Анализ текстов. Генеративные модели

Лекция 1. Введение.

Эль-Айясс Дани Валид

Высшая Школа Экономики

6 сентября 2023

План

- Организационная часть
- Что такое NLP?
- Задачи NLP
- Этапы решения NLP задач
- Признаковые описания документов

Организационная часть

О себе

Дани Эль-Айясс:

- Магистр по направлению «Прикладная математика и информатика», ВМК МГУ (кафедра ММП)
- Исполнительный директор в SberDevices, разрабатываю GigaChat

Контакты:

- Почта: dayyass@yandex.ru
- Телеграм: @dayyass
- LinkedIn: https://www.linkedin.com/in/dayyass/
- GitHub: https://github.com/dayyass

О курсе

GitHub: https://github.com/dayyass/hse_nlp_course

Чат: https://t.me/+eSsRe-CWVu85ZTAy

Оценка:

- Домашняя работа * 0.49 + Самостоятельная работа * 0.21 + Экзамен * 0.3

Домашние задания:

- 4 задания в системе Anytask

Полезные материалы:

- Stanford CS224N: https://web.stanford.edu/class/cs224n/index.html
- Yandex NLP Course: https://github.com/yandexdataschool/nlp_course
- Список литературы: https://www.hse.ru/edu/courses/835160340

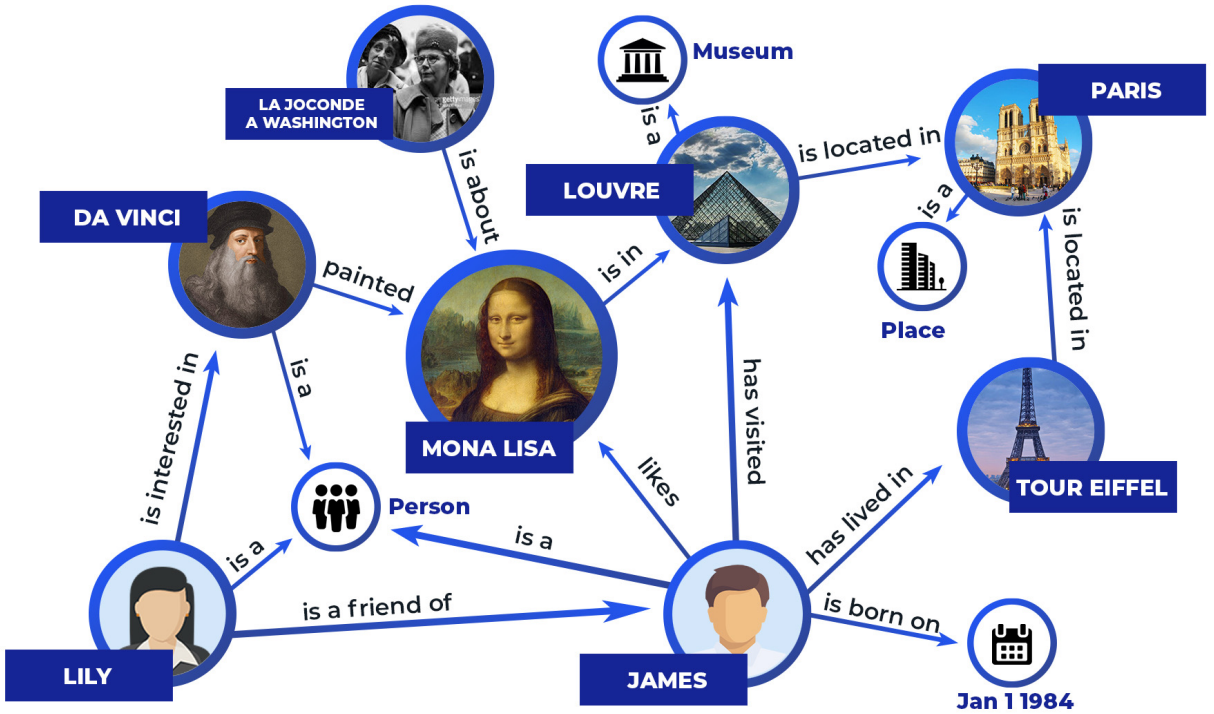
План курса

1. Введение
2. Векторные представления
3. Классификация текстов
4. Классификация последовательностей
5. Языковое моделирование
6. Машинный перевод и трансформеры
7. Предобученные модели
8. Большие языковые модели
9. Инструктивное дообучение и RLHF
10. Суммаризация текстов и вопросно-ответные системы
11. Информационный поиск
12. Мультимодальная обработка текстов

Что такое NLP?

