Анализ текстов

Белов Андрей

Семинары для ФИВТ МФТИ, 2017

Этапы извлечения информации

- Разбиение на предложения и токены
- Морфологический анализ
- Синтаксический анализ
- Семантический анализ
- Прагматический анализ

Примеры проблем, возникающих на этих этапах

- Разбиение на предложения и токены
  - Сокращения, инициалы, тире-дефис, URL, смайлики
- Морфологический анализ
  - Омонимия: три, печь
- Синтаксический анализ
  - Омонимия: Microsoft купила Google
- Семантический анализ
  - Омонимия, Анафорические связи
- Прагматический анализ

Признаки слов

- Слово
- Регистр: Капитализовано, ПОЛНОСТЬЮ, СмеШанный, паттерн XxXx
- Пунктуация: оканчивается на точку (г.), имеет внутреннюю пунктуацию (г.р.) и т.д
- Цифровой шаблон (19 = ##, 3D4 = #x#)
- Символы (слово из одной буквы, наличие греческих букв)
- Морфология
    - Приставки, суффиксы, корень, окончание
    - Часть речи, род, падеж, …
    - Леммы
- Стемы
- Подстроки, суффиксы, аффиксы

N-граммы

- Сегодня на улице холодно
- bigram: Сегодня на, на улице, улице холодно
- 1-skip-2-gram: Сегодня улице, на холодно

Признаки из списков слов

- Общий словарь
- Словари сущностей (имена, фамилии, организации, знаменитости, города, страны)
- Словарь ключевых слов (“компания”, “ООО”, “LTD”)

Признаки, собранные по документу/документам

- Позиция в предложении, абзаце, документе
- Распространенность слова в тексте
- Использование контекстов слов

Об ABBYY Compreno

- Извлечение информации, классификация текстов, машинный перевод, информационный поиск
- Подробнее об алгоритмах:
  - http://www.dialog-21.ru/digests/dialog2014/materials/pdf/StarostinAS.full.pdf
  - http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf