Relazione esercizio Defs

Nell'ambito della comprensione del linguaggio naturale e della semantica lessicale, l'esercitazione **"Defs"** ha visto la definizione di un algoritmo di valutazione della sovrapposizione lessicale su un insieme di definizioni associate ai seguenti 4 termini, aggregati sulle dimensioni di concretezza/specificità:

- Brick (concreto / specifico)
- Person (concreto / generico)
- Revenge (astratto / specifico)
- Emotion (astratto / generico)

L'obiettivo dell'esercitazione è stato quello di mostrare delle evidenze di quanto fosse difficile scrivere la definizione di un concetto e di quanto fosse difficile essere in accordo su una sola definizione. Per fare emergere questo riscontro il problema è diventato quello di esprimere il numero di definizioni con molti termini simili sul totale delle definizioni e aggregare i valori di similarità sulle 2 dimensioni di concretezza e specificità. Con questo obiettivo, in una prima fase sono state raccolte un insieme di definizioni per i 4 concetti, in seguito alcune di queste sono state scremate e rimosse dall'insieme in quanto presentavano elementi di circolarità diretta o poiché troppo riduttive. Il file "**definizioni.csv**" nella cartella "**data**" dei file di progetto contiene tutte le definizioni considerate.

La sovrapposizione lessicale, come criterio di calcolo della similarità tra le definizioni, è stata applicata a seguito di una fase di preprocessing dei dati. Il file "**DataPreprocessing.py**" riporta le funzioni usate per pulire le definizioni di tutti i simboli di punteggiatura e delle stopwords presenti. Le parole restanti sono state lemmatizzate per favorire una maggiore sovrapposizione.

Seguendo il flusso di esecuzione del file **main.py**, una volta raccolte e organizzate in un dizionario, le definizioni dei concetti sono state prima preprocessate poi confrontate a coppie in base alla sovrapposizione lessicale. La funzione "**bag_of_words_similarity**" ricava un punteggio di similarità come rapporto tra il numero di sovrapposizioni tra le liste di termini in input e la lunghezza del vettore con meno elementi, in questo modo lo score viene normalizzato sulla coppia. Da una prima analisi si nota che molte tra le similarità a coppie calcolate sono pari a 0, alcune invece hanno un valore compreso tra 0 e 0.5 e pochissime un punteggio superiore a 0.5. Si è deciso di imporre un valore di threshold per le similarità minori di 0.5 in quanto rappresentanti di definizioni troppo diverse fra loro. In questo modo vengono considerate solo le coppie simili.

Infine le similarità sono state aggregate sulle due dimensioni di concretezza e specificità calcolando per ogni termine il rapporto tra il numero di accoppiamenti di definizioni con similarità maggiore di 0.5 e il totale di tutte le coppie fra tutti i termini.

Risultati

Come ci aspettavamo, definire per un concetto astratto/specifico risulta molto complicato. In generale i concetti concreti/specifici sono i più facili da definire, in questo caso il concetto che ha rilevato la maggior parte delle sovrapposizioni, però, è Person (concreto/generico). Il motivo di tale risultato è da ritrovare analizzando la composizione delle definizioni. Molte di queste presentano parecchi termini in comune e risultano molto corte. I più frequentemente trovati sono: "**human", "being**", "**living**". Ciò comporta un aumento della similarità e, quindi, un maggior numero di coppie con punteggio oltre il valore di threshold. Di seguito riportiamo un estratto dell'output del programma.

```
Number of total couples:  1486
Percentage of the most similar couple for Brick (concrete/specific): 12.0 %
Percentage of the most similar couple for Person (concrete/generic): 15.0 %
Percentage of the most similar couple for Revenge (abstract/specific): 2.0 %
Percentage of the most similar couple for Emotion (abstract/generic): 3.0 %
```