Overview

- Read csv with document data in it
- Split into train/test data
- Train model
- Test model

Initialization

In [167…

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
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")
from nltk.classify import NaiveBayesClassifier

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!
```

Read the document data

In [185…

```
doc_df = pd.read_csv("Data/doc_list1.csv")
doc_df
```

Out[185…

	Unnamed: 0	doc_num	doc_filepath	doc_text	text_cleaned	nltk_words	filt
0	0	0	Data/Environmental/PLAW-104publ70.pdf	PUBLIC LAW 104–70—DEC. 23, 1995\n\n109 STAT. 7...	public law dec stat public law th congressan...	['public', 'law', 'dec', 'stat', 'public', 'la...	['l

	Unnamed: 0	doc_num	doc_filepath	doc_text	text_cleaned	nltk_words	filt
1	1	1	Data/Environmental/PLAW-112publ177.pdf	PUBLIC LAW 112–177—SEPT. 28, 2012 \n\n126 STAT...	public law sept stat public law th congressa...	['public', 'law', 'sept', 'stat', 'public', 'l...	['l
2	2	2	Data/Environmental/PLAW-116publ63.pdf	133 STAT. 1120 \n\nPUBLIC LAW 116–63—OCT. 4, 2...	stat public law oct public law th congressan...	['stat', 'public', 'law', 'oct', 'public', 'la...	['s
3	3	3	Data/Environmental/PLAW-110publ288.pdf	PUBLIC LAW 110–288—JULY 29, 2008 \n\nCLEAN BOA...	public law july clean boating act of swalcilb...	['public', 'law', 'july', 'clean', 'boating', ...	['l
4	4	4	Data/Environmental/PLAW-108publ425.pdf	PUBLIC LAW 108–425—NOV. 30, 2004\n\nTIJUANA RI...	public law nov tijuana river valley estuary a...	['public', 'law', 'nov', 'tijuana', 'river', '...	['l 'n
...	...	...	...	...	...	...	
134	134	134	Data/NonEnvironmental/PLAW-114publ38.pdf	PUBLIC LAW 114–38—JULY 28, 2015 \n\n129 STAT. ...	public law july stat public law th congressa...	['public', 'law', 'july', 'stat', 'public', 'l...	['l
135	135	135	Data/NonEnvironmental/PLAW-115publ281.pdf	PUBLIC LAW 115–281—DEC. 1, 2018 \n\n132 STAT. ...	public law dec stat public law th congressan...	['public', 'law', 'dec', 'stat', 'public', 'la...	['l
136	136	136	Data/NonEnvironmental/PLAW-115publ280.pdf	132 STAT. 4190 \n\nPUBLIC LAW 115–280—NOV. 29,...	stat public law nov public law th congressan...	['stat', 'public', 'law', 'nov', 'public', 'la...	['s
137	137	137	Data/NonEnvironmental/PLAW-116publ52.pdf	133 STAT. 1076 \n\nPUBLIC LAW 116–52—AUG. 23, ...	stat public law aug public law th congressan...	['stat', 'public', 'law', 'aug', 'public', 'la...	['s
138	138	138	Data/NonEnvironmental/PLAW-116publ107.pdf	133 STAT. 3292 \n\nPUBLIC LAW 116–107—JAN. 17,...	stat public law jan public law th congressan...	['stat', 'public', 'law', 'jan', 'public', 'la...	['s

139 rows × 10 columns

In [186…
```
doc_df.shape
```

Out[186… (139, 10)

Split Training vs Testing Data

In [187…
```
train_filter = doc_df['doc_num']%2 == 0
test_filter = ~train_filter
#train_filter
#test_filter
training_df = doc_df.loc[train_filter]
testing_df = doc_df.loc[test_filter]
```

In [188…
```
training_df.shape
```

Out[188… (70, 10)

In [189…
```
testing_df.head()
```

Out[189…

	Unnamed: 0	doc_num	doc_filepath	doc_text	text_cleaned
1	1	1	Data/Environmental/PLAW-112publ177.pdf	PUBLIC LAW 112–177—SEPT. 28, 2012 \n\n126 STAT...	public law sept stat public law th congressa..
3	3	3	Data/Environmental/PLAW-110publ288.pdf	PUBLIC LAW 110–288—JULY 29, 2008 \n\nCLEAN BOA...	public law july clean boating act of swalcilb..
5	5	5	Data/Environmental/PLAW-105publ156.pdf	(cid:80) (cid:85) (cid:66) (cid:76) (cid:73) (cid:6...	cidcidcidcidcidcidcidcidcidcidcidcidcidcidc..
