



Программа Перезапуск

Модуль ML.

Занятие 1

Преподаватель: Марат Гарафутдинов

Преподаватель:

Марат Гарафутдинов

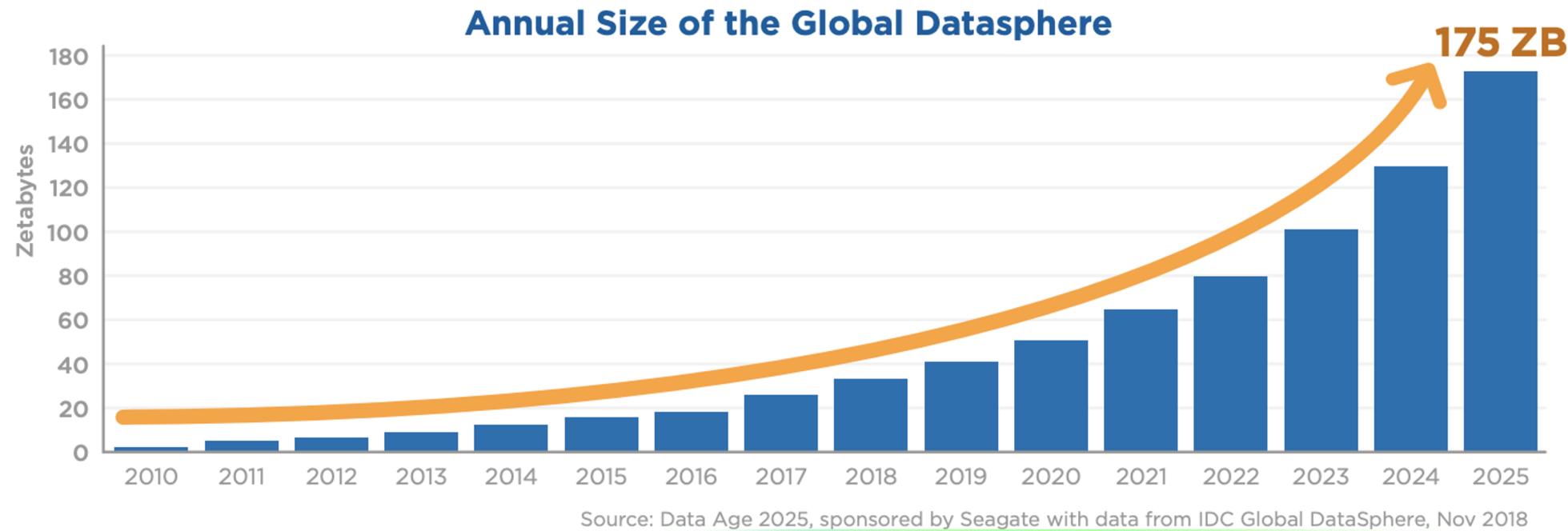
- Более 6 лет в анализе данных
 - Head of DA & Research at Mamba
 - ex Senior Data Analyst at Sber
 - Ex: QIWI Group, NTechLab, Delta Solutions etc.
 - Альма-матер: НИЯУ МИФИ (А & ИФЭБ)
 - Kaggle Competitions Expert
-
- **Telegram: @fffffmistty**
(предпочтительно)
 - **WhatsApp: +7 916 250 97 36**
 - **[linkedin.com/in/marat-g-13218712a/](https://www.linkedin.com/in/marat-g-13218712a/)**



Формат обучения

- На занятие обсуждаем, смотрим, участвуем и задаем вопросы
- После каждого занятия идет ДЗ
- Успеваемость в процессе модуля контролируется
- Всегда, если что-то непонятно, то можно написать в Telegram/чат –
это приветствуется и абсолютно необходимо на начальном этапе обучения