Текстовые данные

- Большая часть данных в мире представлена в текстовом виде
- Текстовые данные могут быть:
  - структурированными (графы знаний, базы данных)



https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/

Текстовые данные

- Большая часть данных в мире представлена в текстовом виде
- Текстовые данные могут быть:
  - структурированными (графы знаний, базы данных)
  - неструктурированными (сырые тексты)

```
Введение в обработку естественного языка.
Под авторством Сидорова Ивана Петровича.
Вступление.
Настоящее пособие предназначено для ...
```

```
Чек
Магазин канцелярских товаров
1. Шариковая ручка (син) 23 руб.
2. Тетрадь клет (12 л) 5 руб.
...
```

Текстовые данные

- Большая часть данных в мире представлена в текстовом виде
- Текстовые данные могут быть:
  - структурированными (графы знаний, базы данных)
  - неструктурированными (сырые тексты)
  - частично структурированными (JSON, XML)

```
{
    "type": "учебник",
    "title": "Введение в обработку естественного языка.",
    "author": "Сидоров Иван Петрович",
    "introduction": "Настоящее пособие предназначено для \dots
",
    \dots
}

```

Естественный язык

- Естественный язык – способ общения между людьми
- Можем противопоставить его *формальным* и *искусственным* языкам:
  - Языки программирования – Programming Language Processing (PLP)
- Немного формализма:
  - Язык – это множество допустимых цепочек символов из некоторого алфавита
  - Текст – это цепочка, построенная по некоторым правилам
  - Алфавит – это множество символов, из которых строятся тексты
  - Каждая цепочка должна нести некоторую информацию (на деле не всегда)

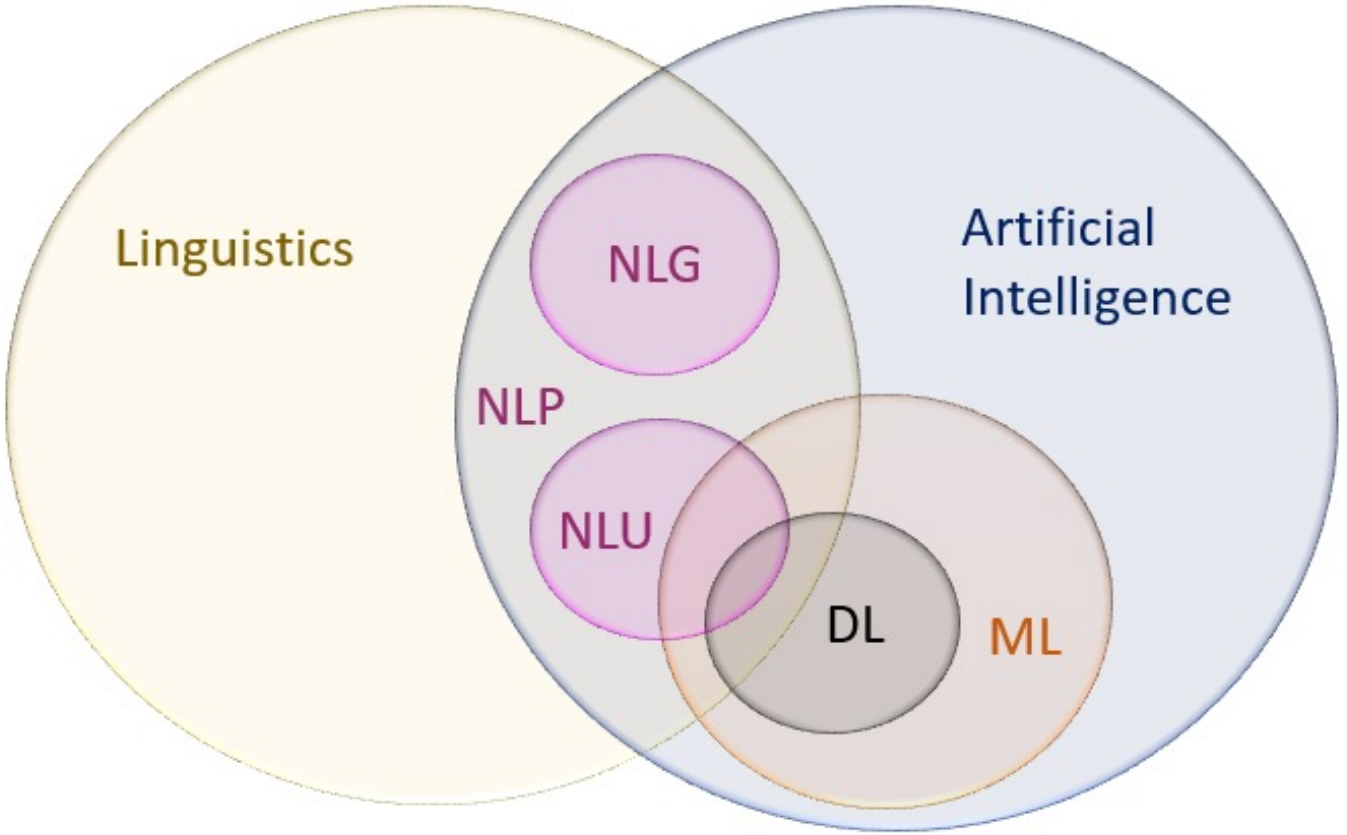
- «Глокая куздра штеко будланула бокра и кудрячит бокренка» (Л.В. Щерба, 1930-е)

Правила языка

- Выстраивается некоторая иехрархия:
  - Графематические – как разделять слова и предложения между собой
  - Морфологические – как строить и изменять слова
  - Синтаксические – как согласовывать словоформы друг с другом
  - Семантические – как применять все предыдущие правила, чтобы сообщить необходимую информацию
  - Стилистические – «уместность» словоупотребления в конкретной ситуации
  - И т.п.

