# Progetto di Linguistica Computazionale

A.A. 2016/2017

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 Hillary Clinton e di Donald Trump, di almeno 5000 token ciascuno. I corpora devono essere creati selezionando i discorsi di Clinton da https://www.hillaryclinton.com/speeches/ e di Trump da https://www.donaldjtrump.com/media/category/speeches 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 token;
- la lunghezza media delle frasi in termini di token;
- la grandezza del vocabolario all'aumento 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) all'aumento del corpus per porzioni incrementali di 1000 token (1000 token, 2000 token, 3000 token, etc.);
- 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-( |.|+|,| ) ).*

*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 token più frequenti escludendo la punteggiatura;
  - i 20 bigrammi di token più frequenti che non contengono punteggiatura, articoli e congiunzioni;
  - i 20 trigrammi di token più frequenti che non contengono punteggiatura, articoli e congiunzioni;
- estraete ed ordinate i 20 bigrammi composti da Aggettivo e Sostantivo (dove ogni token deve avere una frequenza maggiore di 2):
  - 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;
- le due frasi con probabilità più alta. Dove la probabilità della prima frase deve essere calcolata attraverso un modello di Markov di ordine 0 mentre la seconda con un modello di Markov di ordine 1, i due modelli devono usare le statistiche estratte dal corpus che contiene

le frasi; Le frasi devono essere lunghe almeno 10 token e ogni token deve avere una frequenza maggiore di 2;

- dopo aver individuato e classificato le Entità Nominate (NE) presenti nel testo, estraete:
  - i 20 nomi propri di persona più frequenti (tipi), ordinati per frequenza;
  - 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@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.