Ročníkový projekt LS 2007/08, ZS 2008/09

Zadání ročníkového projektu

Téma: Určení smysluplnosti české věty

Anotace: Cílem rp je implementace automatické procedury, která v reálném čase rozezná, je-li shluk textových řetězců českou větou. Automatická procedura bude využívat výstupy externích automatických modulů, které se týkají tvarosloví (slovní druhy a jejich kategorie - rod, číslo, pád, osoba aj.). Vyřešení tohoto úkolu je využitelné ve fulltextovém vyhledávání, kde je mj. důležité rozeznat shluk klíčových slov určených pro zmatení vyhledávače od věty určené uživateli

Student: Vladimír Rovenský, e-mail: v.rovensky@lit.cz

Vedoucí: Barbora Vidová Hladká, ÚFAL MFF UK, *e-mail*: hladka@ufal.mff.cuni.cz

---

ŘEŠENÍ PROJEKTU

Časový harmonogram

LS 2007/2008

1. nastudovat formát CSTS, prostudovat ukázková data z Českého akademického korpusu 2.0 (viz přiložené CD-ROM)
2. seznámit se s nástrojem tool_chain,
3. zpracovat data z CD-ROM nástrojem tool_chain (tokenizace,morfologická analýza, tagování)
4. navrhnout proceduru pro testování smysluplnosti české věty (kterou může být i otázka) na základě tvaroslovných informací poskytnutých
    a. morfologickou analýzou,
    b. tagováním
5. sepsat podrobnou specifikaci,
6. implementovat pilotní verzi, tj. navrženou proceduru, která pracuje s tagováním,
7. testovat proceduru na testovací množině smysluplných i nesmysluplných českých vět,

ZS 2008/2009

8. implementovat navrženou převodní proceduru, která pracuje s morfologickou analýzou,
9. testovat proceduru na testovací množině smysluplných i nesmysluplných českých vět,
10. porovnat úspěšnosti obou procedur,
11. sepsat závěrečnou dokumentaci
    a. instalační příručka - popis instalace a spuštění programu
    b. uživatelská příručka - popis ovládání programu
    c. programátorská dokumentace - postup překladu programu, popis implementace (co, v čem, jak), popis netriviálních algoritmů

Studijní materiály

Tokenizace, morfologická analýza, tagování -
http://ufal.mff.cuni.cz/morce/cac/?chapter=3#nastroje-zprac

Formát CSTS

- stručný popisy - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#data-format
- podrobná dokumentace - http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/ch03.html#a-data-formats-csts

Ukázka souborů v reprezentaci CSTS – viz přiložené CD-ROM; soubory obsahují ručně doplněné morfologické značky

Nástroj tool_chain

- stručný popis - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#nastroje-zprac
- dále viz přiložené CD-ROM

Obsah podrobné specifikace

1. Detailní popis problematiky, podle něhož by jiný programátor napsal "tentýž" program.
2. Návrh struktury programu (moduly, knihovny, vzájemná provázanost).
3. OS, jazyk, vývojové prostředí.

Obsah pilotní verze
Testovací procedura spolupracující s tagováním.

Obsah závěrečné dokumentace

1. Instalační příručka = popis instalace a spuštění programu.
2. Uživatelská příručka = popis ovládání programu.
3. Programátorská dokumentace = postup překladu programu, popis implementace (co, v čem, jak), popis netriviálních algoritmů.

ORGANIZAČNÍ ZÁLEŽITOSTI

Termíny
30. června 2008 – odevzdání pilotní verze

Projektová wiki stránka – https://wiki.ufal.ms.mff.cuni.cz/user:hladka:vladimir-rovensky