Overview

- Read labelled data files and build a dataframe with one data file parsed per row

Initialization

In [19]:
```
import pandas as pd
import numpy as np
import matplotlib
import os
import re
import pdfminer as pdfm

from io import StringIO

from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser

import nltk
nltk.download("punkt")

import string
from collections import Counter
```

```
[nltk_data] Downloading package punkt to /Users/emilyng/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Functions

Create list of useless_words

In [2]:
```
nltk_stopwords = nltk.corpus.stopwords.words("english")
# dont_stop = [] # Put any words that we want to keep that are in nltk_stopwords here,
punct_list = list(string.punctuation)
# dont_stop_punct = []

#mick_list = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o'
#             'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O'
#             '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', 'id', '2d', 'one', 'two
#           ]

tmp = nltk_stopwords + punct_list # + mick_list
useless_words = tmp
# useless_words = tmp minus the dont* variables
#useless_words
```

Text Prep Functions

- get text from pdf
- clean the text (remove garbage characters)
- filter out less useful words
- count words

In [3]:
```
def get_text_from_pdf(pdf_path) :
    output_string = StringIO()
    with open(pdf_path, 'rb') as in_file:
        parser = PDFParser(in_file)
        doc = PDFDocument(parser)
        rsrcmgr = PDFResourceManager()
        device = TextConverter(rsrcmgr, output_string, laparams=LAParams())
        interpreter = PDFPageInterpreter(rsrcmgr, device)
        for page in PDFPage.create_pages(doc):
            interpreter.process_page(page)

    return output_string.getvalue()
```

In [29]:
```
# This should be replaced using regular expressions and can be significantly enhanced f
# Almost a placeholder function right now, just removing newlines.
def clean_text(input_string) :
    str1 = input_string.replace(" \n", "")
    str2 = str1.replace("\n", "")
    str3 = str2.replace("\\x0c", "")
    str4 = re.sub('[^A-Za-z0-9 ]+', '', str3)
    str5 = re.sub('\d+', '', str4)
    final_string = str5.lower()
    return final_string
```

In [14]:
```
def filter_words(input_words) :
    filtered_words = []
    for word in input_words :
        append_it = True
        if word in useless_words :
            #print(f"useless word {word}")
            append_it = False
        elif len(word) == 1 :
            #print(f"word length 1 {word}")
            append_it = False
        elif word.isdigit() :
            #print(f"number {word}")
            append_it = False
        elif word[0] == chr(167) :
            #print(f"section symbol {word}")
            append_it = False
        if append_it :
            filtered_words.append(word)
    return filtered_words
```

In [15]:
```
def count_words(word_list) :
    word_counter = Counter(word_list)
```

```
    wc_rev_sort = sorted(word_counter.items(), key=lambda pair: pair[1], reverse=True)
    return wc_rev_sort
```

In [16]:

```
def build_bow2(words) :
    dict = {}
    for word in words :
        append_it = True
        if word in useless_words :
            #print(f"useless word {word}")
            append_it = False
        if len(word) == 1 :
            #print(f"word length 1 {word}")
            append_it = False
        if word.isdigit() :
            #print(f"number {word}")
            append_it = False
        if word[0] == chr(167) :
            #print(f"section symbol {word}")
            append_it = False
        if append_it :
            dict[word] = 1
    return dict
```

Create DataFrame for ML Use

In [31]:

