Автоматическая обработка естественного языка
Введение

Екатерина Владимировна Еникеева
protoev@yandex.ru
https://github.com/named-entity/hse-nlp

6 сентября 2022

Преподаватели

Екатерина Владимировна

Еникеева

protoev@yandex.ru

Александр Викторович

Орлов

alexander.orlov98@gmail.com

О чем этот курс

АОЕЯ / AOT = Natural Language Processing (NLP)

особое внимание к русскоязычным данным

- 3 курс:
  - базовые статистические и нейросетевые модели
  - методы АОТ по уровням (NLP pipeline)
  - морфология, синтаксис, фонетика (?)
  - оценка качества
- 4 курс: прикладные задачи АОТ («семантика», АОТ на уровне целого текста)

Основные активности

- Все занятия практические:
  - Лекция + обсуждение + код
  - Семинар с самостоятельной работой
- Чтение статей (на английском)
- Семинар с обсуждением статей
- Домашки
- Финальный проект

Оценка

- домашки (2) 40% ~10 дней
- квизы по статьям (2) + ридинг-семинар 20% ~неделя
- проект (1) 40% ~месяц

Что нужно для оценок 9/10:

- задания со * в домашках и квизах
- оформление проекта

Финальный проект

Корпус-менеджер (поиск по корпусу с лингвистической разметкой):

- Поэтапная разметка корпуса
- Организация поиска
- Представление результатов
- Более подробное ТЗ – в начале октября
- Команда до 4 человек

Ранний период NLP

- С чего всё начиналось? Какая первая NLP-шная задача встала перед программистами?

Ранний период NLP

Машинный перевод и идея AI (искусственного интеллекта)

- **1940-е** – тест Тьюринга
- **1947** – Warren Weaver – идея статистического перевода
- **1954** – Джорджтаунский эксперимент – перевод по правилам
- **1958** – первая Всесоюзная конференция по МП
- **1966** – доклад ALPAC, AI Winter