Обработка естественного языка (NLP)

- Положение NLP среди наук по анализу и обработке данных:



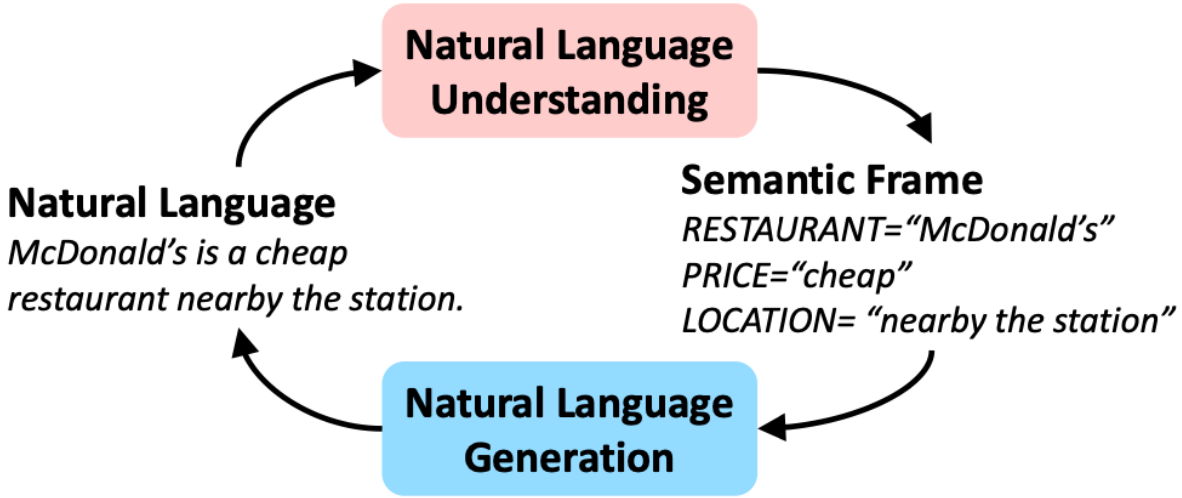
https://medium.com/analytics-vidhya/natural-language-processing-nlp-based-chatbots-7b2436428256

Структура NLP

- Внутри NLP условно выделяются два направления:
  - понимание языка (NLU)
  - генерация языка (NLG)

- Текст –> NLU –> смысл –> NLG –> текст

- Смежные области:
  - распознавание (ASR)
  - генерация (TTS) речи



https://arxiv.org/abs/1905.06196

Пирамила NI P

P.S. В самом низу графематический уровень

Особенности NLP

- Базовая структурная единица языка — слово
  - Даже вне контекста оно несет полезную информацию
  - У слов есть различные словоформы в зависимости от контекста
  - Многозначность слова (полисемия, омонимия)

- Текст без дополнительной разметки имеет внутреннюю структуру, определяемую языком на разных уровнях:
  - текст (порядок реплик)
  - предложение (синтаксис)
  - слова и словосочетания (морфология, синтаксис)

- Наличие огромных массивов сырых текстов и структуры в них позволяет обучать большие общеязыковые модели

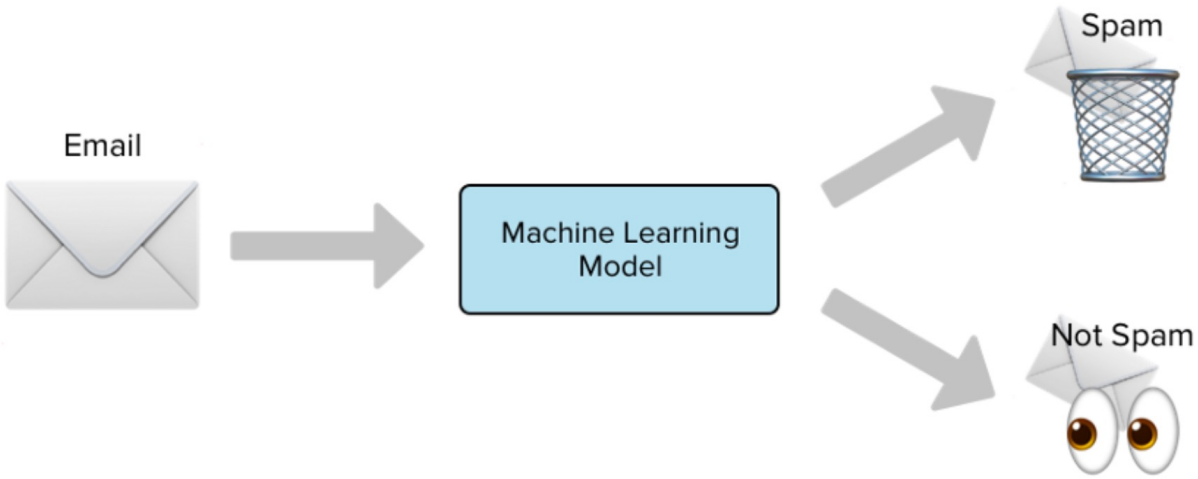
- Существует много лингвистических ресурсов, которые помогают в различных задачах обработки текстов

Задачи NLP

Задача классификации текстов

- Задачи NLP можно формулировать с технической и продуктовой точек зрения.

