческое общение (по большей части) представляет собой текст на естественном языке.
- Культурная продукция представляет собой (в основном) текст на естественном языке.

Представьте, если бы система могла автоматически считывать и «понимать» всё это.

NLP применяется повсюду вокруг нас.



По данным Google, функция автозаполнения

- Экономит 200 лет времени, затрачиваемого на набор текста, каждый день
- Сделала возможным использование мобильных устройств

NLP применяется повсюду вокруг нас.

Вы:
Напишите лимерик о красоте и силе глубокого обучения.

Google ИИ:
Массив из нейронов и веса
Раскрыл нам красоты процесса.
В глубоких слоях
Повержен был страх —
Прозрела машина в три среза!

NLP обладает огромным потенциалом для создания более интеллектуальных продуктов и услуг.

Эта, казалось бы, простая возможность охватывает широкий спектр применений.

Примеры применения: Классификация текста

Классификация текста для

- Анализ настроения
- Маршрутизации
- Намерения
- Фильтрации
- ...

Примеры применения: Извлечение текста

Извлечение данных из текста в свободной форме

- Финансовые данные компании из новостной статьи
- Имя и контактная информация клиента из чата
- Коды заболеваний и лекарств из медицинских записей врача
- ...

Примеры применения: Резюмирование текста

Свести в краткое длинный текст

- Маркированные списки
- Аннотации
- Заголовки
- ...

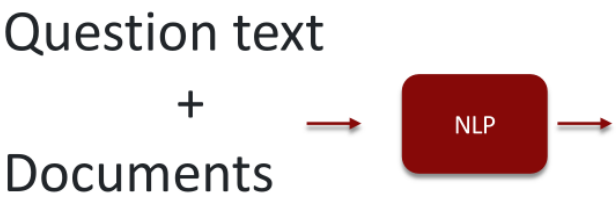
Примеры применения: Генерация текста

- Маркетинговые тексты
- Электронные письма для продаж
- Обзоры рынка
- Описания вакансий
- Публикации в социальных сетях
- Эссе для поступления в колледж
- ...

Примеры применения: Генерация кода

Код, реализующий ввод текста

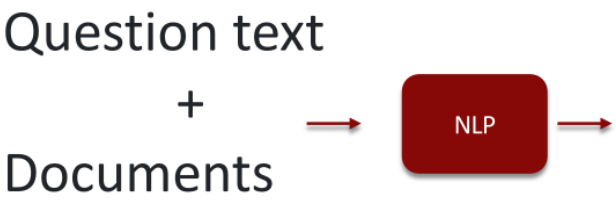
Примеры применения: система вопросов и ответов.



Чат-боты для:

- Медицинско-юридических вопросов
- Колл-центров
- Соблюдения нормативных требований
- Заполнения форм
- Автоматизации рабочих процессов
- ...

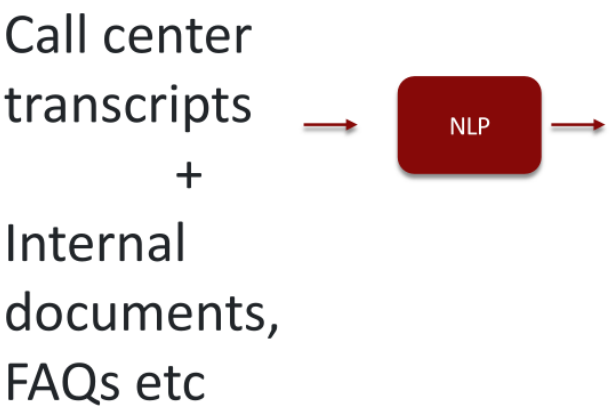
Примеры применения: система вопросов и ответов.



Чат-боты для:

- Медицинско-юридических вопросов
- Колл-центров
- Соблюдения нормативных требований
- Заполнения форм
- Автоматизации рабочих процессов
- ...

Пример области применения: Оптимизация колл-центров



- Основные причины недовольства клиентов
- Какие меры кажутся эффективными?
- Чем отличаются лучшие агенты поддержки от остальных?
- Как оценить взаимодействие этого агента с клиентом X?
- Как изменить сценарий колл-центра в зависимости от ситуации?
- Как обучать агента в режиме реального времени?
- ...

Потенциал NLP сегодня широко признан в общественном дискурсе благодаря стремительному росту популярности больших языковых моделей (LLM).

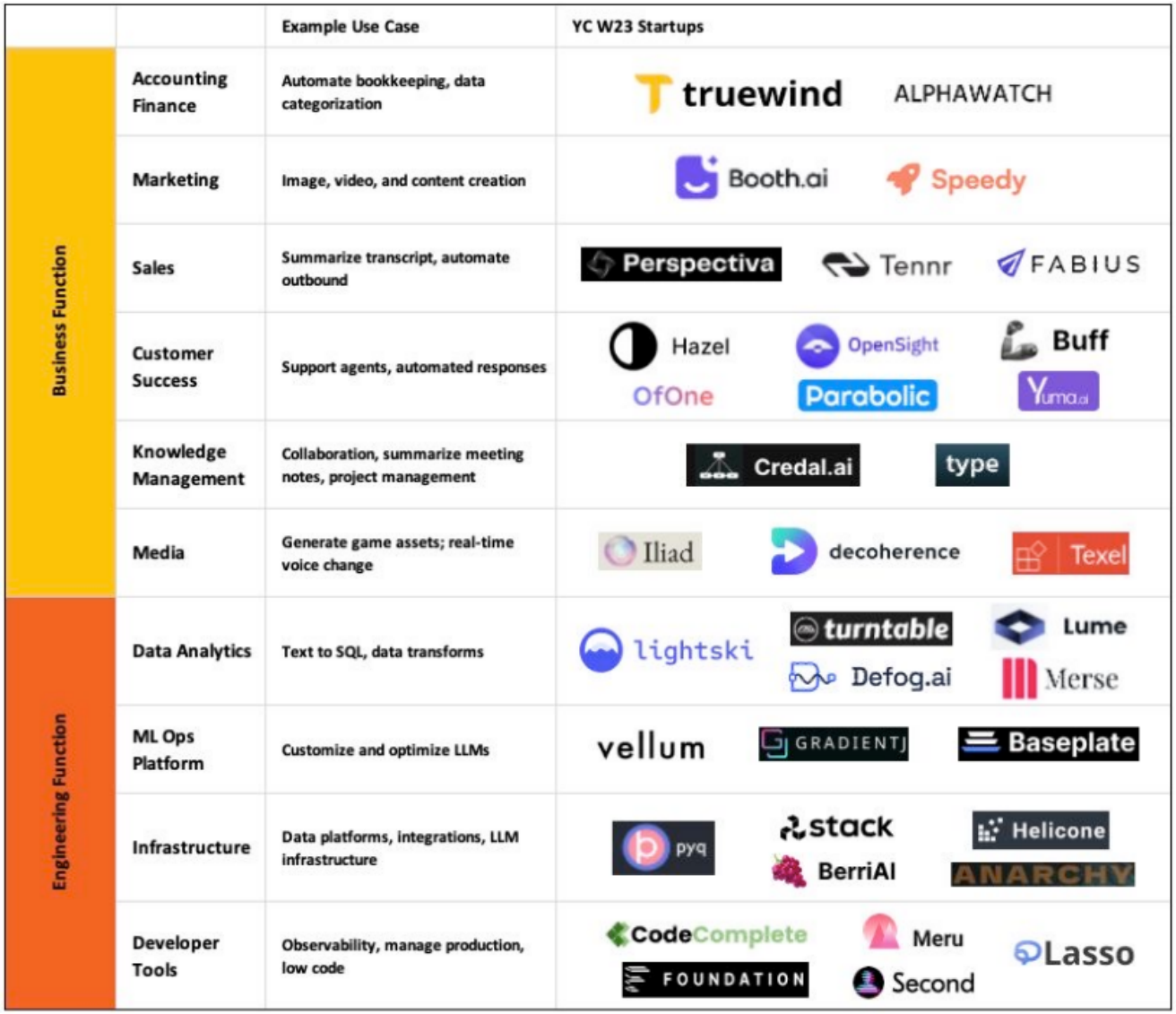
https://www.anthropic.com/index/introducing-claude