Зачем это нужно



Количество данных растет экспоненциально

Каждый год +20%-25%

Зачем это нужно

- Индустрия анализа данных растет >15% в год

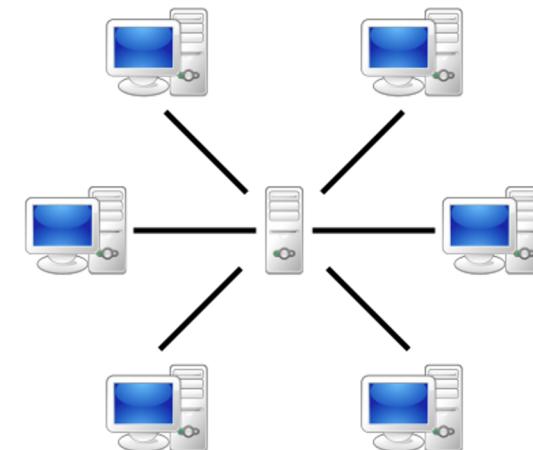
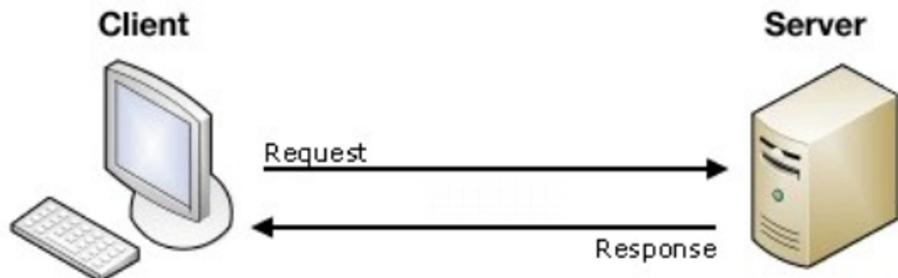
Зачем это нужно

- Индустрия анализа данных растет >15% в год
- Сейчас самые разные индустрии включают навыки работы с данными (юристы врачи)

Зачем это нужно

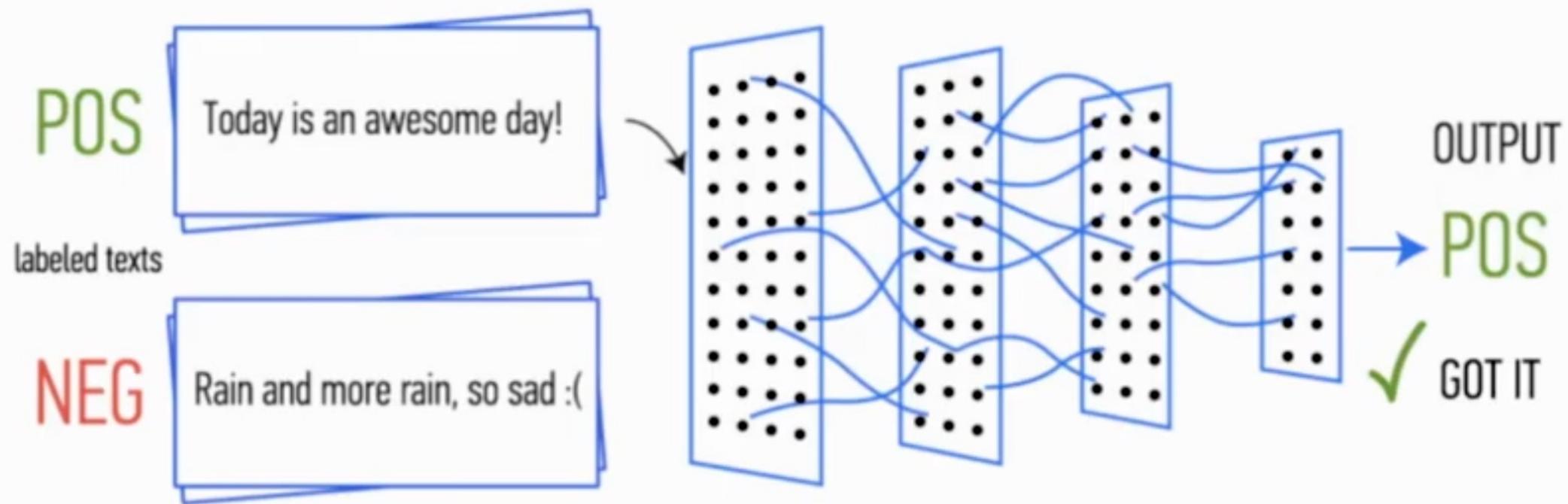
- Индустрия анализа данных растет >15% в год
- Сейчас самые разные индустрии включают навыки работы с данными (юристы врачи)
- Аналитик данных – одна из самых востребованных работ в современном мире [link]

