Обработка текстов на естественном языке
//
Natural Language Processing

Елена Тутубалина
кафедра интеллектуальных технологий поиска
Высшая школа ИТИС

Содержание курса

- Методы, ресурсы, приложения
- Базовые инструменты
- Лекции + практические задания + семестровое задание

Что требуется знать?

- линейная алгебра
- теория вероятностей, статистика
- машинное обучение
- лингвистика на «школьном уровне»

+

- навыки программирования

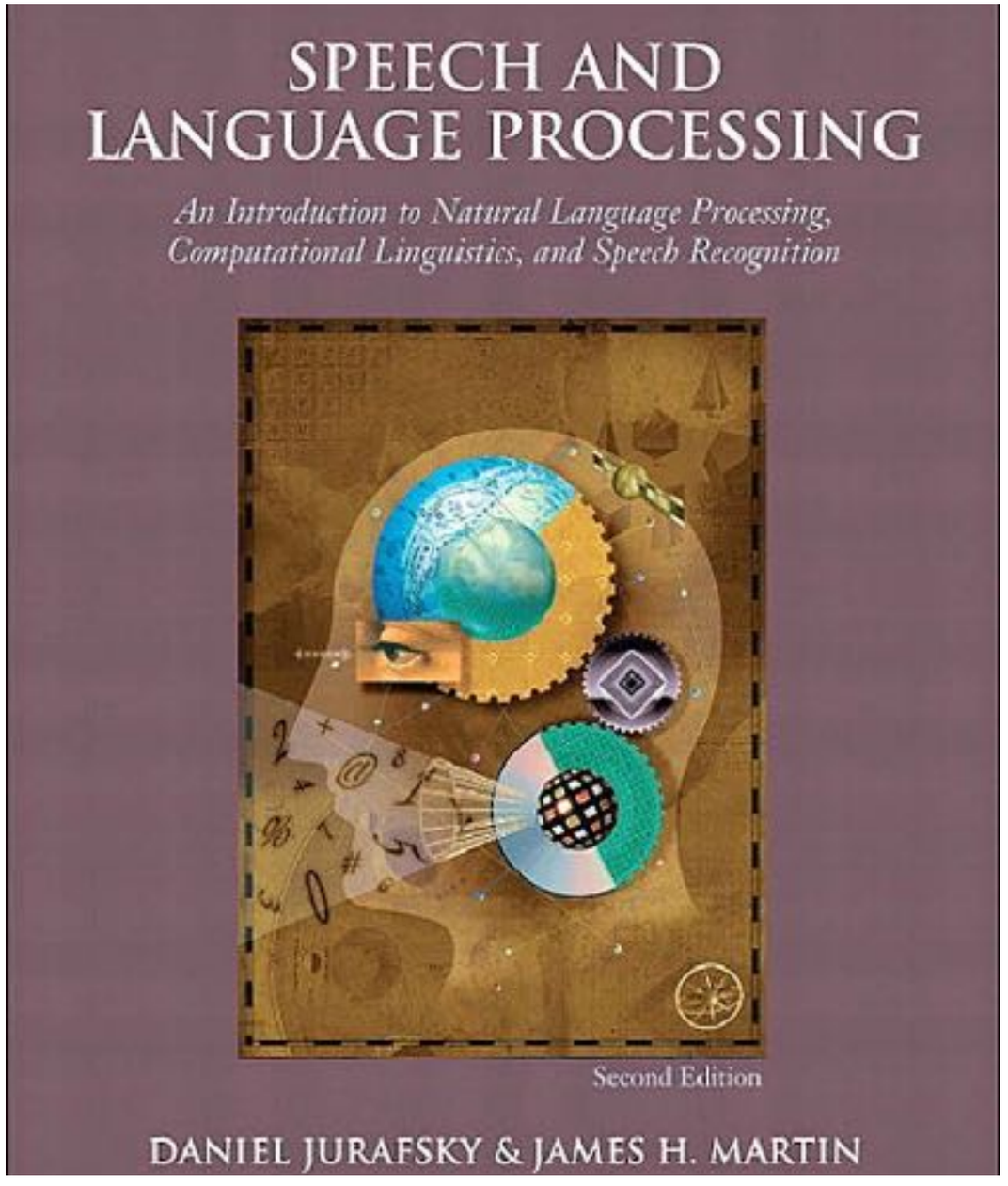
РЕСУРСЫ

Литература

Speech and Language Processing by Daniel Jurafsky and James H. Martin

https://www.cs.colorado.edu/~martin/slp2.html

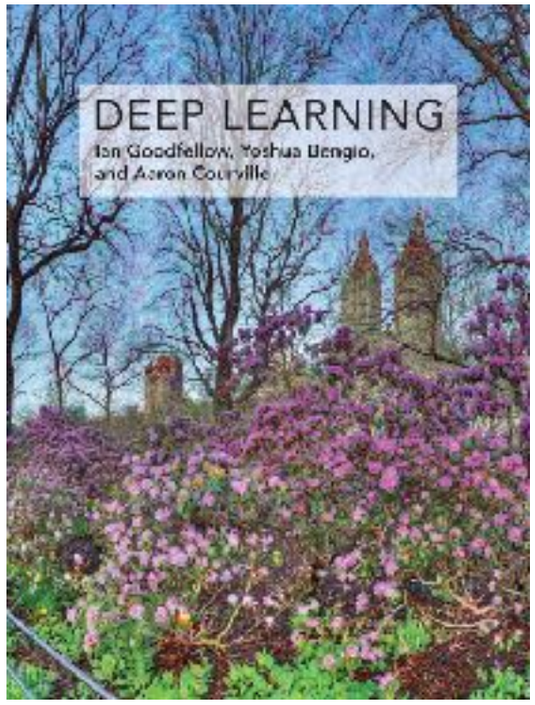
https://web.stanford.edu/~jurafsky/slp3/

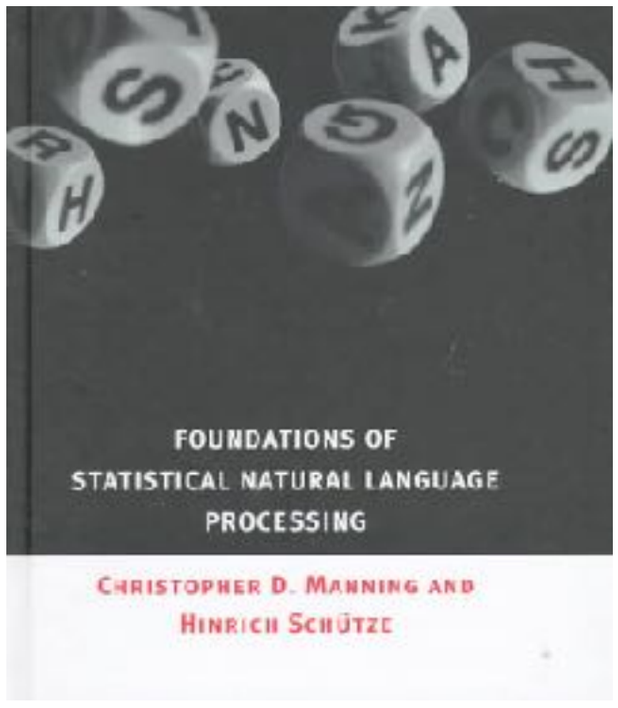


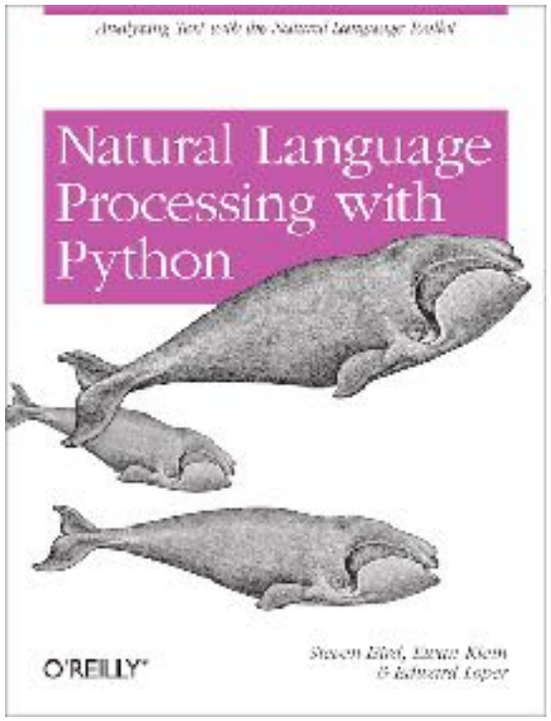
Глубокое обучение.
Погружение в мир
нейронных сетей.

