Relazione
Esercitazione 1

Defs

Introduzione

La seguente esercitazione ha visto la definizione di un algoritmo di valutazione della sovrapposizione lessicale per un insieme di definizioni (sulle dimensioni di concretezza/astrattezza e specificità/genericità) associate ai seguenti quattro termini

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

Lo scopo dell'esercitazione è provare che, dare una definizione di un termine è più complicato di quanto si possa pensare, specialmente se il termine è di natura astratta e/o generica. Una prova di questo aspetto è subito osservabile dando un veloce sguardo alle definizioni annotate, dove si può notare come siano parecchio diverse tra di loro.

Con questo obiettivo, in una prima fase, sono state raccolte un insieme di definizioni per i quattro 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 `resources` contiene le definizioni considerate.

NOTE:

- Dei quattro termini sono state annotate 32 definizioni
- La sovrapposizione lessicale, come criterio di calcolo della similarità tra le definizioni, è stata applicata a seguito di una fase di pre-processing dei dati.
- Le definizioni erano state numerate in maniera discontinua nel csv (si passava da 27 a 29, ecc..)

Struttura del codice ed implementazione

Per lo svolgimento di questa esercitazione l'intero codice è presente nel file **`main.py`** `che` implementa i seguenti metodi

- `parse_input()`

  metodo si occupa di leggere e parsificare il contenuto della risorsa contenente le definizioni (*definizioni.csv*).

  Le definizioni vengono pre-processate attraverso una pipeline di pulizia consistente nei seguenti task

  - rimozione della punteggiatura
  - lemmatizzazione
  - rimozione stop words

  Successivamente le definizioni (ormai ridotte in *bag of words*) vengono processate per ottenere una lista di tuple della forma ('parola', frequenza) e ordinate in senso decrescente in modo tale da avere le parole più utilizzate in cima.

  Infine, viene popolato un dizionario con chiave il termine e valori la lista prima citata.

  `Es.` `{'Emotion': [('feeling', 12),('human', 8),('feel', 8)…]}`

- `lemmatization(sentence)`

  metodo che si occupa della lemmatizzazione della frase passata in input

- `process_defs(defs)`

  metodo che prende in input una lista e si occupa della pipeline di pulizia, prima descritta

- `similarity(dict)`

  metodo che si occupa di calcolare la similarità dato in input il dizionario prima descritto. Per il calcolo della similarità si è analizzato l'overlap delle parole più frequenti in comune nelle definizioni. Sono state per cui individuate le prime cinque parole più frequenti nelle definizioni, dopodiché è stata calcolata la media, dividendo il numero di volte in cui comparivano per il totale delle definizioni.
  Infine, le medie parziali ottenute vengono sommate e divise per la loro cardinalità, per ottenere la media totale.

Risultati

Di seguito i risultati ottenuti

	Astratto	Concreto
Generico	*Emotion* 24%	*Person* 28%
Specifico	*Revenge* 26%	*Brick* 53%

Valore medio per Astratto	Valore medio per Concreto
25%	40%

Conclusioni

I risultati confermano quanto si poteva immaginare: dare definizioni per i termini è più complicato di quanto si possa pensare; infatti, aumentando la genericità e l'astrattezza di un termine si è ottenuto un agreement inter-annotator (similarità) minore.

Analizzando manualmente le definizioni fornite si è potuto constatare che alcune di esse, dopo l'operazione di pre-processing, siano state ridotte ad insiemi di pochi termini, oltre a segnalare che alcune definizioni risultavano già assenti all'intero del file csv (erano presenti dei blank spaces). Questo sicuramente non ha agevolato le operazioni di calcolo

Inoltre, in questa esercitazione, misurare la similarità andando ad analizzare l'overlap delle parole in comune delle definizioni, non sarebbe stato possibile proprio perché non sono state trovate parole che fossero presenti in tutte le definizioni. Questo è dovuto al fatto che il file di partenza risulta annotato da un numero non sufficiente di persone.

In generale dovrebbe essere più semplice dare definizioni per un termine più concreto e specifico, che in questo caso corrisponde a *brick*, infatti, è stata individuata una similarità del 53%. La situazione infatti peggiora per il termine più generico *person*. Mentre per *emotion* e *revenge* abbiamo una similarità ancora più bassa siccome sono termini astratti rispetto ai precedenti.