Удаленный сервер

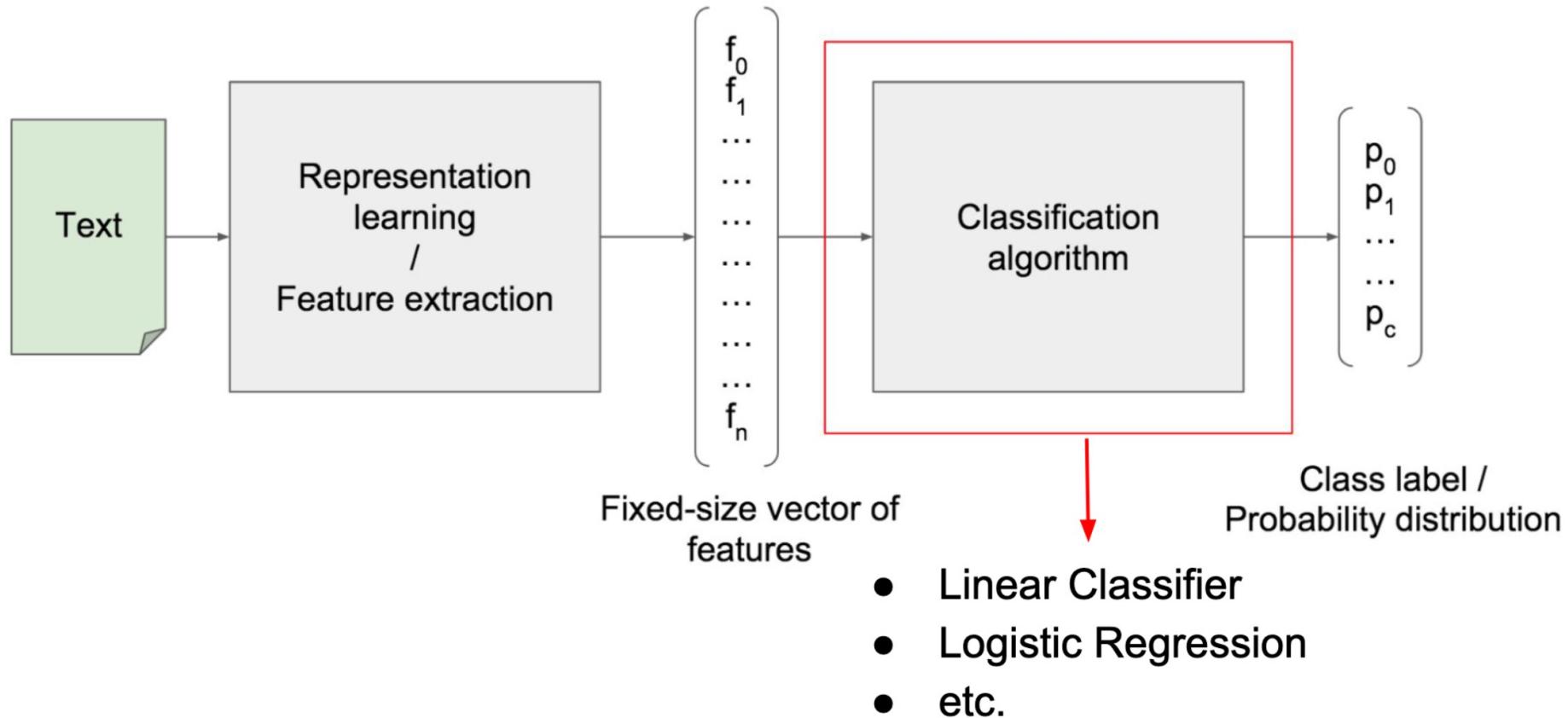


Удаленный сервер – это компьютер, к которому вы можете получить доступ посредством глобальной сети.

NLP (Natural Language Processing)



NLP (Natural Language Processing)

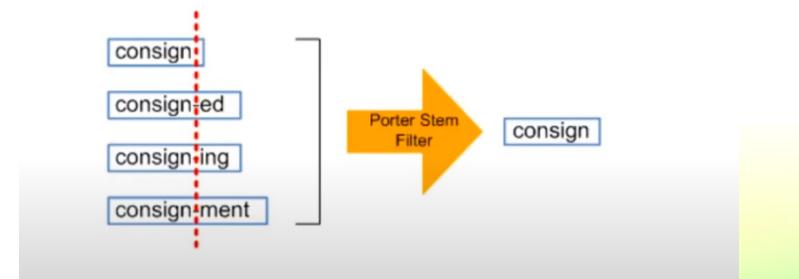


Text processing

Поскольку многие слова в тексте являются копиями других слов, просто находясь в различных формах (например, кот и коты пишутся по-разному, но имеют один и тот же смысл), то для упрощения модели применяется предобработка текста. Вот два основных способа:

Stemming: процесс нахождения основы слова (не обязательно корня) по заданным правилам.

Примеры: котик → кот; data → dat ;

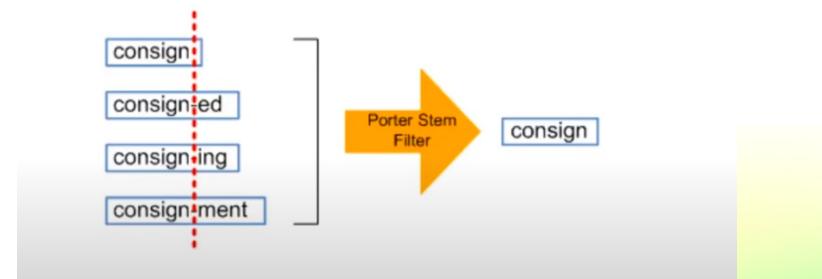


Text processing