Чатботы

они же:
chatterbots, виртуальные собеседники, диалоговые системы

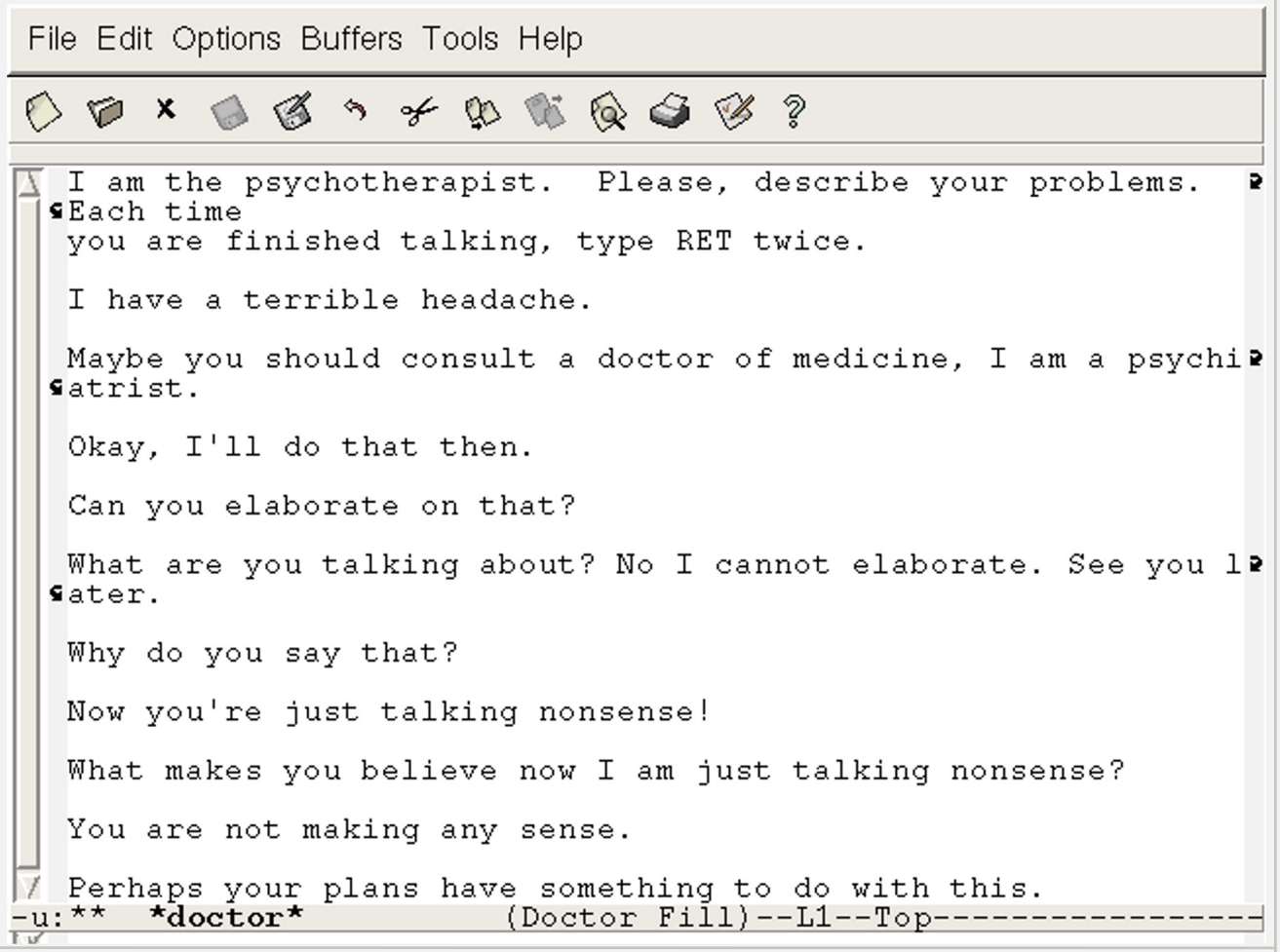
1964-1966 – ELIZA

1972 – PARRY

Как работает?

Правила

нейро



Natural Language Understanding

- **1970-e –** Conceptual Dependency Theory (R. Schank)

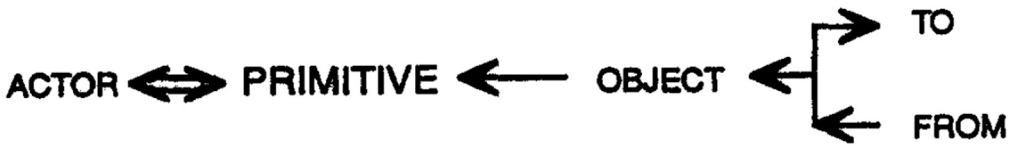
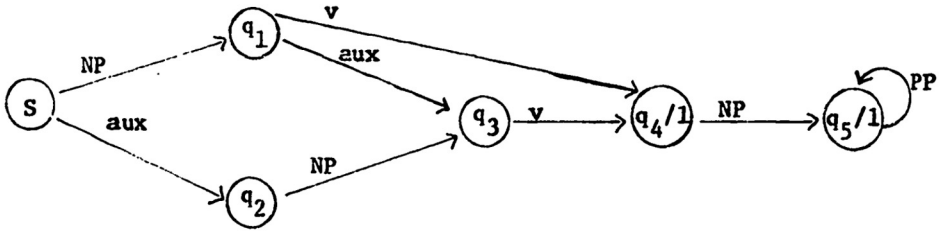


Figure 2. Basic form of a conceptual dependency graph.

Figure 3. Representation of "John gave Mary a book."

- Augmented Transition Network (W.A. Woods)



Задачи моделирования (1)

- синтаксис
  - supervised и unsupervised
  - rule-based — формальные грамматики
- семантика
  - онтологии, тезаурусы
  - дистрибутивные модели — классические ДСМ, word embeddings
  - семантические роли, фреймы

Появление статистических методов

- **конец 1980х-1990е –** внедрение статистических методов в различные направления NLP:
  - распознавание речи (speech recognition)
  - морфологический анализ (POS-tagging)
  - коллокации
  - классификация текстов