🏆 LMSYS Chatbot Arena Leaderboard

Rank	Model	Arena Elo
1	GPT-4-1106-preview	1254
2	GPT-4-0125-preview	1253
3	Bard (Gemini Pro)	1218
4	GPT-4-0314	1191
5	GPT-4-0613	1164
6	Mistral Medium	1152
7	Claude-1	1150
8	Qwen1.5-72B-Chat	1147
9	Claude-2.0	1132
10	Gemini Pro (Dev API)	1122
11	Claude-2.1	1120
12	Mixtral-8x7b-Instruct-v0.1	1120
13	GPT-3.5-Turbo-0613	1118
14	Gemini Pro	1115
15	Yi-34B-Chat	1111

https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

«золотая лихорадка» стартапов, на создание продуктов и услуг на основе NLP.



		Example Use Case	YC W23 Startups
Business Function	Accounting Finance	Automate bookkeeping, data categorization	truewind, ALPHAWATCH
	Marketing	Image, video, and content creation	Booth.ai, Speedy
	Sales	Summarize transcript, automate outbound	Perspectiva, Tennr, FABIUS
	Customer Success	Support agents, automated responses	Hazel, OpenSight, Buff, OfOne, Parabolic, Yuma.ai
	Knowledge Management	Collaboration, summarize meeting notes, project management	Credal.ai, type
	Media	Generate game assets; real-time voice change	Iliad, decoherence, Texel
Engineering Function	Data Analytics	Text to SQL, data transforms	lightski, turntable, Lume, Defog.ai, Merse
	ML Ops Platform	Customize and optimize LLMs	vellum, GRADIENTJ, Baseplate
	Infrastructure	Data platforms, integrations, LLM infrastructure	pyq, stack, Helicone, BerriAI, ANARCHY
	Developer Tools	Observability, manage production, low code	CodeComplete, Meru, Lasso, FOUNDATION, Second

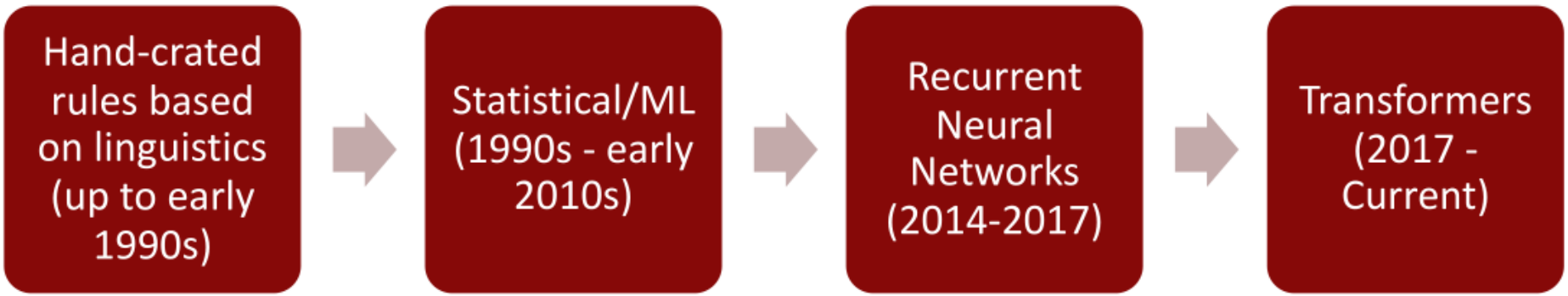
Корпоративные поставщики спешат добавить функции NLP в свои продукты.

Salesforce Announces Einstein GPT, the World's First Generative AI for CRM

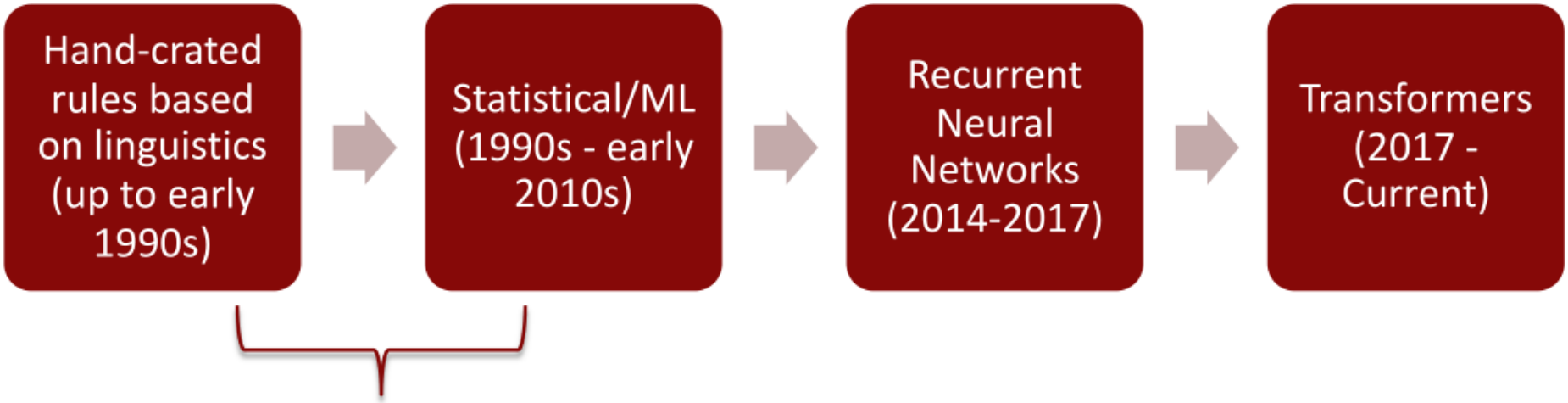
https://www.salesforce.com/news/press-releases/2023/03/07/einstein-generative-a/

Путь прогресса в NLP – Как мы сюда попали?

