

Основы обработки текстов

лекция 1

О курсе

- Лекторы: **Турдаков Денис Юрьевич**
 - Майоров Владимир Дмитриевич
 - Трифонов Владислав Дмитриевич
 - Недумов Ярослав Ростиславович
- Лекции каждый вторник в **10.30 в Zoom**
 - предполагаются минимальные знания
 - линейной алгебры,
 - теории вероятности и математической статистики
 - программирования
 - не все имеют одинаковые знания
 - предполагается, что студенты могут быстро учиться

План на сегодня

- Подробнее о курсе и практикуме
- Язык программирования Python
- Проблемы обработки текстов

Часть 1

О курсе

- Курс состоит из
 - лекций,
 - практикума и
 - итогового экзамена
- Язык программирования Python 3
- Вся информация: <http://tpc.at.ispras.ru>

Практическая часть

- Одна из открытых задач обработки текстов
 - В этом году: одна из задач разметки последовательностей
- Веб-интерфейс для проверки и задание будут доступны в начале октября

Часть 2

Python

«Читаемость имеет значение»

PEP 20 - The Zen of Python

- Код читается гораздо чаще, чем пишется
 - PEP 8 -- Style Guide for Python Code
 - Используйте средства автоматической проверки на соответствие PEP 8 (Pylint, PyFlakes, ...) или комбинированные средства (Pylama, Flake8)
- Комментарии помогают
 - PEP 257 -- Docstring Conventions
 - Автоматическая генерация документации (Sphinx)
 - Автогенерация тестов (Doctest)

Python

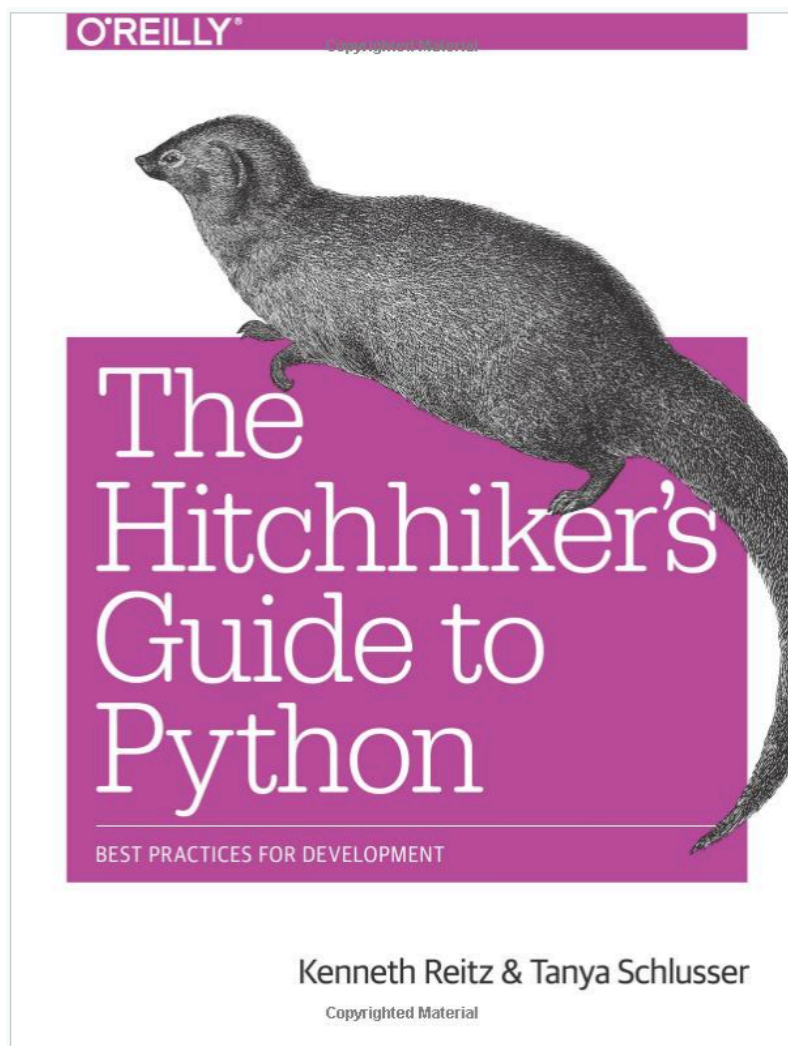
- Значимые пробелы

```
if x==1:  
    print 'x is 1'  
    print 'внутри блока'  
print 'вне блока'
```

- Используйте IDE: PyCharm, Jupyter и т.д.

