Strukturyzacja słownika języka polskiego do pseudo-XML.

Antoni Piechnik

prowadzący: **dr inż. Marek Gajęcki**

22 września 2009

Spis treści

1 Wstęp

1.1 Cel Projektu

Głownym celem projektu było stworzenie narzędzia do konwersji wypisu ze słownika języka polskiego PWN do struktury typu XML o zadanej przez prowadzącego. Projekt miał zaznajomić nas zarówno z formą słownika jak również problemami związanymi z jego konwersją, na które napotkaliśmy.

1.2 Wizja

Projekt zaprojektowany został jako skrypt napisany w dynamicznie typowanym języku Python, który nie tylko pozwala na sprawne przetwarzanie dużych danych, ale również na wygodną pracę z tekstem. Skrypt składa sie z dwóch większych modułów,które razem odpowiadają za załadowanie tekstu słownika a następnie zparsowanie go do odpowiednich struktur.

2 Struktura haseł słownikowych

2.1 Ogólne założenia

Hasła w słowniku języka polskiego PWN z założenia powinny zawierać:

- wyraz definiowany
- informacje dotyczące jego odmiany (niekoniecznie)
- definicję
- zdania przykładowe (niekoniecznie)
- przykłady zastosowań wyrazu w idiomach, nazwach specjalnych etc.
- pochodzenie słowa

Przykładem takiego słowa jest np.:

```
abdominalny
1. anat. <brzuszny>
Tyfus abdominalny.
Abdominalny oddech.
```

```
2. zool. <odwłokowy>
Nogi abdominalne.
n.-łc.
```

2.2 Struktury napotkane w slowniku

Hasła w słowniku są w zdecydowanej większości poprawnie ustrukturyzowane, jednak istnieje wiele przypadków nie do końca poprawnych strukturalnie definicji, co sprawiało bardzo dużo kłopotów przy implementacji skryptu.

3 Zadana struktura słownika

Zadana przez prowadzącego struktura słownika w postaci konkretnych tagów pseudo-xml, które praktycznie wystarczały do ustrukturyzowania większości z haseł zadanych w pliku wejściowym ze słownikiem. Wyróżniono następujące tagi:

- wyraz
- def
- przyklad
- pochodzenie

W celu rozszerzenia przykładów o te zupełnie osobno definiowane, autor projektu dodał nowy tag przyklad_def

Po zastosowaniu skryptu w.wym. wpis w słowniku powinien zamienić się w następujący:

```
<wyraz>abdominalny</wyraz>
<def>(anat.) brzuszny</def>
<przyklad>Tyfus abdominalny</przyklad>
<przyklad>Abdominalny oddech</przyklad>
<pochodzenie>n.-łc.</pochodzenie>

<wyraz>abdominalny</wyraz>
<def>(zool.) odwłokowy</def>
<przyklad>Nogi abdominalne</przyklad>
<pochodzenie>n.-łc.</pochodzenie>
```

4 Struktura projektu

4.1 Technologie

Skrypt został w całości napisany w języku Python, nie tylko ze względu na wygodę programowania w dynamicznie typowanych językach, ale szczególnie z powodu niesłychanie potężnego narzędzia do obsługi wyrażeń regularnych, modułu re oraz string.

4.2 Rozwiązania

Skrypt wczytuje dane, rozdziela poszczególne wpisy a następnie stara się przewidzieć położenie konkretnych informacji we wpisie bazując m.in. na tym, czy jest on numerowany, czy definicja znajduje się w nagłówku wpisu. Skrypt następnie korzysta z przygotowywanych metod wykrywających czy konkretne elementy wpisu są definicją, przykładem, przykłado-definicją czy też opisem pochodzenia słowa

4.3 Problemy

Najwięcej problemów napotkano z powodu niejednoznacznej struktury podanego pliku ze słownikiem, przez co nie można było zastosować utrwalonych schematów file-carvingu.

5 Rezultaty

Skrypt przetwarza słownik w parę naście sekund. W celu wydobycia informacji o jego (przynajmniej pozornej) poprawności wprowadzono prosty system statystyk, które m.in. badają ile słów nie zostało przetworzonych w wyniku niedokładnej implementacji.

Dotychczas otrzymywaliśmy wyniki na poziomie ok 80% przetworzonych haseł słownikowych, zw większości z bardzo pozytywnym skutkiem. Poniżej zamieszczony jest przykładowy plik wynikowy.

6 Linki

System kontroli wersji Git http://git-scm.com/

Python http://www.python.org/