Путь прогресса в NLP



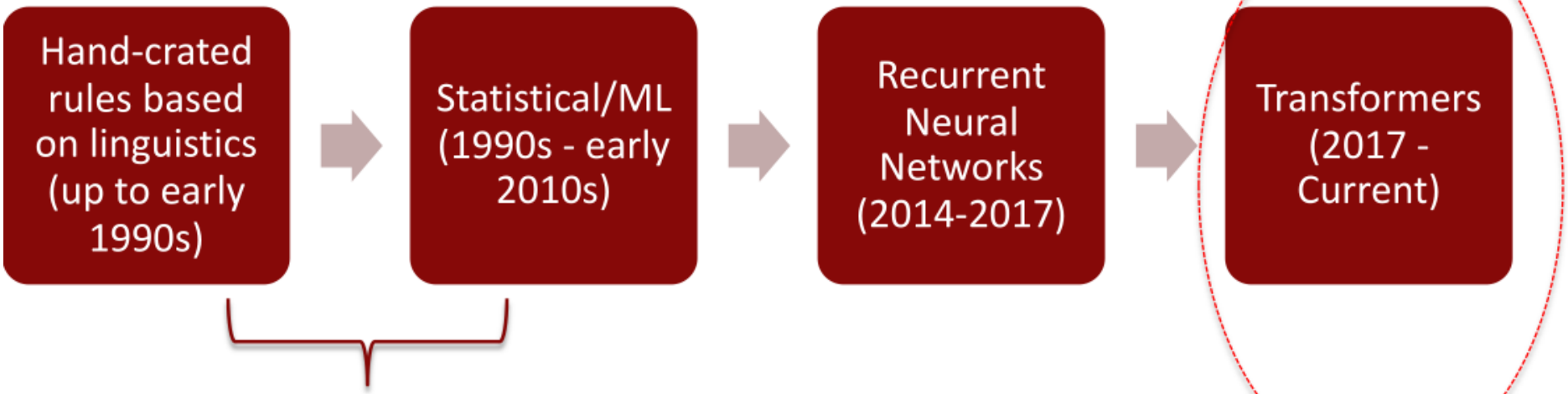
Путь прогресса в NLP



"Every time I fire a linguist, the performance of the speech recognizer goes up."

Frederick Jelinek

Путь прогресса в NLP



"Every time I fire a linguist, the performance of the speech recognizer goes up."

Frederick Jelinek

We will *leapfrog* to this in HODL!

Обзор проблемы с высоты птичьего полета.

Обзор проблемы с высоты птичьего полета.

$$x = text$$

$$y = text, labels, numbers...$$

$$w = weights$$

$$f(x, w) = a\ deep\ neural\ network$$

Обзор проблемы с высоты птичьего полета.

Ключевые вопросы:

Обзор проблемы с высоты птичьего полета.

Ключевые вопросы:

- Как представить x. Сегодня мы сосредоточимся на этом.

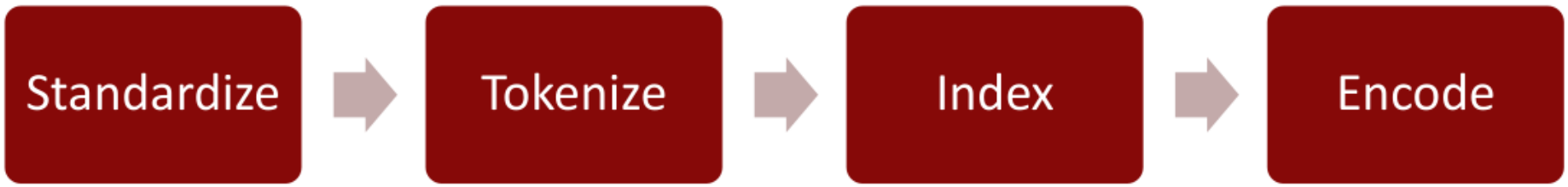
Обзор проблемы с высоты птичьего полета.

Ключевые вопросы:

- Как представить x. Сегодня мы сосредоточимся на этом.
- (На следующей неделе) Какая архитектура нейронной сети лучше всего подходит для обработки текста?

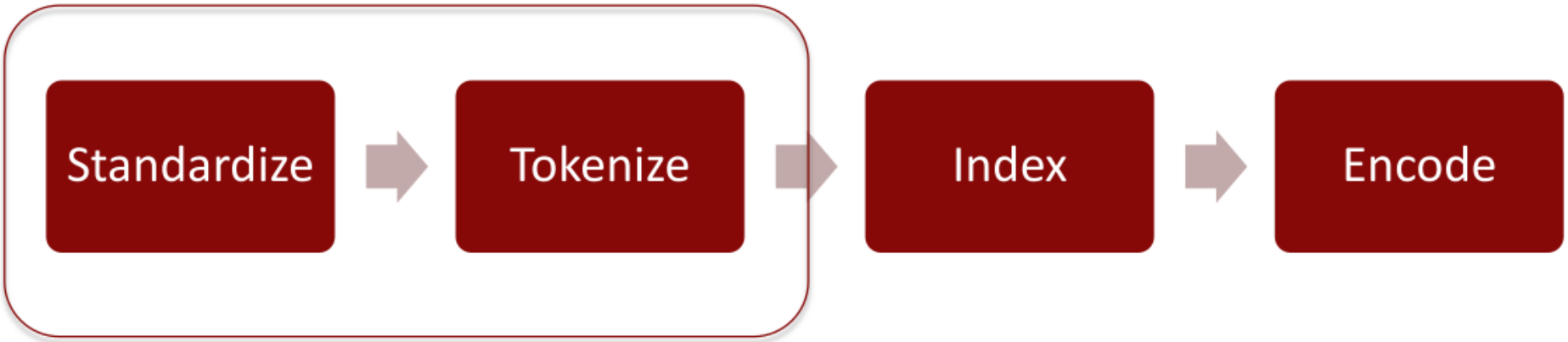
Основы обработки

Базовая предварительная обработка



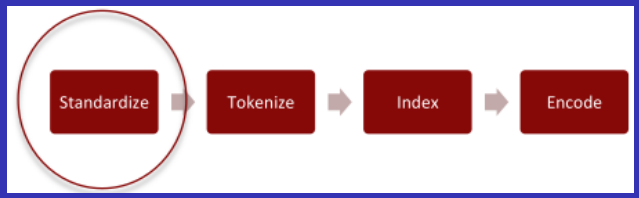
Этот процесс называется векторизацией текста.

Базовая предварительная обработка



Сначала мы выполняем эти два шага для каждого предложения в нашем обучающем наборе данных.

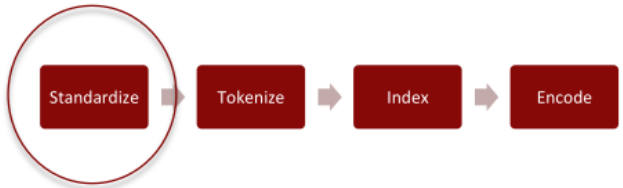
Базовая предварительная обработка



Стандартизация

- Удаление заглавных букв, часто знаков препинания и диакритических знаков (почти всегда)
- Удаление «стоп-слов», например, a, the, it, ... (часто)
- Стемминг (например, ate, eaten, eating, eaten > [eats]) (иногда)

Базовая предварительная обработка



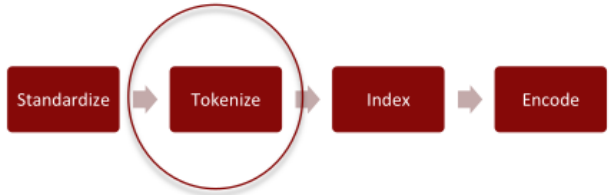
Стандартизация

- Удаление заглавных букв, часто знаков препинания и диакритических знаков (почти всегда)
- Удаление «стоп-слов», например, a, the, it, ... (часто)
- Стемминг (например, ate, eaten, eating, eaten > [eats]) (иногда)

Hola! What do you picture when you think of traveling to Mexico? Sipping a real margarita while soaking up the sun on a laid-back beach in Puerto Vallarta?

