Построение тезауруса предметной области при помощи Википедии

Бочаров Виктор

Задача

- составить список имён собственных
- с указанием типов сущностей
- и отношений между ними

Ожидаемый результат

- Ювентус — это футбольный клуб
- Олимпико — это стадион
- Турин — это город
- Италия — это страна
- Ювентус находится в Турине
- Олимпико находится в Турине
- Турин находится в Италии
- Олимпико — домашний стадион Ювентус
- Старая Синьора — синоним Ювентус

Зачем?

- Поиск текстов на спортивную тематику
- До какой степени снимать объектную омонимию?
- Использовать в шаблонах
- Качество шаблонов vs качество тезауруса

Объектная омонимия

- Спартак (футбольный клуб, Варна) — Болгария
- Спартак-Цхинвали — Грузия
- Спартак (футбольный клуб, Кострома) — Россия
- Спартак (футбольный клуб, Куба) — Азербайджан
- Спартак (футбольный клуб, Москва) — Россия)
- Спартак (футбольный клуб, Рыхнов-над-Кнежной) — Чехия
- Спартак Златибор Вода — Суботица (Сербия)
- Спартак (футбольный клуб, Сумы) — Украина
- Спартак (футбольный клуб, Тамбов) — Россия
- Спартак (футбольный клуб, Цхинвал) — Южная Осетия
- Спартак (футбольный клуб, Щёлково) — Московская область (Россия)
- Спартак-Нальчик — Россия

Инструменты

- perl
- grep
- bash
- sort, uniq
- bzcat
- xargs

Исходный код

- http://code.google.com/p/clschool/
  - source/browse/
    - trunk/wiki_thesaurus_building/

Данные

- ruwiki-20110619-pages-articles.xml.bz2

```
<mediawiki ...>
    <siteinfo> ... </siteinfo>
    <page>
        <title>Заголовок статьи</title>
        <id>Уникальный номер статьи</id>
        <revision>
          <id>Уникальный номер правки</id>
          <timestamp>Дата / время совершения правки</timestamp>
          <contributor>
            <username>Логин автора</username>
            <id>Уникальный идентификатор автора</id>
          </contributor>
          <comment>Комментарий по поводу правки</comment>
          <text xml:space="preserve"> полный текст статьи </text>
        </revision>
    </page>
</mediawiki>
```

Особенности Википедии

- Постоянно меняется
- XML, UTF-8
- Wiki-разметка
- Размер (1.2Гб в bz2)

Викиразметка

- Документация
  - http://ru.wikipedia.org/wiki/Википедия:Вики-разметка
- Plain text
- Вики-форматирование
- Вики-ссылки
- Вики-шаблоны
- Редиректы
- Вики-категории

Вики-ссылки

В Лондоне хороший [[общественный транспорт]].

В Лондоне хороший общественный транспорт.

Лондон располагает хорошим [[общественный транспорт| общественным транспортом]].

Лондон располагает хорошим общественным транспортом.

Вики-ссылки

Окончания сливаются со ссылкой: [[ген]]ы, в [[2008 год]]у

Окончания сливаются со ссылкой: гены, в 2008 году

Вики-ссылки

Автоматически скрывается заключённое в круглых скобках:

[[царство (биология)|]].

Автоматически скрывается заключённое в круглых скобках:

царство.

Вики-ссылки

[[Медведи на улицах Москвы]] — это страница, которая ещё не создана.

Медведи на улицах Москвы — это страница, которая ещё не создана.

Вики-шаблоны

- http://ru.wikipedia.org/wiki/Шаблоны
- http://ru.wikipedia.org/wiki/Википедия:
  - Механизм_шаблонов
- {{Имя шаблона}}
- {{Имя шаблона|Аргумент 1|Аргумент 2|...}}
- {{Имя шаблона
  |Название аргумента 1 = значение аргумента 1
  |Название аргумента 2 = значение аргумента 2
  |...}}

Вики-шаблоны

{{lang-lt|Lietuva}}	лит. Lietuva
{{main|История Литвы}}	Основная статья: **История Литвы**
{{другие значения|Мамонт (значения)}}	У этого термина существуют и другие значения, см. Мамонт (значения).