```
# We will populate these series
doc_num = pd.Series([], name='doc_num', dtype='int')
doc_filepath = pd.Series([], name='doc_filepath', dtype='str')
doc_text = pd.Series([], name='doc_text', dtype='str')
text_cleaned = pd.Series([], name='text_cleaned', dtype='str')
nltk_words = pd.Series([], name='nltk_words', dtype='str')
filtered_words = pd.Series([], name='filtered_words', dtype='str')
word_counts = pd.Series([], name='word_counts', dtype='str')
env_label = pd.Series([], name='env_label', dtype='str')
bow = pd.Series([], name='bow', dtype='str')

doc_ctr = 0

# Process the Environmental data
envdir = 'Data/Environmental'
for file_nm in os.listdir(envdir) :
    filepath = envdir + '/' + file_nm
    doc_str = get_text_from_pdf(filepath)
    txt_cln = clean_text(doc_str)
    nltk_wds = nltk.word_tokenize(txt_cln)
    filt_words = filter_words(nltk_wds)
    wc = count_words(filt_words)
    word_counts[doc_ctr] = wc
    bow2 = build_bow2(filt_words)

    # break # for debugging all text cleanups

    doc_num[doc_ctr] = doc_ctr
    doc_filepath[doc_ctr] = filepath
    doc_text[doc_ctr] = doc_str
    text_cleaned[doc_ctr] = txt_cln
```

```
    nltk_words[doc_ctr] = nltk_wds
    filtered_words[doc_ctr] = filt_words
    env_label[doc_ctr] = 'Environmental'
    bow[doc_ctr] = bow2
    doc_ctr += 1

# Process the Non-Environmental Data - same loop, should get into a function later
envdir = 'Data/NonEnvironmental'
for file_nm in os.listdir(envdir) :
    filepath = envdir + '/' + file_nm
    doc_str = get_text_from_pdf(filepath)
    txt_cln = clean_text(doc_str)
    nltk_wds = nltk.word_tokenize(txt_cln)
    filt_words = filter_words(nltk_wds)
    wc = count_words(filt_words)
    word_counts[doc_ctr] = wc
    bow2 = build_bow2(filt_words)

    doc_num[doc_ctr] = doc_ctr
    doc_filepath[doc_ctr] = filepath
    doc_text[doc_ctr] = doc_str
    text_cleaned[doc_ctr] = txt_cln
    nltk_words[doc_ctr] = nltk_wds
    filtered_words[doc_ctr] = filt_words
    env_label[doc_ctr] = 'NonEnvironmental'
    bow[doc_ctr] = bow2
    doc_ctr += 1

# Assemble the final data frame
doc_df = doc_num.to_frame().\
         join(doc_filepath).\
         join(doc_text).\
         join(text_cleaned).\
         join(nltk_words).\
         join(filtered_words).\
         join(word_counts).\
         join(bow).\
         join(env_label)
print(doc_df)
```

```
     doc_num                                 doc_filepath  \
0          0      Data/Environmental/PLAW-104publ70.pdf
1          1     Data/Environmental/PLAW-112publ177.pdf
2          2      Data/Environmental/PLAW-116publ63.pdf
3          3     Data/Environmental/PLAW-110publ288.pdf
4          4     Data/Environmental/PLAW-108publ425.pdf
..       ...                                       ...
134      134   Data/NonEnvironmental/PLAW-114publ38.pdf
135      135  Data/NonEnvironmental/PLAW-115publ281.pdf
136      136  Data/NonEnvironmental/PLAW-115publ280.pdf
137      137   Data/NonEnvironmental/PLAW-116publ52.pdf
138      138  Data/NonEnvironmental/PLAW-116publ107.pdf

                                             doc_text  \
0    PUBLIC LAW 104-70—DEC. 23, 1995\n\n109 STAT. 7...
1    PUBLIC LAW 112-177—SEPT. 28, 2012 \n\n126 STAT...
2    133 STAT. 1120 \n\nPUBLIC LAW 116-63—OCT. 4, 2...
3    PUBLIC LAW 110-288—JULY 29, 2008 \n\nCLEAN BOA...
4    PUBLIC LAW 108-425—NOV. 30, 2004\n\nTIJUANA RI...
..                                                 ...
134  PUBLIC LAW 114-38—JULY 28, 2015 \n\n129 STAT. ...
135  PUBLIC LAW 115-281—DEC. 1, 2018 \n\n132 STAT. ...
```

```
136  132 STAT. 4190 \n\nPUBLIC LAW 115–280—NOV. 29,...
137  133 STAT. 1076 \n\nPUBLIC LAW 116–52—AUG. 23, ...
138  133 STAT. 3292 \n\nPUBLIC LAW 116–107—JAN. 17,...