hola what do you picture when you [thinks] of [travels] to mexico [sips] real margarita while [soaks] up sun on laidback beach in puerto vallarta

Базовая предварительная обработка



Токенизация

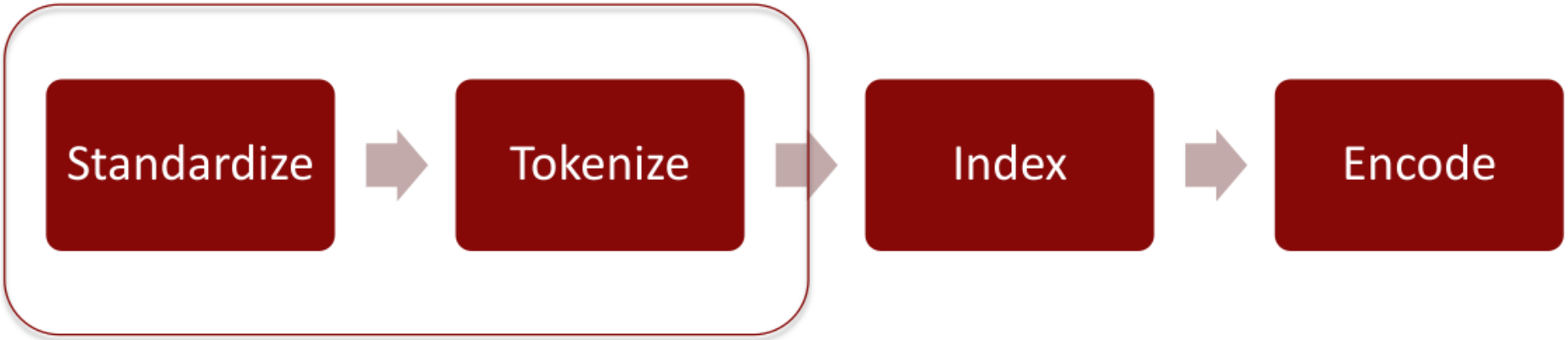
- Как правило, каждую строку разделяют по пробелам, то есть каждое слово является токеном.
- [Решение по дизайну] определить, сколько последовательных слов составляют токен.

hola what do you picture when you [thinks] of [travels] to mexico [sips] real margarita while [soaks] up sun on laidback beach in puerto vallarta

“hola”, “what”, “do”, “you”, “picture”, “when”, “you”, “[thinks]”, “of”, ”[travels]”, “to”, “mexico”, “[sips]”, “real”, “margarita”, “while”, “[soaks]”, “up”, “sun”, “on”, “laidback”, “beach”, “in”, “puerto”, “vallarta”

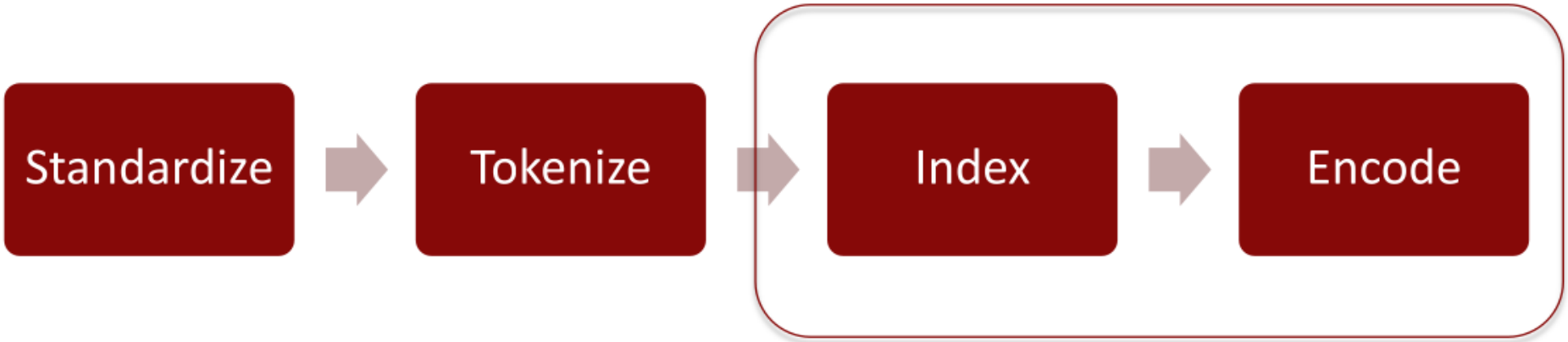
Описанные нами стандартизация и токенизация являются хорошим вариантом по умолчанию для многих задач обработки естественного языка, но имеют недостатки, особенно для задач генерации текста. Современные LLM используют другие схемы (например, кодирование пар байтов), которые мы опишем позже.

Базовая предварительная обработка



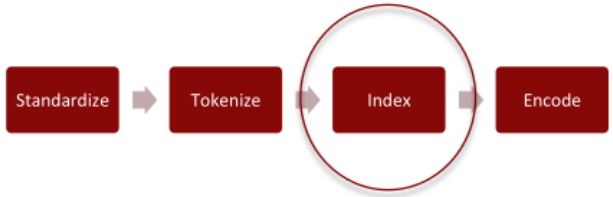
- Если это сделать для каждого предложения в нашем обучающем наборе данных, мы получим список отдельнгых токенов, = нашему словарю.

Базовая предварительная обработка



- Если это сделать для каждого предложения в нашем обучающем наборе данных, мы получим список отдельнгых токенов, = нашему словарю.
- Теперь переходим к третьему и четвертому этапам. На этих этапах мы работаем только со словарём.

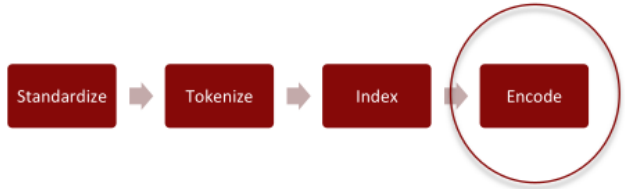
Базовая предварительная обработка



Индексирование: Каждому уникальному токену в словаре присваивается уникальное целое число

Token	Integer
<UNK>	0*
a	1
aardvark	2
...	
zebra	50000

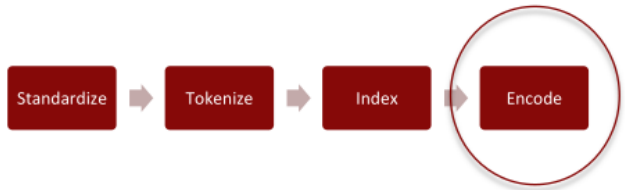
Базовая предварительная обработка



Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

		→	
a	1	→	Vector
	2	→	
...			
zebra	50000	→	Vector

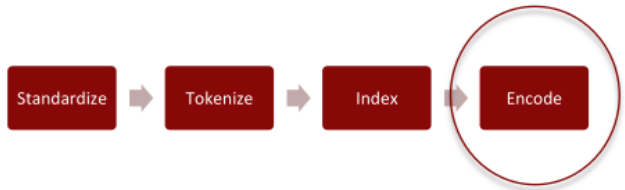
Базовая предварительная обработка



Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

- Самый простой способ сделать это —

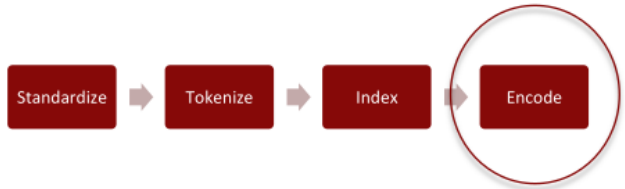
Базовая предварительная обработка



Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

- Самый простой способ сделать это — one-hot encoding.