Поскольку многие слова в тексте являются копиями других слов, просто находясь в различных формах (например, кот и коты пишутся по-разному, но имеют один и тот же смысл), то для упрощения модели применяется предобработка текста. Вот два основных способа:

Stemming: процесс нахождения основы слова (не обязательно корня) по заданным правилам.

Примеры: котик → кот; data → dat ;



Lemmatization: процесс приведения слова к его нормальной форме

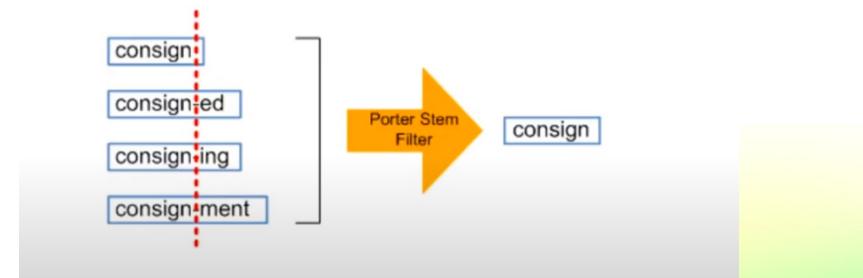
Примеры: бежал → бежать; меня → я ;

Text processing

Поскольку многие слова в тексте являются копиями других слов, просто находясь в различных формах (например, кот и коты пишутся по-разному, но имеют один и тот же смысл), то для упрощения модели применяется предобработка текста. Вот два основных способа:

Stemming: процесс нахождения основы слова (не обязательно корня) по заданным правилам.