Применение методов АОТ

Применение методов АОТ

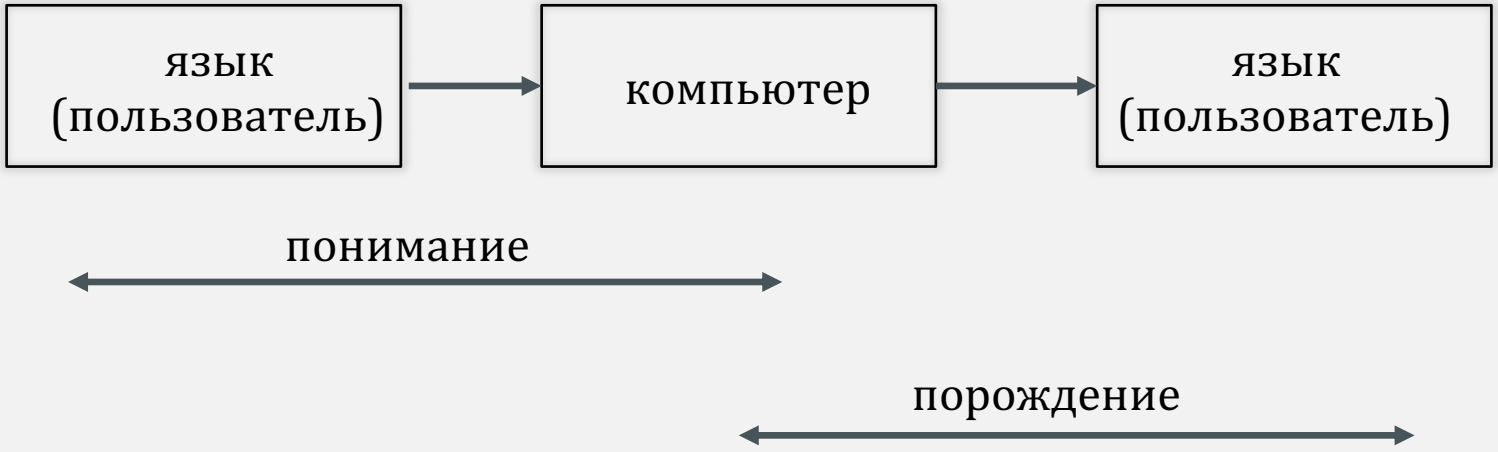
- информационный поиск
- автодополнение (клавиатуры, email); исправление опечаток
- извлечение именованных сущностей, фактов
- автоматическое реферирование; антиплагиат
- оценка тональности, извлечение мнений
- классификация текстов; выделение подтем в документе
- вопросно-ответные системы
- распознавание и синтез речи
- …

Задачи моделирования (1)

- синтаксис
  - supervised и unsupervised
  - rule-based — формальные грамматики
- семантика
  - онтологии, тезаурусы
  - дистрибутивные модели — классические ДСМ, word embeddings
  - семантические роли, фреймы

Задачи моделирования (2)

- понимание (NLU) vs. порождение (NLG) текста



-> связь АОТ и искусственного интеллекта (ИИ, AI – Artificial Intelligence)

NLP challenges

- Почему вообще интересно заниматься NLP и не все задачи до сих пор решены?