Базовая предварительная обработка

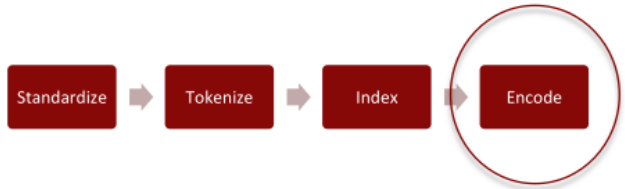


Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

- Самый простой способ сделать это — one-hot encoding.

$$\text{<UNK>} \longrightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{a} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \cdots$$

Базовая предварительная обработка



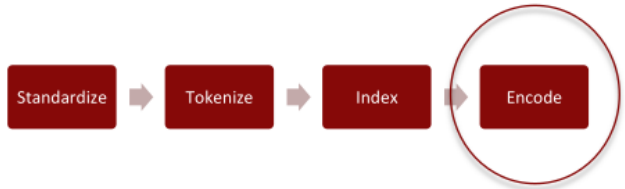
Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

- Самый простой способ сделать это — one-hot encoding.

$$\text{<UNK>} \longrightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{a} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \cdots$$

- Размерность вектора кодирования равна количеству различных токенов в тексте + один для <UNK>.

Базовая предварительная обработка



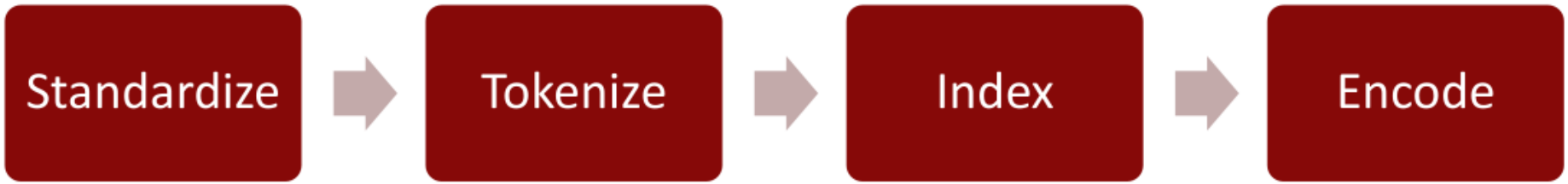
Кодирование: Каждому целому числу в нашем словаре мы присваиваем вектор.

- Самый простой способ сделать это — one-hot encoding.

$$\text{<UNK>} \longrightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{a} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \cdots$$

- Размерность вектора кодирования равна количеству различных токенов в тексте + один для <UNK>.
- Это называется размером «словаря».

Базовая предварительная обработка



На данном этапе

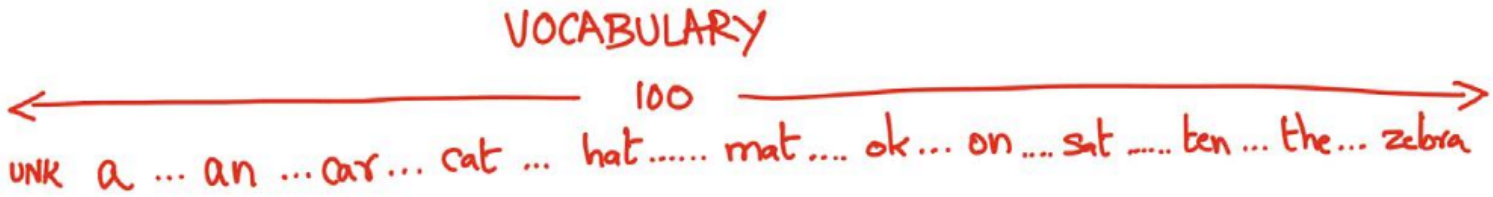
- мы создали словарь на основе обучающего корпуса, и
- каждому уникальному токену в нашем словаре был присвоен one-hot вектор.

Базовая предварительная обработка завершена.

Далее: Как подготовить новое входное предложение для «подачи» в DNN.

Далее: Как подготовить новое входное предложение для «подачи» в DNN.

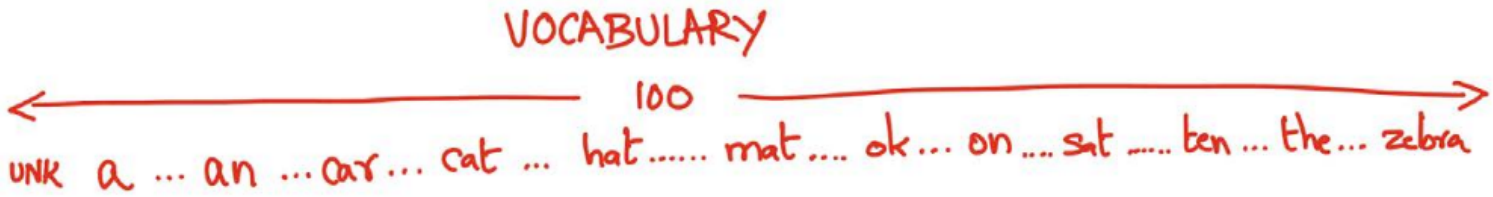
- Допустим, мы завершили STIE[1] на обучающем корпусе, и размер нашего словаря составляет 100 слов.



[1] изменили на нижний регистр, удалили знаки препинания, оставили стоп-слова как есть, не использовали стемминг

Далее: Как подготовить новое входное предложение для «подачи» в DNN.

- Допустим, мы завершили STIE[1] на обучающем корпусе, и размер нашего словаря составляет 100 слов.

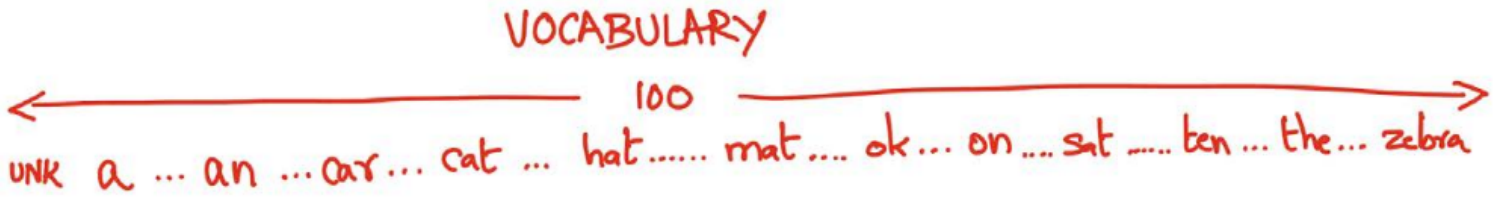


- На вход поступает текстовая строка — «The cat sat on the mat» — и мы обрабатываем её с помощью STIE.

[1] изменили на нижний регистр, удалили знаки препинания, оставили стоп-слова как есть, не использовали стемминг

Далее: Как подготовить новое входное предложение для «подачи» в DNN.

- Допустим, мы завершили STIE[1] на обучающем корпусе, и размер нашего словаря составляет 100 слов.



