Relazione esercitazione - Defs

L'obiettivo di questa esercitazione è stato quello di creare un algoritmo che fosse in grado di partire da un insieme di definizioni assegnate a 4 concetti differenti (Emotion, Person, Revenge e Brick) e risolvere questi punti:

1) Per ogni concetto, calcolare la similarità tra le definizioni create sfruttando la sovrapposizione lessicale.
2) Aggregazione dei risultati sulle dimensioni concretezza/specificità.
3) (Aggiunta non richiesta) Ulteriore sperimentazione in cui è stata calcolata la similarità tra i concetti proposti.

Il punto 1) è stato risolto nel seguente modo:

E' stata eseguita una prima fase di **pre-processing** per ogni frase presente nel dataset chiamato "definizioni.xlsx" contenente appunto le definizioni dei vari concetti. In questa fase, per ogni frase è stata rimossa la punteggiatura, sono state rimosse eventuali stop words e stringhe vuote, sono state rimosse eventuali circolarità dirette nelle definizioni, tutte le parole sono state rese completamente in minuscolo e successivamente è stata applicata la lemmatizzazione per favorire maggiormente gli overlaps.

Per ottenere la similarità tra tutte le definizioni di un certo concetto è stata creata una funzione chiamata "**calcolo_somma_frequenze_delle_prime_n_parole_piu_frequenti**" che sostanzialmente si preoccupa di creare un dizionario ordinato che conterrà la frequenza di ogni parola presente nelle definizioni del concetto iniziale. Chiaramente ogni parola viene considerata al più una sola volta in ogni definizione. La funzione citata pocanzi una volta creato il dizionario, **seleziona le prime n parole più frequenti** e restituisce la somma di tutte le frequenze di queste parole. Dopodichè per ottenere lo score di similarità finale tra le definizioni si utilizza la formula seguente:

$$\mathbf{sim(defs\ Concetto)} = \frac{\mathbf{somma\_frequenze\_prime\_n\_parole\_più\_frequenti}}{\mathbf{num\_tot\_definizioni} * \mathbf{n}}$$

ove:

num_tot_definizioni = numero totale definizioni del concetto.

n = 5 nel nostro caso poiché sperimentalmente è risultato essere il valore più appropriato.

num_tot_definizioni * n = valore che considera il caso in cui **tutte** le definizioni contengano le n parole più frequenti.

In questo modo la similarità assumerà valori in [0,1] e assumerà valore 1 solamente se tutte le prime n parole più frequenti saranno presenti in ogni definizione del concetto.

Punto1) - Risultati di similarità ottenuti per ogni concetto:

	Astratto	**Concreto**
Generico	Emotion -> 0.256	Person -> 0.23
Specifico	Revenge -> 0.275	Brick -> 0.537

Punto 2) - Risultati di aggregazione:

Successivamente sono stati calcolati i valori di aggregazione sulle dimensioni di concretezza e specificità attraverso dei valor medi, i cui risultati sono mostrati nell'immagine sottostante:

Aggregazione concetti concreti Generico-Specifico:
(sim(Person) + sim(Brick)) /2

Aggregazione concetti astratti Generico-Specifico:
(sim(Emotion) + sim(Revenge)) /2

Aggregazione concetti generici Astratto-Concreto:
(sim(Emotion) + sim(Person)) /2

Aggregazione concetti specifici Astratto-Concreto:
(sim(Revenge) + sim(Brick)) /2

Ove:
sim(word) = sovrapposizione lessicale tra tutte le definizioni di word.

```
dizionario_similarita_di_aggregazione:
{'Emotion-Revenge': 0.265625, 'Person-Brick': 0.384375, 'Revenge-Brick': 0.40625, 'Emotion-Person': 0.24375}
```

Come mostrato dai risultati ottenuti dall'aggregazione sembra confermato il fatto che ci sia molta più sovrapposizione media tra concetti concreti e specifici rispetto a quelli astratti e generici.

Punto 3) – Similarità tra coppie di concetti:

Infine, è stato eseguito un ulteriore esperimento di calcolo di similarità (sempre sfruttando l'overlap lessicale) tra ogni coppia di concetti citata sopra in modo da ottenere sostanzialmente, per ogni coppia di concetti considerata, il valore di similarità tra i due, sempre in base alle definizioni di partenza.

Per poter ottenere il valore di similarità tra due concetti sono stati prima considerati tutti i singoli termini lemmatizzati presenti nelle definizioni di entrambi i concetti e dopodichè tramite la funzione chiamata "**calcolo_score_di_sovrapposizione_tra_i_due_concetti_correnti**" è stata calcolata la similarità andando a contare sostanzialmente il numero di termini in comune tra i due concetti. Infine, il valore ottenuto, è stato diviso per la dimensione dell'insieme di termini più piccolo tra i due concetti iniziali in modo da normalizzare il valore di similarità finale tra 0 e 1.

I risultati ottenuti sono i seguenti:

```
dizionario_similarita_coppie_di_concetti:
{'Emotion-Revenge': 0.37662337662337664, 'Person-Brick': 0.020833333333333332, 'Revenge-Brick': 0.08333333333333333, 'Emotion-Person': 0.2203898305084745}
```

Si può notare come secondo le definizioni date, la coppia di concetti Emotion-Revenge risulta avere la similarità più alta rispetto alle altre (0.376), seguita dalla coppia Emotion-Person (0.228). Questo risultato può essere considerato tutto sommato abbastanza ragionevole, in quanto Emotion probabilmente è molto simile a Revenge poiché quest'ultima può essere comunque considerata una sorta di emozione. Emotion e Person potremmo pensare che abbiano una similarità comunque maggiore, rispetto ad esempio a Person e Brick, in quanto una persona essendo un essere umano è in grado di provare emozioni e questa è sicuramente una peculiarità che comunque lo distingue dagli altri esseri viventi.