Примеры: котик → кот; data → dat ;



Lemmatization: процесс приведения слова к его нормальной форме

Примеры: бежал → бежать; меня → я ;

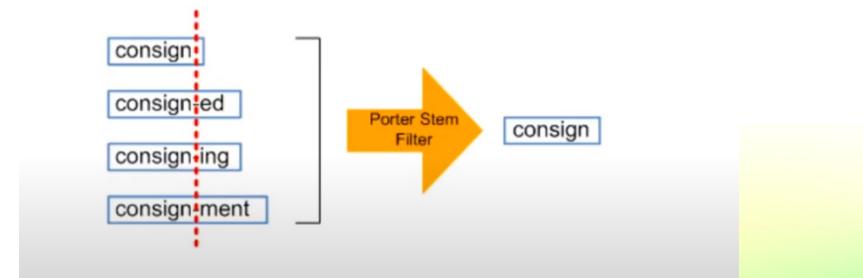
Также в процессе необходимо обработать отдельно разный регистр букв, пунктуацию, числа, стоп-слова (в первую очередь от самой задачи)

Text processing

Поскольку многие слова в тексте являются копиями других слов, просто находясь в различных формах (например, кот и коты пишутся по-разному, но имеют один и тот же смысл), то для упрощения модели применяется предобработка текста. Вот два основных способа:

Stemming: процесс нахождения основы слова (не обязательно корня) по заданным правилам.

Примеры: котик → кот; data → dat ;



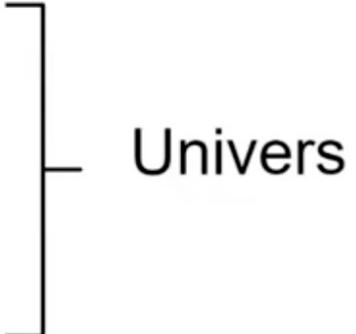
Lemmatization: процесс приведения слова к его нормальной форме

Примеры: бежал → бежать; меня → я ;