7	7	7	Data/Environmental/PLAW-116publ62.pdf	133 STAT. 1118 \n\nPUBLIC LAW 116–62—OCT. 4, 2...	stat public law oct oct s alaska remotegene..

	Unnamed: 0	doc_num	doc_filepath	doc_text	text_cleaned
9	9	9	Data/Environmental/PLAW-111publ191.pdf	124 STAT. 1278 \n\nPUBLIC LAW 111–191—JUNE 15,...	stat public law june public law th congressa..

In [190…
```
training_df.head()
```

Out[190…

	Unnamed: 0	doc_num	doc_filepath	doc_text	text_cleaned	nltk_words	filtered_w
0	0	0	Data/Environmental/PLAW-104publ70.pdf	PUBLIC LAW 104–70—DEC. 23, 1995\n\n109 STAT. 7...	public law dec stat public law th congressan...	['public', 'law', 'dec', 'stat', 'public', 'la...	['public', 'dec', 'public
2	2	2	Data/Environmental/PLAW-116publ63.pdf	133 STAT. 1120 \n\nPUBLIC LAW 116–63—OCT. 4, 2...	stat public law oct public law th congressan...	['stat', 'public', 'law', 'oct', 'public', 'la...	['stat', 'pu 'law', 'public
4	4	4	Data/Environmental/PLAW-108publ425.pdf	PUBLIC LAW 108–425—NOV. 30, 2004\n\nTIJUANA RI...	public law nov tijuana river valley estuary a...	['public', 'law', 'nov', 'tijuana', 'river', '...	['public', 'nov', 'tiju 'rive
6	6	6	Data/Environmental/PLAW-114publ182.pdf	PUBLIC LAW 114–182—JUNE 22, 2016 \n\nFRANK R. ...	public law june frank r lautenberg chemical s...	['public', 'law', 'june', 'frank', 'r', 'laute...	['public', 'june', 'f 'lautenb
8	8	8	Data/Environmental/PLAW-111publ378.pdf	124 STAT. 4128 \n\nPUBLIC LAW 111–378—JAN. 4, ...	stat public law jan public law th congressan...	['stat', 'public', 'law', 'jan', 'public', 'la...	['stat', 'pu 'law', 'public

Train the Model

In [10]:
```
def create_nb_input(doc_df, label) :
    ctr = 0
```

```
    for word_list in doc_df['bow'].tolist() :
#         print(type(word_list))
#         print(word_list)
#         break
        if ctr == 0 :
            nb_input = [(word_list, label)]
        else :
            nb_input.append((word_list, label))
        ctr += 1
    return nb_input
```

In [11]:

```
filter = training_df['env_label'] == "Environmental"
env_training_df = training_df.loc[filter]
notenv_training_df = training_df.loc[~filter]
```

In [12]:

```
nb_env_input = create_nb_input(env_training_df, 'Environmental')
nb_notenv_input = create_nb_input(notenv_training_df, 'NotEnvironmental')
print(nb_env_input[0][0])
```

```
{'public': 1, 'law': 1, '112-177—sept': 1, 'stat': 1, '1327public': 1, '112-177112th':
1, 'congressan': 1, 'actto': 1, 'reauthorize': 1, 'federal': 1, 'insecticide': 1, 'fungi
cide': 1, 'rodenticide': 1, 'act.be': 1, 'enacted': 1, 'senate': 1, 'house': 1, 'represe
ntatives': 1, 'ofthe': 1, 'united': 1, 'states': 1, 'america': 1, 'congress': 1, 'assemb
led': 1, 'section': 1, '1.': 1, 'short': 1, 'title.this': 1, 'act': 1, 'may': 1, 'cite
d': 1, 'pesticide': 1, 'registration': 1, 'improve-ment': 1, 'extension': 1, '.sept': 1,
's.': 1, 'pesticideregistrationimprovementextension': 1, 'actof': 1, '2012.7': 1, 'usc':
1, 'note.sec': 1, '2.': 1, 'improvement': 1, 'maintenance': 1, 'fees.—': 1, 'fees.—secti
on': 1, 'u.s.c': 1, '136a-1': 1, 'amended—': 1, 'paragraph': 1, 'subparagraph': 1, 'stri
king': 1, 'aggregateamount': 1, 'follows': 1, 'end': 1, 'inserting': 1, 'aggregate': 1,
'amount': 1, '27,800,000': 1, 'fiscal': 1, 'years': 1, 'through2017.': 1, 'ii': 1, 'clau
se': 1, 'shall': 1, 'andall': 1, 'thatfollows': 1, 'semicolon': 1, 'andinserting': 1, '1
