Linguistica computazionale

(A/A 2024-25)

Seconda prova in itinere del 19 dicembre 2024

Cognome e nome:

Matricola:

Istruzioni per la consegna:
- ridenominare il file del compito con *CognomeNomeMatricola* (es. RossiMario001293.doc)
- inviare il file a: alessandro.lenci@unipi.it

---

1. Illustrate cosa è e come è strutturato FrameNet (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

2. Spiegate come funziona un Naive Bayes Classifier (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

3. Spiegate cosa è un Language Model (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

4. Un modello di NER viene valutato su un corpus che contiene 400 nomi propri di cui 150 di classe PER e 250 di classe LOC. Il modello classifica 220 nomi propri come PER di cui 110 corretti e classifica 180 nomi come LOC di cui 160 corretti. i.) Costruite la matrice di confusione per la classe PER, ii) calcolate l'accuratezza globale del NER e iii) calcolate Precision, Recall e F-measure per la classe LOC.
(PUNTI: 4).

C = 400
GS PER = 150
GS LOC = 250

	GS PER	GS NON-PER
NER PER	110	110
NER NON-PER	40	140

PER 220
TP = 110

LOC 180
TP = 160
FN = 250-160 = 90
FP = 180-160 = 20

Accuracy = tot_output_corretti/corpus = (110+160)/400 = 0.675

Precision_LOC = TP/(TP+FP) = 160/180 = 0.89
Recall_LOC = TP/(TP+FN) = 160/250 = 0.64

F1_LOC = (2*P*R)/(P+R) = 0.75

5. La parola *stampa* può essere un nome (*stampa*_N) o un verbo (*stampa*_V). In un corpus di addestramento annotato con le POS, *stampa* ricorre 420 volte, di cui 300 taggato come verbo e le restanti come nome. Nello stesso corpus, *stampa*_N è preceduto dall'articolo *una* 45 volte, mentre

*stampa*_V 3 volte. Calcolate il tag più probabile da assegnare a *stampa* nella frase *Maria ha comprato una stampa*. (PUNTI: 5).

F(stampa) = 420
F(stampa_V) = 300
F(stampa_N) = 120

P(stampa_V) = 300/420 = 0.71
P(stampa,N) = 120/420 = 0.29
F(stampa_N, una) = 45

F(stampa_V, una) = 3

P(una|stampa_N) = F(stampa_N, una)/ F(stampa_N) = 45/120 = 0.375
P(una|stampa_V) = F(stampa_V, una)/ F(stampa_V) = 3/300 = 0.01

P(I|O) = P(I) * P(O|I)

P(stampa_N|una) = P(stampa_N) * P(una|stampa_N) = 0.29 * 0.375 = 0.11

P(stampa_V|una) = P(stampa_V) * P(una|stampa_V) = 0.71 * 0.01 = 0.0071

P(stampa_N|una) > P(stampa_V|una) quindi nella frase l'etichetta più probabile per "stampa" è N.

6. In un corpus C (|C| = 1000) ci sono 6 documenti. La parola *di* ricorre 4 volte nei documenti D1 e D2, 7 volte nei documenti D3 e D5, 10 volte nel documento D6. La parola *palla* ricorre invece 15 volte nel documento D2 e 2 volte nel documento D1. Calcolate l'entropia delle due parole nel corpus. (PUNTI: 5)

C = 1000
F(di,D1) = F(di,D2) = 4
F(di, D3) = F(di, D5) = 7
F(di,D4) = 0
F(di, D6) = 10
F(di) = 32

F(palla, D1) = 2
F(palla,D2) = 15
F(palla) = 17

H(w) = -(sommatoria(Pw*log2(Pw)))
H(di) = -(4/32*log2(4/32)+ 4/32*log2(4/32) + 7/32*log2(7/32) + 0 + 7/32*log2(7/32)+ 10/32*log2(10/32)) = 2.23

H(palla) = -(2/17*log2(2/17) + 15/17*log2(15/17)) = 0.52

7. Un test corpus contiene 100 frasi di 20 parole ciascuna. Due language model A e B assegnano le seguenti probabilità al test corpus: P_A= 0.009, P_B = 0.060. Misurate la bontà dei modelli utilizzando la cross-entropy e indicate qual è il modello migliore (PUNTI: 4):

P(A) =0.009
P(B) = 0.06
N corpus =100* 20 = 2000

Cross entropy = -1/N *log2(P(modello))

CE(A)= -1/2000 * log2(0.009) = 0.003

CE(B)= -1/2000 *log2(0.06) = 0.002

CE(B) < CE(A) quindi B è il modello migliore.