Information Retrieval

COMP 479 Project 1
Text preprocessing with NLTK DEMO

Lin Ling (40153877)

A report submitted in partial fulfilment of the requirements of Comp479.

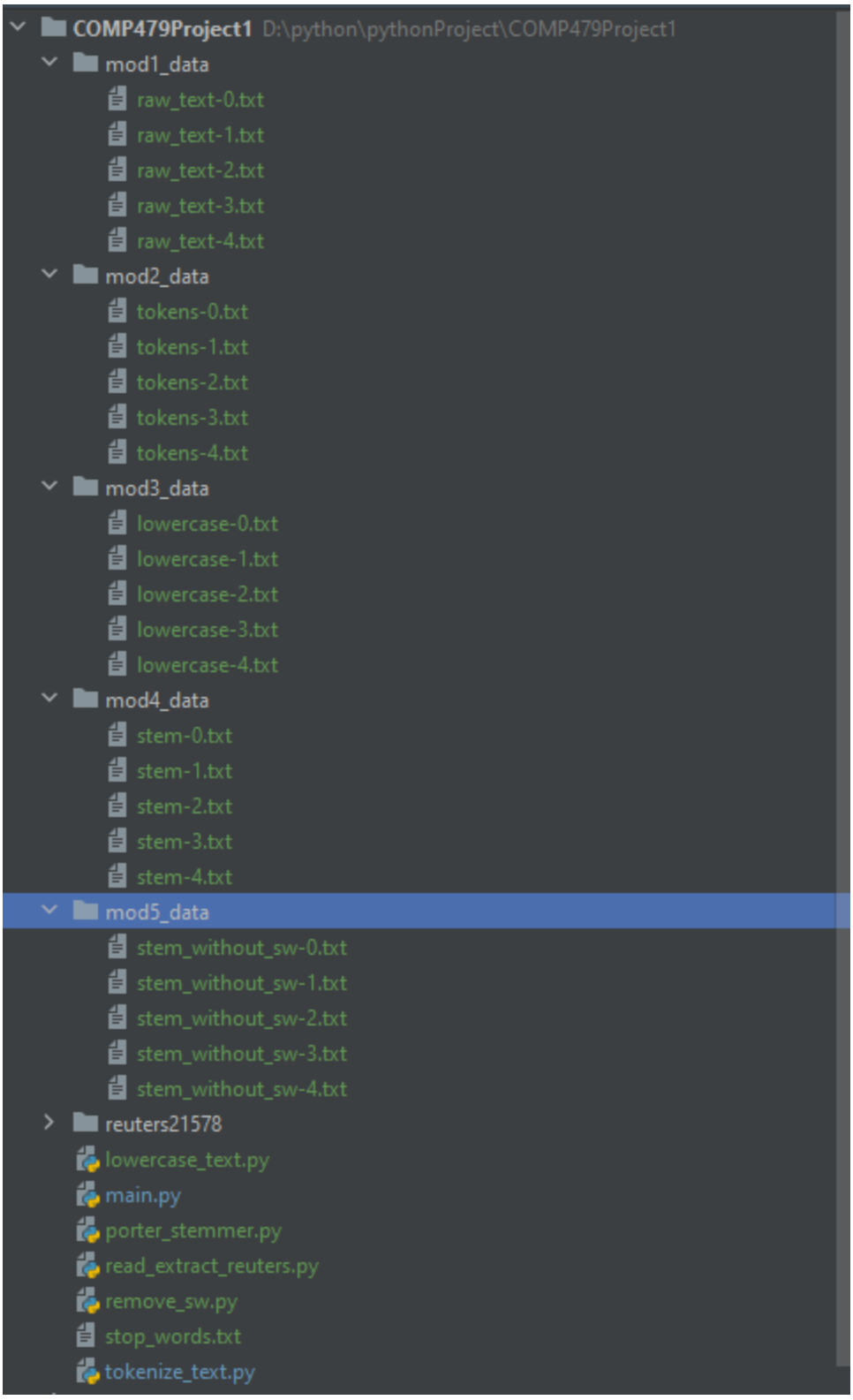
Concordia University

Source Code: main.py

test.py main.py

```python
import read_extract_reuters
import tokenize_text
import lowercase_text
import porter_stemmer
import remove_sw

TOP_FILES_NUM = 5
ORIGINAL_FILE_PATH = 'reuters21578'
ORIGINAL_FILE_TYPE = '.sgm'
MODULE_FILE_TYPE = '.txt'
MODULE_ONE_FILE_PATH = 'mod1_data'
MODULE_TWO_FILE_PATH = 'mod2_data'
MODULE_THREE_FILE_PATH = 'mod3_data'
MODULE_FOUR_FILE_PATH = 'mod4_data'
MODULE_FIVE_FILE_PATH = 'mod5_data'


# First Module:  read the Reuter's collection and extract the raw text of each article from the corpus
print("Starting read the Reuter's collection ...")
read_extract_reuters.output_file(ORIGINAL_FILE_PATH, TOP_FILES_NUM, ORIGINAL_FILE_TYPE, MODULE_ONE_FILE_PATH)
print("Output the handled files in mod1_data directory.")

# Second Module: tokenize
print("Starting read the raw text files ...")
tokenize_text.output_file(MODULE_ONE_FILE_PATH, TOP_FILES_NUM, MODULE_FILE_TYPE, MODULE_TWO_FILE_PATH)
print("Output the handled files in mod2_data directory.")

# Third Module: make all text lowercase
print("Starting read the tokens files ...")
lowercase_text.output_file(MODULE_TWO_FILE_PATH, TOP_FILES_NUM, MODULE_FILE_TYPE, MODULE_THREE_FILE_PATH)
print("Output the handled files in mod3_data directory.")

# Fourth Module: apply Porter stemmer
print("Starting read the lowercase files ...")
porter_stemmer.output_file(MODULE_THREE_FILE_PATH, TOP_FILES_NUM, MODULE_FILE_TYPE, MODULE_FOUR_FILE_PATH)
print("Output the handled files in mod4_data directory.")

# Fifth Module: remove stop words
print("Starting read the stem files ...")
with open('stop_words.txt', 'r') as f:
    stop_words = f.read().splitlines()
remove_sw.output_file(MODULE_FOUR_FILE_PATH, TOP_FILES_NUM, MODULE_FILE_TYPE, MODULE_FIVE_FILE_PATH, stop_words)
print("Output the handled files in mod5_data directory.")
```

