La quinta esercitazione prevede due parti: una di semantic word similarity e una di sense identification.

Semantic word similarity:

1. Individuazione di 50 coppie di termini dal file *it.test.data.txt* sulla base del cognome (utilizzare la funzione presente nel notebook *semeval_mapper.ipynb*).
2. Annotazione con punteggio di semantic similarity le 50 coppie di termini. Il criterio da utilizzare è il seguente (https://tinyurl.com/y6f8h2kd):

   4: Very similar *- The two words are synonyms (e.g., midday-noon).*
   3: Similar *- The two words share many of the important ideas of their meaning but include slightly different details.They refer to similar but not identical concepts (e.g., lion-zebra).*
   2: Slightly similar *- The two words do not have a ver y similar meaning, but share a common topic/domain/function and ideas or concepts that are related (e.g., house-window).*
   1: Dissimilar *- The two items describe clearly dissimilar concepts, but may share some small details, a far relationship or a domain in common and might be likely to be found together in a longer document on the same topic (e.g., software-keyboard).*
   0: Totally dissimilar and unrelated *- The two items do not mean the same thing and are not on the same topic (e.g., pencil-frog).*

   L'output della prima consegna è un file (in formato *.tsv*) di 50 linee, ciascuna contenente un numero in $[0,4]$, con la seguente struttura

   ```
   [word1, word2, similarity]
   ```

3. La valutazione dei punteggi annotati dovrà essere condotta in rapporto alla similarità ottenuta utilizzando i vettori NASARI (versione embedded, file *mini_NASARI.tsv*, nel materiale della lezione). La similarità tra due termini si determina tramite massimizzazione della cosine similarity

$$cos - sim(\overrightarrow{V1}, \overrightarrow{V2}) = \frac{\overrightarrow{V1} \cdot \overrightarrow{V2}}{||\overrightarrow{V1}||\,||\overrightarrow{V2}||}$$

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

Sense identification:

1. Il secondo compito consiste nell'individuare i sensi selezionati nel giudizio di similarità.
   La domanda che ci poniamo è la seguente: *quali sensi abbiamo effettivamente utilizzato quando abbiamo assegnato un valore di similarità a una coppia di termini (per esempio, società e cultura)?*
2. L'output di questa parte dell'esercitazione consiste in 2 Babel synset ID e dai termini del synset. Il formato dell'output ha la seguente struttura

   ```
   #Term1 Term2 BS1 BS2 Terms_in_BS1 Terms_in_BS2
   ```

3. Viene fatta un'annotazione automatica basata sulla coppia di sensi di babelnet che massimizza la funzione di Cosine similarity e ottenuta una lista di babelnet ids viene confrontata con la lista di sensi annotata manualmente, ottenendo una percentuale di uguaglianza.

1.2 Svolgimento

Semantic word similarity

Le coppie individuate in base al cognome (Pellegrino) sono:

```
Pellegrino     :        coppie nell'intervallo 51-100
```

Di seguito un estratto dell'annotazione manuale:

```
biotopo biologia                2.8
magma vulcano                   3.5
brainstorming telescopio        0
livello punteggio               3
centesimo affare                1.5
partito politico associazione   2.8
tsunami mare                            3.5
struzzo frutteto                1
```

Il calcolo della similarità tra due termini in base ai vettori Nasari viene fatto con la funzione `calculateMaxSimilarity(babel id1, babel id2, nasari)` per la quale i babel ids vengono calcolati utilizzando il file che associa le parole ai vari id di babel connessi ad esse. (*SemEval17_IT_senses2synsets.txt*), tramite la funzione `findIds(term, sense to synsets file)`. La funzione oltre a restituire il valore massimo di similarità calcolata su tutti i sensi delle coppie di parole, ritorna anche due indici che vengono utilizzati per salvare i due sensi che massimizzano la cosine similarity.

Infine, vengono estratte le liste contenenti i punteggi di similarità normalizzando la lista annotata manualmente e viene calcolato l'indice di correlazione tra i due punteggi in base ai coefficienti di Pearson e di Spearman, ottenendo il seguente risultato:

```
The Pearson correlation in my work is equal to: 0.7686800467559665
The Spearman correlation in my work is equal to: 0.6868222961931585
```

Sense identification

Dopo aver annotato manualmente, con l'utilizzo di http://live.babelnet.org/, un file nella forma:

```
[word1, word2, id_sense_word1, id_sense_word2, Terms_in_BS1, Terms_in_BS2]
```

Utilizzando la funzione `getTableResult(personal ids, automatic ids)` vengono estratti sia gli id babelnet annotati manualmente, sia quelli ottenuti attraverso la funzione precedentemente discussa.
Viene creata dunque una tabella contenente sia le scelte fatte manualmente che quelle fatte automaticamente e viene calcolata la percentuale di id uguali, ossia quelli che rappresentano la scelta comune.

Il risultato ottenuto è:

```
The percentage of same Babelnet synset ids in my work is equal to: 55.1 %
```