TODO

- Documentation

Demo File

+ Tests

In [1]:

```
# ** IMPORTS

# stop words from nltk for demo purposes
from nltk.corpus import stopwords

# create set of english stop words from nltk
nltk_stop_words = set(stopwords.words('english'))

# import tests
import asserts

# import pipeline

# step 1
from pipeline import read_smg

# step 1.1
from pipeline import segment_documents

# step 1.2
from pipeline import extract_html_symbols

# step 1.3
from pipeline import extract_punctuation
from pipeline import extract_punctuation_keep_digits

# step 2
from pipeline import tokenize_doc_str

# step 3
from pipeline import case_fold_tokens

# step 4
from pipeline import stem_tokens

# step 5
from pipeline import filter_out_stop_words
```

STEP 1: Read Files --> Output Strings

In [2]:

```
# set file path
file_paths  = ['reuters21578/reut2-000.sgm']

# return generator from reader
read_gen = read_smg(file_paths)

# extract first file
file_num, file_content = next(read_gen)

# run tests
```

```
asserts.reader_tests(file_num, file_content)

# sample output
print('file num: ', file_num)
print('file content, first 1000 chars: ', file_content[:1000])
```

```
file num:  00
file content, first 1000 chars:  <!DOCTYPE lewis SYSTEM "lewis.dtd">
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5544" NEWID="1">
<DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>cocoa</D></TOPICS>
<PLACES><D>el-salvador</D><D>usa</D><D>uruguay</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;C T
&#22;&#22;&#1;f0704&#31;reute
u f BC-BAHIA-COCOA-REVIEW   02-26 0105</UNKNOWN>
<TEXT>&#2;
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE>    SALVADOR, Feb 26 - </DATELINE><BODY>Showers continued throughout the week i
n
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review.
    The dry period means the temporao will be late this year.
    Arrivals for the week ended February 22 were 155,221 bags
of 60 kilos making a cumulative total for the season of 5.93
mln against 5.81 at the sa
```

STEP 1.1: Input Files --> Output Documents

In [3]:

```
# return generator from doc segmenter
doc_gen = segment_documents(file_content)

# extract first doc
doc_id, doc_str = next(doc_gen)

# run tests
asserts.segmenter_test(doc_id, doc_str)

# sample output
print('doc_id: ', doc_id)
print('doc_str, first 1000 chars: ', doc_str[:1000])
```

```
doc_id:  1
doc_str, first 1000 chars:  Showers continued throughout the week in
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review.
    The dry period means the temporao will be late this year.
    Arrivals for the week ended February 22 were 155,221 bags
of 60 kilos making a cumulative total for the season of 5.93
mln against 5.81 at the same stage last year. Again it seems
that cocoa delivered earlier on consignment was included in the
arrivals figures.
    Comissaria Smith said there is still some doubt as to how
much old crop cocoa is still available as harvesting has
practically come to an end. With total Bahia crop estimates
around 6.4 mln bags and sales standing at almost 6.2 mln there
are a few hundred thousand bags still in the hands of farmers,
middlemen, exporters and processors.
    There are doubts as to how much of this cocoa would be fit
for export
```

STEP 1.2, 1.3: Input Document --> Output Document (without HTML Symbols, Punctuation, Numbers, or Linebreaks)

In [4]:

```
print('doc_str, last 1000 chars: ', doc_str[-1000:])
```

```
doc_str, last 1000 chars:  o 4,450 dlrs and at
2.27 and 2.28 times New York Sept and Oct/Dec at 4,480 dlrs and
2.27 times New York Dec, Comissaria Smith said.
    Destinations were the U.S., Covertible currency areas,
Uruguay and open ports.
    Cake sales were registered at 785 to 995 dlrs for
March/April, 785 dlrs for May, 753 dlrs for Aug and 0.39 times
New York Dec for Oct/Dec.
    Buyers were the U.S., Argentina, Uruguay and convertible
currency areas.
    Liquor sales were limited with March/April selling at 2,325
and 2,380 dlrs, June/July at 2,375 dlrs and at 1.25 times New
York July, Aug/Sept at 2,400 dlrs and at 1.25 times New York
Sept and Oct/Dec at 1.25 times New York Dec, Comissaria Smith
said.
    Total Bahia sales are currently estimated at 6.13 mln bags
against the 1986/87 crop and 1.06 mln bags against the 1987/88
crop.
    Final figures for the period to February 28 are expected to
be published by the Brazilian Cocoa Trade Commission after
carnival which ends midday on February 27.
 Reuter
&#3;
```

as seen above, html symbol is present at the end, as well as line breaks, punctuation and numbers

