Progetto di Linguistica Computazionale

A.A. 2021/2022

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 di almeno 5000 parole, ognuno rappresentativo di un diverso genere testuale (per esempio testi giornalistici, prosa letteraria, blog, social media, articoli scientifici, etc). I corpora devono essere salvati come file di testo semplice con codifca utf-8. 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 di frasi e di token;
- la lunghezza media delle frasi in termini di token e dei token (escludendo la punteggiatura) in termini di caratteri;
- il numero di hapax sui primi 1000 token;
- la grandezza del vocabolario e la ricchezza lessicale calcolata attraverso la Type Token Ratio (TTR), in entrambi i casi calcolati all'aumentare del corpus per porzioni incrementali di 500 token;
- distribuzione in termini di percentuale dell'insieme delle parole piene (Aggettivi, Sostantivi, Verbi, Avverbi) e delle parole funzionali (Articoli, Preposizioni, Congiunzioni, Pronomi).

Programma 2 - Per ognuno dei due corpora estraete le seguenti informazioni:

- estraete ed ordinate in ordine di frequenza decrescente, indicando anche la relativa frequenza:
  - le 10 PoS (Part-of-Speech) più frequenti;
  - i 10 bigrammi di PoS più frequenti;
  - i 10 trigrammi di PoS più frequenti;
  - i 20 Aggettivi e i 20 Avverbi più frequenti;

- estraete ed ordinate i 20 bigrammi composti da Aggettivo e Sostantivo e dove ogni token ha una frequenza maggiore di 3:
  - con frequenza massima, indicando anche la relativa frequenza;
  - con probabilità condizionata massima, indicando anche la relativa probabilità;
  - con forza associativa (calcolata in termini di Local Mutual Information) massima, indicando anche la relativa forza associativa;

- estraete le frasi con almeno 6 token e più corta di 25 token, dove ogni singolo token occorre almeno due volte nel corpus di riferimento:
  - con la media della distribuzione di frequenza dei token più alta, in un caso, e più bassa nell'altro, riportando anche la distribuzione media di frequenza. La distribuzione media di frequenza deve essere calcolata tenendo in considerazione la frequenza di tutti i token presenti nella frase (calcolando la frequenza nel corpus dal quale la frase è stata estratta) e dividendo la somma delle frequenze per il numero di token della frase stessa;

    - con probabilità più alta, dove la probabilità deve essere calcolata attraverso un modello di Markov di ordine 2. Il modello deve usare le statistiche estratte dal corpus che contiene le frasi;

- dopo aver individuato e classificato le Entità Nominate (NE) presenti nel testo, estraete:
    - i 15 nomi propri di persona più frequenti (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 ben commentati scritti in Python;
c. i file di testo contenenti l'output ben formattato dei programmi.

Date di consegna del progetto:
il progetto deve essere consegnato per posta elettronica a felice.dellorletta@ilc.cnr.it, alessio.miaschi@phd.unipi.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.