Import five modules in a main.py to test the functionality of each part of the pipeline. All the read and write information like the number of files and the directory and path of the input files and output files is put as parameter in output_file method.



After implementation, each module output the files to their certain directory to easy manually proofread the result of steps in the pipeline.

main.py | remove_sw.py | lowercase_text.py | porter_stemmer.py

```python
import os


def remove_stop_words(path, words):
    with open(path, "r") as f:
        lines = f.read().splitlines()
        start_len = len(lines)
    lines = [line for line in lines if not line in words]
    end_len = len(lines)
    print(f'read {path} successfully.')
    print(f'remove {start_len - end_len} words')
    return lines


def output_file(input_path, file_num, file_ends, output_path, words):
    count = 0
    while count < file_num:
        file_list = os.listdir(input_path)
        file_list = [file for file in file_list if file.endswith(file_ends)]
        for file in file_list[:file_num]:
            stem_list = remove_stop_words(input_path + "/" + file, words)
            with open(output_path + '/' + 'stem_without_sw-' + str(count) + '.txt', "a") as f:
                f.writelines("%s\n" % stem for stem in stem_list)
                print(f'write to {output_path} successfully.')
                count = count + 1
```

Put logs in each function's certain position to check the data flow in each step and easy to debug.

Run: main

```
D:\Users\janel\anaconda3\python.exe D:/python/pythonProject/COMP479Project1/main.py
Starting read the stem files ...
read mod4_data/stem-0.txt successfully.
remove 20345 words
write to mod5_data successfully.
read mod4_data/stem-1.txt successfully.
remove 18934 words
write to mod5_data successfully.
read mod4_data/stem-2.txt successfully.
remove 17773 words
write to mod5_data successfully.
read mod4_data/stem-3.txt successfully.
remove 20747 words
write to mod5_data successfully.
read mod4_data/stem-4.txt successfully.
remove 20900 words
write to mod5_data successfully.
Output the handled files in mod5_data directory.
```

Do the test cases by unit test for each module by each function

```python
import unittest

from read_extract_reuters import read_split_article, extract_article_text
from tokenize_text import tokenize
from lowercase_text import lowercase
from porter_stemmer import porter_stemmer
from remove_sw import remove_stop_words


class TestModuleOne(unittest.TestCase):
    def test_split_article(self):
        """
        Test that it can separate the whole text into each article
        """
        path_0 = "reuters21578/reut2-000.sgm"
        path_21 = "reuters21578/reut2-021.sgm"
        result = read_split_article(path_0)
        self.assertEqual(len(result), 1000)
        result = read_split_article(path_21)
        self.assertEqual(len(result), 578)

    def test_extract_article(self):
        """
        Test that it can extract the body content for each article
        """
        article_2 = """ <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="16319" NEWID="999">
<DATE> 3-MAR-1987 09:16:02.85</DATE>
<TOPICS><D>money-fx</D><D>interest</D></TOPICS>
<PLACES><D>uk</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
```

Test results:

```
D:\Users\janel\anaconda3\python.exe "C:\Users\janel\AppData\Local\JetBrains\PyCharm Community Edition 2022.1.2\plugins\python-ce\helpers\pycharm\_jb_pytest_runner.py" --path "D:/0
Testing started at 9:41 p.m. ...
Launching pytest with arguments D:/0.Information Retrieval/Project/P1/COMP479Project1/test.py --no-header --no-summary -q in D:\0.Information Retrieval\Project\P1\COMP479Project1

============================= test session starts =============================
collecting ... collected 6 items

test.py::TestModuleOne::test_extract_article
test.py::TestModuleOne::test_split_article
test.py::TestModuleTwo::test_tokenize PASSED                     [ 16%]PASSED                      [ 33%]read reuters21578/reut2-000.sgm successfully.
read reuters21578/reut2-021.sgm successfully.

test.py::TestModuleThree::test_lowercase PASSED                            [ 50%]read mod1_data/raw_text-0.txt successfully.

test.py::TestModuleFour::test_porter_stemmer PASSED                         [ 66%]read mod2_data/tokens-0.txt successfully.

test.py::TestModuleFive::test_remove_stop_words PASSED                     [ 83%]read mod3_data/lowercase-0.txt successfully.
PASSED                [100%]read mod4_data/stem-0.txt successfully.
remove 3699 words


============================== 6 passed in 4.39s ==============================
```