Assignment 2

due 11/3/2022

You always promised yourself that you would read *Pride and Prejudice* one day and this is your chance! You can find the entire text of Austen's tome in `pride_and_prejudice.docx`. [1] The book is tough going in places, but fear not!, we can get Python to do most of the heavy lifting and give us a summary (in the guise of an analysis of word frequencies).

Write a Python program named `a2.py` that includes a function `analyze(docfile)` that reads the text contained in the file named `docfile` (presumed to be a .docx file) and calculates the frequency of occurrence of each word appearing therein. The frequency of a word is the ratio of the number of appearances of that word and the total number of words in the document.

This assignment will involve the use of two packages `pyexcel` and `docx` encountered in Lecture 8. You may also use packages `re` and `os`, but no others. These packages should be installed on the machines in WGB 1.10. I will provide guidance on how to install these on your PC separately.

The output should be written into a workbook containing a single two-column work sheet (words and their frequencies) in decreasing order of frequency. Include only those words with a frequency of 0.001 or greater. The name of the work book file should be derived from that of the file containing the text in the following manner:

```
a_book_name.docx    ==>    a_book_name_word_stats.xlsx
```

i.e by tacking "_word_stats" onto the end of the name and adding an ".xlsx" extension.

1. Use the text from the file `pride_and_prejudice.docx` and use the `docx` package to read through it. The documentation for this package is available here:

   `python-docx.readthedocs.io/en/latest/`

   You may find it useful to create a truncated version of this file (say with just the first chapter) to use in the initial testing.

2. Write the output into a workbook file using the `pyexcel` package. The documentation for this package is available here:

   `docs.pyexcel.org/en/latest/`

   The work book should have a single sheet named "Word Frequency Stats". The work sheet should contain two columns, housing the words and their frequencies respectively. The sheet should not contain any column headings.

[1] The text is taken from the Project Gutenberg website (`www.gutenberg.org`) that holds a large collection of out-of-copyright works in various formats. The .docx format used here was derived from the .txt version from that site. The legal notices at the start and end of the downloaded file have been placed in a separate file (`pride_and_prejudice_notice.txt`) for convenience, though any use of the material is bound by the terms outlined in that document.

3. Ignore case: treat “the” and “The” as the same. Treat words sharing a common stem such as “cat” and “cats” as distinct. Similarly “cats”, “cat’s” and “cats’” should all be regarded as distinct words. Some spurious “words” (i.e. non-words) may feature in the text. Don’t worry about these.
4. Recall that Python’s `os.path` package has some useful functions for manipulating file names.

Your submission must conform to the conditions specified below.

1. Code must conform tho the usual programming style strictures.
2. Code must strictly adhere to specified naming. The file must be named `a2.py`. The function must be named `analyze` (note `-yze` rather than `-yse`). The naming of the file, worksheet etc. must be exactly as specified.
3. No extraneous code should be included in the file i.e. code outside the `analyze` function apart from any necessary imports and constant definitions. Any test code should be confined to a separate file (not to to be submitted).
4. Code should not include:
    - Any input statements
    - Any hard-coded file path names

    Assume that the `pride_and_prejudice.docx` is located in the same directory as your `a2.py` file.