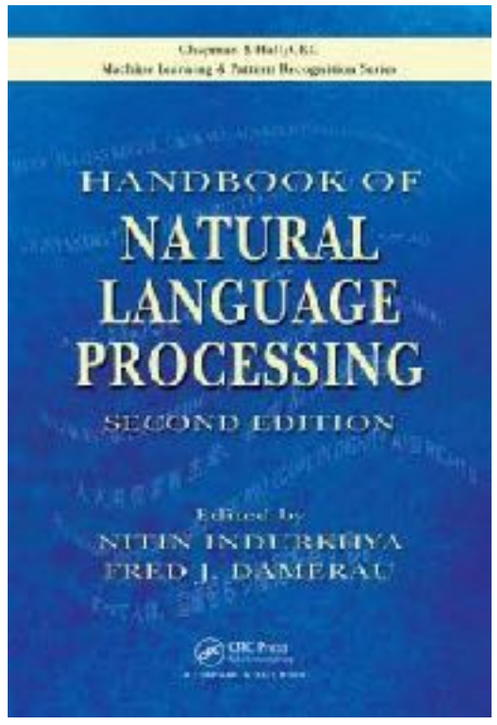
Сергей Николенко, А.
Кадурин, Е. Архангельская

https://www.piter.com/
product/glubokoe-
obuchenie

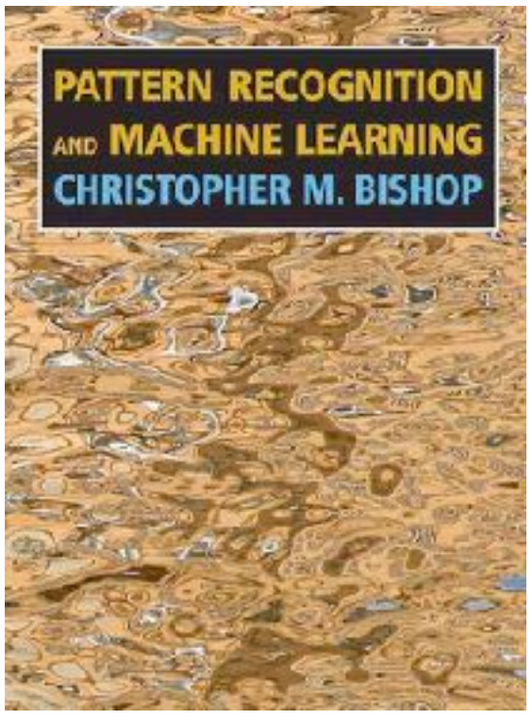

DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

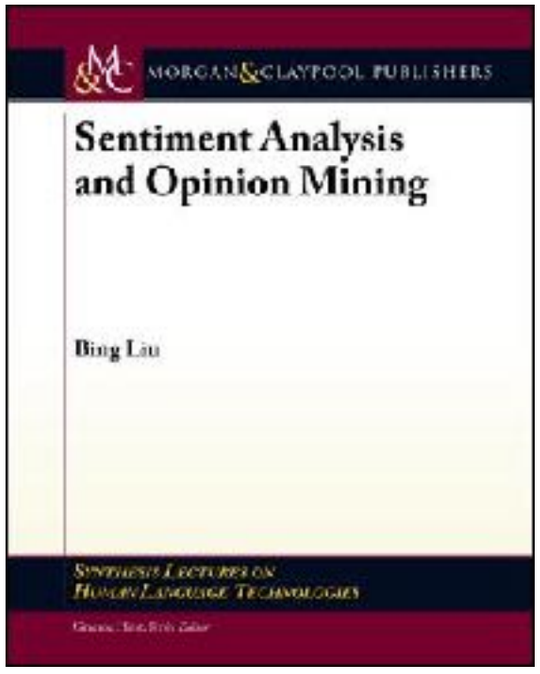

FOUNDATIONS OF
STATISTICAL NATURAL LANGUAGE
PROCESSING
CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE

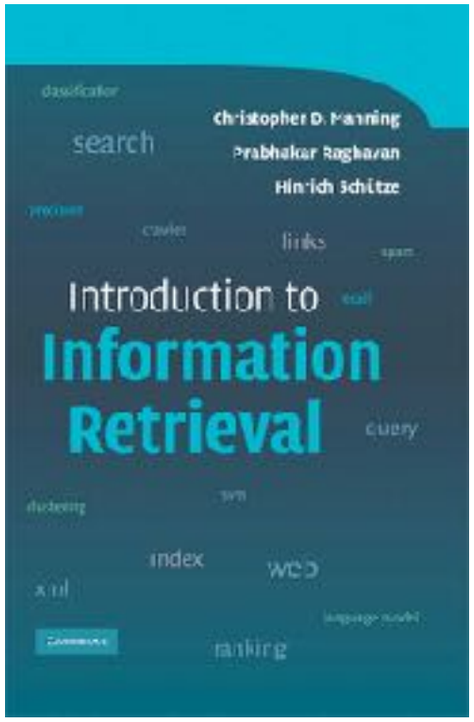

Natural Language
Processing with
Python
O'REILLY


HANDBOOK OF
NATURAL
LANGUAGE
PROCESSING
SECOND EDITION
Edited by
NITIN INDURKHYA
FRED J. DAMERAU


Н.В. Лукашевич
ТЕЗАУРУСЫ
В ЗАДАЧАХ
ИНФОРМАЦИОННОГО
ПОИСКА
Издательство
Московского университета