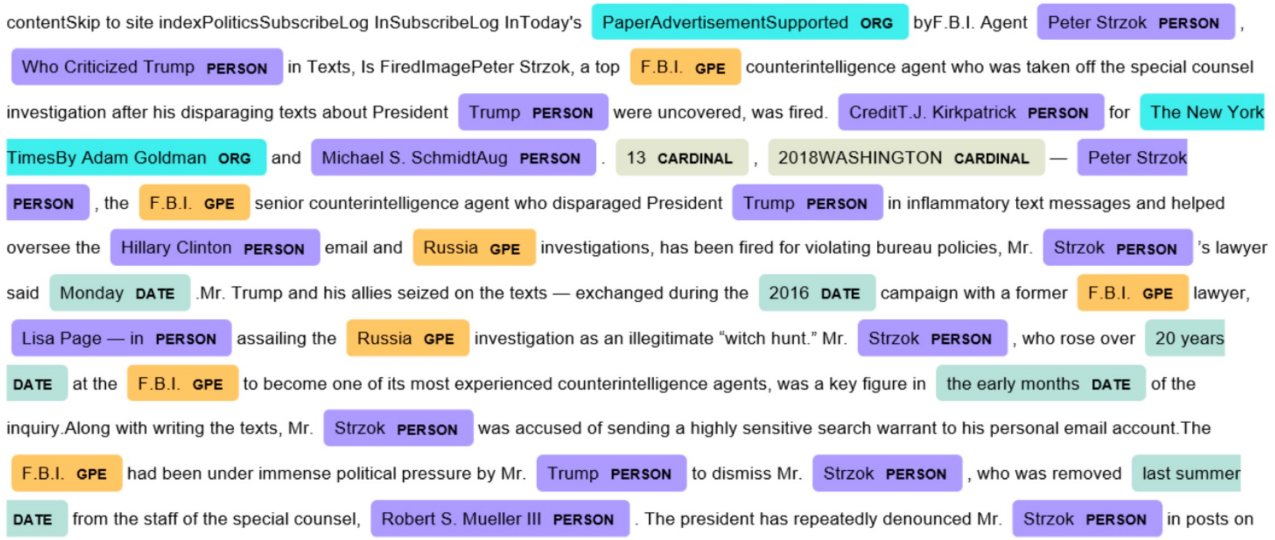
- Классификация - одна из основных задач в NLP, лежит в основе многих продуктовых задач:
  - Анализ тональности
  - **Фильтрация спама**
  - Определение намерений
  - Категоризация новостей и статей



https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

Задача разметки последовательностей

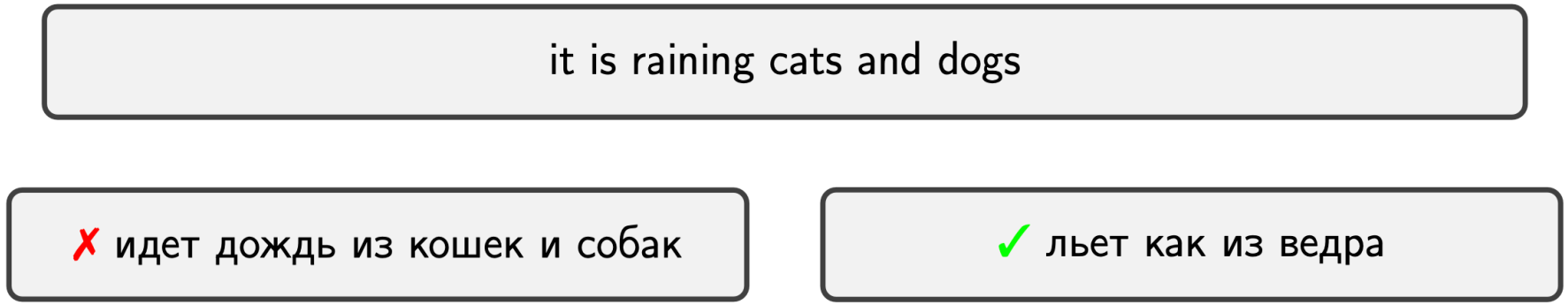
- Извлечение информации
  - **Распознавание именованных сущностей**
- Частеречная разметка
- Разрешение кореференции



