Relazione esercitazione – False friends

L'obiettivo di questa esercitazione è stato quello di creare un algoritmo che fosse in grado di **individuare coppie di termini della lingua inglese che potessero essere considerati dei "false friends"**. Poiché in questo caso, consideriamo una sola lingua, possiamo considerare come false friends due termini che dal punto di vista della forma sono quasi identici ma che differiscono notevolmente nel significato che assumono.

Le risorse principali utilizzate per questa esercitazione sono state:

- **Wordnet**, per ottenere i possibili sensi associati ad un certo termine.

- **Semcor,** che invece è un corpus già utilizzato nella seconda parte del corso. Esso contiene al proprio interno circa 37000 frasi annotate a diversi livelli (per ogni parola in ogni frase contiene anche il synset di WN di riferimento), che è stato utilizzato per ottenere le frasi nelle quali andare a cercare le possibili parole che potrebbero essere false friends.

L'algoritmo sviluppato è composto dai seguenti passi:

1) L'utilizzo di una procedura chiamata "**individuazione_coppie_parole_synset**" che si preoccupa di andare ad individuare tutte le coppie **parola-synset** partendo da un insieme di frasi casuali prese da Semcor. Tutte le coppie trovate (senza duplicati) vengono serializzate. Il numero di frasi considerate per l'estrazione è stato 3000.

2) Una volta terminata la procedura descritta al passo 1), viene deserializzata la lista di coppie parola-synset e subito dopo viene chiamata la funzione "**get_coppie_parole_con_dist_giusta**" che prendendo in input la lista appena citata, si preoccuperà di trovare tutte le possibili coppie di parole che dal punto di vista della **distanza edit risultano essere molto simili tra loro.** La **distanza edit** è semplicemente il numero di operazioni di: **rimozione, inserimento e modifica** che sono necessarie per trasformare una parola in un'altra. Il numero massimo di operazioni richieste è un parametro che viene scelto a priori e in questa esercitazione il valore di soglia deciso è stato 2, quindi sono state selezionate tutte le coppie di parole che avevano una distanza edit minore di 2. La funzione utilizzata per calcolare l'edit è presente nella libreria **nltk** ed è chiamata "**edit_distance**". Ogni parola prima di essere confrontata con un'altra è stata lemmatizzata.

3) Dopodichè, una volta ottenute tutte le coppie di termini con edit distance minore di 2, per ciascuna coppia di parole sono stati eseguiti i seguenti passi:

    3.1) Sono stati presi da Wordnet tutti i possibili synsets associati a ciascuna parola della coppia.
    3.2) Per il task di Word Similarity, è stata utilizzata la misura di similarità chiamata "**Wu&Palmer**". Il motivo di tale scelta è dovuto al fatto che durante le sperimentazioni fatte in un'esercitazione con il professor Radicioni, essa è risultata essere la metrica che restituiva risultati di similarità più vicini a quelli forniti dagli esseri umani.

Similarity Wu&Palmer:

$$\boldsymbol{sim(c1, c2)} = \frac{\boldsymbol{2 * depth(LCS)}}{\boldsymbol{depth(c1) + depth(c2)}}$$

Ove:

LCS = Lowest Common Subsumer (è il primo antenato comune, presente come synset in Wordnet, tra i sensi c1 e c2).

depth(x) = è una funzione che misura la distanza tra la radice di Wordnet e il synset x.

3.3) Il valore di similarità tra ogni coppia di termini (w1, w2) è stato ottenuto eseguendo questo calcolo:

$$\boldsymbol{sim(w1, w2)} = \underset{\boldsymbol{c1 \in S(w1),\ c2 \in S(w2)}}{\boldsymbol{MAX}} [sim(c1,c2)]$$

dove:

S(w1) = insieme di tutti i synsets associati al termine w1.

S(w2) = insieme di tutti i synsets associati al termine w2.

c1 = possibile synset associato al termine w1.

c2 = possibile synset associato al termine w2.

4. Se, il valore di similarità calcolato, **è al di sotto di una certa soglia che è stata fissata a 0.20**, allora, poiché questo vuol dire che le due parole dal punto di vista semantico sono abbastanza diverse tra loro, la coppia di termini viene memorizzata all'interno di una lista. Alla fine, dopo aver valutato tutte le varie coppie di parole da considerare, l'algoritmo stampa la lista di coppie di parole della lingua inglese che possono essere considerate false friends.

Risultati ottenuti

Il numero totale di false friends trovati dall'algoritmo è **120**, essi sono presenti nel file chiamato "**False friends – 120**".

