INLS 560
Programming for Information Professionals

Assignment 3
Text Analysis

Due Sunday, July 14, end of day
25 points

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

Joan Boone
jpboone@email.unc.edu

Text Analysis

This assignment incorporates all of the fundamental programming concepts: decision and repetition structures, functions, files, lists, dictionaries, and strings.

Objectives

- To demonstrate your understanding of how to apply these concepts, with emphasis on strings, lists, and dictionaries
- To solve a practical problem: analyze unstructured text to determine word frequencies that characterize the content of a document

Overview

Your program will analyze unstructured text by representing it as lists of words (or strings), and then count the number of occurrences of each word in the text. These counts, or frequencies, are stored in a dictionary, and do not include words that appear in the stopwords list.

There are many 3rd party Python libraries that will perform these tasks for you with little code. However, the primary purpose is not just to generate the correct output, but rather to apply Python programming concepts to solve the problem.

Refer to the Text Analysis slides for additional information on the techniques to be used in this assignment.

Text Analysis

Requirements

The `main()` function calls the following functions:

- `create_stopwords_list()` function reads the `stopwords.txt` file, creates a list of words from the contents, and returns the list
- `calculate_word_frequencies(stopwords_list)` function reads the `wizard-of-oz.txt` file, tokenizes the content, calculates the frequency for each word (not including stopwords), and returns the dictionary of word frequencies
- `display_results(word_frequencies)` function displays the total word count, total stopword count, and the list of words with frequencies greater than or equal to 100. NOTE: if your total word/stopword count is off by 1 or 2, that is OK!

In the A3 directory there are 3 files to be used for this assignment:

- A draft version of the program, `text_analysis_draft.py` with comments that provide additional information and requirements
- `wizard-of-oz.txt` contains the plain text version of *The Wonderful Wizard of Oz* (Source: Project Gutenberg)
- `stopwords.txt` contains the stopwords

Text Analysis

Additional requirements

- The expected output is shown on the next slide
- Your program must include appropriate exception handlers
- Your program must not generate any Python errors
- Use descriptive variable names
- Include at least one comment to describe the purpose of the program
- Do not use `break`, `pass`, `continue`, `exit()`, or `quit()` statements
- Do not use Python's `Counter` object, or the file `read` or `readlines` methods

Submitting your assignment

- Name your program: *your_last_name_text_analysis.py*
- Before attaching your program to your email, please rename your program file to *your_last_name_text_analysis.py.**txt***

Email to me at ***jpboone@email.unc.edu***

Expected Output

```
Total word count: 39648
Total stopword count: 25758

Words with frequencies >= 100
WORD                    FREQUENCY
dorothy                  347
scarecrow                219
woodman                  176
lion                     173
oz                       163
great                    142
tin                      140
witch                    125
asked                    114
green                    101
```