NLP challenges

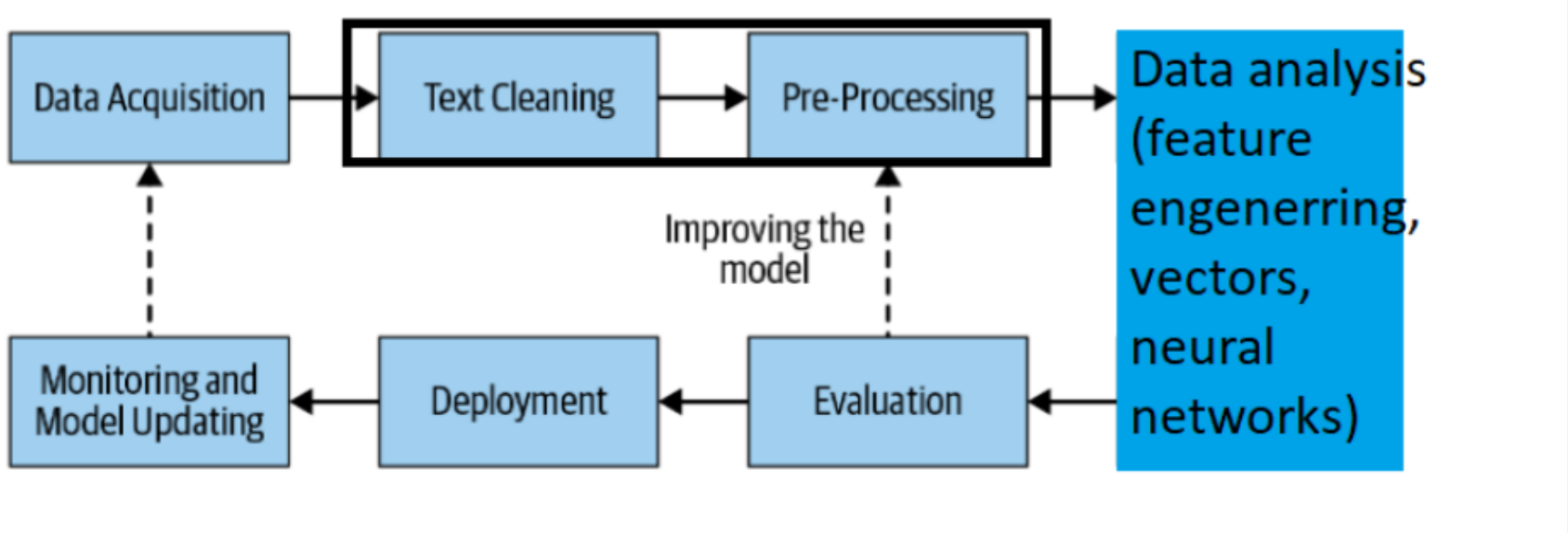
- **неоднозначность** языка на всех уровнях (linguistic ambiguity): 1 форма – N значений
- **синонимия** всех уровнях: 1 значение может выражаться N разними способами
- стилистическое разнообразие
- продуктивность (неологизмы)
- идиоматичность, некомпозициональность
- low-resourced languages

Методы

- rule-based (основанные на правилах, требуют экспертизы)
- **статистические** (требуют данных)
  - классические
  - основанные на машинном обучении
- гибридные

Почти во всех задачах state-of-the-art (SOTA) – нейронные сети

Пайплайн


Data Acquisition
Text Cleaning
Pre-Processing
Data analysis (feature engenerring, vectors, neural networks)
Improving the model
Monitoring and Model Updating
Deployment
Evaluation

1. Сбор данных

Есть готовые корпуса, а есть парсинг. Для разных задач нужны разные тексты.

Где нам взять данные для того, чтобы.... (и какие доп. проблемы мы можем найти, какие вопросы задать?)

1. Обучить нейросеть определять тональность высказывания (проще говоря, является ли текст положительным/хвалебным, негативным или нейтральным)?
2. Создать программу, которая будет определять, написан ли текст в формальном стиле или нет?
3. Обучить нейросеть определять хейтспич?
4. Создать функцию, которая убирает из текста стоп-слова в малом языке?
5. Создать спеллчекер?
6. спелчекер без морфологии?

Данные: корпуса / датасеты

Тесты + сегментация + метаданные + разметка

- Корпуса одного языка
  - Brown corpus, British National Corpus, Penn Treebank
  - Национальный корпус русского языка (НКРЯ)
- Параллельные и многоязычные:
  - Europarl, UN Corpus, Opus
- Под специфические задачи
  - Twitter US Airline Sentiment …

Данные: тегсеты

- Английский и мультиязычные:
- Stanford NLP
- Universal Dependencies
- Русский
- Соревнования «Диалога» (Ru-Eval)
- НКРЯ (Mystem), pymorphy / OpenCorpora

Этапы обработки текста

Сегментация
(тексты, абзацы, предложения)

```
In [1]: text = "Mr. Smith bought ticket to San Francisco. He was very happy."
        text.split('.')

Out[1]: ['Mr', ' Smith bought ticket to San Francisco', ' He was very happy', '']

In [2]: import nltk

In [3]: nltk.sent_tokenize(text)

Out[3]: ['Mr. Smith bought ticket to San Francisco.', 'He was very happy.']
```

