Assignment 1: NLTK Exercise

COMP4901K and MATH 4824B

Fall 2020

Prerequisites

You need to install the NLTK packages:

```
pip3 install --upgrade nltk
```

Check the .ipynb file in Tutorial 1 for more details.

1 Assignment

You should complete the script:

- `assignment1.py`

by fulfilling the following questions.

Q1 Finish codes to perform the following tasks on the corpus named "`austen-sense.txt`" from the project Gutenberg electronic text archive.

Hint: you could refer to section 4 in the Tutorial 1 NLTK.ipynb for how to retrieve the corpus

1. Print the number of word tokens in the corpus.
2. Print the size of the vocabulary (number of unique word tokens).
3. Print the tokenized words of the first sentence in the corpus using nltk.word_tokenize.

Q2 Finish codes to perform the following tasks on the `brown` corpus.

Hint: you could refer to section 4 in the Tutorial 1 NLTK.ipynb for how to retrieve the corpus. And you could check nltk.FreqDist function for this question.

1. Print the top 10 most common words in the `romance` category.
2. Print the word frequency of the following words: `[ring,activities,love,sports,church]` in the `romance` and `hobbies` categories respectively.

Q3 Finish codes to perform the following tasks using `WordNet`.

1. Print all synonymous lemma words of the word `‘dictionary’` using lemma_names method of synsets.
2. Print all hyponyms of the word `‘dictionary’`.
3. Use one of the predefined similarity measures to score the similarity of the following pairs of synsets and rank the pairs in order of decreasing similarity.
   (`right_whale.n.01`, `novel.n.01`)
   (`right_whale.n.01`, `minke_whale.n.01`)
   (`right_whale.n.01`, `tortoise.n.01`)

Hints: predefined similarity measures can be found in ***http://www.nltk.org/howto/wordnet.html***

2 Submission

You need to submit two files, program output and your python script. After you finish the assignments, make sure you include your name and student ID in the beginning of your code.

```
# author: Your_name
# student_id: Your_student_ID
```

Copy all the program output to a text file named `StudentID_assignment1_output.txt`. Rename your python script solution as `StudentID_assignment1.py`. Zip them altogether and submit it to Canvas.