In [5]:

```
# remove html symbols
doc_str_no_html = extract_html_symbols(doc_str)

# option to keep digits
doc_str_no_punc = extract_punctuation_keep_digits(doc_str_no_html)

# option to remove digits
doc_str_no_punc_no_dig = extract_punctuation(doc_str_no_html)

print('DOC STR NO PUNC, LAST 1000 CHARS: ', doc_str_no_punc[-1000:])
print('\n')
print('DOC STR NO PUNC NO DIGIT, LAST 1000 CHARS: ', doc_str_no_punc_no_dig[-1000:])
```

```
DOC STR NO PUNC, LAST 1000 CHARS:  t 4400 and 4415 dlrs Aug Sept at 4351 to 4450 dlrs and
at 227 and 228 times New York Sept and Oct Dec at 4480 dlrs and 227 times New York Dec Co
missaria Smith said     Destinations were the US Covertible currency areas Uruguay and op
en ports     Cake sales were registered at 785 to 995 dlrs for March April 785 dlrs for M
ay 753 dlrs for Aug and 039 times New York Dec for Oct Dec     Buyers were the US Argenti
na Uruguay and convertible currency areas     Liquor sales were limited with March April
selling at 2325 and 2380 dlrs June July at 2375 dlrs and at 125 times New York July Aug S
ept at 2400 dlrs and at 125 times New York Sept and Oct Dec at 125 times New York Dec Com
issaria Smith said     Total Bahia sales are currently estimated at 613 mln bags against
the 1986 87 crop and 106 mln bags against the 1987 88 crop     Final figures for the peri
od to February 28 are expected to be published by the Brazilian Cocoa Trade Commission af
ter carnival which ends midday on February 27  Reuter


DOC STR NO PUNC NO DIGIT, LAST 1000 CHARS:  arch April sold at   and  dlrs     April May
butter went at  times New York May June July at  and  dlrs Aug Sept at  to  dlrs and at
and  times New York Sept and Oct Dec at  dlrs and  times New York Dec Comissaria Smith sa
id     Destinations were the US Covertible currency areas Uruguay and open ports     Cake
sales were registered at  to  dlrs for March April  dlrs for May  dlrs for Aug and  times
New York Dec for Oct Dec     Buyers were the US Argentina Uruguay and convertible currenc
y areas     Liquor sales were limited with March April selling at  and  dlrs June July at
dlrs and at  times New York July Aug Sept at  dlrs and at  times New York Sept and Oct De
```

```
c at  times New York Dec Comissaria Smith said     Total Bahia sales are currently estima
ted at  mln bags against the   crop and  mln bags against the   crop     Final figures fo
r the period to February  are expected to be published by the Brazilian Cocoa Trade Commi
ssion after carnival which ends midday on February   Reuter
```

In [6]:

```
# run tests
asserts.extractor_test(doc_str_no_punc_no_dig)
```

STEP 2: Tokenize

In [7]:

```
doc_str = doc_str_no_punc_no_dig

# tokenize document
token_tuples = tokenize_doc_str(doc_id, doc_str)

# print sample of tokens
for token_tuple in token_tuples[:15]:
    print(token_tuple)

# run tests
asserts.tokenizer_test(token_tuples)
```

```
('1', 'Showers')
('1', 'continued')
('1', 'throughout')
('1', 'the')
('1', 'week')
('1', 'in')
('1', 'the')
('1', 'Bahia')
('1', 'cocoa')
('1', 'zone')
('1', 'alleviating')
('1', 'the')
('1', 'drought')
('1', 'since')
('1', 'early')
```

STEP 3: Case Fold

In [8]:

```
# case-fold tokens
token_tuples = case_fold_tokens(token_tuples)

# print sample of case-folded tokens
for token_tuple in token_tuples[:15]:
    print(token_tuple)

# run tests
asserts.case_folder_test(token_tuples)
```

```
('1', 'showers')
('1', 'continued')
('1', 'throughout')
('1', 'the')
('1', 'week')
('1', 'in')
('1', 'the')
('1', 'bahia')
('1', 'cocoa')
('1', 'zone')
('1', 'alleviating')
('1', 'the')
('1', 'drought')
```

```
('1', 'since')
('1', 'early')
```

STEP 4: Stemming

In [9]:

```
# stem tokens
token_tuples = stem_tokens(token_tuples)

# print sample of case-folded tokens
for token_tuple in token_tuples[:15]:
    print(token_tuple)

# run tests
asserts.stemmer_test(token_tuples)
```

```
('1', 'shower')
('1', 'continu')
('1', 'throughout')
('1', 'the')
('1', 'week')
('1', 'in')
('1', 'the')
('1', 'bahia')
('1', 'cocoa')
('1', 'zone')
('1', 'allevi')
('1', 'the')
('1', 'drought')
('1', 'sinc')
('1', 'earli')
```

STEP 5: Stop Words

In [10]:

```
# filter stop words from tokens
token_tuples = filter_out_stop_words(token_tuples, nltk_stop_words)

# print sample of stop word filtered tokens
for token_tuple in token_tuples[:15]:
    print(token_tuple)

asserts.stop_word_test(token_tuples, nltk_stop_words)
```

```
('1', 'shower')
('1', 'continu')
('1', 'throughout')
('1', 'week')
('1', 'bahia')
('1', 'cocoa')
('1', 'zone')
('1', 'allevi')
('1', 'drought')
('1', 'sinc')
('1', 'earli')
('1', 'januari')
('1', 'improv')
('1', 'prospect')
('1', 'come')
```