Этапы обработки текста

Токенизация
(слова, токены, стоп-слова)

```
In [5]: en_sentence = nltk.sent_tokenize(text)[0]
        en_sentence.split()

Out[5]: ['Mr.', 'Smith', 'bought', 'ticket', 'to', 'San', 'Francisco.']

In [6]: ru_sentence = "Мистер Смит купил билет до Сан-Франциско."
        ru_sentence.split()

Out[6]: ['Мистер', 'Смит', 'купил', 'билет', 'до', 'Сан-Франциско.']
```

??? Аналитические формы, компаунды, коллокации

```
In [7]: nltk.word_tokenize(en_sentence)

Out[7]: ['Mr.', 'Smith', 'bought', 'ticket', 'to', 'San', 'Francisco', '.']
```

Этапы обработки текста

Лемматизация / стемминг

```
In [21]: import pymorphy2
         m = pymorphy2.MorphAnalyzer()
         for t in nltk.word_tokenize(ru_sentence):
             print(m.parse(t)[0].normal_form)
```

```
мистер
смит
купить
билет
до
сан-франциско
.
```

Лемма ~ лексема ~ начальная форма
Стем ~ основа ~ усеченная словоформа

Этапы обработки текста

Морфологический анализ (~POS-tagging)

```
In [23]: for t, p in nltk.pos_tag(nltk.word_tokenize(en_sentence)):
             print(t, p)

Mr. NNP
Smith NNP
bought VBD
ticket NN
to TO
San NNP
Francisco NNP
. .
```

Этапы обработки текста

Морфологический анализ (~POS-tagging)

In [24]:
```
import pymorphy2
m = pymorphy2.MorphAnalyzer()
for t in nltk.word_tokenize(ru_sentence):
    print(m.parse(t)[0])
```

```
Parse(word='мистер', tag=OpencorporaTag('NOUN,anim,masc sing,nomn'), normal_form='мистер', score=1.0, methods_stack=
((DictionaryAnalyzer(), 'мистер', 52, 0),))
Parse(word='смит', tag=OpencorporaTag('NOUN,anim,masc,Sgtm,Surn sing,nomn'), normal_form='смит', score=0.333333, meth
ods_stack=((DictionaryAnalyzer(), 'смит', 29, 0),))
Parse(word='купил', tag=OpencorporaTag('VERB,perf,tran masc,sing,past,indc'), normal_form='купить', score=1.0, method
s_stack=((DictionaryAnalyzer(), 'купил', 680, 1),))
Parse(word='билет', tag=OpencorporaTag('NOUN,inan,masc sing,accs'), normal_form='билет', score=0.666666, methods_stac
k=((DictionaryAnalyzer(), 'билет', 34, 3),))
Parse(word='до', tag=OpencorporaTag('PREP'), normal_form='до', score=1.0, methods_stack=((DictionaryAnalyzer(), 'до',
24, 0),))
Parse(word='сан-франциско', tag=OpencorporaTag('NOUN,inan,masc,Sgtm,Fixd,Geox sing,loct'), normal_form='сан-франциск
о', score=0.416666, methods_stack=((DictionaryAnalyzer(), 'сан-франциско', 31, 5),))
Parse(word='.', tag=OpencorporaTag('PNCT'), normal_form='.', score=1.0, methods_stack=((PunctuationAnalyzer(score=0.
9), '.'),))
```

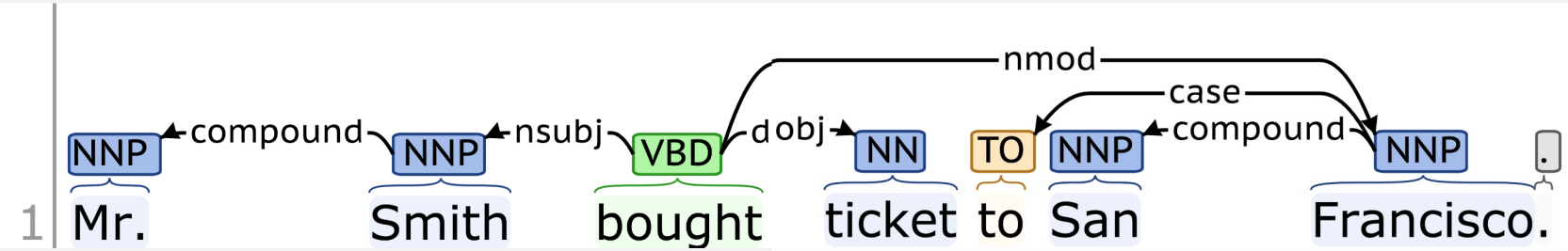
Этапы обработки текста

Разрешение неоднозначности (лемм / тегов)

