Progetto di Linguistica Computazionale

A.A. 2019/2020

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:

Scegliete e scaricate in formato di testo semplice utf-8 due libri a scelta tra quelli che trovate nel sito del Progetto Gutenberg (http://www.gutenberg.org/ebooks/).

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

Programma 1 - Confrontate i due testi sulla base delle seguenti informazioni statistiche:

- il numero totale di frasi e di token;
- la lunghezza media delle frasi in termini di token e la lunghezza media delle parole in termini di caratteri;
- la grandezza del vocabolario e la distribuzione degli hapax all'aumentare del corpus per porzioni incrementali di 1000 token (1000 token, 2000 token, 3000 token, etc.);
- il rapporto tra Sostantivi e Verbi;
- le 10 PoS (Part-of-Speech) più frequenti;
- estraete ed ordinate i 10 bigrammi di PoS:
  - con *probabilità condizionata* massima, indicando anche la relativa probabilità;
  - con *forza associativa* massima (calcolata in termini di Local Mutual Information), indicando anche la relativa forza associativa.

Programma 2 - Per ognuno dei due corpora estraete i dieci nomi propri di persona più frequenti e per ognuno dei nomi la lista delle frasi che lo contengono. Restituite come informazioni per ogni nome proprio:

- la frase più lunga e la frase più breve che lo contengono;
- analizzando solo le frasi dove compare il nome proprio, estraete ed ordinate in ordine di frequenza decrescente, indicando anche la relativa frequenza:
  - i 10 Luoghi più frequenti;
  - le 10 Persone più frequenti;
  - i 10 Sostantivi più frequenti;
  - i 10 Verbi più frequenti;
  - le Date, i Mesi e i Giorni della settimana estratti attraverso l'utilizzo di espressioni regolari;
  - la frase lunga minimo 8 token e massimo 12 con probabilità più alta. La probabilità deve essere calcolata attraverso un modello di Markov di ordine 0. Il modello deve usare le distribuzioni di frequenza estratte dall'intero corpus del libro.

Risultati del progetto:

perché il progetto sia giudicato idoneo, devono essere consegnati:

a. i due file di testo contenenti i corpora;
b. i programmi ben commentati 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@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.