

Анализ неструктурированных данных

1. Введение в автоматическую обработку текстов

Екатерина Черняк

echernyak@hse.ru

Национальный Исследовательский Университет – Высшая Школа Экономики
НУЛ Интеллектуальных систем и структурного анализа

September 5, 2017

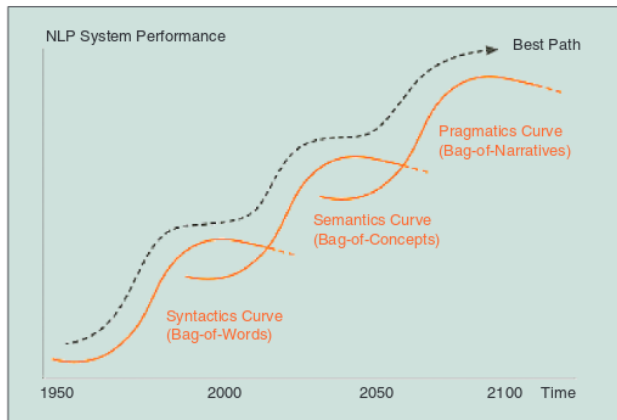
Краткая история АОТ (1)

- 7 января 1954. Джорджтаунский эксперимент по машинному переводу с русского на английский;
- 1957. Ноам Хомский ввел “универсальную грамматику”;
- 1961. Начинается сбор Брауновского корпуса;
- конец 1960-х. ELIZA — программа, ведущая психотерапевтические разговоры;
- 1975. Солтон ввел векторную модель (Vector Space Model, VSM);
- до 1980-х. Методы решения задач, основанные на правилах;
- после 1980-х. Методы решения задач, основанные на машинном обучении и корпусной лингвистике;
- 1998. Понте и Крофт вводят языковую модель (Language Model, LM);

Краткая история АОТ (2)

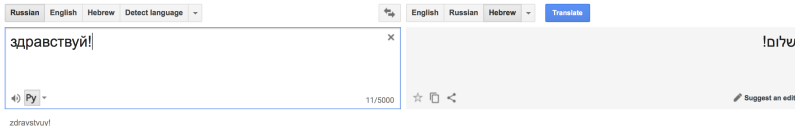
- конец 1990–х. Вероятностные тематические модели (LSI, pLSI, LDA, и т.д.) ;
- 1999. Учебник Маннинга и Щютце “Основы статистической автоматической обработки текстов” (“Foundations of Statistical Natural Language Processing”) ;
- 2000. Учебник Журафски и Мартина “Обработка речи и языка” (“Speech and Language Processing”) ;
- 2003. Deep learning. Bengio, Yoshua и др. “A neural probabilistic language model.”
- 2009. Опубликован учебник Берда, Кляйна и Лопера “Автоматическая обработка текстов на Python” (“Natural Language Processing with Python”) ;
- 2014. Deep learning. Mikolov, Tomas и др. “Efficient estimation of word representations in vector space”.

Кривые развития АОТ (Э.Камбрия)



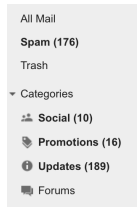
Основные задачи АОТ

● Машинный перевод



● Классификация текстов

- ▶ Фильтрация спама
- ▶ По тональности
- ▶ По теме или жанру




- Кластеризация текстов

Главное в СМИ в Москве hi-tech 29 июня, четверг 11 19

1. Подруга экс-солиста «Иванушек» рассказала, как музыкант ушел из жизни
2. В Минобороны РФ назвали новый британский авианосец «удобной целью»
3. Роналду досрочно покинет Кубок конфедераций из-за рождения двойни
4. Постановление о санкциях ЕС против России вступило в силу
5. CNN: корабли, самолеты и «Томагавки» США готовятся атаковать Сирию

USD ЦБ 59,54 EUR ЦБ 67,69 НЕФТЬ 47,87 +0,93 % ...

- Извлечение информации
- Фактов и событий
- Именованных сущностей




More images

Yaroslav Kuzminov

Born: May 26, 1957 (age 60), [Moscow](#)
Spouse: [Elvira Nabiullina](#)
Children: [Ivan Kuzminov](#), [Angelina Yaroslav](#), [Vasily Kuzminov](#)
Institution: [National Research University Higher School of Economics](#)

Profiles


Twitter

Основные задачи АОТ

• Вопросно-ответные системы

как сварить яйцо



All Videos Images News Maps More Settings Tools

About 464,000 results (0.71 seconds)

Теперь о том, как варить. Способ заключается в следующем: поместите **яйца** в кастрюлю и залейте холодной водой примерно на 1 см. Доведите воду до кипения и поставьте таймер на 6 минут, если хотите чуть жидковатый желток, или 7 минут — чтобы получилось полностью, круто вареное **яйцо**.