15,500': 1, 'years2013': 1, 'period': 1, '184,800': 1, '2013through': 1, '2017.': 1, 'ii
i': 1, 'subclause': 1, '70,600': 1, '122,100': 1, 'iv': 1, 'human': 1, 'redesignating':
1, 'subparagraphs': 1, 'respectively': 1, 'humans': 1, 'andinsertingswalcilbuphtiwdorp1n
vx7thksdnoesuarkdverdate': 1, 'mar': 1, '12:50': 1, 'oct': 1, 'jkt': 1, 'po': 1, 'frm':
1, 'fmt': 1, 'sfmt': 1, '\\publaw\\publ177.112': 1, 'publ177': 1, '1328public': 1, 'vi':
1, 'fol-lowing': 1, 'fee': 1, 'reduction': 1, 'certain': 1, 'small': 1, 'businesses.—':
1, 'definition.—in': 1, 'term': 1, 'qualified': 1, 'business': 1, 'entity': 1, 'means':
1, 'corporation': 1, 'partnership': 1, 'unincorporated': 1, 'that—': 1, 'fewer': 1, 'emp
loyees': 1, '3-year': 1, 'prior': 1, 'mostrecent': 1, 'billing': 1, 'cycle': 1, 'aver-ag
e': 1, 'annual': 1, 'global': 1, 'gross': 1, 'revenue': 1, 'sourcesthat': 1, 'exceed':
1, '10,000,000': 1, 'holds': 1, 'registra-tions': 1, 'paragraph.': 1, 'waiver.—except':
1, 'provided': 1, 'administrator': 1, 'waive': 1, 'percent': 1, 'feeunder': 1, 'applicab
le': 1, 'first': 1, 'registrationof': 1, 'para-graph.': 1, 'limitation.—the': 1, 'notgra
nt': 1, 'waiver': 1, 'smallbusiness': 1, 'determines': 1, 'thatthe': 1, 'formed': 1, 'ma
nipulated': 1, 'primarilyfor': 1, 'purpose': 1, 'qualifying': 1, 'waiver.': 1, 'vii': 1,
'redesignated': 1, 'paragraphs': 1, 'para-graphs': 1, 'conforming': 1, 'amendments.—':
1, 'subsection': 1, 'sub-section': 1, 'sentence': 1, 'third': 1, 'sixth': 1, 'sentence
s': 1, 'bystriking': 1, 'place': 1, 'appearsand': 1, 'fun-gicide': 1, '136w-8': 1, 'isam
ended—': 1, 'itappears': 1, 'clauses': 1, 'bb': 1, 'eachplace': 1, 'appears': 1, '—swalc
ilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'applicable.': 1, 'revenues': 1, 'rev-': 1,
'andenue': 1, 'prohibition': 1, 'tolerance': 1, 'fees.—sec-tion': 1, 'food': 1, 'drug':
1, 'cosmetic': 1, '346a': 1, 'amended': 1, 'september': 1, '30,2012': 1, 'reregistratio
n': 1, 'expedited': 1, 'processing': 1, 'fund.—': 1, 'source': 1, 'use.—section': 1, 'fe
deralinsecticide': 1, 'enhance': 1, 'information': 1, 'sys-tems': 1, 'capabilities': 1,
'improve': 1, 'tracking': 1, 'pesticideregistration': 1, 'decisions': 1, 'placeit': 1,
'—reregistration': 1, 'andpriated': 1, 'funds': 1, 'offset': 1, 'costs': 1, 'portion':
1, 'appro-': 1, 'similar': 1, 'applications.—section': 1, 'androdenticide': 1, 'matter':
1, 'preceding': 1, '1/8': 1, '1/7': 1, '1/9': 1, 'new': 1, 'application': 1, 'enhancemen
ts': 1, 'technology': 1, 'review': 1, 'applica-tions.—section': 1, 'following': 1, 'syst
```

emsfor': 1, 'applications.—': 1, 'general.—for': 1, 'use': 1, '800,000': 1, 'amounts': 1, 'made': 1, 'available': 1, 'adminis-trator': 1, 'fundfor': 1, 'activities': 1, 'described': 1, 'activities.—the': 1, 'amountsmade': 1, 'expeditedprocessing': 1, 'fund': 1, 'systemscapabilities': 1, 'office': 1, 'programs': 1, 'enhancetracking': 1, 'shallinclude—': 1, 'electronic': 1, 'of—': 1, 'submissions': 1, 'status': 1, 'conditional': 1, 'registrations': 1, 'enhancing': 1, 'databaseinformationregarding': 1, 'endangered': 1, 'species': 1, 'assessments': 1, 'registra-tion': 1, 'forswalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1330public': 1, 'implementing': 1, 'capability': 1, 'electronicallyreview': 1, 'labels': 1, 'submitted': 1, 'actions': 1, 'acquiring': 1, 'capabilityto': 1, 'electronically': 1, 'assess': 1, 'evaluate': 1, 'confidential': 1, 'state-ments': 1, 'formula': 1, 'actions.': 1, 'asredesignated': 1, 'carry': 1, 'outthe': 1, 'goals': 1, 'established': 1, 'purposes': 1, 'establishedunder': 1, 'service': 1, 'insecti-cide': 1, 'schedule': 1, 'covered': 1, 'applications': 1, 'fees.