Python

- Правильная установка и использование
 - Прочитайте про `pip` и `virtualenv`
- Прочитайте о возможностях языка, и попробуйте их использовать в своих проектах
 - Но прежде чем отсылать решение подумайте, действительно ли стало проще читать ваш код?



<https://docs.python-guide.org>

Python и обработка текстов

- NLTK
- <http://www.nltk.org>
- NLTK book

```
import nltk
text = "Hello world!"
tokens = nltk.word_tokenize(text)
print tokens

> ['Hello', 'world', '!']
```

Python и машинное обучение

- scikit-learn
- <http://scikit-learn.org>

```
from sklearn.naive_bayes import GaussianNB
x = [[0,0],[1,1]]
y = [0,1]

classifier = GaussianNB()
trained_classifier = classifier.fit(x,y)
predicted_value = trained_classifier.predict([0.6,0.6])

> [1]
```

- Keras, PyTorch, TensorFlow

См. спецкурс:
<https://mlcourse.at.ispras.ru/>

Часть 3

Классические задачи обработки текстов

- Информационный поиск (IR)
- Извлечение информации (IE)
- Вопросно-ответные системы (QA)
- Классификация и кластеризация
- Автоматическое аннотирование и реферирование
- Диалоговые системы
- Машинный перевод

Приложения обработки текстов

Уровни обработки текстов

- Морфологический
 - I'm - I am
 - кошка-кошки, дно-?
- Синтаксический
 - Мне один черный кофе и один сладкий булка...
- Семантический
 - Сколько китайского шелка было экспортировано в Западную Европу в конце 18 века?
 - лексическая и композиционная семантика
- Прагматический (дискурс)
 - Сколько тогда было штатов в США?
 - установление кореферентности (coreference resolution)

Многозначность

- Ключевая проблема обработки текстов
- Я траву **косил косой**,
Дождик вдруг пошел **косой**.
Бросил я тогда **косить**
И на Стешу стал **косить**.
Ну а Стеша, ох, краса,
Как огонь её **коса**!

Многозначность

- Морфологическая

- часть речи
- мой (-- нос, -- руки)
- look (look at me, have a look)

Алгоритмы определения
частей речи (part of
speech tagging)

- Синтаксическая

- мужу изменять нельзя
- мать любит дочь
- Flying planes can be dangerous

Синтаксический
разбор (parsing)

Многозначность

- Лексическая (семантическая)

- Омонимия (ключ)

- полисемия (платформа)

- семантическая многозначность (лиса)

разрешение
лексической
многозначности (word
sense disambiguation)

- Прагматическая

- Огонь! (в армии или в комнате)

- You have a green light

Многозначность и перевод

- Help для Windows 95

... Мышь может неадекватно реагировать на щелчок по почкам. Но не спешите! Это могут быть физические проблемы, а не клоп Окон 95.

Почистите вашу мышь.

Отсоедините ее поводок от компьютера, вытащите гениталий и промойте его и ролики внутренностей спиртом. Снова зашейте мышь.

Проверьте на переломы поводка.

Подсоедините мышь к компьютеру.

Приглядитесь к вашей прокладке (подушке) - она не должна быть источником мусора и пыли в гениталии и роликах.

Поверхность прокладки не должна стеснять движения мыши.

...

Сложность языка

- Естественный язык:
 - многозначен на всех уровнях
 - сложное, едва уловимое использование контекста для передачи значения
 - включает знания и рассуждения о мире
- Но обработка естественного языка может быть иногда очень простой
 - использование грубых признаков часто позволяют достичь очень хороших результатов

Понимание

- Тест Тьюринга
- Китайская комната
- ELIZA (1966)
 - „I am unhappy.“
 - „How long have you been unhappy?“
 - „Six months. Can you help me?“
 - „What makes you think I help you?“
 - „You remind me of my sister.“
 - „Can you tell me more about your sister?“
 - „I like teasing my sister.“
 - „Does anyone else in your family like teasing your sister?“
 - „No. Only me.“
 - „Please go on.“

ELIZA

User: Men are all alike.