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


MORGAN&CLAYPOOL PUBLISHERS
Sentiment Analysis
and Opinion Mining
Bing Liu


Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze
Introduction to
Information
Retrieval

Курсы

- Coursera: Dan Jurafsky & Chris Manning (2012) — http://web.stanford.edu/~jurafsky/NLPCourseraSlides.html

- Coursera: Michael Collins (2013) — http://www.cs.columbia.edu/~mcollins/notes-spring2013.html

- CS224: Natural Language Processing with Deep Learning — http://cs224d.stanford.edu

- «Машинное обучение и анализ данных» (курс К.В. Воронцова, Coursera, machinelearning.ru)

- Введение в обработку естественного языка (курс П. Браславского, stepik.org)

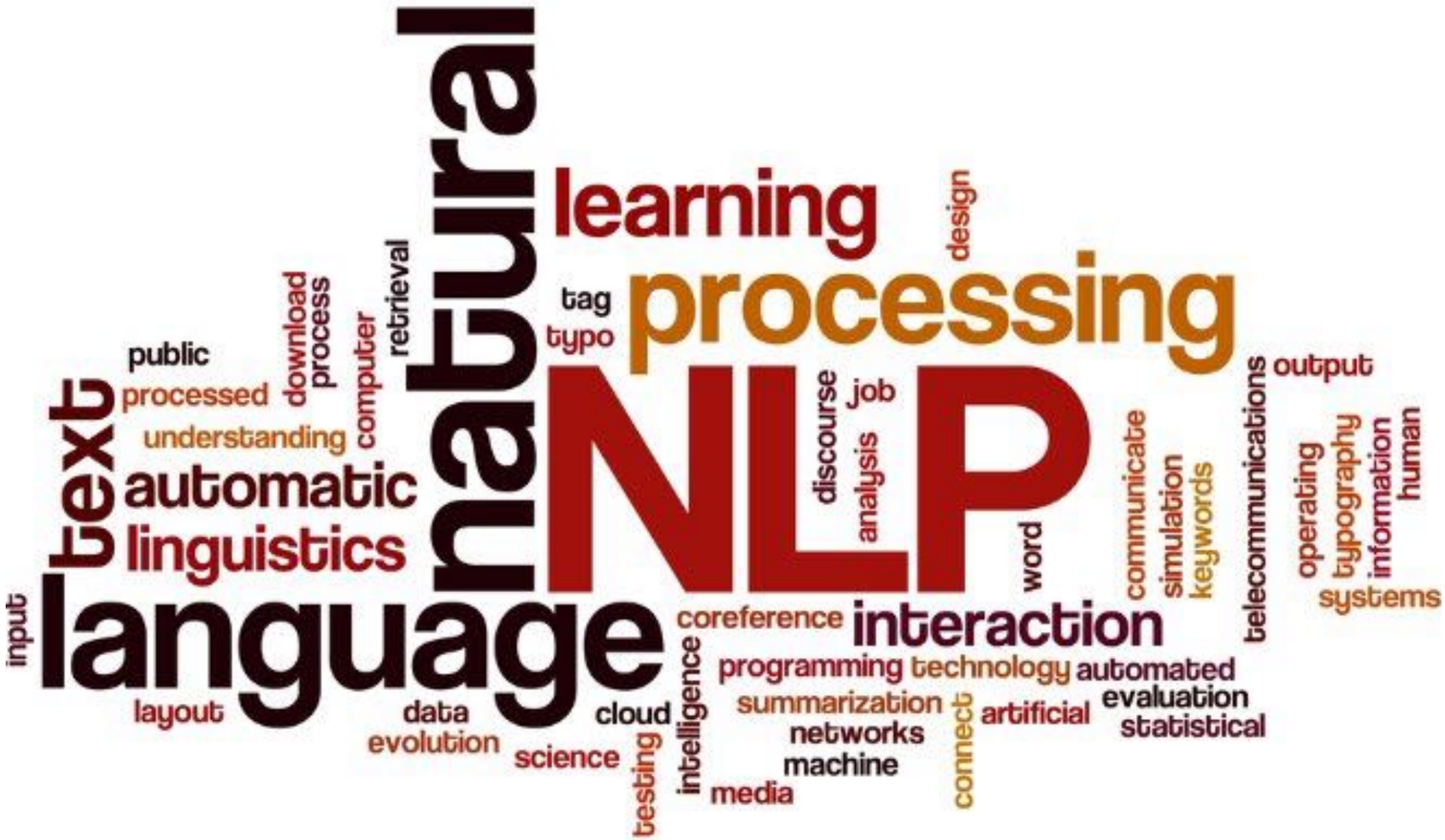
- OpenDataScience Machine Learning course https://github.com/Yorko/mlcourse_open