—subject': 1, 'scheduleof': 1, 'correspondingregistration': 1, 'fees': 1, 'table': 1, 'division': 1, 'activeingredientsepano.newcrno.actiondecisionreviewtime': 1, 'months': 1, 'registra-tionservice': 1, 'r0101new': 1, 'active': 1, 'ingredient,24569,221r020218569,221r040318419,502food': 1, 'ingredient': 1, 'reducedrisk': 1, 'experi-mental': 1, 'permitapplication': 1, 'establishtemporary': 1, 'ap-plication': 1, 'credit': 1, 'feetoward': 1, 'in-gredient': 1, 'applicationthat': 1, 'swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'activeingredients—continuedepano.newcrno.r0604decisionreviewtime': 1, '21395,467r070516395,467r090616293,596actionnew': 1, 'non-food': 1, 'outdoor': 1, 'out-door': 1, 'reduced': 1, 'risk': 1, 'experimentaluse': 1, 'permit': 1, 'applica-tion': 1, 'be-fore': 1, 'forregistration': 1, 'credit45': 1, 'towardnew': 1, 'indoor': 1, 'experimental': 1, 'usepermit': 1, 'r110720219,949r120814219,949r121918165,375swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1332public': 1, 'activeingredients—continuedepano.newcrno.actiondecisionreviewtime': 1, 'r12210': 1, 'enriched': 1, 'isomer': 1, 'of18287,643r12318427,991registered': 1, 'mixed-iso-mer': 1, 'seed': 1, 'treatment': 1, 'includes': 1, 'agriculturaland': 1, 'non-agriculturalseeds': 1, 'residues': 1, 'notexpected': 1, 'raw': 1, 'agri-cultural': 1, 'commodities': 1, 'experimental': 1, 'per-mit': 1, 'sub-mitted': 1, 'to-ward': 1, 'r125new12': 1, 'ingredient,16293,596extension': 1, 'decision': 1, 'time': 1, 'would': 1, 'otherwise': 1, 'satur-day': 1, 'sunday': 1, 'holiday': 1, 'extended': 1, 'nextbusiness': 1, 'day.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'requests': 1, 'uses': 1, 'and/or': 1, 'nonfood': 1, 'contained': 1, 'inany': 1, 'arecovered': 1, 'base': 1, 'fooduse': 1, 'retain': 1, 'asthe': 1, 'must': 1, 'received': 1, 'agency': 1, 'one': 1, 'package': 1, 'feefor': 1, 'category': 1, 'covers': 1, 'maximum': 1, 'five': 1, 'products': 1, 'additional': 1, 'product': 1, 'inertapproval': 1, 'applicationpackage': 1, 'subject': 1, 'reg-istration': 1, 'inert': 1, 'approval': 1, 'allsuch': 1, 'associated': 1, 'together': 1, 'besubject': 1, 're-view': 1, 'case': 1, 'untilthat': 1, 'approved': 1, 'subsequent': 1, 'applicationfor': 1, 'another': 1, 'containing': 1, 'oran': 1, 'amendment': 1, 'proposed': 1, 'labeling': 1, 'deemed': 1, 'ac-tive': 1, 'feeand': 1, 'ofa': 1, 'useswill': 1, 'reviewtime': 1, 'neither': 1, 're-quested': 1, 'required': 1, 'ap-plicant': 1, 'applicant': 1, 'initiative': 1, 'support': 1, 'aftercompletion': 1, 'technical': 1, 'deficiency': 1, 'screening': 1, 'itselfa': 1, 'assessed': 1, 'fullregistration': 1, 'action': 1, 'involves': 1, 'amendedlabel': 1, 'date': 1, 'theagency': 1, 'provide': 1, 'draft': 1, 'accepted': 1, 'label': 1, 'includ-ing': 1, 'changes': 1, 'differ': 1, 'applicant-submitted': 1, 'relevant': 1, 'supporting': 1, 'data': 1, 'reviewed': 1, 'notify': 1, 'ei-ther': 1, 'agrees': 1, 'terms': 1, 'accept-ed': 1, 'issuedas': 1, 'final': 1, 'agency-stamped': 1, 'agree': 1, 'toone': 1, 'bythe': 