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 di: Luigi Di Caro

Esercitazione 2

L'esercitazione richiede di formulare un algoritmo che, dato un insieme di definizioni espresse sotto forma di singola frase, restituisca il synset che meglio corrisponda a tali definizioni, similarmente a quanto un essere umano comune sa fare.

Algoritmo

Dopo la lettura del file e l'agglomerazione di tutte le definizioni di una colonna in un unico insieme, si è eseguito l'algoritmo progettato al fine di trovare un *synset* che rappresenti al meglio possibile tale insieme di definizioni.
L'algoritmo è diviso in due parti:

- *preparazione*: si raccolgono le informazioni relative alle definizioni per generare un contesto, simile a quello di disambiguazione
- *ricerca*: si scansionano iterativamente i *synset* delle definizioni, formando la *frontiera*, per cercare quello più aderente alle definizioni date.

Inizializzazione

Durante l'inizializzazione si forma un insieme di parole, il cosiddetto *bag pesato*, estratte dai vari *synset* di quelle che formano le frasi delle definizioni, e si associa un peso ad ognuna di esse.
Il calcolo è relativamente complesso, sommando le seguenti componenti:

- al nome del *synset* presente nelle frasi ed i sinonimi viene associato 4.
- alle parole nelle definizioni e ai nomi degli iperonimi viene associato 2.
- al resto 1, ossia ad iponimi, alle parole negli esempi, ai meronimi ed agli olonimi.
- dopodichè, si aggregano tutti i pesi calcolati per ogni parola, qualora questa fosse stata individuata in più definizioni distinte, in una unica lista di pesi qualora tale parola fosse stata individuata più volte.
- si trasformano le eventuali liste di pesi in un unico valore tramite un calcolo, che può essere la media dei valori o, come avviene, la sommatoria degli stessi.

In questo modo si forma il cosiddetto *genus*.

Similarmente, si forma la *differentia* per ogni *synset* di ogni antinomia dei *synset* delle parole presenti unicamente nelle definizioni, ma limitando la "fonte" di parole al solo nome e definizione dell'antinomia. In questo modo andiamo ad esplorare la tassonomia di WordNet cercando di tagliare i percorsi che contengono risultati non adeguati alla bag prodotta.
Infine, si forma la *frontiera*, ossia una struttura FIFO di nomi di *synset* da valutare.

Ricerca

Estraendo una parola alla volta dalla frontiera, per ogni *synset* relativo, si calcola il *bag* relativo e si computa il *weighted overlap* con esso e quello ricavato inizialmente dalle definizioni. Ad esso si sottrae il *weighted overlap* del *bag* sopracitato con il *differentia*. Se il valore di *similarità* così ottenuto ed approssimato superasse quello del *synset* migliore fin'ora trovato (inzialmente zero) allora il *synset* corrente sarà considerato il migliore.

Il *weighted overlap* si calcola tra due *bag* sommando i pesi delle parole tra di loro condivise, ossia sommando i pesi dell'intersezione. Nessuna normalizzazione viene apportata.

Risultati

Output atteso:		***Output ottenuto***:
Justice	→	principle.n.01
Patience	→	time.n.01
Greed	→	quality.n.01
Politics	→	state.n.04
Food	→	thing.n.01
Radiator	→	heat.n.03
Vehicle	→	place.n.02
Screw	→	point.n.20

Codice condiviso

Sono state definite delle piccole librerie, una per file ed accorpate sotto il package utilities, per meglio definire ed isolare gli algoritmi di questa e delle future esercitazioni. Di seguito sono elencate le funzionalità utilizzate, raggruppate per libreria di definizione.

- *functions*: l'unica funzione qui definita ed utilizzata è quella di similarità *weighted similarity*, simile ed ispirata alla funzione *similarity overlap*. Il risultato non è normalizzato sulla minore delle cardinalità ed il numeratore non è la cardinalità dell'intersezione, bensì la somma dei pesi di ogni parola in essa contenuta. Considerato i contesti in cui viene utilizzata e la maggiore precisione e facilità di confronto dei numeri interi rispetto a quelli in virgola mobile, nonché la differenza di prestazioni nelle operazioni associate, si è preferito mantenere il risultato come numero intero non normalizzato.
- *cacheInfoExtraction*: solo la classe *CacheSynsetsBag* viene importata. Tale classe permette di ottenere dei *synset* associati ad una parola testuale e racchiude, come *cache*, una mappa. Nel caso in cui il *synset* non sia presente nella mappatura, si richiamano le *API* di *Wordnet*.
- *synsetInfoExtraction*: questa libreria include varie funzionalità per elaborare insiemi di parole e di *synset*

  - *weighted_bag_for_sentence*: dato un insieme di parole od una frase, la quale viene raffinata per mezzo della funzione *preprocessing* per ottenere tale insieme, si raccolgono alcune informazioni di ciascuna parola tramite la funzione *weighted_bag_for_word* descritta successivamente. Con essa, si associa a ciascuna parola dell'insieme dato per parametro un insieme di parole pesate, ossia si produce una mappatura. Tali mappature sia sono conservate individualmente sia sono unite in una unica mappatura per mezzo della funzione *merge_weighted_bags* descritta successivamente.
  - *SynsetInfoExtractionOptions*: classe che racchiude quali informazioni devono essere estratte da un *synset* e quali pesi sono ad esse associati. Le informazioni sono: nome, parole nella definizione e negli esempi, sinonimi, iperonimi, iponimi, meronimi ed olonimi. I pesi sono tre ed hanno valori a crescita esponenziale di base 2: 1, 2 e 4.
  - *weighted_bag_for_word*: data una parola, una istanza opzionale di *CacheSynsetsBag* ed una istanza opzionale di *SynsetInfoExtractionOptions*, si estraggono le informazioni pesate dai *synset* associati a tale parola.
  - *merge_weighted_bags*: date due mappature, che associano ad una parola un peso, si ottiene una terza mappatura che ha per dominio l'unione dei domini delle due date. I pesi delle parole presenti in più di una mappatura sono agglomerati e ridotti ad un unico valore finale. Correntemente, l'operazione di riduzione consiste in una sommatoria, seppur la versione iniziale fosse la media aritmetica, nonostante la distribuzione dei pesi.