Mr./NNP Smith/NNP bought/**VBD** ticket/NN to/TO San/NNP Francisco/NNP ./.
Мистер/(NOUN,anim,masc sing,nomn) Смит/(**NOUN,anim,masc,Surn sing,nomn** | **...**) купил/(VERB,perf,tran masc,sing,past,indc) билет/(NOUN,inan,masc sing,nomn | **NOUN,inan,masc sing,accs**) **…**

Этапы обработки текста

Синтаксический анализ (parsing)



Этапы обработки текста

Семантический анализ? (Semantic Role Labeling)

купить: [ARG0: Мистер Смит]

[V: купил]

[ARG1: билет до Сан-Франциско]

Этапы обработки текста

Семантический анализ?

Данные: подготовка

Разметка

(спец. инструменты – BRAT, …) → xml / tsv / …

- Согласованность разметчиков (Cohen's kappa)
- Краудсорсинг

Отбор данных

- Dataset augmentation / distillation

Каппа Коэна мера согласованности между двумя категориальными переменными. Обычно говорят о двух оценщиках, которые распеделяют n наблюдений по s категориям.

$\kappa = \frac{p_o - p_e}{1 - p_e}$,

где p_o — доля полного согласия, а p_e — вероятность случайного согласия.

Для случая s = 2, можно нарисовать следующую таблицу сопряженности:

	s_1	s_2
s_1	a	b
s_2	c	d

В таком случае:

- $p_o = \frac{a+d}{a+b+c+d}$
- $p_e = \frac{(a+b)\times(a+c)+(d+b)\times(d+c)}{(a+b+c+d)^2}$

семинар

Представление данных

Вспомним tf*idf: можно сравнивать между собой тексты и слова (|d|=длина документа)

Суперъяхта — большая прогулочная яхта. Этот термин не имеет формального определения. |d|=10

Ранг неориентированного графа имеет два не связанных друг с другом определения. |d|=11

Осада Парижа — осада Парижа в 1590 году во время Восьмой (и последней) Религиозной войны во Франции. |d|=16

tf * idf	doc_1	doc_2	doc_3
суперъяхта	(1/10) * (3/1)	(0/11) * (3/1)	(0/16) * (3/1)
имеет	(1/10) * (3/2)	(1/11) * (3/2)	(0/16) * (3/2)
определения	(1/10) * (3/2)	(1/11) * (3/2)	(0/16) * (3/2)
осада	(0/10) * (3/1)	(0/11) * (3/1)	(2/16) * (3/1)

Вектора слов

One-hot encoding

```
Суперъяхта — большая прогулочная яхта.
```

Словарь:

```
0 большая
1 прогулочная
2 суперъяхта
3 яхта
4 .
```

	0	1	2	3	4
суперъяхта	0	0	1	0	0
большая	1	0	0	0	0
прогулочная	0	1	0	0	0
яхта	0	0	0	1	0