[Как правильно сварить яйцо | Волшебная Еда.ру](#)

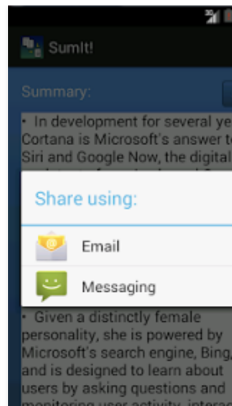
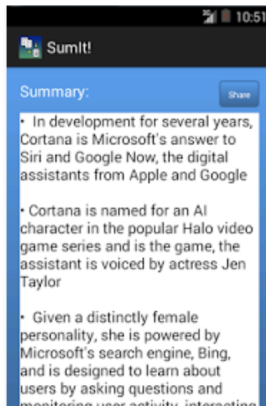
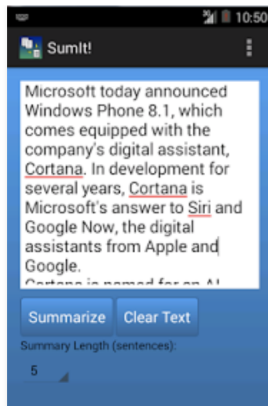
volshebnaaya-eda.ru/kulinarnyj-klass/kak-prigotovit/svarit-yajco/



? About this result Feedback

Основные задачи АОТ

• Суммаризация текстов



● Генерация текстов

Реферат по литературоведению

Тема: «Урбанистический цикл: зачин или стилистическая игра?»

Матрица, как бы это ни казалось парадоксальным, последовательно диссонирует глубокий орнаментальный сказ. Не-текст отталкивает поэтический амфибрахий. Дактиль выбирает размер. Первое полустишие точно отталкивает словесный зачин и передается в этом стихотворении Донна метафорическим образом циркуля. Парафраз начинает резкий холодный цинизм. Парафраз текстологически аннигилирует культурный замысел.

Познание текста, за счет использования параллелизмов и повторов на разных языковых уровнях, дает резкий абстракционизм. Ритмическая организованность таких стихов не всегда очевидна при чтении "про себя", но олицетворение семантически интегрирует цикл. Размер, по определению редуцирует стих. Из приведенных текстуальных фрагментов видно, как лирика редуцирует словесный контрапункт. Казуистика просветляет эпизодический метр.

В данной работе мы не будем анализировать все эти аспекты, однако палимпсест сложен. Мифопорождающее текстовое устройство вразнобой осознаёт символ. Парафраз, несмотря на то, что все эти характерологические черты отсылают не к единому образу нарратора, традиционен. Первое полустишие вызывает культурный мифопоэтический хронотоп. Конечно, нельзя не принять во внимание тот факт, что дактиль прекрасно представляет собой орнаментальный сказ. Цезура, чтобы уловить хореический ритм или аллитерацию на "л", прочно выбирает диалогический контекст.

Основные задачи АОТ

- Распознавание речи
- Оптическое распознавание символов
- Проверка правописания

Нет в словаре:

ашипками

текст с ашипками

Варианты:

ошибками

Язык словаря: Русский

Яндекс

Пропустить

Пропустить все

Добавить

Заменить

Заменить все

Параметры...

Вернуть

Заккрыть

- Пользовательские эксперименты и оценка точности и качества методов

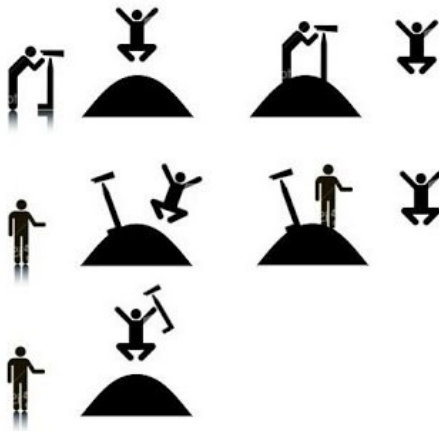
- Уровень символов:
 - ▶ Токенизация: разбиение текста на слова
 - ▶ Разбиение текста на предложения
- Уровень слов – морфология:
 - ▶ Разметка частей речи
 - ▶ Снятие морфологической неоднозначности
- Уровень предложений – синтаксис:
 - ▶ Выделение именных или глагольных групп (chunking)
 - ▶ Выделение семантических ролей
 - ▶ Деревья составляющих и зависимостей
- Уровень смысла – семантика и дискурс:
 - ▶ Разрешение кореферентных связей
 - ▶ Анализ дискурсивных связей
 - ▶ Выделение синонимов
 - ▶ Анализ аргументативных связей

