Relazione TLN sulle esecitazioni Ngrams e Word Sense Disambiguation

Patric Reineri, Alessandro Scicolone

July 6, 2025

1 Esercizio 1: Modeling social media and literary language

1.1 Consegna dell'esercitazione

L'esercizio richiede di:

1. **Raccolta di tweet**: Trovare due diversi set di tweet (ad esempio, due politici? Lingua italiana e/o inglese?). Per ciascun set:
   - Acquisire due modelli linguistici (uno per bi-grammi e uno per tri-grammi)
   - Utilizzare questi modelli linguistici per generare tweet e confrontare i diversi stili
2. **Modello da testo Moby-Dick**:
   - Addestrare un modello (bi-grammi e tri-grammi) dal romanzo epico Moby-Dick (testo completo nel bundle per questa lezione)
   - Utilizzarlo per generare brevi testi
   - Libera scelta della lunghezza dei testi generati

1.2 Raccolta di Tweet

Abbiamo raccolto un insieme di tweet di due personaggi politici italiani: Matteo Salvini e Matteo Renzi risalenti al periodo pre-elezioni italiane del 2022.
Abbiamo scelto due personaggi che hanno idee e pensieri totalmente opposti, in modo tale da avere due stili di tweet ben evidenti.
Il modello Ngrams è stato costruito a partire dalla classe "Context.py" contenente due variabili:

- **context_tokens**: si tratta di un array di token che rappresentano il contesto. Nel Caso di N grammi, avremo N-1 token di contesto considerato.
- **context_counts**: rappresenta un dizionario chiave-valore, dove la chiave è un token e il valore rappresenta il numero di volte in cui il token è preceduto dal contesto.

Grazie a questa classe è possibile ottenere due informazioni fondamentali per il calcolo della probabilità di un token dato un contesto: il numero di volte in cui la sequeza <contesto token> è stata osservata (andando ad accedere al campo value del dizionario) e il numero di volte in cui il contesto è stato osservato andando a sommare i valori del dizionario.
Nel codice è inoltre presente uno smoothing di laplace in modo tale da evitare di avere probabilità nulle causate dalle sequenze che non vengono mai osservate.
E' sta inserita anche una nozione di temperatura, che permette di regolare la probabilità di selezione dei token.

1.2.1 Considerazioni generali e confronto dei due modelli

In entrambi i casi le probabilità ottenute dipendono fortemente dal vocabolario presente nei tweet raccolti. I due politci hanno un vocabolario molto diverso, e questo si riflette nei modelli Ngrams.
Inoltre la dimensione del corpus di tweet varia notevolmente tra i due, infatti per Matteo Salvini sono state raccolte 848 frasi, mentre per Matteo Renzi sono state raccolte 141 frasi.
Questo porta a una differenza significativa nel numero di oggetti Context creati, che sono 16715 per Salvini e 3454 per Renzi nel caso di tri-grams.
Pertanto per il secondo all'aumentare degli N-grammi il numero di contesti osservati diminuisce notevolmente, portando a una maggiore probabilità di sequenze non osservate.

1.3 Modello creato a partire dal testo letterario Moby Dick

A differenza dei tweet, il numero di frasi generate sono molte di più (6725) pertanto abbiamo un vocabolario molto più ampio, ottenendo migliaia di contesti, consentendo una riduzione della probabilità che un certo contesto non venga trovato.
All'aumentare del numero di token che vengono, i token ripetivi, questo è dovuto al processo deterministico di selezione della probabilità massima.
Questo problema è possibile risolverlo specificando una temperatura diversa da 0.
All'aumentare degli n-grammi il processo di conteggio delle frequenze diventa computazionalmente più costoso

2 WSD con DistillBert

L'obiettivo è utilizzare DistilBERT per disambiguare termini polisemici appartenenti a frasi estratte in maniera casuale dal corpus semcor.
Le frasi vengono date in input a DistilBERT, successivamente viene estratto l'ultimo layer hidden per recuperare il vettore contestualizzato del termine polisemico.
Per ciascun senso del termine viene calcolata la cosine similarity tra il vettore medio per un dato senso e il vettore contestualizzato ottenuto a partire dal frase rappresentante il contesto, scegliendo il senso con la massima similarità.
L'accuracy ottenuta è stata di circa il 75% calcolata su un totale di 50 frasi.