```
import nltk
from nltk.corpus import stopwords
import string
from string import punctuation
import os
from os import listdir
from collections import Counter
```

```
def text_to_tokens(filename):
 '''Opens the input text file and
 returns a list of all of its words.'''
 file = open(filename, 'r')
 text = file.read()
 file.close()
 tokens = text.split()
 return tokens
```

```
# go up one level of the directory to be able to get into the data folder
cd ..
```

```
/Users/kamilapalys/Desktop/school/data450/capstone
```

```
# Ensure we're in the correct directory to be able to navigate to the data folder
pwd
```

```
'/Users/kamilapalys/Desktop/school/data450/capstone'
```

```
# example of using the function

filepath = 'data/text/cnn_trump.txt'
text = text_to_tokens(filepath)
len(text)
```

```
1141
```

```
# how all the text files will be looped through
```

```
directory = 'data/text/'
for filename in os.listdir(directory):
    filepath = str(os.path.join(directory, filename))
    # use preprocessing functions
    # somehow append each list of tokens (each article) to a dataframe
```

```
data/text/nyp_affirmative.txt
data/text/bbc_tanks.txt
data/text/cnn_hamas.txt
data/text/wsj_trump.txt
data/text/wp_affirmative.txt
data/text/bbc_balloon.txt
data/text/wsj_affirmative.txt
data/text/nyt_trump.txt
data/text/wp_balloon.txt
data/text/cnn_pentagon.txt
data/text/cnn_tanks.txt
data/text/bbc_hamas.txt
data/text/nbc_tanks.txt
data/text/nyp_hamas.txt
data/text/.DS_Store
data/text/cnn_balloon.txt
data/text/fox_pentagon.txt
data/text/wp_biden.txt
data/text/nyp_tanks.txt
data/text/nbc_hamas.txt
data/text/abc_balloon.txt
data/text/nyt_pentagon.txt
data/text/fox_tanks.txt
data/text/wp_trump.txt
data/text/fox_hamas.txt
data/text/nyp_balloon.txt
data/text/abc_hamas.txt
data/text/abc_santos.txt
data/text/nyp_santos.txt
data/text/fox_affirmative.txt
data/text/wsj_biden.txt
data/text/nyt_biden.txt
data/text/abc_tanks.txt
data/text/nbc_affirmative.txt
data/text/nyt_hamas.txt
```

```
data/text/wsj_tanks.txt
data/text/wp_pentagon.txt
data/text/nyp_pentagon.txt
data/text/bbc_trump.txt
data/text/nyt_affirmative.txt
data/text/cnn_trump.txt
data/text/abc_biden.txt
data/text/cnn_affirmative.txt
data/text/bbc_affirmative.txt
data/text/wsj_hamas.txt
data/text/wsj_pentagon.txt
data/text/nyt_tanks.txt
data/text/nbc_balloon.txt
data/text/fox_biden.txt
data/text/wsj_balloon.txt
data/text/nbc_pentagon.txt
data/text/bbc_santos.txt
data/text/wsj_santos.txt
data/text/nbc_trump.txt
data/text/nyp_trump.txt
data/text/cnn_santos.txt
data/text/abc_pentagon.txt
data/text/fox_santos.txt
data/text/wp_hamas.txt
data/text/fox_trump.txt
data/text/nbc_biden.txt
data/text/nyp_biden.txt
data/text/nyt_balloon.txt
data/text/wp_tanks.txt
data/text/wp_santos.txt
data/text/bbc_biden.txt
data/text/cnn_biden.txt
data/text/abc_trump.txt
data/text/nyt_santos.txt
data/text/bbc_pentagon.txt
data/text/abc_affirmative.txt
data/text/nbc_santos.txt
data/text/fox_balloon.txt
```