Progetto di Linguistica Computazionale

A.A. 2020/2021

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 Joe Biden e di Donald Trump, di almeno 5000 token ciascuno. I corpora devono essere creati selezionando i discorsi di Biden da https://www.rev.com/blog/transcript-tag/joe-biden-transcripts e di Trump da https://www.rev.com/blog/transcript-tag/donald-trump-speech-transcripts e salvandoli in due file di testo semplice 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 delle parole in termini di caratteri;
- la grandezza del vocabolario e la ricchezza lessicale calcolata attraverso la Type Token Ratio (TTR), in entrambi i casi calcolati nei primi 5000 token;
- la distribuzione delle classi di frequenza |V1|, |V5| e |V10| all'aumentare del corpus per porzioni incrementali di 500 token (500 token, 1000 token, 1500 token, etc.);
- media di Sostantivi e Verbi per frase;
- 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-(|.|+|,|)).

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 20 sostantivi e i 20 verbi più frequenti;
  - i 20 bigrammi composti da un Sostantivo seguito da un Verbo più frequenti;
  - i 20 bigrammi composti da un Aggettivo seguito da un Sostantivo più frequenti;

- estraete ed ordinate i 20 bigrammi di token (dove ogni token deve avere una frequenza maggiore di 3):
  - con probabilità congiunta massima, indicando anche la relativa probabilità;
  - 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;

- per ogni lunghezza di frase da 8 a 15 token, estraete la frase con probabilità più alta, dove la probabilità deve essere calcolata attraverso un modello di Markov di ordine 1 usando lo Add-one Smoothing. 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;
  - i 15 nomi propri di luogo 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.