Převod dat slovníku Wordnet do DEBVisDic XML formátu

Jan Tomášek

Závěrečná zpráva

Zpráva obsahuje shrnutí problémů, se kterými se autor potkal během konverze Slovenského a Litevského wordnetu do DEBVisDic formátu.

Nastudování formátů wordnetů, určení mapování

Nejprve bylo nutné nastudovat formát Slovenského resp. Litevského wordnetu, Princeton wordnetu (ten tvoří anglickou část vstupních wordnetů) a DEBVisDic formát wordnetu. Následně bylo určeno, jak budou části záznamu ve zdrojovém wordnetu namapovány do elementů/atributů výsledného formátu.

Ve zdrojových souborech byly objeveny odkazy na synsety, které v daném souboru neexistují - takové se neobjeví ve výsledném formátu, ale jsou logovány do samostatného souboru. Po domluvě se cvičícím jsou zbylé odkazy filtrovány a některé z nich nejsou ukládány z důvodu zbytečné redundance.

Parsování vstupu, XML Binding

Vstupní data jsou parsována za pomocí regulárních výrazů. Původně byly využity 4 regulární výrazy. Při testování se zjistilo, že v jistých případech jsou tyto výrazy značně neefektivní. Následovala jejich optimalizace. Nejdůležitejší částí optimalizace je redukce greedy kvantifikátorů. Za pomoci dodatečných regulárních výrazů jsou ze záznamu vytaženy informace o počtu znaků v určité jeho sekci, toto číslo je poté použito místo greedy kvantifikátorů v dalších regulárních výrazech.

Slovenský a Litevský wordnet mají téměř stejný formát, aplikace je důsledkem toho navržena tak, že nejprve je podle formátu určen první regulární výraz (vzor pro jeden záznam), ostatní regulární výrazy a následné zpracování je přitom společné pro oba wordnety. Jelikož Litevský wordnet v sobě nese i data ze Slovenského wordnetu, aplikace umožní i konverzi z Litevského wordnetu do slovenského DEBVisDic wordnetu.

Původní myšlenka byla taková, že záznamy budou řádek po řádku parsovány a rovnou zapisovány. Kvůli odkazům na neexistující synsety je však nutné mít před samotným zápisem přehled o všech synsetech, které se v souboru vyskytují. Soubor je proto prečten dvakrát, přičemž při prvním čtení je zjištěn formát souboru a zkonstruována množina obsahující ID všech synsetů. Při druhém čtení jsou již parsována všechna potřebná data, která jsou zároveň filtrována podle množiny s ID získané prvním čtením.

Pro parsovaná data byly vytvořeny třídy, které kopírují strukturu DEBVisDic XML. Konverze do XML poté probíhá pomocí JAXB architektury.

Grafické uživatelské rozhraní

V ramci GUI se autor postaral o progress bar a chod Swing workeru.