ELIZA: IN WHAT WAY?

User: They're always bugging us about something other

ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE?

User: My boyfriend made me come here

ELIZA: YOUR BOYFRIEND MADE YOU COME HERE

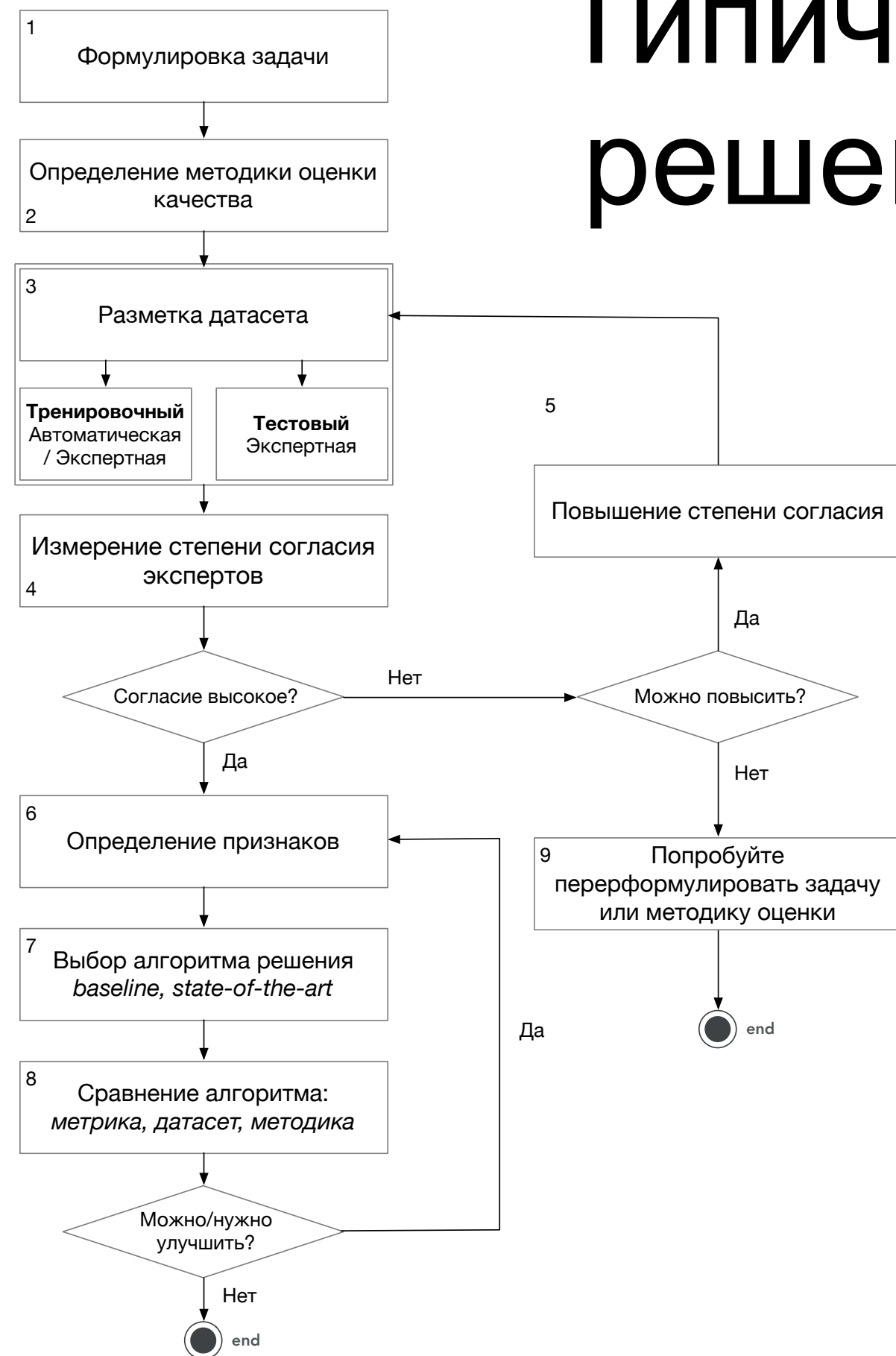
User: He says I'm depressed much of the time

ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED

```
import re
...
sentence=re.sub("I'm", "YOU ARE", sentence)
...
sentence=re.sub(".* YOU ARE (depressed|sad) .*", "I'M SORRY TO HEAR YOU ARE \\1", sentence)
sentence=re.sub(".* all .*", "IN WHAT WAY?", sentence)
sentence=re.sub(".* always .*", "CAN YOU THINK OF A SPECIFIC EXAMPLE?", sentence)
```

Типичная схема решения задач

- На каждом шаге производится *исследовательский поиск*
- Способ оценки качества часто влияет на постановку задачи
- Тестировать алгоритм надо на данных, которые он никогда “не видел”
- Согласие экспертов показывает разрешимость задачи людьми и определяет *верхнюю границу* качества
- Для честного сравнения алгоритмов должны быть зафиксированы датасет и методика
- В реальных задачах шаги 1-5 занимают до 90% всего времени



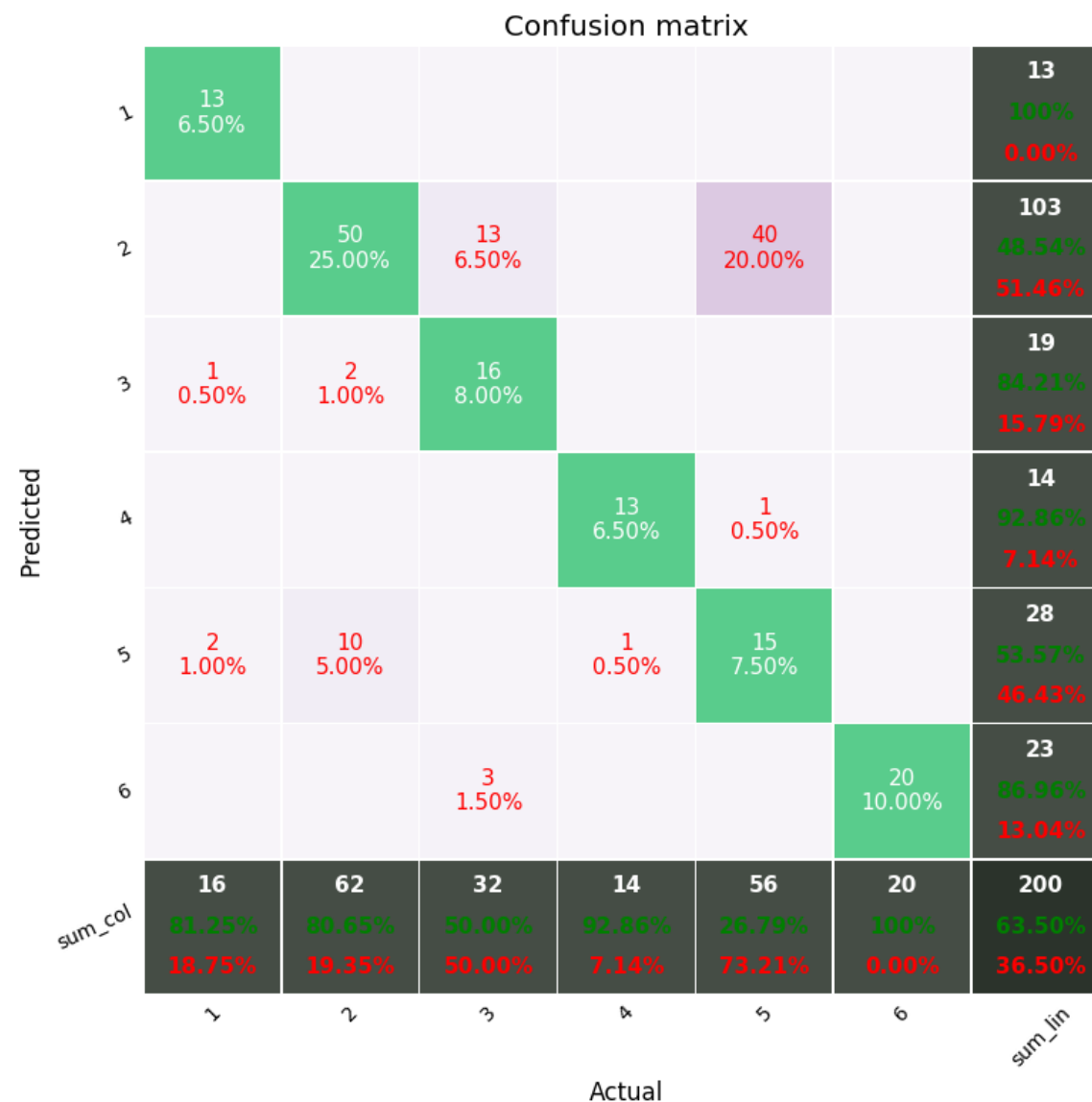
Пример

В начале 2019 года аппарат миссии NASA "Вояджер" (Voyager-2) пересек гелиопаузу и вышел в межзвездное пространство. Его брат-близнец (Voyager-1) сделал это шестью годами ранее. Зонды, запущенные в 1977-м для исследования планет-гигантов, до сих пор работоспособны, запасов радиоактивного топлива хватит до 2030 года.

- Определите ключевые слова текста

Матрица ошибок

- Confusion matrix
- Позволяет увидеть различие в ответах



Каппа Коэна

- Cohen's kappa
- Позволяет измерить согласие двух экспертов по сравнению со случайным совпадением ответов

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

p_o — наблюдаемое согласие

p_e — случайное согласие

Каппа Коэна

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

	Спам	Не спам	Сумма
Спам	20	3	23
Не спам	5	34	39
Сумма	25	37	62

$$p_o = \frac{a + d}{a + b + c + d} = \frac{20 + 34}{62} = 0.87$$

Каппа Коэна

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

	Спам	Не спам	Сумма
Спам	20	3	23
Не спам	5	34	39
Сумма	25	37	62

$$p_{\text{spam}} = \frac{a + b}{a + b + c + d} \cdot \frac{a + c}{a + b + c + d} = \frac{23}{62} \times \frac{25}{62} = 0.15$$

Каппа Коэна

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

	Спам	Не спам	Сумма
