Linguistica computazionale

(A/A 2023-24)

Prova in itinere del 20 dicembre 2023

Testo A
(studenti la cui ultima cifra del numero di matricola è compresa tra 0 e 4)

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. Spiegate la nozione di n-fold crossvalidation (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

2. Illustrate il task di Named Entity Recognition e lo schema di annotazione usato (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

3. Illustrate come si valuta l'affidabilità di un'annotazione linguistica manuale (PUNTI: 4) (**la risposta deve essere lunga min. 5, max 10 righe**)

4. In un test corpus ci sono 820 nomi concreti (classe C) e 740 nomi astratti (classe A). Un sistema di annotazione semantica assegna la classe C a 900 nomi, di cui 400 sono corretti, mentre assegna la classe A a 660 nomi, di cui 350 sono corretti. Costruite la matrice di confusione e calcolate precision, recall e F-measure del sistema per entrambe le classi (PUNTI: 4):

TP_C = 400
TP_A = 350

FP_C = 900 – 400 = 500
FP_A = 660 – 350 = 310

FN _C = 820 – 400 = 420
FN_A = 740 – 350 = 390

P_C = 400/(400+500) = 0.44
P_A = 350/(350+310) = 0.53

R_C = 400/(400+420) = 0.49
R_A = 350/(350+390) = 0.47

F1_C = 2*((0.44*0.49)/(0.44+0.49)) = 0.46

F1_A = 2*((0.53*0.47)/(0.53+0.47)) = 0.50

5. In un corpus di 10.000 parole annotato con le Named Entity, i.) ci sono 1.500 nomi LOC e 1.800 nomi ORG; ii.) 1.700 ORG sono preceduti da un articolo, mentre solo 300 nomi LOC sono preceduti da un articolo. Calcolate qual è la classe più probabile da assegnare al nome *Roma* nella frase *Ho visto la Roma giocare*. (PUNTI: 5).

P(LOC|la Roma) = P(LOC) * P(la Roma|LOC) = 1500/10000 * 300/1500 = 0.15 * 0.2 = 0.03

P(ORG|la Roma) = P(ORG) * P(la Roma|ORG) = 1800/10000 * 1700/1800 = 0.18 * 0.94 = 0.17

La classe più probabile è ORG

6. Dato il contesto "C = *La macchina insegue la ...*", un language model genera la parola seguente con questa distribuzione di probabilità *m*: P(*talpa*|C) = 0.05; P(*dolce*|C) = 0.01; P(*carta*|C) = 0.03; P(*moto*|C) = 0.70; P(*gatta*|C) = 0.20; P(*felpa*|C) = 0.01. Calcolate l'entropia del language model e, assumendo *m* come la distribuzione reale, calcolate la sua cross-entropy rispetto a un modello che assegna a tutte le parole la stessa probabilità (PUNTI: 5).

H(LM) = - (P(*talpa*|C)* log2 P(*talpa*|C) + P(*dolce*|C) * log2 P(*dolce*|C) + P(*carta*|C) * log2 P(*carta*|C) + P(*moto*|C) * log2 P(*moto*|C) + P(*gatta*|C) * log2 P(*gatta*|C) + P(*felpa*|C) * log2 P(*felpa*|C)) = -(0.05*log2(0.05) + 0.01*log2(0.01) + 0.03*log2(0.03) + 0.70*log2(0.70) + 0.20*log2(0.20) + 0.01*log2(0.01)) = 1.32

Nel modello equiprobabile, ogni parola riceve una probabilità 1/6 = 0.17

Cross_entropy (LM, equiProb) = - (0.05*log2(0.17) + 0.01*log2(0.17) + 0.03*log2(0.17) + 0.70*log2(0.17) + 0.20*log2(0.17) + 0.01*log2(0.17)) = 2.56

7. Due language model A e B assegnano a un corpus lungo 1.000 token le seguenti probabilità: P_A= 0.045, P_B = 0.0023. Misurate la bontà dei modelli e indicate qual è il modello migliore (PUNTI: 4):

Si misura la bontà del modello con la perplexity:

Perplexity(A) = 0.045^-(1/1000) = 1.003

Perplexity(B) = 0.0023^-(1/1000) = 1.006

Il modello migliore è A, perché ha la perplexity minore