Progetto di Linguistica Computazionale

A.A. 2012/2013

Linee guida

Obiettivo: realizzazione di due programmi scritti in Python che utilizzino i moduli presenti in Natural Language Toolkit per leggere due file di testo in inglese, annotarli linguisticamente, confrontarli sulla base degli indici statistici richiesti ed estrarne le informazioni richieste.

Fasi realizzative:

Create due corpora in inglese contenenti i discorsi di Obama e di Romney, di almeno 5000 token ciascuno. I corpora devono essere creati selezionando i discorsi di Romney da http://mittromneycentral.com/speeches/ e di Obama da http://www.whitehouse.gov/briefing-room/speeches-and-remarks?page=5 e salvandoli in due file di testo utf-8.

Sviluppate due programmi che prendano in input i due file da riga di comando, che li analizzino linguisticamente fino al Part-of-Speech tagging e che eseguano le operazioni richieste.

Confronti i due testi sulla base delle seguenti informazioni statistiche:

- il numero di token;
- la lunghezza media delle frasi in termini di token;
- la grandezza del vocabolario del testo;
- la ricchezza lessicale calcolata attraverso la *Type Token Ratio* (**TTR**) sui primi 1000 token di ogni corpus;
- il rapporto tra Sostantivi e Verbi (indice che caratterizza variazioni di registro linguistico);
- la *densità lessicale*, calcolata come il rapporto tra il numero totale di occorrenze nel testo di Sostantivi, Verbi, Avverbi, Aggettivi e il numero totale di parole nel testo (ad esclusione dei segni di punteggiatura marcati con POS "," "."):

(|Sostantivi|+|Verbi|+|Avverbi|+|Aggettivi|)/(TOT-(|.|+|,|)).

Per ognuno dei due corpora il programma deve estrarre le seguenti informazioni:

- i primi venti token in ordine di frequenza decrescente, con relativa frequenza;
- le prime 10 PoS (Part-of-Speech) in ordine di frequenza decrescente, con relativa frequenza;
- i primi 10 bigrammi di token (dove ogni token deve avere una frequenza maggiore di 2) composti solo di Aggettivi e Sostantivi ordinati rispetto:
  - alla frequenza decrescente, con relativa frequenza;
  - alla forza associativa (calcolata in termini di Local Mutual Information), con relativa forza associativa;
- la frase con probabilità più alta. La frase deve essere lunga almeno 8 token e ogni token deve avere una frequenza maggiore di 2. La probabilità deve essere calcolata attraverso un modello di Markov di ordine 1 che sfrutta statistiche estratte dal corpus che contiene la frase.
- dopo aver individuato e classificato le Entità Nominate (NE) presenti nel testo, estraete:
  - la lista dei nomi propri di persona (tipi), ordinati per frequenza;
  - la lista dei nomi propri di luogo (tipi), ordinati per frequenza.

Risultati del progetto: perché il progetto sia giudicato idoneo, devono essere consegnati:

a. i due file di testo contenenti i corpora;
b. i programmi scritti in Python;
c. i file di testo contenenti l'output dei programmi.

Date di consegna del progetto: il progetto deve essere consegnato per posta elettronica a felice.dellorletta@ilc.cnr.it e alessandro.lenci@ling.unipi.it almeno una settimana prima dello scritto di ogni appello per poter essere considerato valido per l'appello.

NB: il progetto **DEVE** essere svolto individualmente.