Progetto di Linguistica Computazionale

A.A. 2017/2018

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 token ciascuno, contenenti testi estratti rispettivamente da blog di racconti di viaggio scritti da uomini e donne. Esempi di testi che possono essere utilizzati per costruire il corpus di blog femminili sono scaricabili a questo indirizzo https://www.flipkey.com/blog/2014/11/03/top-25-solo-female-travel-bloggers-to-follow-in-2015/, mentre esempi di testi maschili possono essere trovati qui http://www.huffingtonpost.co.uk/janet-newenham/worlds-top-male-travel-bloggers_b_9048436.html o qui https://www.nomadichustle.com/best-travel-blogs-adventurous-men/. I corpora devono essere salvati in due file di testo semplice in codifica 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 la lunghezza media delle parole in termini di caratteri;
- la grandezza del vocabolario e il numero di hapax all'aumentare del corpus per porzioni incrementali di 1000 token (1000 token, 2000 token, 3000 token, etc.);
- la ricchezza lessicale calcolata attraverso la Type Token Ratio (TTR) sui primi 5000 token;
- la distribuzione (in termini percentuali) di Sostantivi, Aggettivi, Verbi e Pronomi;
- il numero medio di Sostantivi, Aggettivi, Verbi e Pronomi per frase.

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

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

- estraete ed ordinate in ordine decrescente i 10 bigrammi di PoS (Part-of-Speech):
  - con *probabilità congiunta* massima, indicando anche la relativa probabilità;
  - con *probabilità* condizionata massima, indicando anche la relativa probabilità;

- create un'unica lista con i 10 Sostantivi più frequenti contenuti nei blog maschili e i 10 Sostantivi più frequenti dei blog femminili e per ognuno di questi Sostantivi ordinate gli Aggettivi che li precedono rispetto alla forza associativa (calcolata in termini di Local Mutual Information);

- dopo aver individuato e classificato le Entità Nominate (NE) presenti nel testo, estraete:
  - i 20 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 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.