Вики-шаблоны

- Шаблоны — это статьи
- Текст шаблона подставляется в статью

- ru.wikipedia.org/wiki/Шаблон:lang-lt
- [[Литовский язык|лит.]] {{lang1|lt|{{{1}}}}}

Редиректы

- ru.wikipedia.org/wiki/Юве

  –> ru.wikipedia.org/wiki/Ювентус

  (Перенаправлено с Юве)

- ru.wikipedia.org/wiki/Юве:

  #REDIRECT [[Ювентус]]

Инфобоксы

Дворец Спорта «Сокольники»

Местоположение	Москва
Построен	1956
Вместимость	5 530
Домашняя команда	ХК «Спартак» (КХЛ) МХК «Спартак» (МХЛ)

Спартак Москва

Конференция	Западная
Дивизион	Боброва
Основан	22 декабря 1946
Арена	ДС «Сокольники»
Город	Москва
Командные цвета	
Владелец	АКБ «Инвестбанк»
Генеральный менеджер	Андрей Яковенко
Главный тренер	Виктор Пачкалин
Капитан	Бранко Радивоевич
Аффилированные клубы	Крылья Советов (ВХЛ) МХК Спартак (МХЛ)

Инфобоксы

{{Карточка клуба КХЛ
|Название = Спартак Москва
|Цвет фона = red
|Цвет текста = White|
|Эмблема = [[Файл:HC Spartak logo.jpg|250px|Эмблема ХК «Спартак» Москва]]|
|Дивизион =Боброва
|Конференция = Западная
|Основан = [[22 декабря]] [[1946]]
|Арена = [[Сокольники (дворец спорта)|ДС «Сокольники»]]
|Город = {{Флаг|Москва}} [[Москва]]|
|Командные цвета = {{color box|red}} {{color box|white}}|
|Владелец = {{Россия}} [[Инвестбанк|АКБ «Инвестбанк»]]
|Цвета = {{color box|red}} {{color box|white}}|
|Генеральный менеджер = {{Флаг|Россия}} [[Яковенко, Андрей Васильевич|Андрей Яковенко]]
|Главный тренер = {{Флаг|Россия}} [[Пачкалин, Виктор Николаевич|Виктор Пачкалин]]

Как составить список шаблонов?

- bzcat ruwiki-20110619-pages-articles.xml.bz2 | grep -E -o "^{{[^}^|]+" | sort | uniq -c | sort -r -n

- 231490 {{примечания

  220954 {{!

  76743 {{rq

  63922 {{Несвободный файл/ОДИ

  50108 {{nobr

  45532 {{Отчёт о матче

Какие шаблоны нужны?

- Футболист, Хоккеист, Баскетболист, ...
- Стадион
- Карточка ФК, Хоккейный клуб, ...
- Государство
- НП, НП-[А-ЯЁа-яё]+
- ...

Шаг 1: выделить статьи

- mkdir i

  bzcat ruwiki-20110619-pages-articles.xml.bz2 | perl extract_infoboxes.pl i

- i/
  - Баскетболист/
    - 131837-Джордан, Майкл.page
    - 145999-Мэлоун, Карл.page
  - Стадион/
    - 100671-Хайбери (стадион).page
    - 125258-Татнефть-Арена.page

Какие аргументы нужны?

- Для стадионов:
  - название
  - местоположение
  - команда
  - ...
- Для городов:
  - русское название
  - страна
- ...

Стадион / Местоположение

- ls i/Стадион/*.page | xargs -I XX cat "XX" | grep "| местоположение"

- |местоположение = Хайбери, [[Лондон]]
  |местоположение= [[Порту]], [[Португалия]]
  |местоположение=[[Турин]], [[Италия]]
  |местоположение= [[Минск]], [[Белоруссия]]
  |местоположение=[[Турин]], [[Италия]]
  |местоположение = [[Химки]], [[Россия]]

Шаг 2: выделить аргументы

- mkdir result

  ls i/Стадион/ | xargs -I XX bash -c 'cat "i/Стадион/XX" | perl extract_fields.pl stad' > result/tab_stad.txt

- result/
  - tab_stad.txt
  - ...

Формат таблицы

- **id** - «100671»

  title - «Хайбери (стадион)»
- **ref-название** - «»
- **ref-местоположение** - «Лондон»
- **ref-команда** - «Арсенал (футбольный клуб, Лондон)»
- **название** - «Arsenal Stadium
Highbury»
- **местоположение** - «Хайбери, [[Лондон]]»
- **команда** - «[[Арсенал (футбольный клуб, Лондон)|Арсенал]] (1913-2006)»

Шаг 3: Выделить синонимы

- Редиректы
- bzcat ruwiki-20110619-pages-articles.xml.bz2 | perl extract_redirects.pl > tab_redirects.txt

- grep "\[\[Ювентус\]\]" tab_redirects.txt

  193832 Ювентус Турин #REDIRECT [[Ювентус]]

  2233000 Juventus #REDIRECT [[Ювентус]]

  2235049 Vecchia Signora #REDIRECT [[Ювентус]]

  2235050 Старая Синьора #REDIRECT [[Ювентус]]

  2995343 Юве #REDIRECT [[Ювентус]]

Проблемы

- Годится только для структурированных данных
- Лемматизация
- Неполнота

Вопросы?