https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

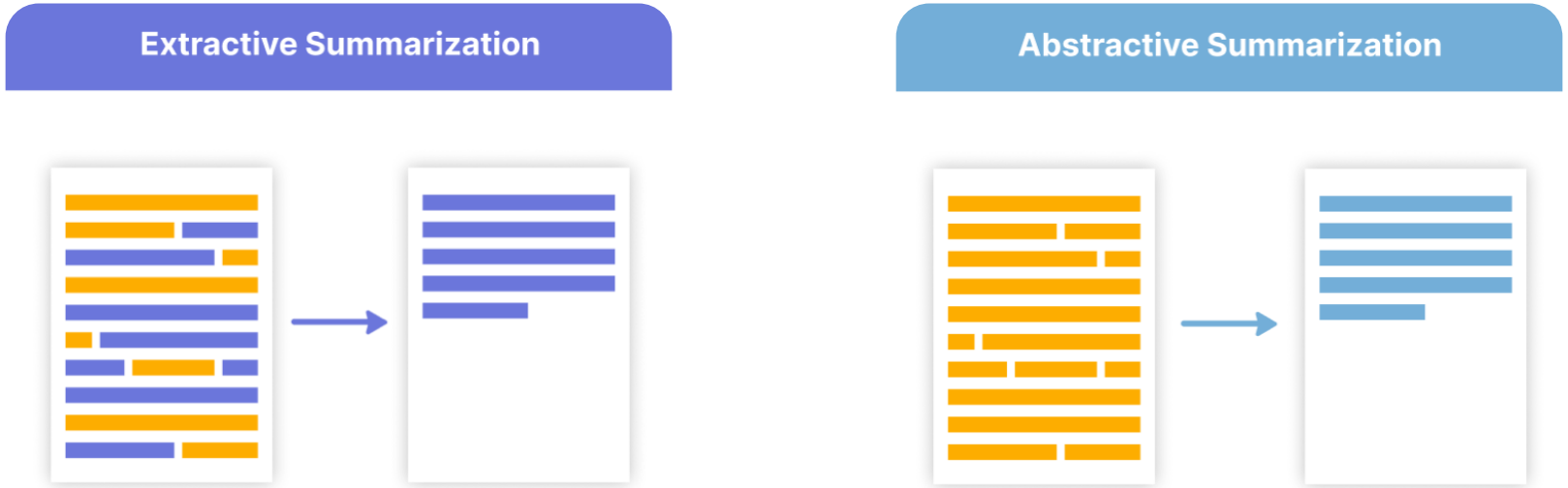
Задача машинного перевода

- Одна из фундаментальных задач NLP, двигатель многих исследований и открытий:
  - Attention
  - Transformers
- Машинный перевод:
  - Статистический
  - Нейронный



Задача суммаризации текстов

- Для текстового документа нужно сгенерировать краткое изложение
- Важна не только передача смысла, но и сохранение важных фактов



https://www.abstractivehealth.com/extractive-vs-abstractive-summarization-in-healthcare

Задача поиска ответов на вопросы

- Вопросно-ответные системы (QA-система) используются для поиска ответов на вопросы, заданные на естественном языке
- QA-системы часто используются в качестве элементов поисковых систем
- Пример: «что такое луна?»

Луна

Естественный спутник

Луна́ — единственный естественный спутник Земли. Самый близкий к Солнцу спутник планеты, так как у ближайших к Солнцу планет их нет. Второй по яркости объект на земном небосводе после Солнца и пятый по величине естественный спутник планеты Солнечной системы. Среднее расстояние между центрами Земли и Луны — 384 467 км. Википедия

