Lab 2 NLTK Exercise

COMP4901K and MATH 4824B

Fall 2018

Prerequisites

You need to install the NLTK packages:

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

You need to download NLTK data:
Run the Python interpreter and type the commands:

```
import nltk
nltk.download()
```

A new window should open, showing the NLTK Downloader. Next, please select the packages or collections you want to download. Then, download data to the default directory. If you did not install the data to the default directory, you need to set the `NLTK_DATA` environment variable to specify the location of the data. Other ways to download ntlk data can be found in `https://www.nltk.org/data.html`.

1 Assignment

You need to download the following file(s) from canvas:

- `lab2_skeleton.py`: the program skeleton.

Q1 Write code to perform the following tasks on the corpus named "`austen-sense.txt`" from the project Gutenberg electronic text archive. Hints: word token means an occurrence of a word, and word type is a vocabulary item.

1. Print the number of word tokens.
2. Print the number of word types.
3. Print all tokens in the first sentence.

Q2 Write code to perform the following tasks on the `brown` corpus.

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 Write code to perform the following tasks using `WordNet`.

1. Print all synonymous words (lemmas) of the word `dictionary`.
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 finished the assignments, make sure you include the header information in the beginning of your code

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

Copy all the program output in to text file named `StudentID_lab2_output.txt`, and submit with your python script solution named `StudentID_lab2.py` to canvas.