Relazione
Esercitazione 3.1

Annotazione di Corpora (semantic similarity)

Introduzione

La prima operazione consiste nell'annotare, con punteggio di semantic similarity, 50 coppie di termini, scelti grazie ad un metodo che in base al proprio cognome restituisce un range di 50 valori, (nel mio caso personale i valori risultanti sono compresi nell'intervallo da 51 a 100) dal file it.test.data.txt tramite i seguenti criteri:

- 4 molto simili -> sinonimi
- 3 simili -> le parole condividono molto, ma non si riferiscono allo stesso concetto (*es* Leone e Zebra)
- 2 leggermente simili -> le parole non hanno un significato molto simile, ma condividono lo stesso argomento, dominio, funzione oppure sono correlati (*es* Casa e Finestra)
- 1 diversi -> le due parole sono diverse, ma potrebbero avere piccoli dettagli che le accomunino; potrebbero essere trovate in un articolo che riguardi un argomento comune (*es* Software e Tastiera)
- 0 completamente diversi -> le due parole non hanno lo stesso significato e appartengono a topic diversi (*es* Biro e Rana)

L'output della prima consegna è un file (in formato tsv) di 50 linee, ciascuna contenente un numero nell'intervallo [0,4].

La valutazione dei punteggi annotati dovrà essere condotta in rapporto alla similarità ottenuta utilizzando i vettori NASARI (versione embedded; file *mini_NASARI.tsv*) massimizzando la cosine similarity al posto della generica funzione `sim(c1, c2)`

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

La valutazione dell'annotazione è condotta calcolando i coefficienti di Pearsons e Separman fra i punteggi annotati a mano e quelli calcolati con la versione embedded di NASARI.

Struttura del codice

Il codice, per entrambe le parti dell'esercitazione, è diviso in due file

- **`utils.py`** file che contiene metodi di utilità tra cui
    - *`parse_mini_nasari`* dove viene parsificato il file "*mini_NASARI.tsv*" e ritorniamo un dizionario `{synset:vector}` dove `chiave=synset` e `valore=vettore`
    - *`parse_it_test_data`* dove parsifichiamo il file *"it.test.data.tsv"* e ristiuiamo una la coppia di termini e una lista contenente il valore annotato per ogni coppia
    - *`parse_bn_syns_annotated`* legge in input il file *"it.test.data.tsv"* e recupera i *babelSynsetIDs* e i termini annotati a mano
    - *`max_nasari_similarity`* cicla sui *babelSynsetIDs* di ciascuna coppia di termini data in input e restituisce la massima similarità (max cosine similarity). Tale metodo quindi si avvale su altri due metodi:
        - *`get_synsets_from_term`* Dato il termine in inpt, ritorna i suoi *babelSynsetIDs* nel file *"SemEval17_IT_senses2synsets.txt"* o la lista vuota se il termine non viene trovato
        - *`get_max_similarity`* dati i synset di due termini restituisce la massima similarità (max cosine similarity) con il metodo *`cosine_similarity`*
    - *`get_best_syns_by_term`* lavora in maniera analoga al metodo prima descritto ma questa volta ritorna i babelSynsetIDs che hanno portato alla massima cosine similarity
    - *`get_terms_by_synset`* metodo che dato un *babelSynset* consente di ricavare i suoi termini grazie alle API di BabelNet

- **`main.py`** file principale dove vengono presi in input i file indicati precedentemente e calcolati gli indici e le misure di valutazione di entrambe le parti dell'esercitazione

Implementazione

Le fasi dell'implementazione sono state le seguenti

1. Annotazione è stato annotato il file con i valori di similarità e con i *BabelNetSynsetID* scelti dal file *SemEval17_IT_senses2synsets.txt*.

2. Ricerca Automatica del Synset per ogni coppia di termini si sono cercati i *BabelNetSynsetID* dal file *SemEval17_IT_senses2synsets.txt* nel e quest'ultimi sono stati cercati nel file *mini_NASARI.tsv* considerando quelli che massimizzavano la *cosine similarity*. I *BabelNetSynsetID* che hanno massimizzato quest'ultima, sono stati i concetti presi come riferimento per l'appunto.

3. Valutazione Similarità tramite le metriche di *Pearson* e *Spearman*, sono state calcolati i valori di similarità tra i valori annotati e quelli trovati al passo precedente.

Risultati

Di seguito i risultati ottenuti con le metriche prima indicate e i valori, annotati e calcolati, di similarità (per ragioni di spazio vengono mostrate le prime righe dell'output).

