Tutorial 3. Computing TF and TF-IDF

This tutorial will guide you through the process of pre-processing text files and compute term frequency (TF) and term frequency–inverse document frequency (TF-IDF).

NOTE: Before you start, you should make sure that Python 2.7 is already installed in your computer (For installation instructions, visit here: http://community.mis.temple.edu/zuyinzheng/pythonworkshop/)

You also need to complete Tutorial 2. Extracting Data from 10-K before starting this tutorial. The text files generated in Tutorial 2 will be used as input files for this tutorial. If you haven't completed Tutorial 2, you can download the text files here:
https://www.dropbox.com/sh/4epko0rs3gp43we/AADB_Vx7vOLlX6g_G5zlc1Fna?dl=0

1 Install the nltk package

You need to have the Python package, nltk (The Natural Language Toolkit), installed in your computer before executing the scripts.

To do so, typing the following command in your command line interface (On Windows it is called "Command Prompt", and on Mac it is called "Terminal"):

```
pip install nltk
```

2 Download 5tfidf.py

Download the Python script 5tfidf.py from the following link. Make sure you save the script in the main folder which contains the subfolder "/txt/" (the subfolder "/txt/" was created in Tutorial 2.)
http://community.mis.temple.edu/zuyinzheng/pythonworkshop/

- The 5tfidf.py performs simple text processing procedures, and compute TF, and TF-IDF values for each text document.

3 Change Working Directory

Find the folder where you have saved the python script in your computer.

Change the working directory to where you put subfolder "/txt/".

To do so:

i) Open the Python script with IDLE.

ii) Find the `os.chdir()` function. The `os.chdir()` function should be in Line 4 of all the four scripts.

iii) Change the parameter in `os.chdir()` function.

For example, I have the subfolder “/txt/” in the folder:
`/Users/alvinzuyinzheng/Dropbox/PythonWorkshop/scripts/`

(Meaning the subfolder path would be
`/Users/alvinzuyinzheng/Dropbox/PythonWorkshop/scripts/txt`)

Therefore, my `os.chdir()` function looks like this:
`os.chdir(/Users/alvinzuyinzheng/Dropbox/PythonWorkshop/scripts/')`

If you have a different folder name, make changes accordingly.

((In Windows, the folder names probably look like this:
`C:\username\Dropbox\python\workshop\Scripts.`
If you are not sure how to find the folder path, check the instructions here: Copy File Folder Path in Mac OS X)

4 Run the 5tfidf.py script

Steps to run the 5tfidf.py script:

i) Double check if you've changed the working directory in the previous step.
i) Open the python script with IDLE.
ii) Click the Run menu and choose “Run Module”.

Once finished, two csv files “tf.csv” and “tfidf.csv” will be created in your working directory. It contains the list of index links extracted from the search result pages.

The files should look like this.
tf.csv: https://www.dropbox.com/s/w6lzxe61zf0494a/tf.csv?dl=0
tfidf.csv: https://www.dropbox.com/s/yjc3lg0g1s6sa4s/tfidf.csv?dl=0