Ресурсы

- ACL, COLING, EACL, NAACL, EMNLP, RANLP... (http://aclweb.org/anthology/)
- SIGIR, CIKM, WSDM, ECIR
- INTERSPEECH
- Диалог (http://www.dialog-21.ru/)
- Research Groups: Google, Facebook, Microsoft, Yandex
- http://www.machinelearning.ru
- Каталог ресурсов для обработки естественного языка — https://nlpub.ru
- habrahabr по тегам «лингвистика», «машинное обучение», «нейронные сети»

Группы vk

- Deep Learning — https://vk.com/deeplearning
- Математическая лингвистика / NLP — https://vk.com/mathlingvo
- Кафедра интеллектуальных технологий поиска — https://vk.com/cllkazan
- RuSSIR — https://vk.com/russir
- Машинное обучение — https://vk.com/mashinnoe_obuchenie_ai_big_data
- Машинное обучение — https://vk.com/mlresearch



Что такое NLP

Чем занимается

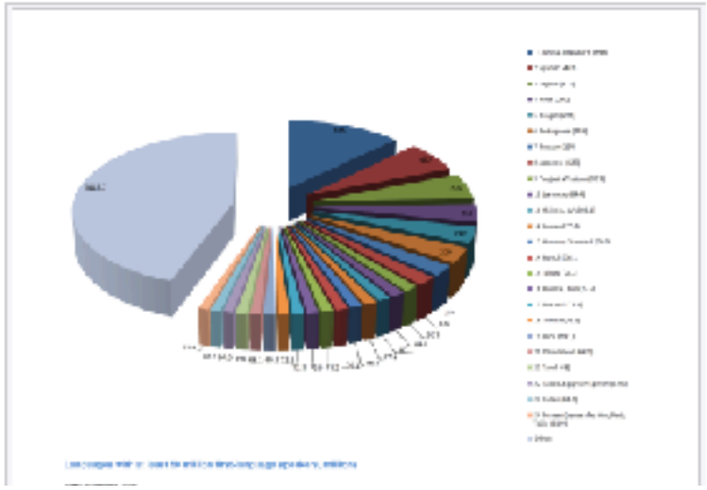
Автоматическая обработка речи/текста с использованием знаний о языке

Русский язык

Языки мирового значения [править | править вики-текст]

Современными международными языками можно считать[5] (в порядке убывания общего количества владеющих языком):

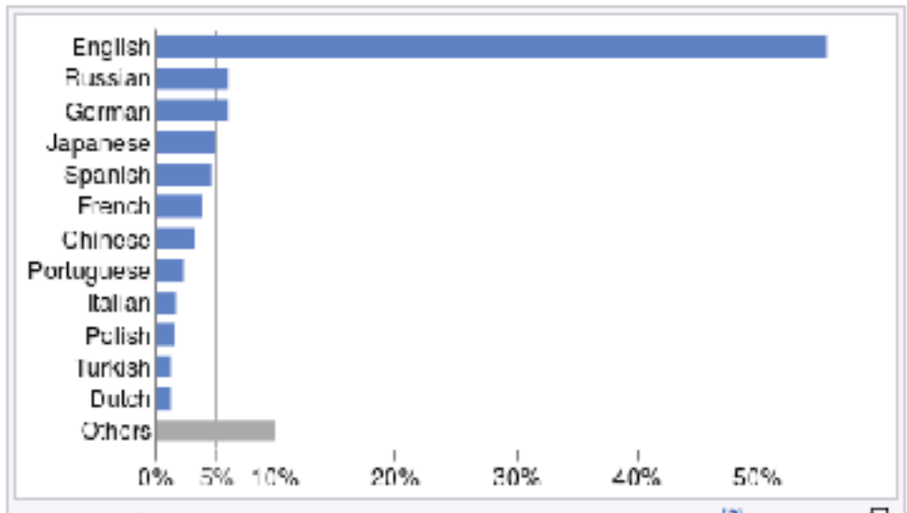
Ранг	Язык	Родной	Второй	Общее число носителей
1	Китайский язык[8]	1,2 миллиарда	до 300 миллионов	до 1,5 миллиарда
2	Английский язык[6]	500 миллионов	до 1 миллиарда	до 1,5 миллиарда
3	Испанский язык[9]	425 миллионов	до 125 миллионов	до 550 миллионов
4	Арабский язык[10]	300 миллионов	до 120 миллионов	до 420 миллионов
5	Русский язык	160 миллионов	до 100 миллионов	до 260 миллионов
6	Португальский язык[11]	230 миллионо		
7	Немецкий язык[12]	120 миллионо		
8	Французский язык[13]	75 миллионо		



Content languages for websites [edit]