Расстояние до Земли: 384 400 км

Ускорение свободного падения: 1,62 м/с²

Радиус: 1 737,1 км

Возраст: 4,53E9 лет

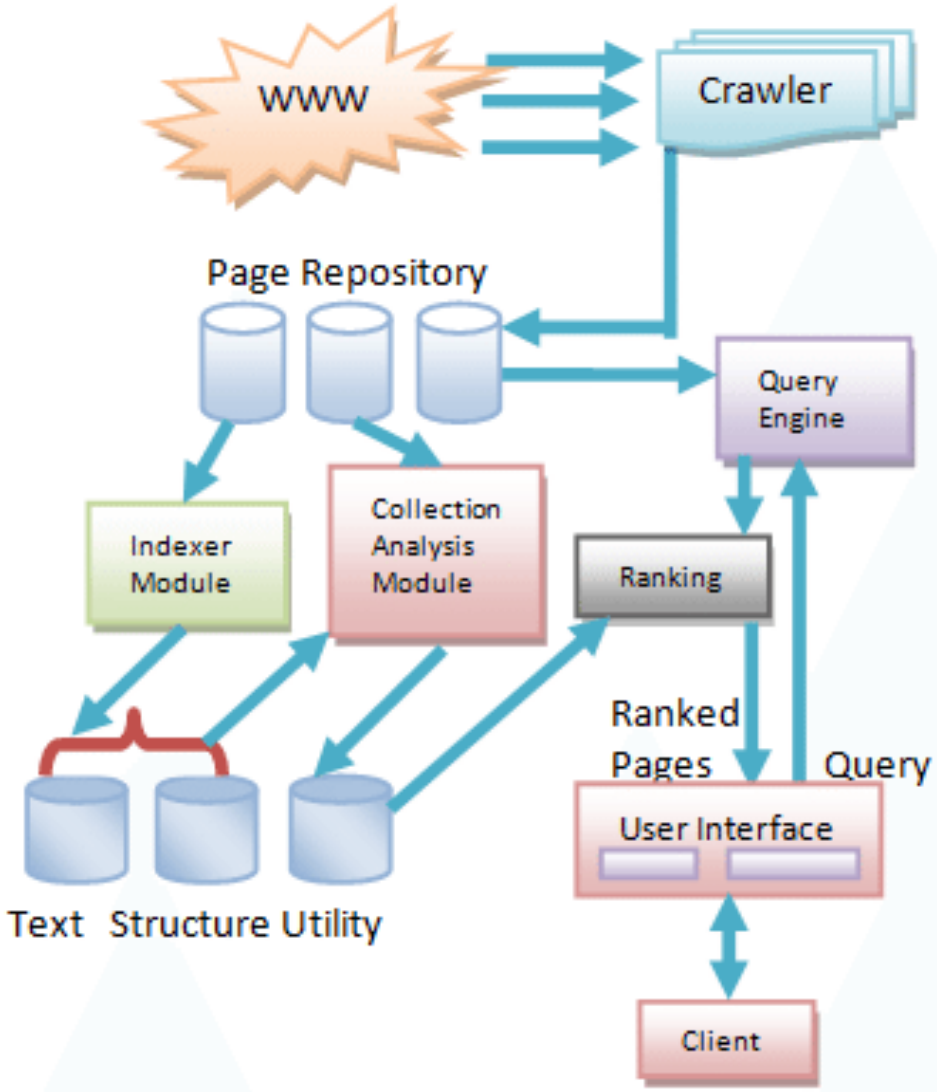
Период обращения: 27 дней

Обращается вокруг: Земля

https://www.google.ru

Задача ранжирования

- Ранжирование решает задачу сортировки объектов по заданному критерию полезности:
  - *информационный поиск* — релевантность страницы сайта пользовательскому запросу
  - *рекомендации* — близость текстовой статьи к текущим интересам пользователя

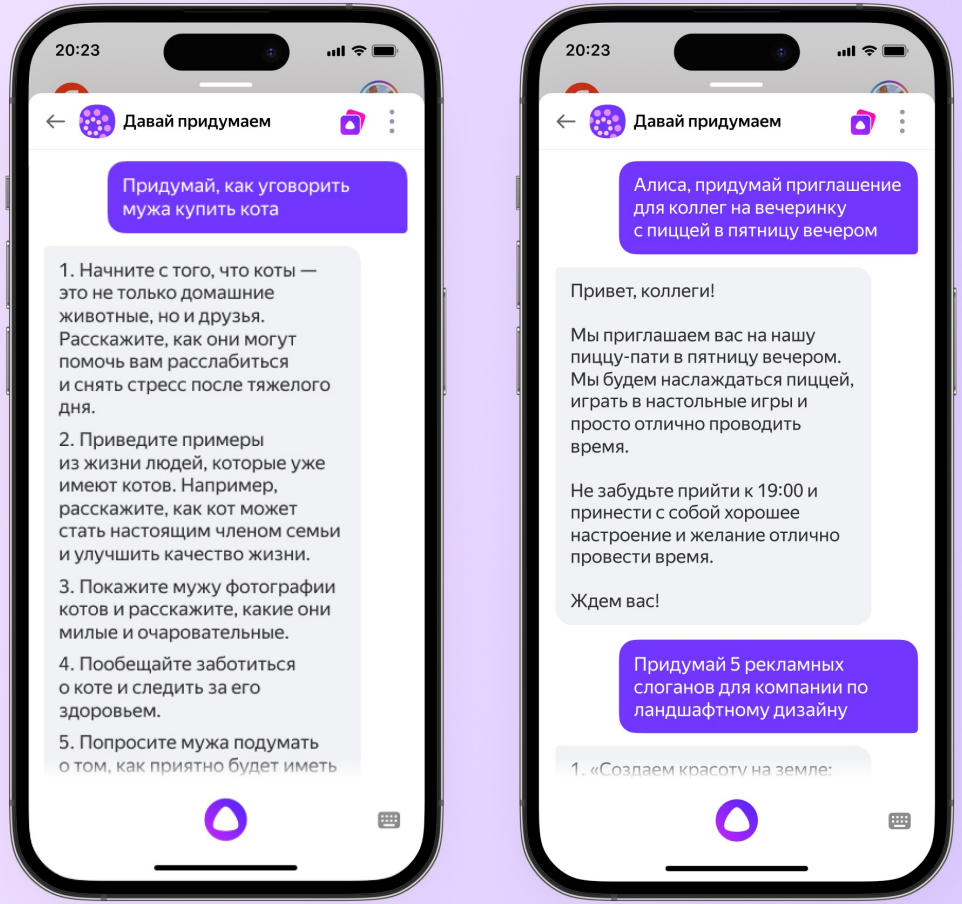


https://www.researchgate.net/figure/Basic-Search-Engine-Architecture-Conceptually-Search-Engine-retrieves-a-URL-according-to_fig1_334645287

Задача ведения диалога

- Диалоговые системы (чат-боты) общаются с человеком на естественном языке

- Хороший пример NLU -> NLG



Этапы решения NLP задач

Этапы решения NLP-задачи

- Всё так же, как и при обработке других типов данных:
  - Выбор верной метрики качества
  - Сбор обучающих и тестовых данных
  - ***Предобработка данных***
  - ***Формирование признакового описания текста***
  - Выбор подхода и класса моделей
  - Обучение моделей и настройка решения

- Предположим, что данные есть в некотором подходящем для работы формате

Инструменты для работы с текстами

- В обработке текстов часто полезны библиотеки общего назначения:
  - re/regex — модули для работы с регулярными выражениями
  - numpy/pandas/scipy/sklearn — базовые библиотеки для анализа данных и ML
  - pytorch — одна из основных библиотек для обучений нейросетей

- Специализированные библиотеки:
  - nltk
  - gensim
  - transformers

Регулярные выражения

- Регулярные выражения появились от т.н. регулярных автоматов (классификация грамматик по Хомскому)
- По факту это некоторый строковый шаблон, на соответствие которому можно проверить текст