                                          text_cleaned  \
0    public law dec   stat public law th congressan...
1    public law sept   stat public law th congressa...
2     stat public law oct  public law th congressan...
3    public law july  clean boating act of swalcilb...
4    public law nov  tijuana river valley estuary a...
..                                                 ...
134  public law july   stat public law th congressa...
135  public law dec   stat public law th congressan...
136   stat public law nov  public law th congressan...
137   stat public law aug  public law th congressan...
138   stat public law jan  public law th congressan...

                                            nltk_words  \
0    [public, law, dec, stat, public, law, th, cong...
1    [public, law, sept, stat, public, law, th, con...
2    [stat, public, law, oct, public, law, th, cong...
3    [public, law, july, clean, boating, act, of, s...
4    [public, law, nov, tijuana, river, valley, est...
..                                                 ...
134  [public, law, july, stat, public, law, th, con...
135  [public, law, dec, stat, public, law, th, cong...
136  [stat, public, law, nov, public, law, th, cong...
137  [stat, public, law, aug, public, law, th, cong...
138  [stat, public, law, jan, public, law, th, cong...

                                         filtered_words  \
0    [public, law, dec, stat, public, law, th, cong...
1    [public, law, sept, stat, public, law, th, con...
2    [stat, public, law, oct, public, law, th, cong...
3    [public, law, july, clean, boating, act, swalc...
4    [public, law, nov, tijuana, river, valley, est...
..                                                 ...
134  [public, law, july, stat, public, law, th, con...
135  [public, law, dec, stat, public, law, th, cong...
136  [stat, public, law, nov, public, law, th, cong...
137  [stat, public, law, aug, public, law, th, cong...
138  [stat, public, law, jan, public, law, th, cong...

                                            word_counts  \
0    [(may, 4), (occupancy, 4), (vehicle, 3), (trip...
1    [(new, 367), (application, 239), (use, 204), (...
2    [(water, 19), (state, 12), (revolving, 9), (dr...
3    [(management, 14), (recreational, 12), (admini...
4    [(may, 9), (act, 7), (paragraph, 7), (law, 6),...
..                                                 ...
134  [(business, 21), (small, 12), (act, 11), (loan...
135  [(national, 5), (december, 4), (act, 4), (floo...
136  [(disaster, 3), (public, 2), (law, 2), (nov, 2...
137  [(act, 5), (title, 5), (aug, 4), (paid, 4), (d...
138  [(veterans, 12), (secretary, 9), (grant, 8), (...

                                                   bow         env_label
0    {'public': 1, 'law': 1, 'dec': 1, 'stat': 1, '...     Environmental
1    {'public': 1, 'law': 1, 'sept': 1, 'stat': 1, ...     Environmental
2    {'stat': 1, 'public': 1, 'law': 1, 'oct': 1, '...     Environmental
3    {'public': 1, 'law': 1, 'july': 1, 'clean': 1,...     Environmental
4    {'public': 1, 'law': 1, 'nov': 1, 'tijuana': 1...     Environmental
..                                                 ...               ...
134  {'public': 1, 'law': 1, 'july': 1, 'stat': 1, ...  NonEnvironmental
135  {'public': 1, 'law': 1, 'dec': 1, 'stat': 1, '...  NonEnvironmental
```

```
136  {'stat': 1, 'public': 1, 'law': 1, 'nov': 1, '...  NonEnvironmental
137  {'stat': 1, 'public': 1, 'law': 1, 'aug': 1, '...  NonEnvironmental
138  {'stat': 1, 'public': 1, 'law': 1, 'jan': 1, '...  NonEnvironmental