Spearman: 0.8057

Pearson: 0.7811

termine 1	termine 2	gold_score	cos_sim_score
biotopo	biologia	1.9	2.91
magma	vulcano	2.3	3.47
brainstorming	telescopio	0.0	1.27
livello	punteggio	1.8	3.85
centesimo	affare	1.0	2.16
partito politico	associazione	2.1	2.86
tsunami	mare	2.0	2.42
struzzo	frutteto	0.0	1.82
cannella	caramella	0.9	3.18
scopa	polvere	1.8	2.95
galassia	astronomo	1.9	3.05
succo	frappè	2.0	3.25
tapparella	tenda	1.9	2.88
criminale	colpevole	2.8	3.47
cancro al pancreas	chemioterapia	2.0	3.63
passato	antecedente	3.8	3.65
nazioni unite	parlamento europeo	1.1	2.65
canzone	esecutore	0.8	2.82
pistola	taccuino	0.0	1.88
acetilcolina	iride	0.3	2.33

Relazione
Esercitazione 3.2

Sense Identification

Introduzione

La seconda parte consiste nell'individuare i sensi selezionati nel giudizio di similarità.

Per cui si è ragionato sul chiedersi: nell'attribuire un valore di similarità a una coppia di termini (per esempio, *società* e *cultura*), quali sensi vengono selezionati?

Si parte dall'assunzione che i due termini funzionino come contesto di disambiguazione l'uno per l'altro.

A differenza di quanto fatto nella prima esercitazione, ora non si è interessati a calcolare il punteggio di similarità fra due termini, ma di individuare i sensi che hanno portato alla massimizzazione di tale punteggio.

Si tratta quindi di eseguire questa operazione

$$c_1, c_2 \leftarrow \underset{c_1 \in s(w_1), c_2 \in s(w_2)}{\arg\max} [sim(c_1, c_2)]$$

Implementazione

Le fasi dell'implementazione sono state le seguenti

1. **Annotare** ai *BabelNetSynsetID* ai termini che si ipotizza massimizzino la similarità (già eseguito nella prima parte dell'esercitazione)

   Per cui ne deriva una struttura costituita da 6 campi:

   - Term1
   - Term2
   - BS1 (dato personalmente)
   - BS2 (dato personalmente)
   - Terms_in_BS1
   - Terms_in_BS2

2. **Ricerca** dei *BabelNetSynsetID* che hanno portato alla max cosine similarity e i loro relativi termini (glossa) tramite chiamate API Babelnet (eseguito nella prima parte).

3. **Valutare** sulla misura di precisione dei *BabelNetSynsetID* annotati e di quelli indentificati (al passo precedente).

Risultati

Di seguito i risultati della misura di Precision sia sui singoli termini che sulle coppie, e un output generale (per ragioni di spazio vengono mostrate le prime righe dell'output e i soli *BabelNetSynsetID*).

Accuratezza tra i singoli termini: 43%

Accuratezza delle coppie di termini: 32%

termine 1	termine 2	gold_score	cos_sim_score	BN syn annotato per term1	BN syn calcolato per term1	BN syn annotato per term2	BN syn calcolato per term2
biotopo	biologia	1.9	2.91	bn:03353031n	bn:03353031n	bn:00010543n	bn:00010543n
magma	vulcano	2.3	3.47	bn:00052703n	bn:00052703n	bn:00079748n	bn:00080211n
brainstorming	telescopio	0.0	1.27	bn:00012707n	bn:00012707n	bn:00069738n	bn:00069738n
livello	punteggio	1.8	3.85	bn:00025965n	bn:01182399n	bn:00041241n	bn:00069755n
centesimo	affare	1.0	2.16	bn:00017162n	bn:00025112n	bn:00014152n	bn:00014139n
partito politico	associazione	2.1	2.86	bn:00060834n	bn:00060834n	bn:14146654n	bn:00020876n
tsunami	mare	2.0	2.42	bn:00078509n	bn:00078509n	bn:00069946n	bn:00069946n
struzzo	frutteto	0.0	1.82	bn:00059688n	bn:00059688n	bn:00041967n	bn:00041967n
cannella	caramella	0.9	3.18	bn:00019142n	bn:00017431n	bn:00015227n	bn:00021707n
scopa	polvere	1.8	2.95	bn:00013352n	bn:00013352n	bn:00029193n	bn:03723815n
galassia	astronomo	1.9	3.05	bn:00032476n	bn:00056292n	bn:00006659n	bn:00006659n
succo	frappè	2.0	3.25	bn:00048532n	bn:00048532n	bn:00055009n	bn:00055009n
tapparella	tenda	1.9	2.88	bn:00068154n	bn:01256008n	bn:00024534n	bn:00007513n
criminale	colpevole	2.8	3.47	bn:05040795n	bn:00023807n	bn:00024341n	bn:00011099n
cancro al pancreas	chemioterapia	2.0	3.63	bn:00060358n	bn:00060358n	bn:00018141n	bn:00018141n
passato	antecedente	3.8	3.65	bn:00060928n	bn:00060929n	bn:00004490n	bn:15063066n
nazioni unite	parlamento europeo	1.1	2.65	bn:00078931n	bn:00078931n	bn:03855948n	bn:03855948n
canzone	esecutore	0.8	2.82	n:00072794n	bn:00072794n	bn:00032171n	bn:00046975n
pistola	taccuino	0.0	1.88	bn:00042808n	bn:01573920n	bn:00058156n	bn:00058157n