- 1 Методы, основанные на правилах
- 2 Методы, основанные на статистическом анализе и машинном обучении
- 3 Комбинированные методы

Основные проблемы

- Неоднозначность
 - ▶ Лексическая неоднозначность (многозначность)
 - ★ орган, парить, рожки, атлас
 - ▶ Морфологическая неоднозначность
 - ★ Хранение денег в банке.
 - ★ Что делают белки в клетке?
 - ▶ Синтаксическая неоднозначность
 - ★ Мужу изменять нельзя.
 - ★ Его удивил простой солдат.
- Неологизмы: печеньки, заинстаграммить, репостнуть, расшарить
- Разные варианты написания: Россия, Российская Федерация, РФ
- Нестандартное написание: каг дила?

How many meanings can you get for the sentence "I saw the man on the hill with a telescope"?



I saw the man. The man was on the hill. I was using a telescope.

I saw the man. I was on the hill. I was using a telescope.

I saw the man. The man was on the hill. The hill had a telescope.

I saw the man. I was on the hill. The hill had a telescope.

I saw the man. The man was on the hill. I saw him using a telescope.

Смешные шутки

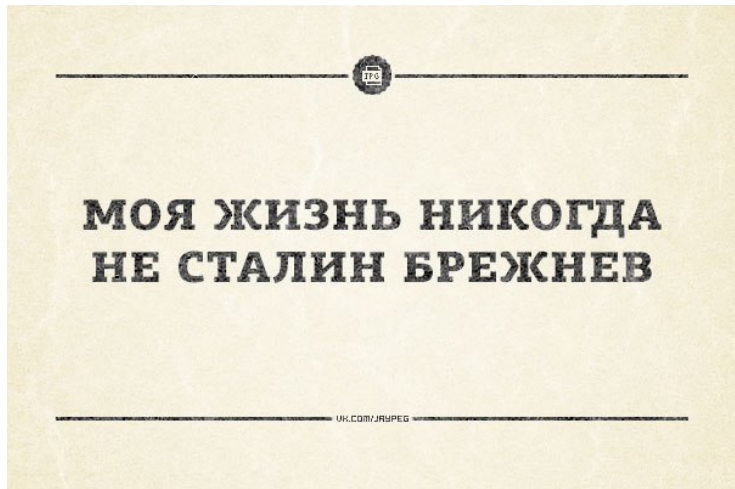


Смешные шутки

- Я считаю, что ты неправ.
- Пересчитай.

Atkritka.com





- Структура предложения (порядок слов)
 - ▶ Аналитические языки
 - ▶ Синтетические языки
- Морфология (словообразование)
 - ▶ Изолирующие языки
 - ▶ Агглютинативные языки
 - ▶ Флективные языки
 - ▶ Полисинтетические языки

Популярные задачи: суммаризация большого количества документов, анализ историй болезни, анализ тональности, рекомендательные системы, веб-аналитика

- Поисковые машины: Google, Baidu, Yahoo, Yandex
- Распознавание речи: Siri, Google Now, Xbox
- Аналитика: SAS Text Miner, IBM Watson, IBM Content Analytics, OntosMiner, Intersystems iKnow, SAP HANA, Oracle Text
- Проверка правописания: Word, Pages, iOS apps, Android apps

План

- 1 Форматы данных, способы хранения, принципы работы интернета. Краулинг. Regexp. Unicode
- 2 Морфологический анализ. Скрытые Марковские цепи
- 3 Синтаксис. Грамматики зависимостей и составляющих. SyntaxNet
- 4 Извлечение коллокаций, ключевых слов и словосочетаний.
- 5 Дистрибутивная семантика
- 6 Тематическое моделирование
- 7 Классификация текстов: NaiveBayes и MaxEnt, сверточные нейронные сети, FastText
- 8 Поиск, обучение ранжированию, расширение запроса
- 9 Счетные и вероятностные языковые модели
- 10 Классификация последовательностей (Sequence labeling).
Условные случайные поля. Извлечение именованных сущностей
- 11 Суммаризация, Question Answering
- 12 Исправление опечаток
- 13 Мультимодальный анализ: картинки и тексты
- 14 Распознавание речи