Комплинг исследование научно-популярных текстов

Анастасия Кузнецова, Анна Лапидус,
Юлия Коломенская, Ксения Самойленко

НИУ ВШЭ

9 ноября 2017

План доклада

1. Цели и задачи проекта
2. Этапы работы
3. Что уже сделано?
4. Ресурсы
5. Дальнейшие действия

План доклада

Цели и задачи проекта

Цель проекта:
Компьютерно-лингвистический анализ научно-популярных текстов на русском языке.
Ожидаемые результаты проекта:

- Создание словаря научно-популярных терминов
- Написание нескольких статей на основе выводов, сделанных в ходе исследования
- Потенциал для использования полученных данных в прикладной сфере, например, для создания компилятора научно-популярных новостей или обзоров

Цели и задачи проекта

Задачи для реализации проекта:

- Проанализировать список русскоязычных научно-популярных ресурсов и определить релевантные для данного исследования
- Собрать данные:
  - Запросить непосредственно у владельцев ресурсов
  - Обкачать сайты (учимся работать с краулером)
- Разметить и структурировать полученные данные (text clusterization)
- Непосредственно анализ данных (text similarities, fact extraction etc)
- Выводы

План доклада

Этапы работы

Модуль 2

(0–1) Сбор данных, анализ и классификация источников

Модуль 3

(2–3) Применение комплинг методов, обработка результатов

Модуль 4

(4) Написание статьи

План доклада

Что уже сделано?

- Создан репозиторий проекта на GitHub
- 04.11 проведена первая беседа с научным руководителем, определена общая стратегия работы над проектом и план работы на 2 модуль
- Участники проекта распределили между собой ресурсы для анализа, результаты исследования выложили на GitHub
- Обсудили возникшие в ходе анализа ресурсов вопросы:
  - Какие тексты включать в исследование?
  - Включать ли новостные статьи?
  - Переводные статьи?
- Обратились за помощью к редакторам научно-популярных сайтов

План доклада

N+1. Контент

1. Новости (на основе *Science*, *Nature* и др. авторитетных изданий)
2. Материалы
   - Лонгриды (оригинальные статьи авторов)
   - Экспертные интервью
3. Блоги
4. Тесты и игры, которые мы рассматривать не будем

N+1. Структура

- Тексты разбиты по рубрикам, разным предметным областям (около 35)
- На главной странице разбиты по тегам, а не по рубрикам
- Едины формат разметки текстов статей
- Тегами в интервью обозначена прямая речь интервьюера и эксперта
- **Проблема:** лонгриды и экспертные интервью никак не отмечены в URL, а просто лежат в разделе *materials*.

ПостНаука. Форматы

- короткие видео-лекции с расшифровками
- FAQ — тексты, созданные на основе интервью, но оформленные как монологии
- talks — интервью «за жизнь»
- longreads — статьи и отрывки из книг
- подборки книг и фильмов
- микроформаты — тесты, мультфильмы

ПостНаука. Структура

- есть рубрикатор по URL
- есть разбивка по темам
- есть разбивка по "проектам"(главы отделены от статей), но это не отображено в урлах
- почти все материалы — «прямая речь», в остальных случаях лишнее можно отсечь за счет специального оформление

Элементы

- много новостей
- но кто их автор — не всегда понятно
- очень много энциклопедических разделов — календарь, библиотека, "масштабы и так далее
- большинство материалов —тексты без каких-либо цитат, отсылок, и комментариев

Предварительный анализ источников

1. **8 источников**
   N + 1, ProScience, Geektimes, Чердак, ПостНаука, Элементы, Полит.ру, Индикатор
2. Контент сайтов
   - Новости науки
   - Longreads - оригинальные статьи
   - Лекции
   - Интервью с экспертами
   - Переводы статей
3. Структура
   - Разная рубрикация
   - Разные способы кодировки значимой информации (рубрики, экспертные комментарии ученых)
   - Комментарии, рейтинги

Вопросы

- Нужно ли использовать переводные статьи? Отрывки из книг?
- Получение данных (редакторы, автоматические обращения к сайтам, API)

План доклада

Что дальше?

Ждем текстов, учимся выкачивать сайты.

Спасибо за внимание!