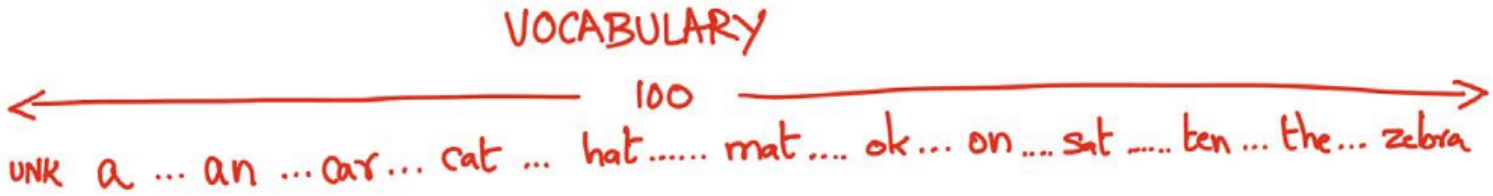
- На вход поступает текстовая строка — «The cat sat on the mat» — и мы обрабатываем её с помощью STIE.

- Результатом является таблица с A строками и B столбцами. Чему равны A и B?

[1] изменили на нижний регистр, удалили знаки препинания, оставили стоп-слова как есть, не использовали стемминг

Далее: Как подготовить новое входное предложение для «подачи» в DNN.

- Допустим, мы завершили STIE на обучающем корпусе, и размер нашего словаря составляет 100 слов.

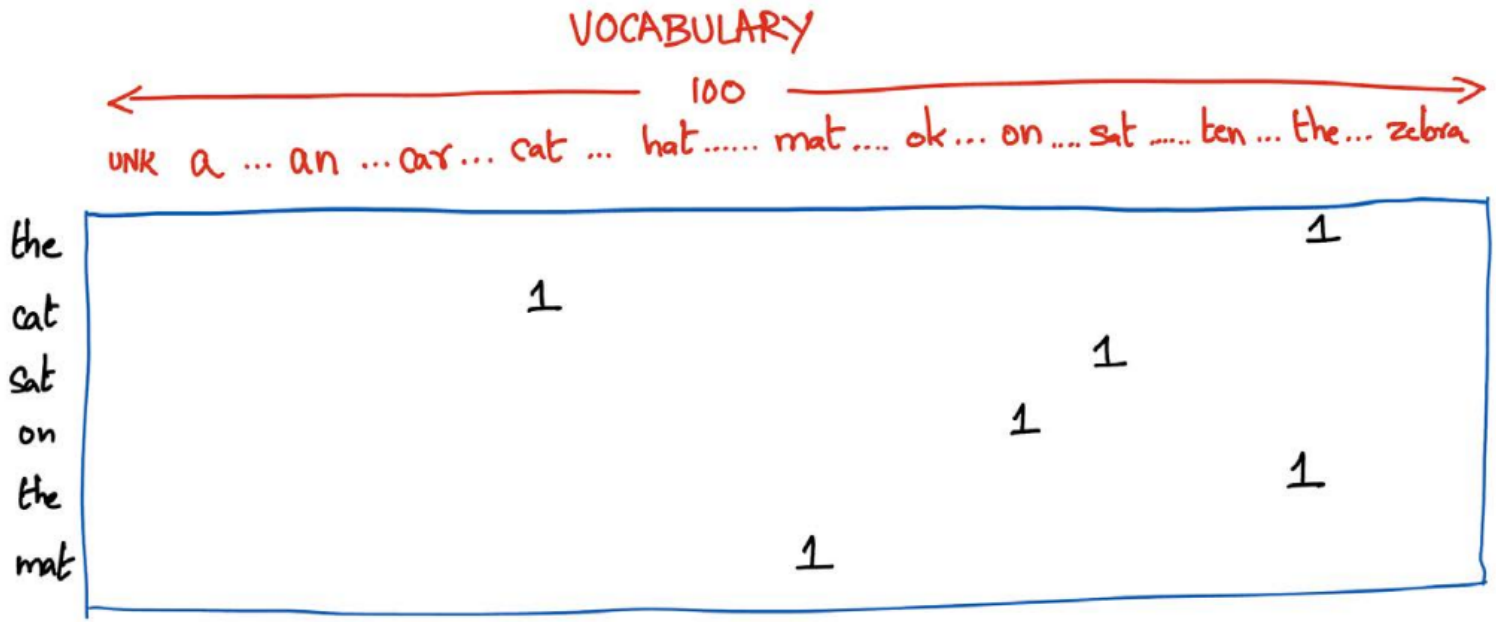


- На вход поступает текстовая строка — «The cat sat on the mat» — и мы обрабатываем её с помощью STIE.

- Ввывод таблицап 6×100

Как подготовить новое входное предложение для «подачи» в DNN.

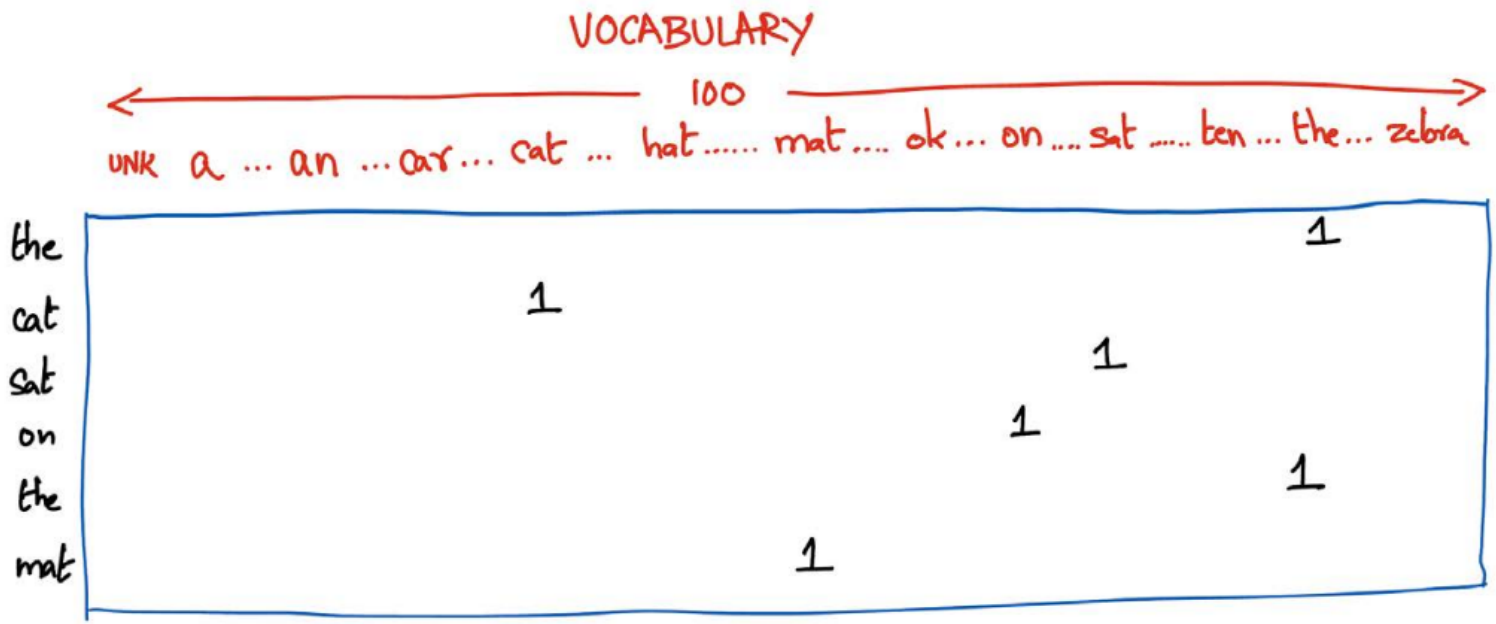
Выходная таблица[2]



[2]Не отображаем нули во избежание беспорядка

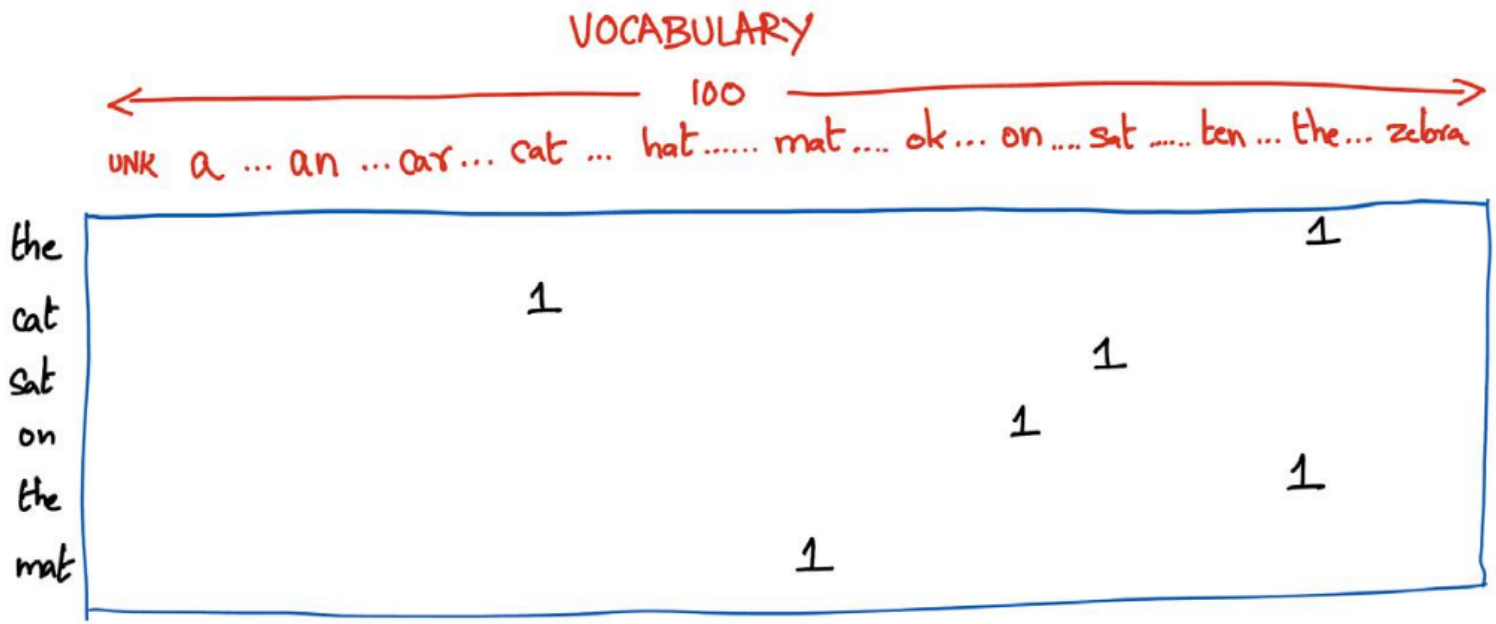
Как подготовить новое входное предложение для «подачи» в DNN.

- Как лучше всего «передать» эту таблицу чисел размером 6 x 100 в нейронную сеть?



Как подготовить новое входное предложение для «подачи» в DNN.

- Как лучше всего «передать» эту таблицу чисел размером 6 x 100 в нейронную сеть?
- Можно ли передать эту таблицу в DNN в неизмененном виде?



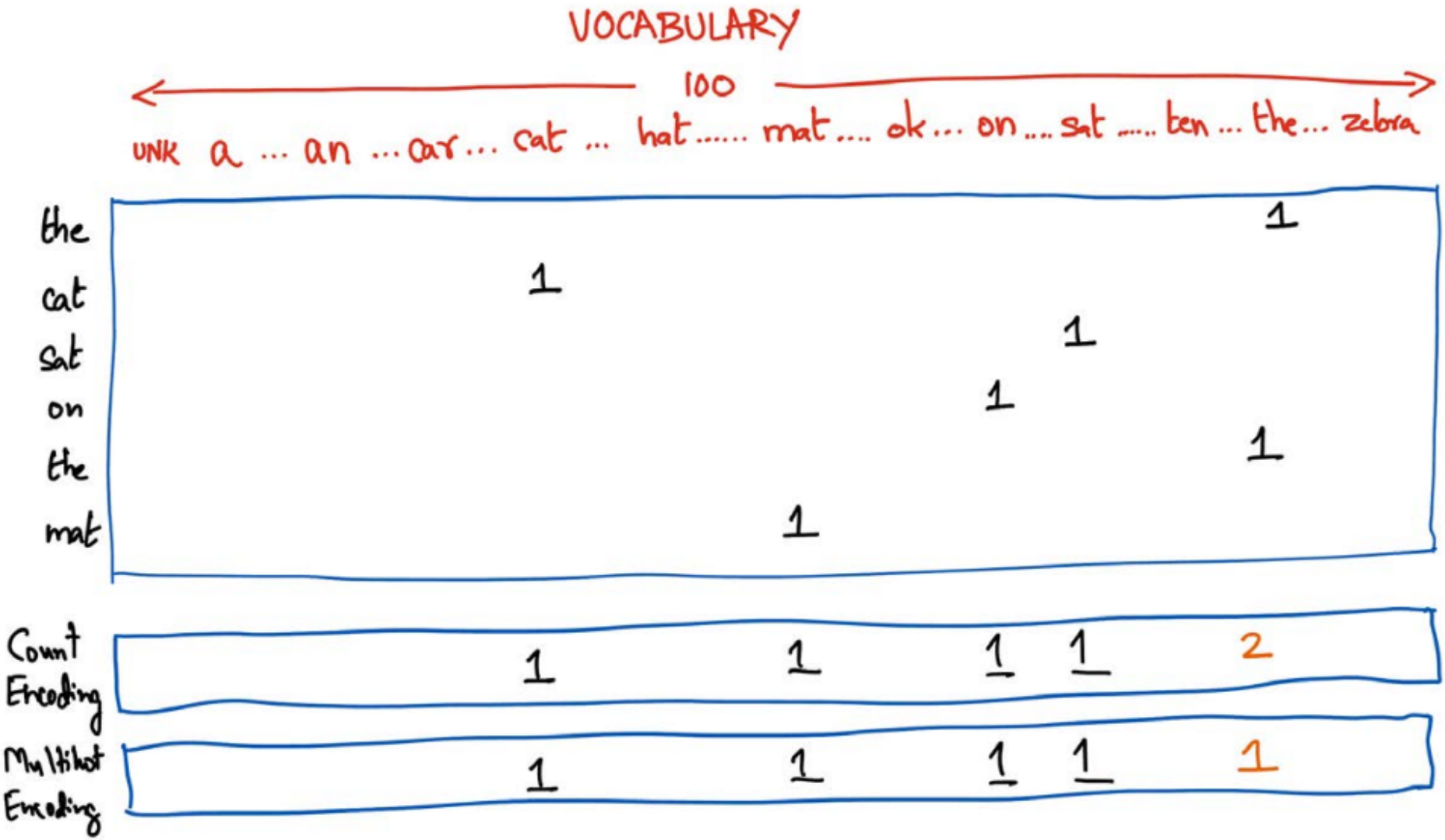
Как подготовить новое входное предложение для «подачи» в DNN.