1, 'resolve': 1, 'dif-ference': 1, 'withdraws': 1, 'without': 1, 'prejudice': 1, 'forsubsequent': 1, 'resubmission': 1, 'forfeits': 1, 'registrationservice': 1, 'cases': 1, 'upto': 1, 'calendar': 1, 'days': 1, 'reach': 1, 'agreement': 1, 'thefinal': 1, 'agency-accepted': 1, 'toall': 1, 'including': 1, 'upon': 1, 'reso-lution': 1, 'differences': 1, 'acceptedfinal': 1, 'registrant': 1, 'within': 1, 'daysfollowing': 1, 'written': 1, 'confirmation': 1, 'ofagreement': 1, 'agency.labeling.deadlines.notification.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1334public': 1, 'usesepano.newcrno.actiondecisionreviewtime': 1, 'r13013first': 1, '21173,644r14014': 1, 'in-1540,518r15015first': 1, 'r16016first': 1, '2116239,684239,684r17017': 1, '1559,976r175new18': 1, 'uses1059,976food/food': 1, 'handling': 1, 'door': 1, 'food/food': 1, 'han-dling': 1, 'cropgroup': 1, 'resulting': 1, 'fromthe': 1, 'conversion': 1, 'ex-isting': 1, 'ormore': 1, 'revised': 1, 'cropgroups': 1, 'duced': 1, 'r18019': 1, 're-1059,976r190r20020': 1, '6or': 1, 'inone': 1, 're-duced': 1, '15359,85610359,856swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'uses—continuedepano.newcrno.actiondecisionreviewtime': 1, 'r21022': 1, 'ex-1244,431r22023': 1, 'ex-617,993perimental': 1, 'es-tablish': 1, 'temporary': 1, 'tol-erance': 1, 'perimental': 1, 'cropdestruct': 1, 'basis': 1, 'nocredit': 1, 'toward': 1, 'newuse': 1, 'credittoward': 1, 'mit': 1, 'whichrequires': 1, 'changesto': 1, 'non-crop': 1, 'destructbasis': 1, 'r23024': 1, 'r24025': 1, '151023,96923,969r25026': 1,

'non-617,993r251new27': 1, 'per-817,993r26028': 1, 'in-door': 1, '1211,577swalcilbuphtiw
dorp1nvx7thksdnoesuarkdverdate': 1, '1336public': 1, 'r27029': 1, 'in-11,577r2718,82096r
2731245,754door': 1, 'seedtreatment': 1, 'limiteduptake': 1, 'crops': 1, 'withestablishe
d': 1, 'toler-ances': 1, 'e.g.': 1, 'soil': 1, 'orfoliar': 1, 'in-cludes': 1, 'lim-ite
d': 1, 'uptake': 1, 'rawagricultural': 1, 'commod-ities': 1, 'cropswith': 1, 'tol-erance
s': 1, 'soilor': 1, 'foliar': 1, 'and/ornon-food': 1, 'r27432': 1, 'seed12274,523extensi
on': 1, '1338public': 1, 'add': 1, 'registeredproduct': 1, 'allitems': 1, 'onepackage':
1, 'newproduct': 1, 'however': 1, 'applicationonly': 1, 'proposes': 1, 'register': 1, 't
hereare': 1, 'amendments': 1, 'associ-ated': 1, 'thenew': 1, 'subsequentto': 1, 'submiss
ion': 1, 'conclusionof': 1, 'separate': 1, 'new-use': 1, 'sepa-rate': 1, 'and/oroutdoo
r': 1, 'appropriate': 1, 'feeis': 1, 'due': 1, 'type': 1, 'longest': 1, 'applies': 1, 'r
equested': 1, 'application.any': 1, 'ini-tiative': 1, 'completion': 1, 'technicaldeficie
ncy': 1, 'screen': 1, 'full': 1, 'forthe': 1, 'application.': 1, 'import': 1, 'andothe
r': 1, 'tolerancesepano.newcrno.actionr28033': 1, 'establish': 1, 'toler-ance': 1, 'r290
34': 1, 'foodusedecisionreviewtime': 1, '21289,4071557,882r29135': 1, 'toler-15347,288an
ces': 1, 'fooduses': 1, 'cropssubmitted': 1, 'pe-titionswalcilbuphtiwdorp1nvx7thksdnoesu
arkdverdate': 1, 'tolerances—continuedepano.newcrno.actiondecisionreviewtime': 1, 'r2923
6': 1, 'amend': 1, 'established1141,124r293r2941248,51012291,060r2951559,976tolerance':
1, 'de-crease': 1, 'increase': 1, 'domestic': 1, 'applicant-initiated37': 1, 'inadverten
t': 1, 'resi-dues': 1, 'crop': 1, 'ap-plicant-initiated38': 1, 'tolerances': 1, 'forinad