Вектора слов

Word embeddings: e.g. Word2vec

! Фиксированный размер вектора

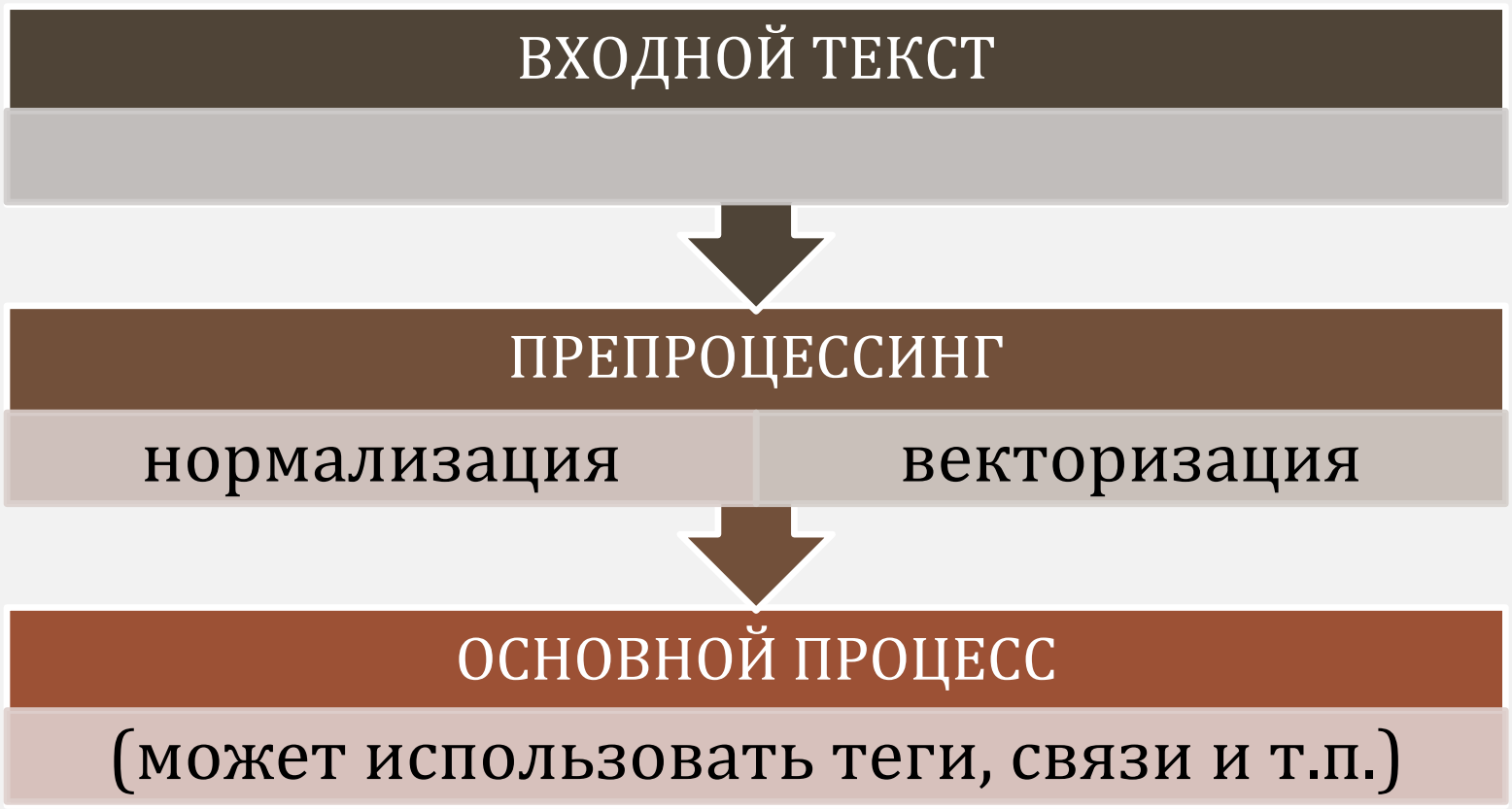
```
суперъяхта: [1.23 3.48 2.59]

большая: [0.74. 4.31 0.98]
```

…

Можно оценивать: близость слов, текстов и тп

NLP Pipeline



семинар

Оценка качества

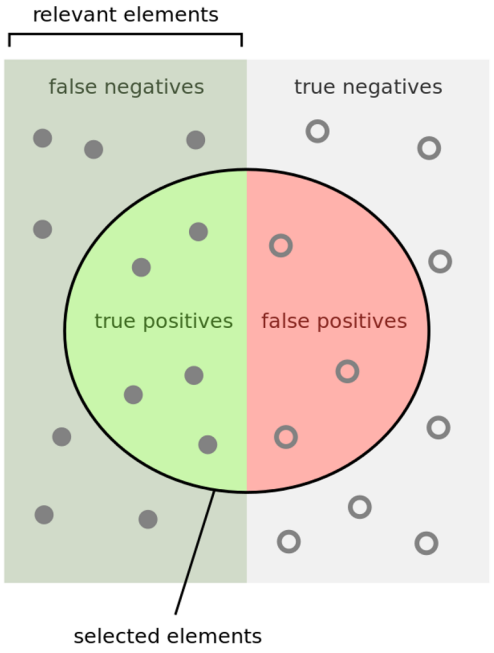
- **внутренняя**

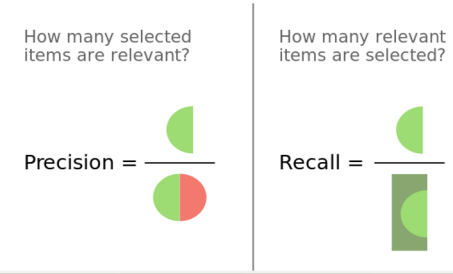
(из IR) точность, полнота, accuracy; специфические метрики