[139 rows x 9 columns]
```

Validation

In [34]:
```
doc_str
```

Out[34]:
```
'133 STAT. 3292 \n\nPUBLIC LAW 116–107—JAN. 17, 2020 \n\nPublic Law 116–107 \n116th Cong
ress \n\nAn Act \n\nJan. 17, 2020 \n\n[H.R. 2385] \n\nTo permit the Secretary of Veteran
s Affairs to establish a grant program to conduct \ncemetery research and produce ed
ucational materials for the Veterans Legacy \nProgram. \n\nBe it enacted by the
Senate and House of Representatives of \n\nthe United States of America in Congress
assembled, \n\n38 USC 2400 \nnote. \n\nSECTION 1. GRANTS FOR CEMETERY RESEARCH AND
THE PRODUC-\n\nTION OF EDUCATIONAL MATERIALS. \n\n(a) GRANTS AUTHORIZED.— \n\n(1) IN G
ENERAL.—The Secretary of Veterans Affairs may \nestablish a grant program to c
onduct cemetery research and \nproduce educational materials for the Veterans L
egacy Pro-\ngram. \n\n(2) ELIGIBLE RECIPIENTS.—The Secretary may award a \n\ngrant
under this section to any of the following entities: \n(A) An institution of higher lear
ning. \n(B) A local education agency. \n(C) A non-profit entity that the Secretary
determines \n\nhas a demonstrated history of community engagement. \n\n(D) Another rec
ipient the Secretary determines to be \n\nappropriate. \n(3) USE OF FUNDS.—A recipi
ent of a grant under this section \n\nmay use the grant amount to— \n\n(A) conduct res
earch related to national, State, or \n\nTribal veterans' cemeteries; \n\n(B) prod
uce education materials that teach about the \nhistory of veterans interred in
national, State, or Tribal \nveterans' cemeteries; and \n\n(C) promote community enga
gement with the histories \nof veterans interred in national, State, or Tribal v
eterans' \ncemeteries. \n(4) MAXIMUM AMOUNT.—A grant awarded under this section \n\nmay
not exceed $500,000. \n(b) REGULATIONS.—If the Secretary establishes a grant program \nu
nder this section, the Secretary shall prescribe regulations \nregarding— \n\n(1)
the evaluation of applications for grants under the \n\nprogram; and \n\n(2) admi
nistration of the program. \n\n(c) REPORT REQUIRED.—Not later than 2 years after
the Sec-\nretary establishes a grant program under this section, the Secretary \nshall
submit to the committees on Veterans' Affairs of the House \n\nDeterminations.
\n\nEvaluation. \n\nDetermination. \n\n \n\nF\nD\nP\nS\nW\nA\nL\n \nB\nU\nP\n \nh\nt\ni
\n\n \n\nw\nD\nO\nR\nP\n2\n8\n0\nR\n0\n2\n5\nP\nA\nL\n \nn\no\n \ny\nn\na\nb\na\nR\n\nl
\n\nVerDate Sep 11 2014 06:11 Feb 04, 2020 Jkt 099139 PO 00107 Frm 00001 Fmt 6580 Sfmt
6581 E:\\PUBLAW\\PUBL107.116 PUBL107\n\n\x0cPUBLIC LAW 116–107—JAN. 17, 2020 \n\n133 STA
T. 3293 \n\nof Representatives and the Senate a report regarding the determina-\ntion o
f the Secretary whether the grant program is a financially \neffective means to
promote the purposes in subsection (a)(3). \n\n(d) DEFINITIONS.—In this section: \n\n(1)
The term ''Veterans Legacy Program'' means the pro-\ngram of the National Cem
etery Administration that is respon-\nsible for providing engagement and educat
ional tools and \nopportunities to the public regarding the service and sacrif
ice \nof veterans interred in national, State, or Tribal veterans' ceme-\nteries. \n\n
(2) The term ''institution of higher learning'' has the \nmeaning given that
term in section 3452(f) of title 38, United \nStates Code. \n\n(3) The term
''local educational agency'' has the meaning \ngiven that term in section 8101
of the Elementary and Sec-\nondary Education Act of 1965 (20 U.S.C. 7801). \n\nAppro
ved January 17, 2020. \n\nLEGISLATIVE HISTORY—H.R. 2385: \n\nHOUSE REPORTS: No. 116–179