- С синтаксисом можно ознакомиться на странице выбранного инструмента, но основные правила одинаковы, например:
  - . — означает наличие одного любого символа
  - [a-zA-Z0-9] — означает множество символов из заданного диапазона
  - +, *, ? — показывают, что следующий перед ними символ или последовательность символов должны повториться >= 1 раз (r+), >= 0 раз ((xa-)*) и 0/1 раз (r?)

https://habr.com/ru/articles/545150/

Предобработка текстов

- Пусть дана коллекция текстовых документов - текст представляет собой одну строку и алфавитных и неалфавитных символов
- Обрабатывать его в таком виде неудобно, сперва нужно выделить числовые признаки
- Базовые шаги предобработки:
  1. токенизация
  2. приведение к нижнему регистру
  3. удаление пунктуации
  4. удаление стоп-слов
  5. фильтрация слов по частоте/длине/регулярному выражению
  6. нормализация слов - лемматизация или стемминг

Токенизация

- Токенизацию можно производить по словам и/или предложениям
- Используются как подходы, основанные на правилах, так и ML-модели
- В nltk есть много различных токенизаторов, например RegexpTokenizer и sent_tokenize
- Часто слова грубо выделяют разделением по пробелам с помощью метода split

```
text = 'Набор слов, составляющий какое-то предложение.'
print(text.split(' '))
#['Набор', 'слов,', 'составляющий', 'какое-то', 'предложение.']
```

Регистр и пунктуация

- Есть задачи, в которых пунктуация и регистр несут важную информацию
- Это важно для определения границ предложений, для решения задачи выделения именованных сущностей
  - в комнату вошел ***лев*** и, потянувшись, достал из кармана сигару
  - ***лев*** обитает в саванне, в арктике не обитает
- В задаче анализа тональности существенное значение имеют смайлы (текстовые или символы в Unicode):
  - ❌ Одежда у вас в магазине очень своеобразная***:/***
  - ✅ Одежда у вас в магазине очень своеобразная***:)***

Нормализация слов

- Слова в тексте могут иметь различную формы, часто такая информация скорее мешает, чем помогает анализу
- Для нормализации применяется один из подходов:
  - *лемматизация* (pymorphy2, pymystem3) – приводит слова к нормальной форме
  - *стемминг* (реализации в nltk) – стемминг приводит слова к псевдооснове (убирает окончания и формообразующие суффиксы)

```
1 import pymorphy2
2 text = 'я хотел бы поговорить с вами'.split(' ')
3 lemmatizer = pymorphy2.MorphAnalyzer()
4 print([lemmatizer.parse(t)[0].normal_form for t in text])
5 # ['я', 'хотеть', 'бы', 'поговорить', 'с', 'вы']
```

Фильтрация словаря

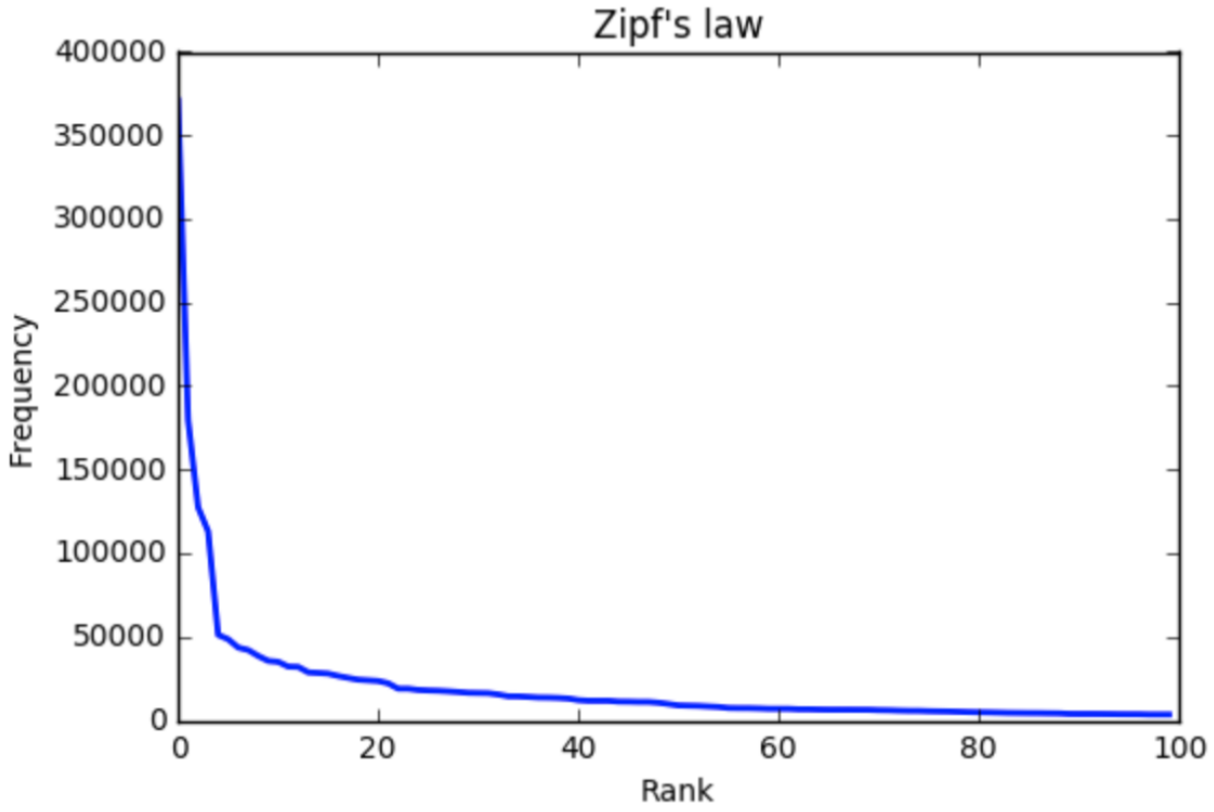
- Часто из текстов нужно удалять лишние слова
- Обычно это стоп-слова - очень редкие и очень частые слова
- К стоп-словам относятся союзы, предлоги, модальные глаголы, местоимения, вводные слова
- Большой набор стоп-слов есть в nltk:

```
from nltk.corpus import stopwords
sw = set(stopwords.words('russian'))

for w in ['я', 'хотеть', 'бы', 'поговорить', 'с', 'вы']:
    if w not in sw:
        print(w, end=' ')
# хотеть поговорить
```