vertent': 1, 'appli-cation': 1, 'applicant-ini-tiated39': 1, 'onerotational': 1, 're-spo
nse': 1, 'specificrotational': 1, 'applicant-ini-tiated40': 1, 'forresidues': 1, 'rotati
onalcrops': 1, 'response': 1, 'aspecific': 1, 'rotationalcrop': 1, 'petition': 1, 'submi
ttedin': 1, 'applicant-initiatedr29615359,856swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdat
e': 1, '1340public': 1, 'es-11246,74442': 1, 'established1353,120r297newr298newr299newta
blished': 1, 'decrease': 1, 'in-crease': 1, 'peti-tion': 1, 'im-port': 1, 'applicant-ini
ti-atedtolerance': 1, 'amend-ed': 1, 'requiringscience': 1, 'ad-dition': 1, 'appli-cant-
initiated': 1, 'tablished': 1, 'orimport': 1, 'ofamended': 1, 're-quiring': 1, 'scienc
e': 1, 'addition': 1, 'tothose': 1, 'withthe': 1, 'applicant-initi-ated': 1, 'es-13258,7
40extension': 1, 'agency.labeling.deadlines.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdat
e': 1, '1342public': 1, 'productsepano.newcrno.actiondecisionreviewtime': 1, 'r30044':
1, 'similar41,434combination': 1, 'already': 1, 'registered': 1, 'identical': 1, 'orsubs
tantially': 1, 'similarin': 1, 'composition': 1, 'anduse': 1, 'registeredsource': 1, 'ac
ute': 1, 'tox-icity': 1, 'efficacy': 1, 'crp–': 1, 'chem-istry': 1, 'cite-alldata': 1,
'citation': 1, 'se-lective': 1, 'citationwhere': 1, 'ownsall': 1, 'orapplicant': 1, 'sub
mitsspecific': 1, 'authorizationletter': 1, 'dataowner': 1, 'alsoincludes': 1, 're-packa
ge': 1, 'registeredend-use': 1, 'manufac-turing-use': 1, 'productthat': 1, 'requires':
1, 'datasubmission': 1, 'datamatrix': 1, 'products—continuedepano.newcrno.actiondecision
reviewtime': 1, 'r30145': 1, 'similar41,720combination': 1, 'selectivedata': 1, 'fordat
a': 1, 'productchemistry': 1, 'and/oracute': 1, 'toxicity': 1, 'and/orpublic': 1, 'healt
h': 1, 'pest': 1, 'ef-ficacy': 1, 'appli-cant': 1, 'allrequired': 1, 'anddoes': 1, 'spe-
cific': 1, '1344public': 1, 'r31046': 1, 'end-use': 1, 'manu-74,807facturing-use': 1, 'p
roductwith': 1, 'includesproducts': 1, 'containingtwo': 1, 'reg-istered': 1, 'ingredi-en
ts': 1, 'previously': 1, 'com-bined': 1, 're-quires': 1, 'datapackage': 1, 'rdonly': 1,
'dataand/or': 1, 'waivers': 1, 'ofdata': 1, 'chemistryand/or∑': 1, 'and/or∑': 1, 'chil
d': 1, 'resistant': 1, 'pack-aging': 1, 'r314new47': 1, 'product86,009containing': 1, 't
wo': 1, 'ingredientsnever': 1, 'com-bination': 1, 'formu-lated': 1, 'iden-tical': 1, 'su
bstantiallysimilar': 1, 'labelsof': 1, 'currently': 1, 'productswhich': 1, 'separately':
1, 'con-tain': 1, 'respectivecomponent': 1, 'in-gredients': 1, 'requiresreview': 1, 'pac
k-age': 1, 'rd': 1, 'and/orwaivers': 1, 'foronly': 1, '1346public': 1, 'products—continu
edepano.newcrno.actionr315newdecisionreviewtime': 1, '98,00048': 1, 'non-foodanimal': 1,
'withsubmission': 1, 'target': 1, 'animalsafety': 1, 'studies': 1, 'animal': 1, 'safet
y': 1, 'studiesand/or∑': 1, 'newphysical': 1, 'form': 1, 'inscience': 1, 'divisions': 1,
'amanufacturing-useproduct': 1, 'r3201211,996r33150': 1, 'repack': 1, 'of32,294swalcilbu
phtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'r33251': 1, 'manufacturing-use24256,883produc
t': 1, 'registeredactive': 1, 'un-registered': 1, 'ofactive': 1, 'com-pletely': 1, 'gene
ricdata': 1, 'rdand': 1, 'withunregistered': 1, 'sourceof': 1, 'datareview': 1, 'physica
lform': 1, 'etc': 1, 'cite-all': 1, 'orselective': 1, 'r333new52': 1, 'mup': 1, 'or1017,
993swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1348public': 1, 'r334new53': 1, 'o
r1117,993end': 1, 'ingre-dient': 1, 'requiresscience': 1, 'physical': 1, 'selective': 1,
'ci-tation': 1, 'extension.labeling.deadlines.notification': 1, 'day': 1, 'using': 1, 'y
et': 1, 'pending': 1, 'considered': 1, 'unregistered': 1, 'activeingredient': 1, 'agenc
y.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'toregistrationepano.newcrno.actiond
ecisionreviewtime': 1, 'r34054': 1, 'requiring43,617r345new55': 1, 'amending': 1, 'ani-7
8,000r35056': 1, 'requiring911,996data': 1, 'withinrd': 1, 'toprecautionary': 1, 'labels

tatements': 1, 'mal': 1, 'oftarget': 1, 'safetydata': 1, 'rei': 1, 'orppe': 1, 'phi': 1, 'userate': 1, 'number': 1, 'ofapplications': 1, 'addaerial': 1, 'ormodify': 1, 'gw/sw': 1, 'advi-sory': 1, 'statement': 1, 'unregisteredsource': 1, 'r351newr352new57': 1, 'adding': 1, 'a811,996811,99658': 1, 'al-ready': 1, 'method': 1, 'ofsupport': 1, 'notapply': 1, 'applicantowns': 1, '1350public': 1, 'toregistration—continuedextension.labeling.deadlines.notification.epano.newcrno.r371action59': 1, 'include': 1, 'ex-tending': 1, 'stime': 1, 'decisionreviewtime': 1, '69,151': 1, 'day.fast-track': 1, 'epa-initiated': 1, 'charged': 1, 'registrant-initiated': 1, 'fast-track': 1, 'amendmentsare': 1, 'completed': 1, 'timelines': 1, 'specified': 1, 'fifra': 1, 'section3': 1, 'reg-istrant-initiatedtheantimicrobials': 1, 'timelinesspecified': 1, 'initiated': 1, 'no-tification': 1, 'pr': 1, 'notices': 1, 'notice': 1, '98–10': 1, 'continueunder': 1, 'serv-ice': 1, 'requiring': 1, 'aresubject': 1, 'fees.amendmentshandledby': 1, 'actionsepano.newcrno.actiondecisionreviewtime': 1, 'r12460': 1, 'ruling': 1, 'on62,294r27261': 1, 'study': 1, 'pro-32,294preapplication': 1, 'studywaivers': 1, 'applicant-initiatedtocol': 1, 'excludes': 1, 'dart': 1, 'pre-registration': 1, 'con-ference': 1, 'rapid': 1, 'dntprotocol': 1, 'pro-tocol': 1, 'needing': 1, 'hsrbreviewr275newr37062': 1, 'rebuttal': 1, 're-viewed': 1, 'protocol': 1, 'initiated63': 1, 'cancer': 1, 'reassessment': 1, 'applicant-initiated32,294181 79,818': 1, 'day.extension.': 1, 'antimicrobials': 1, 'a3806424104,187food': 1, 'exemption': 1, 'a39065food': 1, 'tol-24173,644erance': 1, '1352public': 1, 'activeingredients—continuedepano.newcrno.actiona40066': 1, 'mm': 1, '1886,823a41067': 1, '21173,644uses': 1, 'thanfifra': 1, 'a42068': 1, '1857,882a43069': 1, '2086,823a43170': 1, '1260,638uses': 1, 'low-risk': 1, 'low-toxicityfood-grade': 1, 'efficacytesting': 1, 'publichealth': 1, 'claims': 1, 're-quired': 1, 'glpand': 1, 'dis/tss': 1, 'ad-approvedstudy': 1, 'agency.labeling.deadline.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1354public': 1, 'a440712128,942first': 1, 'establishtolerance': 1, 'a45072first': 1, 'ex-emption': 1, '211586,82311,5771528,94275': 1, '15173,652a48076': 1, 'non-17,36577': 1, 'non-food104,190establish': 1, 'fifra§2': 1, '99a460a470a471newa481newa49078': 1, 'usesother': 1, '1528,942swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'uses—continueda50080': 1, 'non-11,577epano.newcrno.actiona491new79': 1, 'otherthan': 1, 'a501newa510a511newfood': 1, '15173,6529969,4621211,5771269,462': 1, 'day.extension.