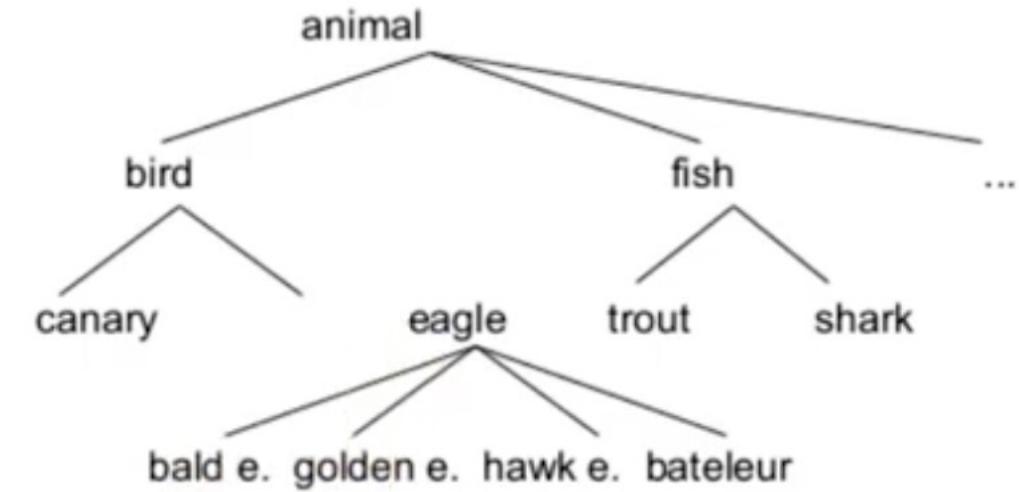
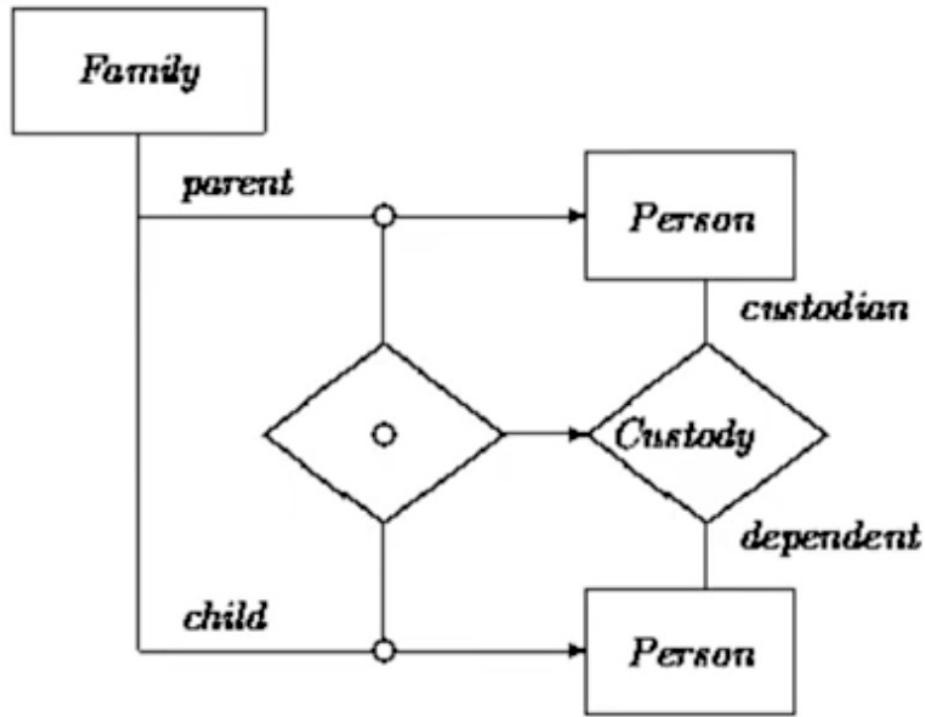
Также в процессе необходимо обработать отдельно разный регистр букв, пунктуацию, числа, стоп-слова (в первую очередь от самой задачи)

Text processing

Проблема предобработки:
Overstemming

- University
 - Universal
 - Universities
 - Universe
- 
- Univers

Lemmatization: NLTK (WordNet/RusNet)



Text processing: remove

- Символы (пунктуация итд)
- Сокращения (итд)
- Теги
- Стоп-слова (это, бы)

Text processing: tools

- **NLTK**: stemming, lemmatizer
- BeautifulSoup (wotking with HTML)
- RegEx (re)
- PyMorphy2
- DIY instruments

Text processing: tools

- **NLTK**: stemming, lemmatizer
- BeautifulSoup (wotking with HTML)
- RegEx (re)
- PyMorphy2
- DIY instruments

Text processing: tools

- **NLTK**: stemming, lemmatizer
- BeautifulSoup (wotking with HTML)
- RegEx (re)
- PyMorphy2
- DIY instruments

Bag-of-Words (BoW)

- Для начала составляется словарь всех или наиболее часто употребляемых слов исходного датасета.
- Затем каждому тексту ставится в соответствие вектор длины словаря, где на i -ой позиции записывается количество вхождений i -ого слова.
- Такой подход уже позволяет сравнивать тексты, например при помощи косинусной меры.
- Однако у него есть множество недостатков:
- теряется информация о порядке слов;
- вектора представлений слишком большие и разреженные;
- вектора представлений не нормализованы.

Bag-of-Words (BoW)

the dog is on the table

0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the



Демонстрация и решение упражнений