LAB 2-3

The aim of this lab is to test the similarity between two sentences using online lexical database WordNet. The students can refer to the original paper of Mihalcea et al. (Corpus-based and Knowledge-based Measures of Text Semantic Similarity), appeared in AAAI 2006. See, (https://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf)

1. Study Section 5 of Chapter 2 of NLTK online book, and try to reproduce the coding examples and try to use your own examples of wording to identify the synsets, hyponyms, hypernyms, and various semantic similarity between two words of your choice.
2. Identify the synsets of the word "car" and rank them in the order of their frequency of occurrence (most common synset first, less common synset at the end). For this purpose, you may use the coding:

```
car = wn.synsets('car', 'n')[0]     # Get the most common synset
print car.lemmas()[0].count()       # Get the first lemma
```

3. Now consider two sentences T1 and T2, each constituted with a set of tokens. For this purpose, study expression (1) of the aforementioned Mihalcea et al.'s paper above (see below).
4. Use a set of pair of sentences of your choice. Try to start with sentences that have close semantic meaning to pairs that are very disparate from each other, and notice how the semantic similarity of the two sentences varies.
5. Test the various available word-to-word semantic similarity (e.g., Wu and Palmer, Resnik, path length) and how they contribute to changing overall semantic similarity of the pair of sentences.

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{T_1\}} (maxSim(w, T_2) * idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in \{T_2\}} (maxSim(w, T_1) * idf(w))}{\sum_{w \in \{T_2\}} idf(w)}\right)$$