- Как лучше всего «передать» эту таблицу чисел размером 6 x 100 в нейронную сеть?
- Можно ли передать эту таблицу в DNN в неизмененном виде?
- Сложность: каждое входящее предложение может содержать разное количество слов, то есть иметь разную длину. Было бы неплохо иметь входные данные фиксированной длины.
- А что, если мы «объединим» векторы?

Как подготовить новое входное предложение для «подачи» в DNN.

- Как лучше всего «передать» эту таблицу чисел размером 6 x 100 в нейронную сеть?
- Можно ли передать эту таблицу в DNN в неизмененном виде?
- Сложность: каждое входящее предложение может содержать разное количество слов, то есть иметь разную длину. Было бы неплохо иметь входные данные фиксированной длины.
- А что, если мы «объединим» векторы?
  - Сложим векторы. Это называется “count encoding”.
  - "OR"векторы. Это назыается “multi-hot encoding”.

Пример: Count and Multi-hot Encoding



Как подготовить новое входное предложение для «подачи» в DNN.

- Как лучше всего «передать» эту таблицу чисел размером 6 x 100 в нейронную сеть?
- Можно ли передать эту таблицу в DNN в неизмененном виде?
- Сложность: каждое входящее предложение может содержать разное количество слов, то есть иметь разную длину. Было бы неплохо иметь входные данные фиксированной длины.
- А что, если мы «объединим» векторы?
  - Сложим векторы. Это называется “count encoding”.
  - "OR"векторы. Это назыается “multi-hot encoding”.
- Этот метод агрегирования называется моделью “мешка слов”

Есть ли у подхода «мешок слов» какие-либо недостатки?

Есть ли у подхода «мешок слов» какие-либо недостатки?

- Мы теряем смысл, заложенный в порядке слов (т.е., теряем «последовательность»).

Есть ли у подхода «мешок слов» какие-либо недостатки?

- Мы теряем смысл, заложенный в порядке слов (т.е., теряем «последовательность»).
- Если словарь очень длинный, каждый ввод — независимо от количества токенов — будет представлять собой вектор, длина которого равна размеру словаря.

Есть ли у подхода «мешок слов» какие-либо недостатки?

- Мы теряем смысл, заложенный в порядке слов (т.е., теряем «последовательность»).
- Если словарь очень длинный, каждый ввод — независимо от количества токенов — будет представлять собой вектор, длина которого равна размеру словаря.
  - Эту проблему можно частично смягчить, выбирая только наиболее часто встречающиеся слова.
  - Это увеличивает количество весов, которые модель должна изучить, а следовательно, и время вычислений, а также риск переобучения.

Задание 1 по NLP

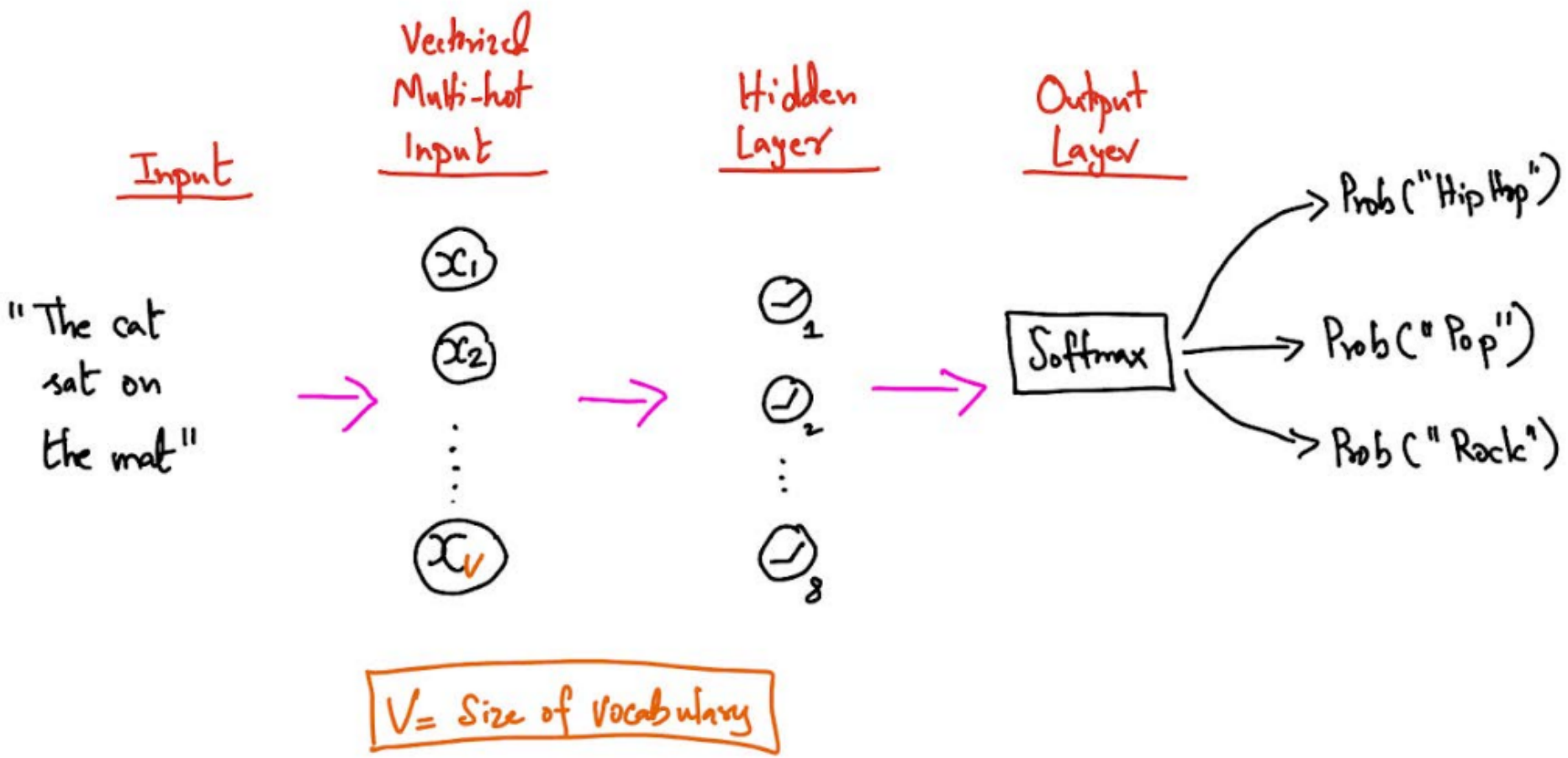
Применение: Прогнозирование жанров

I grew up on the crime side, the New York Times side
Stayin' alive was no jive
Had secondhands, Mom's bounced on old man
So then we moved to Shaolin land

I walked through the door with you
The air was cold
But something about it felt like home somehow
And I, left my scarf there at your sisters house

Можете ли вы отнести каждый из приведенных выше куплетов к жанрам хип-хоп, рок или поп?

Какой самый простой классификатор на основе NN мы можем построить?



Colab (предварительная обработка текста, мешок слов и биграммы)

Ссылка на colab