(Comm. on Veterans' Affairs). \nCONGRESSIONAL RECORD, Vol. 165 (2019): \n\nOct. 15, cons
idered and passed House. \nDec. 19, considered and passed Senate. \n\nÆ \n\n \n\nF\nD\nP
\nS\nW\nA\nL\n \nB\nU\nP\n \nh\nt\ni\n\n \n\nw\nD\nO\nR\nP\n2\n8\n0\nR\n0\n2\n5\nP\nA\nL
\n \nn\no\n \ny\nn\na\nb\na\nR\n\nl\n\nVerDate Sep 11 2014 06:11 Feb 04, 2020 Jkt 09913
9 PO 00107 Frm 00002 Fmt 6580 Sfmt 6580 E:\\PUBLAW\\PUBL107.116 PUBL107\n\n\x0c'
```

In [32]:
```
txt_cln
```

Out[32]:
' stat public law jan public law th congressan actjan hr to permit the secretary of ve
terans affairs to establish a grant program to conductcemetery research and produce
educational materials for the veterans legacyprogrambe it enacted by the senat
e and house of representatives ofthe united states of america in congress assembled
usc notesection grants for cemetery research and the production of educational
materialsa grants authorized in generalthe secretary of veterans affairs mayestabl
ish a grant program to conduct cemetery research andproduce educational materi
als for the veterans legacy program eligible recipientsthe secretary may award
agrant under this section to any of the following entitiesa an institution of higher lea
rningb a local education agencyc a nonprofit entity that the secretary determines
has a demonstrated history of community engagementd another recipient the secretary
determines to beappropriate use of fundsa recipient of a grant under this sectionmay u
se the grant amount toa conduct research related to national state ortribal veter
ans cemeteriesb produce education materials that teach about thehistory of vete
rans interred in national state or tribalveterans cemeteries andc promote communit
y engagement with the historiesof veterans interred in national state or tribal
veteranscemeteries maximum amounta grant awarded under this sectionmay not exceed b regu
lationsif the secretary establishes a grant programunder this section the secretary
shall prescribe regulationsregarding the evaluation of applications for grants
under theprogram and administration of the programc report requirednot later than
years after the secretary establishes a grant program under this section the secretar
yshall submit to the committees on veterans affairs of the housedeterminations
evaluationdeterminationfdpswalbuphtiwdorprpalnoynabarlverdate sep feb jkt po frm
fmt sfmt epublawpubl publpublic law jan stat of representatives and the senate a rep
ort regarding the determination of the secretary whether the grant program is a
financiallyeffective means to promote the purposes in subsection ad definitionsin this s
ection the term veterans legacy program means the program of the national ce
metery administration that is responsible for providing engagement and educatio
nal tools andopportunities to the public regarding the service and sacrificeof
veterans interred in national state or tribal veterans cemeteries the term institutio
n of higher learning has themeaning given that term in section f of title
unitedstates code the term local educational agency has the meaninggiven that
term in section of the elementary and secondary education act of usc approved
january legislative historyhr house reports no comm on veterans affairscongressional r
ecord vol oct considered and passed housedec considered and passed senatefdpswalbupht
iwdorprpalnoynabarlverdate sep feb jkt po frm fmt sfmt epublawpubl publ'

And write it out

In [33]:
```
doc_df.to_csv("Data/doc_list1.csv")
```

In [ ]:
```

```