Основы обработки текстов

лекция 1

О курсе

- Лектор: **Турдаков Денис Юрьевич**
- Лекции каждую **среду в 10.30 ауд. 523**
  - 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

Обработка текстов

Часть 1

О курсе

- Курс состоит из
  - –лекций,
  - –практической части и
  - –итогового экзамена
- Язык программирования Python
- Вся информация: http://tpc.at.ispras.ru

Практическая часть

- Одна из открытых задач обработки текстов
  - В этом году: выявление терминов и определение лексической многозначности
- Веб-интерфейс для проверки и задание будут доступны в начале октября

Обработка текстов

Часть 2

Python

- Значимые пробелы

```
if x==1:
    print ‘x is 1’
    print ‘внутри блока’
print ‘вне блока’
```

Python

- Конструкторы списков и словарей

```
number_list=[1,2,3,4]
string_list=['a','b','c','d']
mixed_list=['a',2,'c',7]
```

- Словарь (ключ/значение)

```
ages={'John':34, 'Sarah':20, 'Max':24}
```

- Доступ к элементам []

```
string_list[3]  # 'c'
ages['Sarah'] # '20'
```

Python

- Трансформация списков
  - [выражение for переменная in список]
  - [выражение for переменная in список if условие]

```
l1=[1,2,3,4,5,6,7,8,9]
print([v*10 for v in l1 if v>4])
 > [50, 60, 70, 80, 90]
```

- Функции:
  - zip

```
text = "a b c d"
text_splitted = text.split()
print(text_splitted)
> [‘a’, ‘b’, ‘c’, ‘d’]
print(list(zip(text_splitted[:-1], text_splitted[1:])))
> [(‘a’, ‘b’), (‘b’, ‘c’), (‘c’, ‘d’)]
```

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]
```

Обработка текстов

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

Что нужно знать о тексте?

- Рассмотрим приложение
  - **Siri**: интеллектуальный ассистент на iPhone

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)
- Прагматическая
  - –Огонь! (в армии или в комнате)
  - –You have a green light

разрешение лексической многозначности (word sense disambiguation)

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.“

Текущее состояние

- Разговорные агенты используются некоторыми авиакомпаниями
- Можно отдавать голосовые команды устройствам (телефон, в автомобиле...)
- Многоязыковой информационный поиск Google
- Перевод страниц Google
- Компании занимающиеся анализом текстов позволяют анализировать мнения и предпочтения людей

Новые взгляд на старые проблемы

- Информационный взрыв и масштабируемость (big data)
- Обработка сообщений в социальных сетях и Интернете в целом
- Автоматическое извлечение знаний из текста

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
- Журналы: Computational Linguistics, Natural Language Engineering, Speech & Language Processing
- 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
- Курс Stanford NLP: http://see.stanford.edu/

Следующая лекция

- регулярные выражения
- конечные автоматы