Marco Ottina 795130
Davide Sgaramella 938632

A.A. 2019/2020

Università degli Studi di Torino – polo Scienze della Natura
Laurea Magistrale in Informatica
Corso: Tecnologie del Linguaggio Naturale
Parte terza - professor Luigi Di Caro

Esercitazione 1

Come da consegna, sono stati analizzati quattro insiemi definizioni relativi ad altrettanti concetti (Building, Molecule, Freedom, Compassion) ed è stato assegnato un valore a ciascun insieme che rappresenti il grado di similarità delle stesse considerate a 2 a 2.
In particolare, per ogni insieme di definizioni si sono valutate tutte le possibili coppie utilizzando una funzione di similarità. La valutazione di un insieme di definizioni è un numero reale compreso tra 0 ed 1. Usiamo come funzione di similarità il *coefficiente di sovrapposizione*, illustrato in figura.

$$O(A,B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

Le definizioni sono state pre-processate mediante un'apposita funzione che prevedeva l'eliminazione della punteggiatura, delle *stop words,* così da eliminare parole prive di contenuto o che ne apportino in quantità irrisorie, questo per facilitare il confronto tra due definizioni. Questo passaggio è seguito dal processo di *lemmatizzazione* delle parole rimanenti.

Risultati

Dopo aver calcolato i valori di similarità per le definizioni, sono stati filtrati i valori che non superavano la soglia fissato arbitrariamente a 0.5. Il risultato di questa operazione è poi stato normalizzato sul numero totale di coppie possibili.
Nelle seguenti tabelle sono riportati i risultati ottenuti.

Soglia esclusiva	***Astratto***	***Concreto***
Generico	2.3%	11.7 %
Specifico	1.8%	7.6 %

Soglia inclusiva	***Astratto***	***Concreto***
Generico	8.8%	29.2%
Specifico	4.1%	17.0%

Analizzando manualmente le definizioni fornite si è potuto constatare che alcune di esse, dopo l'operazione di pre-processing, sono state ridotte ad insiemi di pochi termini (meno di quattro) od anche un singolo termine.

Inoltre, si è effettuato un test utilizzando la similarità di Jaccard al post dell'overlap. Tuttavia, non si sono ottenuti risultati soddisfacenti tanto quanto quelli ottenuti con l'altra misura d similarità.