Informationsvisualisierung

Frank Hasenbalg, 571087

Visualisierung von Wikipediatexten

Installation

`sudo apt-get install python3-setuptools python3-dev build-essential`

`sudo easy_install3 pip`

`sudo pip3 install -U nltk`

`python3` >>> `import nltk` >>> `nltk.download()` >>> `l` >>> `all`

`sudo pip3 install flask`

`sudo pip3 install Flask-WTF`

`sudo pip3 install HTMLparser`

getestet mit Ubuntu 16.04 LTS

Benutzung

`cd ...interface`

`python3 main.py`

Texttheorie

Textfunktion

- Informationsfunktion(zb. kein appell)[104pp]

Struktur

Deskriptive Themenentfaltung:

- Teil-Ganzes- oder Enthaltensein-Struktur

- Beschreibung wesentlicher Merkmale (auch Quantitative Merkmale)
- Durchgehende Wiederaufnahmestruktur [Brinker, Linguistische Textanalyse, 63pp, 3.Auflage, 1992]

- Wiederaufnahme kann expliziet(Elefant) und implizit(aus dem Kontext) sein(zb. er)[27pp]

- Text ist zusammenhaengend ueber Wiederaufnahmestruktur

- Thema ist der Gegendstand eines Textes (Referenztraeger) und was darueber ausgesagt wird[54pp]

- Wiederaufnahmestruktur spiegelt dominierende Referenztraeger wieder

Aeussere Stuktur

Texte

- besteht aus Teiltexten (Ueberschriften-Absaetzgefuege)
- Teiltext, Kapitel, Paragraphe, Abschnitte und Absaetze
- Teiltext -> Subthema
- linear aufgebaut(im Gegensatz zu hypertext)
- Thema kann Baumstruktur haben, der Text nicht

Absaetze

- Strukturieren den Text Sinneinheitsweise
- Strukturieren den Text optisch

Saetze

- bringen Referenztraeger in Zusammenhang.

- Geschlossene Einheit

- Satz ist nicht unbedingt das, was zwischen 2 Punkten steht.(Nebensaetze)

Woerter

- werden durch leer- oder Satzzeichen voneinander getrennt
- Worstamm Affixe: Prefix Suffix Circumfix Duplifix Infix Interfix Transfix Simulfix Suprafix Disfix
- Eigenstaendige bedeutung

- Besteht aus Silben

- Ich suche dominante Referenztraeger(Hochfrequenz), mit expliziter Wiederaufnahme.

Vorgehensweise

- text.html runterladen
- Javascript raus
- Titel finden und merken
- Tokens fuer Absatzenden setzen
- alle Tags raus
- alle Fussnoten raus

- alle HTML-Kommentare raus

- Tokenizer "Natural Language Toolkit" Author: Steven Bird

- Position von Absaetzen und Satzenden .(!?)

- Satzzeichen und Tokens raus aus Tokenizerliste
- Lemmatizer/Stemmer
- Stopwords raus

Affix	Example	Schema	Description
Prefix	un-do	prefix-stem	Appears befo the stem
Suffixoid[1]/semi-suffix[2]	cat-like	stem-suffixoid	Appears after the stem, but i only partially bound to it
Infix	Minne⟨fecking'⟩ sota	st⟨infix⟩em	Appears withi a stem — common in Borneo-Philippines languages
Circumfix	en⟩light⟨en	circumfix⟩ stem⟨circumfix	One portion appears befor the stem, the other after
Interfix	speed-o-meter	stema-interfix-stemb	Links two ster together in a compound
			Incorporates a

Duplifix	money~shmoney	stem~duplifix	reduplicated portion of a stem (may occur before, after, or within the stem)
Transfix	Maltese: k⟨i⟩t⟨e⟩b "he wrote" (compare root ktb "write")	s⟨transfix⟩te⟨transfix⟩m	A discontinuous affix that interleaves within a discontinuous stem
Simulfix	mouse → mice	stem\simulfix	Changes a segment of a stem
Suprafix	produce (noun)produce (verb)	stem\suprafix	Changes a suprasegmental feature of a stem
Disfix	Alabama: tipli "break up" (compare root tipasli "break")	st⟩disfix⟨m	The elision of a portion of a stem

- Die populaersten Worter (Referenztraegen) werden mit idf ausgegeben(Histogram)
- Die Saetze(nltk.sent_tokenize) werden auf gemeinsames Vorkommen von Referenztraegen untersucht(Force Diagram)
- Dispersion Plot(Referaenztraegerverteilung) -> Zeight Positionen der Woerter auf einer Art Landkarte
- Daraus folgt minimap (Verteilung der Dominanten Referenztraeger)