Di seguito vengono riportati **tutti** i false friends trovati dall'algoritmo e il loro valore di similarità:

```
['published', 'publisher', 0.18181818181818182],
['four', 'for', 0],
['used', 'sed', 0],
['pulled', 'pulley', 0.13333333333333333],
['many', 'may', 0.18181818181818182],
['many', 'mary', 0.18181818181818182],
['led', 'sed', 0],
['brother', 'bother', 0.18181818181818182],
['aunt', 'taunt', 0.13333333333333333],
['set', 'sed', 0],
['the', 'tie', 0],
['victory', 'victor', 0.15384615384615385],
['long', 'lung', 0.16666666666666666],
['sat', 'oat', 0.18181818181818182],
['bloat', 'boat', 0.18181818181818182],
['cattle', 'battle', 0.09523809523809523],
['lung', 'hung', 0.16666666666666666],
['lung', 'flung', 0.15384615384615385],
['lung', 'clung', 0.16666666666666666],
['find', 'finn', 0.18181818181818182],
['out', 'oat', 0.18181818181818182],
['navy', 'wavy', 0.18181818181818182],
['see', 'sed', 0],
['lady', 'lay', 0.18181818181818182],
['lady', 'lacy', 0.18181818181818182],
['came', 'cafe', 0.16666666666666666],
['had', 'dad', 0.13333333333333333],
['wide', 'wife', 0.18181818181818182],
['thanking', 'thinking', 0.18181818181818182],
['form', 'for', 0],
['join', 'goin', 0],
['know', 'knox', 0.16666666666666666],
['sped', 'sed', 0],
['law', 'lawn', 0.14285714285714285],
['tricked', 'ticked', 0.16666666666666666],
['got', 'god', 0.18181818181818182],
['lawn', 'jawn', 0],
['went', 'cent', 0.18181818181818182],
['feat', 'felt', 0.18181818181818182],
['feat', 'eat', 0.16666666666666666],
['feat', 'neat', 0.18181818181818182],
```

```
    ['with', 'wish', 0],
    ['won', 'ion', 0.18181818181818182],
    ['royale', 'royal', 0],
    ['mary', 'marry', 0.15384615384615385],
    ['dad', 'bad', 0.13333333333333333],
    ['dad', 'did', 0.13333333333333333],
    ['dad', 'dead', 0.13333333333333333],
    ['word', 'sword', 0.125],
    ['sea', 'sed', 0],
    ['own', 'owl', 0.13333333333333333],
    ['later', 'liter', 0.18181818181818182],
    ['gain', 'goin', 0],
    ['try', 'tray', 0.15384615384615385],
    ['creed', 'creek', 0.14285714285714285],
    ['creed', 'cried', 0.16666666666666666],
    ['care', 'car', 0.15384615384615385],
    ['care', 'cafe', 0.15384615384615385],
    ['knee', 'kneel', 0.16666666666666666],
    ['backed', 'backer', 0.16666666666666666],
    ['haste', 'hasty', 0.18181818181818182],
    ['fist', 'fast', 0.16666666666666666],
    ['fist', 'list', 0.15384615384615385],
    ['fist', 'fit', 0.16666666666666666],
    ['bought', 'bough', 0.13333333333333333],
    ['eat', 'oat', 0.16666666666666666],
    ['eat', 'ear', 0.16666666666666666],
    ['ever', 'fever', 0.18181818181818182],
    ['finn', 'fine', 0.18181818181818182],
    ['done', 'doe', 0.13333333333333333],
    ['add', 'adc', 0],
    ['hear', 'heir', 0.18181818181818182],
    ['hear', 'ear', 0.18181818181818182],
    ['ion', 'ton', 0.125],
    ['sworn', 'sword', 0.15384615384615385],
    ['going', 'goin', 0],
    ['for', 'far', 0],
    ['for', 'fox', 0],
    ['for', 'fort', 0],
    ['for', 'fog', 0],
    ['boy', 'buy', 0.18181818181818182],
    ['boy', 'boo', 0.15384615384615385],
    ['like', 'pike', 0.18181818181818182],
    ['composer', 'composed', 0.16666666666666666],
    ['peer', 'per', 0],
    ['vast', 'vas', 0.18181818181818182],
    ['car', 'war', 0.15384615384615385],
    ['car', 'far', 0.16666666666666666],
    ['quiet', 'quirt', 0.19047619047619047],
    ['sit', 'skit', 0.18181818181818182],
    ['guest', 'guess', 0.15384615384615385],
    ['economics', 'economic', 0.15384615384615385],
    ['poet', 'poem', 0.14285714285714285],
    ['robe', 'role', 0.18181818181818182],
    ['neutralism', 'neutralist', 0.1111111111111111],
    ['earn', 'ear', 0.16666666666666666],
    ['per', 'pea', 0],
    ['per', 'peru', 0],
    ['per', 'pen', 0],
    ['ton', 'torn', 0.18181818181818182],
    ['lurked', 'lured', 0.13333333333333333],
    ['near', 'ear', 0.18181818181818182],
    ['coin', 'goin', 0],
    ['tutor', 'tumor', 0.125],
    ['wipe', 'wife', 0.15384615384615385],
    ['send', 'sed', 0],
    ['swore', 'sword', 0.14285714285714285],
    ['familiar', 'familar', 0],
    ['red', 'sed', 0],
    ['quirt', 'quit', 0.15384615384615385],
    ['gust', 'just', 0.15384615384615385],
    ['thanked', 'shanked', 0.14285714285714285],
    ['whisked', 'whiskey', 0.14285714285714285],
    ['manufactured', 'manufacturer', 0.18181818181818182],
    ['lamp', 'damp', 0.16666666666666666],
    ['sed', 'bed', 0],
    ['sed', 'fed', 0],
    ['sed', 'sex', 0],
    ['stir', 'stair', 0.15384615384615385],
    ['inner', 'sinner', 0.18181818181818182]
]
```