Relazione esercizio “False Friends”

L’esercitazione ha visto la definizione di un algoritmo di individuazione di parole **false friends**. La definizione generale di un false friend è quella di due parole del lessico di lingue diverse che presentano una veste lessicale molto simile, se non identica in alcuni casi, ma che individuano significati completamente differenti. Nel caso di una sola lingua, la definizione di false friends si restringe semplicemente a due termini quasi omonimi che condividono molti caratteri in comune ma che differiscono di molto nel significato. Per questa esercitazione si è deciso di lavorare su una sola lingua, l’inglese, sfruttando la risorsa lessicale di wordNET per accedere ai diversi sensi dei termini.

Per comprendere se due parole sono o meno false friends si è deciso di guardare la loro distanza di edit, ovvero il minimo numero di operazioni di inserimento, rimozione, modifica, per trasformare una stringa in un’altra. In seguito sono state valutare le similarità, di modo che termini con similarità bassa e distanza di edit bassa sono dei buoni candidati per essere dei false friends.

Avendo già affrontato nella seconda parte del corso un’esercitazione con il corpus **SemCor** si è deciso di sfruttare nuovamente tale risorsa, usata in questo contesto solo per accedere ad una lista di parole di contenuto della lingua inglese e per verificare la presenza di false friends. SemCor è un corpus in lingua inglese composto da 352 testi annotati semanticamente, con un totale di **37176 frasi**, che provengono dal **Brown corpus**. Attualmente, Semcor rappresenta il più ampio dataset annotato a mano con i synset di wordnet.
I file del corpus sono documenti XML, nei quali un tag **<p pnum="1">** segna l’inizio di un paragrafo del documento e un tag **<s snum="1">** individua l’inizio di una frase all’interno del paragrafo.
Ad esempio, le seguenti righe individuano 3 frasi distinte:

- <wf cmd="ignore" pos="DT">The</wf>
- <wf cmd="done" pos="JJ" lemma="nuclear" wnsn="1" lexsn = " 3:00:00 ::" > nuclear</wf>
- <wf cmd="done" pos="NN" lemma="war" wnsn="2" lexsn = "1:26:00::" > war</wf>

Non tutte le frasi del corpus presentano lo stesso contenuto informativo. Ad esempio, le **stopword** come “The”, nella prima frase, non hanno associato un attributo “lemma” o un attributo “wnsn” che individua il rispettivo synset di wordnet. In generale, queste mancanze sono presenti sia per le parole di contenuto, sia per le parole di funzionalità.
Ai fini di questa esercitazione sono state considerate solo le frasi annotate con le indicazioni del POS e del riferimento a wordnet. Il file “**false_friends.py**” contiene la definizione di 3 funzioni usate ai fini di questa esercitazione:

- La funzione **“open_random_semcor”** restituisce un sottoinsieme di frasi scelte in modo causale, prelevate dall’intero nucleo di documenti del corpus. Il numero di frasi da recuperare è dato in input come parametro.
- Una seconda funzione, **“read_nouns_from_semcor”,** legge da una lista di frasi tratte da un corpus semcor i lemmi associati alle parole di contenuto. Tali informazioni verranno usate ai fini dell'individuazione dei false friends.

- Infine la funzione “**random_couples**” a partire dalle parole restituite dalla precedente fu calcola tutti i possibili accoppiamenti di termini, rimuove le coppie duplicate e filtra solo quelle con una distanza di edit minore di un valore **“edit_threshold”** passato come parametro. Infine calcola la similarità di **wu and palmer** tra i termini delle coppie rimaste e restituisce quelle con una similarità minore di un certo valore “**similarity_threshold”**

Risultati

Il programma all’atto dell’esecuzione richiede 3 parametri da riga di comando:

- SENT = numero di frasi da prelevare dal corpus Semcor.
- EDIT_THRESHOLD = valore soglia limite per la distanza di edit
- SIM_THRESHOLD = valore soglia limite per la similarità tra termini

Facendo diversi tentativi si è notato che la migliore configurazione è

- Non più di 20 frasi altrimenti la ricerca impiega molto tempo
- un valore di edit distance pari a 2 (tutte le parole che differiscono massimo 2 caratteri)
- un valore di similarità minore di 0.3

Sono stati condotti diversi test, qui di seguito riportiamo alcuni risultati. La selezione delle frasi avviene su porzioni causali del corpus. Alcune di queste sono ricche di parole di contenuto, altre meno. Per tale ragione non sempre accade che vengano restituiti molte coppie.