Спам	20	3	23
Не спам	5	34	39
Сумма	25	37	62

$$p_{\text{no spam}} = \frac{c + d}{a + b + c + d} \cdot \frac{b + d}{a + b + c + d} = \frac{39}{62} \times \frac{37}{62} = 0.38$$

Каппа Коэна

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

$$p_e = p_{\text{spam}} + p_{\text{no spam}} = 0.15 + 0.38 = 0.53$$

$$p_o = \frac{a + d}{a + b + c + d} = \frac{20 + 34}{62} = 0.87$$

$$\kappa = 1 - \frac{1 - 0.87}{1 - 0.53} = 0.72$$

Интерпретация

0.01 – 0.20	Незначительное
0.21 – 0.40	Удовлетворительное
0.41 – 0.60	Умеренное
0.61 – 0.80	Существенное
0.81 – 1.00	Почти идеальное

- Не всегда работает (в частности каппа может быть отрицательной)
- Нужно смотреть одновременно с матрицей ошибок
- <https://idostatistics.com/cohen-kappa-free-calculator/>

Задание

- Что делать, если оценщиков больше двух?
 - Самостоятельно изучить статистику “Каппа Флейса (Fleiss' kappa)”
 - Объяснить как с помощью нее оценить согласие для задачи определения ключевых слов (потенциальный вопрос на экзамене)

Резюме

- Хороший способ понять проблемы обработки текстов - сделать систему машинного перевода, вопросно-ответную систему или разговорного агента
- Обработка текста основана на формальных моделях
- Основы обработки текста лежат в компьютерных науках, математике, лингвистике, электротехнике и психологии
- Сейчас - удивительное время, когда революционные разработки используются повсеместно

Дополнительные ресурсы

- Конференции: ACL, EACL, COLING, CoNLL, EMNLP, Диалог
- <http://www.aclweb.org/anthology-new/>
- Книги:
 - D. Jurafsky, J.H. Martin. Speech and Language processing.
 - C. Manning, H. Schutze. Foundations of Statistical Natural Language Processing
 - Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep learning. MIT Press. 2016
- Курс 2016: <https://www.youtube.com/playlist?list=PL5cBzMOPJgCUn6TbfhqilyToW5lScOdd3>

Следующая лекция

- Методы классификации текстов
 - Логистическая регрессия
 - Машины опорных векторов (SVM)
 - Скрытая марковская модель (HMM, MEMM)
 - Условные случайные поля (CRF)
- Задача распознавания и классификации именованных сущностей (NERC)