Estimated percentages of the top 10 million websites using various content languages as of 4 March 2017:[2]

Rank	Language	Percentage
1	English	51.6%
2	Russian	6.6%
3	Japanese	5.6%
4	German	5.6%
5	Spanish	5.1%
6	French	4.1%
7	Portuguese	2.6%
8	Italian	2.3%
9	Chinese	2.0%
10	Polish	1.7%
11	Turkish	1.6%



Content languages for websites as of 12 March 2014[2]

Особенности русского языка

- Флективный язык
- Более свободный синтаксис
- Мало ресурсов

Термины

- Computational linguistics / математическая/компьютерная лингвистика
- Natural language processing / обработка естественного языка / автоматическая обработка текстов
- Natural language engineering, human language technology
- Прикладная лингвистика
- Speech and language processing
- Speech recognition and synthesis / распознавание и синтез речи

- Междисциплинарная область:
  - Computer Science
  - лингвистика
  - логика
  - психология
  - ML
  - философия
  - etc.
- Значение NLP
  - Связь языка и сознания
  - Объем текстовых /речевых данных
  - Мобильные технологии
  - Многоязычие

Разделы знаний о языке

- Фонетика
- Морфология
- Синтаксис
- Семантика
- Прагматика

Инструменты vs. Приложения

Основные приложения

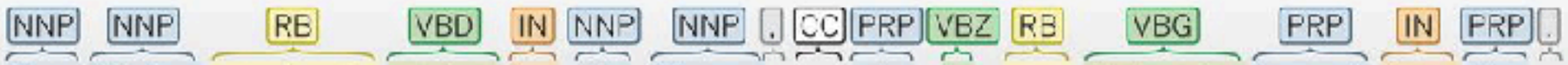
- Диалоговые системы / системы общения
  conversational agents / dialog systems
- Вопросно-ответные системы / question answering
- Информационный поиск / information retrieval
- Машинный перевод / machine translation

Основные приложения

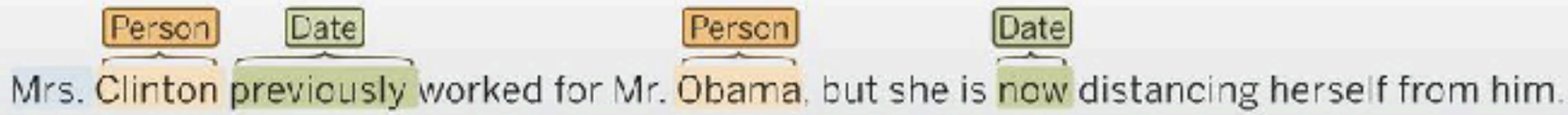
- Извлечение информации / information extraction
- Анализ тональности / sentiment analysis
- Автоматическое реферирование / automatic summarization
- Обучение языку / language learning

Syntactic Tasks

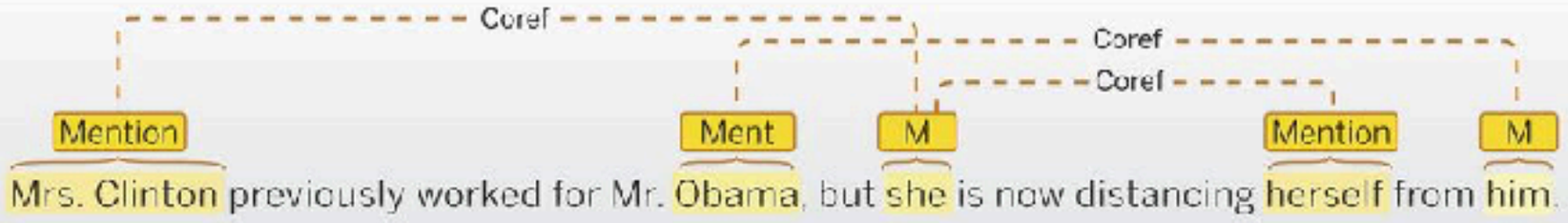
Part of speech:



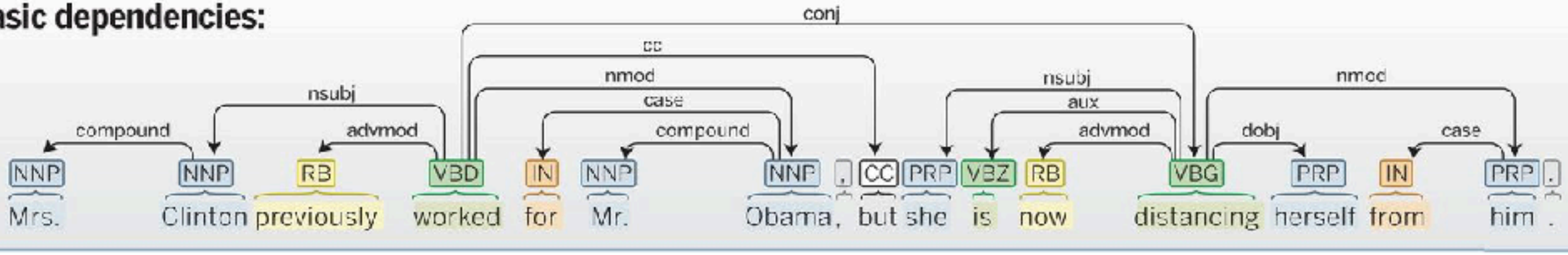
Named entity recognition:



Co-reference:



Basic dependencies:



Information Extraction

Subject: **curriculum meeting**
Date: January 15, 2012
To: Dan

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris

Create new Calendar entry

"Find me an Italian restaurant in New York City."
Action · Food type · search type · location

"And what's the weather there tomorrow?"
search type · city name · time

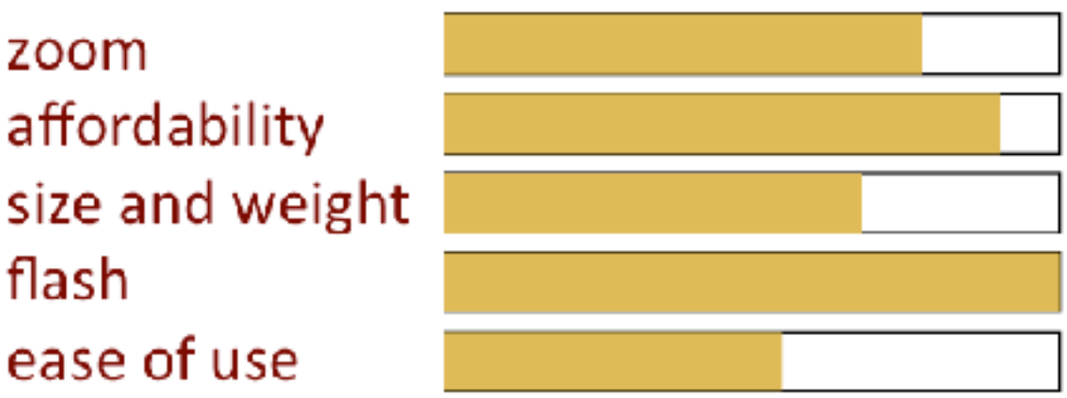
Question Answering: IBM's Watson

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

→ Bram Stoker

Information Extraction & Sentiment Analysis

Attributes:



Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

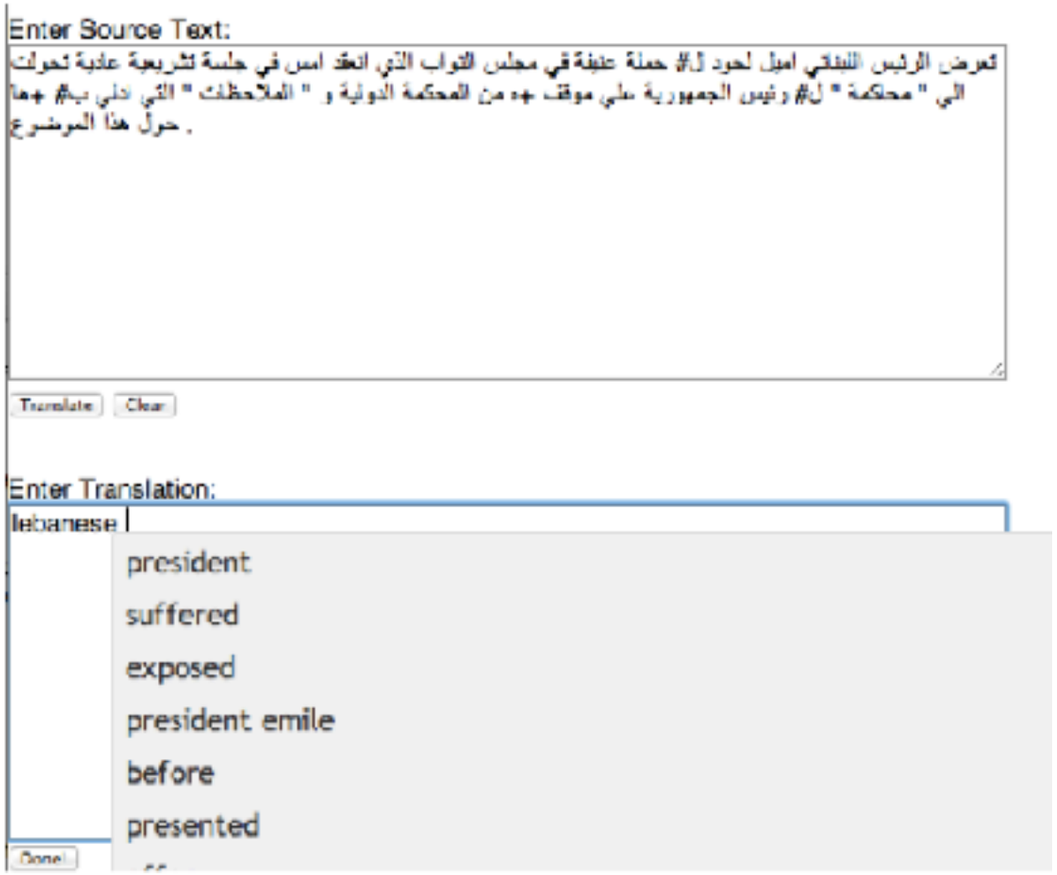
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators



NLP problems

- Well-defined syntactic problems with semantic complications:
  - part-of-speech tagging
  - morphological segmentation
  - stemming and lemmatization
  - sentence boundary disambiguation and word segmentation
  - named entity recognition
  - word sense disambiguation
  - syntactic parsing
  - coreference resolution

- Well-defined semantic problems:
  - language modeling;
  - sentiment analysis;
  - relationship/fact extraction;
  - question answering;
- Text generation problems, usually not so very well defined:
  - text generation per se;
  - automatic summarization;
  - machine translation; dialog and conversational models

Сложность NLP

- Неоднозначность (ambiguity)
- Многие задачи можно рассматривать как задачи снятия неоднозначности (disambiguation):
  - Печь – существительное или глагол?
  - Лук – овощ, оружие или фотография?
  - Только рупор капитана //их к отплытью призовет. – призовет капитана или рупор капитана?
  - Скрипка, лиса или скрип колеса

Методология

- Правила
- Статистика
- Основные модели: конечные автоматы, системы на основе правил, логика, вероятностные модели, векторное представление
- Основные методы: поиск в пространстве состояний (динамическое программирование), машинное обучение

КРАТКАЯ ИСТОРИЯ

1940-е и 1950-е

- Язык изучают разные науки: радиотехника, информатика/кибернетика, лингвистика, психология, философия
- Теория формальных языков, КСГ (CFG)
- Теория информации
  - Канал с помехами (noisy channel), теория кодирования

1957-1970

- Ноам Хомский
- ОЕЯ в рамках ИИ
  - «игрушечные» системы на правилах
- Байесовские методы (определение авторства)
- Брауновский корус (Brown corpus) – 1M слов (1964)

1970-1983

- Статистический подход (HMM, noisy channel, ...)
- Логика
- Понимание языка (от синтаксиса к семантике)
- Моделирование дискурса

1983-1993

- Конечные автоматы (FSA) в морфологии, фонологии и синтаксисе

- Подходы «от данных» (data-driven), новые стандарты в оценке (evaluation)

- Генерация речи

1994-1999

- Вероятностные методы ++
  - Частеречная разметка (POS tagging), синтаксический анализ (parsing), разрешение анафоры (anaphora resolution), ...
- Приложения +
- Веб

2000-2008

- Доступные данные
- Мероприятия по оценке
- Взаимодействие с сообществом ML
- Высокопроизводительные системы
- Подходы «без учителя» (unsupervised) – topic modeling, LDA
- Статистический машинный перевод (SMT)

2008-2016

- Deep learning ++
- Приложения ++
- Индустрия

NLP в СССР и России

- Теория «Смысл <=>Текст» (Игорь Мельчук, 1960-е гг.)
  - Машинный перевод (ЭТАП), синтаксический анализ
- Информационный поиск
  - Рамблер, Яндекс, Mail.Ru
- Information Extraction: Интегрум, Медиалогия, Крибрум, ...
- Машинный перевод: ЭТАП, ПРОМТ, Яндекс, ABBYY
- Корпусы: НКРЯ, opencorpora
- Инструменты: АОТ, pymorphy
- Инициативы по оценке: РОМИП, Dialog evaluation

Summary

- Natural language processing is a field at the intersection of computer science, artificial intelligence, linguistics

- Goal: for computers to process or «understand» natural language in order to perform tasks that are useful

- Fully understanding and representing the meaning of language is an illusive goal

- Perfect language understanding is AI-complete

- Text mining is the process of extracting information or discovering knowledge automatically from texts

Задания

- Составить список доступных корпусов текстов (e.g., дамп русскоязычной Википедии)

- Почитать про открытые соревнования алгоритмов разных лет:
  - Dialog evaluation (http://dialog-21.ru)
  - SemEval
  - ShARe/CLEF eHealth tasks
  - etc.