Фильтрация словаря

- Слишком частые или редкие слова тоже могут оказаться вредными
- Такие слова могут и мешать обучению модели, и увеличивать затраты ресурсов памяти и времени счета
- При обработке коллекции стоит проверить выполнение закона Ципфа:



Признаковые описания документов

Признаковые описания документов

- Обычно в ML данные представляют собой матрицу «объекты-признаки»:

Номер автомобиль	Тип топлива	Мощность двигателя	...	Масса
1	Бензин	120	...	1700
...	...	...	...	...
N	Дизель	160	...	2100

- Для текстов тоже нужно как-то получить такую матрицу

Модель мешка слов

- Можно проверять наличие всех возможных слов из некоторых словаря:

Номер текста	Содержит «абрикос»	. . .	Содержит «яблоко»
1	0	. . .	1
. . .	. . .	. . .	. . .
N	1	. . .	0

- Пусть значением признака будет не наличие слова, а число его вхождений в документ («мешок слов»):

Номер текста	Вхождений «абрикос»	. . .	Вхождений «яблоко»
1	0	. . .	23
. . .	. . .	. . .	. . .
N	2	. . .	0

TF-IDF

- Представление «мешка слов» часто используется при обработке текстов, но частота встречаемости слов не самый информативный признак
- Идея: хотим выделить слова, которые часто встречаются в данном тексте, и редко — в других текстах - используем значения TF-IDF:

$$v_{wd} = tf_{wd} \times \log \frac{N}{df_w}$$

- tf_{wd} — доля слова w в словах документа d
- df_w — число документов, содержащих w
- N — общее число документов

«Мешок слов» и TF-IDF в Python

```
1  from sklearn.feature_extraction.text import CountVectorizer
2  from sklearn.feature_extraction.text import TfidfVectorizer
3
4  c_vectorizer, t_vectorizer = CountVectorizer(), TfidfVectorizer()
5  corpus = [
6       'This is the first document.',
7       'This is the second second document.',
8       'And the third one.',
9       'Is this the first document?',
10 ]
11 X_c = c_vectorizer.fit_transform(corpus)
12 X_t = t_vectorizer.fit_transform(corpus)
```

Коллокации

- N-граммы — устойчивые последовательности из N слов, идущих подряд («машина опорных векторов»)
- Коллокация — устойчивое сочетание слов, не обязательно идущих подряд («Он сломал своему противнику руку»)
- Часто коллокациями бывают именованные сущности (но не всегда)
- Методы получения N-грамм:
  - на основе частот встречаемости (sklearn, nltk)
  - на основе морфологических шаблонов (Томита, YARGY-парсер)
  - с помощью ассоциации и статистических критериев на основе частот совместных встречаемостей (nltk, TopMine)
  - иные подходы (RAKE, TextRank)

Меры ассоциации биграмм

- Поточечная совместная информация (Pointwise Mutual Information, PMI):

$$\mathrm{PMI}(w_1, w_2) = \log \frac{f(w_1, w_2)}{f(w_1) f(w_2)}$$

- T-Score (по сути — тест Стьюдента):

$$\mathrm{T}_{\mathrm{score}}(w_1, w_2) = \frac{f(w_1, w_2) - f(w_1) f(w_2)}{\sqrt{f(w_1, w_2)/N}}$$

  - w_i — слово
  - $f(\cdot)$ — частота слова или биграммы

Меры ассоциации биграмм

- В обоих случаях проверяется гипотеза независимости появления пары токенов (слов или N-грамм)
- Чем выше значение критерия, тем скорее пара токенов является устойчивым сочетанием
- Можно обобщить на произвольные коллокации

Итоги занятия

- NLP — очень востребованная и активно развивающаяся область на стыке машинного обучения, анализа данных и лингвистики
- Существуют разнообразные постановки задач обработки текстов, технические и бизнесовые
- Работа с текстами почти всегда требует тщательного изучения и аккуратной предобработки данных
- Можно использовать разнообразные признаковые описания, базовыми являются представления «мешка слов» и TF-IDF