Implementare l'algoritmo di Lesk (non utilizzare l'implementazione esistente, e.g. NLTK).

1. Disambiguare i termini polisemici all'interno delle frasi del file *sentences.txt*; oltre a restituire i synset ID del senso (appropriato per il contesto), il programma deve riscrivere ciascuna frase in input sostituendo il termine polisemico con l'elenco dei sinonimi eventualmente presenti nel synset.
2. Estrarre 50 frasi dal corpus SemCor (corpus annotato con i synset di WN) e disambiguare almeno un sostantivo per frase. Calcolare l'accuratezza del sistema implementato sulla base dei sensi annotati in SemCor.

1.2 Svolgimento 1° Parte

La prima parte di questa esercitazione prevede di determinare il miglior senso associato al termine e modificare la frase in input sostituendo la parola di cui andiamo a fare il lesk con i suoi (se presenti) sinonimi.
La funzione `lesk(context, keyword)` prende in input il termine e la frase completa (`sentence`). Il metodo `bagOfWord(sentence)` calcola il contesto con un approccio bag of words, andando a rimuovere le stop words (prese dal file stop_words_FULL), la punteggiature e lemmatizzando tutti i termini rimanenti. Il risultato è una lista di parole. Successivamente, per ogni senso associato alla parola da disambiguare, viene estratto il contesto del senso, ossia una lista di termini presenti nella definizione e negli esempi presenti nel synset (calcolati tramite `sense.definition()` e `sense.examples()`), che vengono poi processati con la funzione `bagOfWord(sentence)`. La sovrapposizione (*overlap*) tra i due contesti viene calcolata come la lunghezza dell'intersezione tra le due liste di termini (*context_words* e *signature*). Il miglior senso è quello che ha il valore di overlap maggiore. Per determinare i sinonimi di un synset si utilizza la funzione di NLTK `lemma_names()`).

Il risultato, presente nel file *output Word Disambiguation.txt* è una tabella, di cui di seguito viene mostrato un estratto:

Original Sentences	**Ambiguous Word**	**Choosen Synset**	**New Sentence**
the key broke in the lock	*key*	*key.v.02*	*the ['key'] broke in the lock*

1.3 Svolgimento 2° parte

La seconda parte dell'esercitazione prevede la disambiguazione di 50 parole (selezionate casualmente) tra le

prime 50 frasi del corpus annotato SemCor. Dal corpus vengono estratte le prime 50 frasi già annotate con i vari sensi (quindi semanticamente), quelle annotate con le parti del discorso (quindi sintatticamente), che serviranno per scegliere il termine da disambiguare solo tra i *Nouns* ed infine le frasi non taggate, che serviranno per il lesk e per la tabella dei risultati. La disambiguazione viene effettuata con l'algoritmo di lesk della prima parte dell'esercitazione e il metodo `semCorDisambiguation()` calcola anche l'accuratezza del metodo restituendo la percentuale di sensi che vengono individuati uguali a quelli presenti nell'annotazione di SemCor. Vengono effettuate 50 prove sulle frasi selezionate con diversi termini e la percentuale di accuratezza media ottenuta è del 41,32%.

Sentence	**Ambiguous Word**	**Choosen Synset**	**SemCor Synset**
It urged that the city `` take steps to remedy '' this problem .	*problem*	*problem.n.01*	*problem.n.02*