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

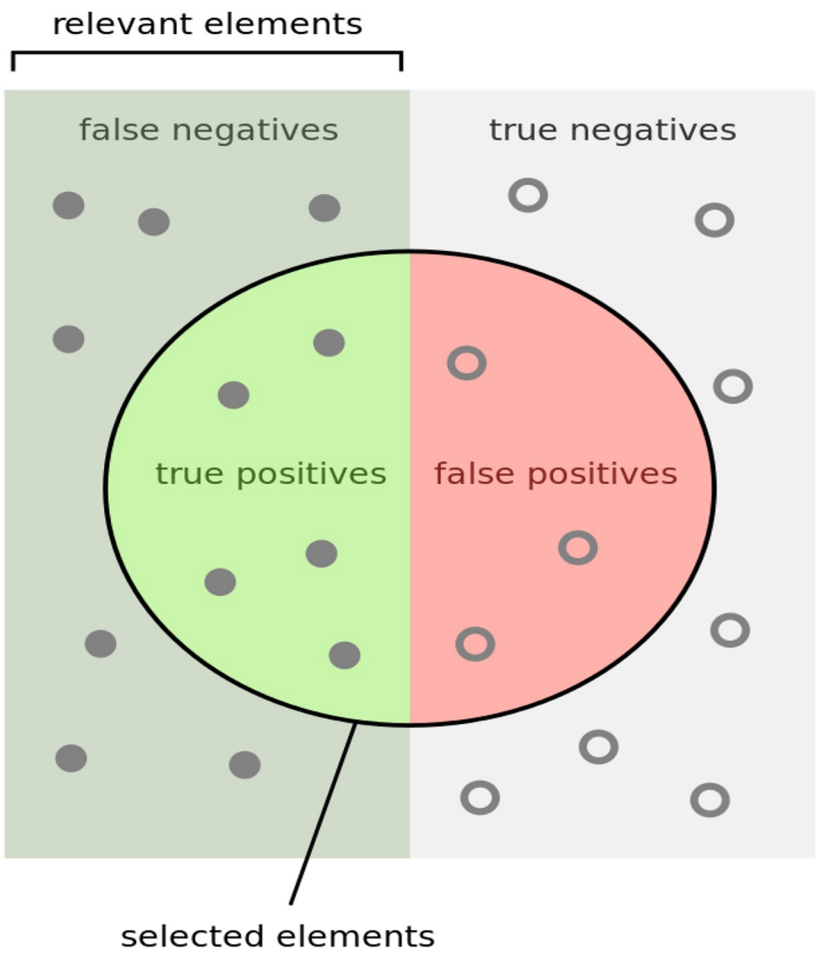
- **внешняя**

в более высокоуровневых приложениях





Оценка качества



How many selected

How many relevant

- F1-score is **harmonic mean of precision and recall score** and is used as a metrics in the scenarios where choosing either of precision or recall score can result in compromise in terms of model giving high false positives and false negatives respectively.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Золотой стандарт

= golden standard / benchmark

Проблемы:

- репрезентативность, сбалансированность
- выбор экспертов
- приближенность к реальным данным

семинар

Практическая вставка про токенизацию

https://colab.research.google.com/drive/1ne5HCczQSwCh6m0wj2VJvOA8Rbh7-fxk?usp=sharing

- https://colab.research.google.com/drive/1w02iQpc18eQcToG1vKJoheVdblGBqFqk#scrollTo=Dh0CDmMmiScF blanked
  https://colab.research.google.com/drive/1BqXRwoi_qJnmVYUgTsfa1SKHDat5xROk#scrollTo=Dh0CDmMmiScF done

Спасибо!

Вопросы?