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1356public': 1, 'epa': 1, 'rules': 1, 'newly': 1, 'require': 1, 'clearanceunder': 1, 'ffdca': 1, 'anti-microbial': 1, 'subjectto': 1, 'clearance': 1, 'clearanceof': 1, 'effective': 1, 'therule': 1, 'agency.time': 1, 'period.labeling.deadlines.notification.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'application.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1358public': 1, 'productsand': 1, 'amendmentsepano.newcrno.a530actiondecisionreviewtime': 1, '41,15984': 1, 'identicalor': 1, 'substantially': 1, 'simi-lar': 1, 'compositionand': 1, 'nodata': 1, 'onlyproduct': 1, 'chemistrydata': 1, 'whenapplicant': 1, 'owns': 1, 'submits': 1, 'specificauthorization': 1, 'letterfor': 1, 'owner': 1, 'cat-egory': 1, 'also': 1, 'includes100': 1, 'ofregistered': 1, 'ormanufacturing-useproduct': 1, 'requiresno': 1, 'submissionnor': 1, 'matrix': 1, 'amendments—continuedepano.newcrno.a531actiondecisionreviewtime': 1, '41,65485': 1, 'selec-tive': 1, 'citationonly': 1, 'prod-uct': 1, 'chemistry': 1, 'datacitation': 1, 'except': 1, 'forproduct': 1, 'a53254,631a54087': 1, 'usesonly': 1, '54,631swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1360public': 1, 'amendments—continuedepano.newcrno.actiondecisionreviewtime': 1, 'a55088': 1, '74,631a56089': 1, 'manufacturing-use1217,365uses': 1, 'non-fqpa': 1, '493,47411,996a57090label': 1, 'a572new91': 1, 'amend-ment': 1, 'as-sessment': 1, 'sciencebranch': 1, 'ppe': 1, 'oruse': 1, 'rate': 1, 'activeingredient.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1361labeling.deadlines.notification': 1, 'agency.fast-track': 1, 'reg-theistrant-initiatedantimicrobials': 1, 'permits': 1, 'actionsepano.newcrno.actiona52092': 1, '95,789swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1362public': 1, 'actions—continuedepano.newcrno.actiondecisionreviewtime': 1, 'a52193': 1, 'health32,250a52294': 1, 'health1211,025efficacy': 1, 'ad': 1, 'perad': 1, 'internal': 1, 'guid-ance': 1, 'efficacyprotocol': 1, 'proc-ess': 1, 'code': 1, 'in-clude': 1, 'studyprotocol': 1, 'devices': 1, 'mak-ing': 1, 'pesticidal': 1, 'applicant-initiated': 1, 'tier': 1, '1efficacy': 1, 'outside': 1, 'bymembers': 1, 'effi-cacy': 1, 'reviewexpert': 1, 'panel': 1, 'codewill': 1, 'healthefficacy': 1, 'reviewfor': 1, 'makingpesticidal': 1, 'ap-plicant-initiated': 1, 'tier2swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'ingredient,18138,91696': 1, 'ingredient,1883,594a524newa525newa526newa527newexperimental': 1, 'requirestolerance': 1, 'ingredientapplication': 1, 'fol-lows': 1, 'exemption.credit': 1, 'outdooruse': 1, 'offee': 1, 'indooruse': 1, 'ingredient,1586,82398': 1, 'ingredient,1558,000swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1364public': 1, 'actions—continuedepano.newcrno.actiona528newa529newa523new99': 1, 'reviewor': 1, 'assessment': 1, 'i.e.': 1, 'toxicology': 1, 'orexposure': 1, 'protocols': 1, '1520,26010,36599101': 1, 'protocol11,025swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'actions—continuedepano.newcrno.actiona571new102': 1, 'refinedecological': 1, 'and/orendangered': 1, 'applicant-ini

tiateddecisionreviewtime': 1, '1886,823': 1, 'agency.extension.labeling.deadlines.notifi
cation.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1366public': 1, 'biopesticide
s': 1, 'pollution': 1, 'preven-tion': 1, 'microbial': 1, 'biochemical': 1, 'pes-ticide
s': 1, 'ingredientsepano.newcrno.actionb580b590103': 1, 'toestablish': 1, 'toleranceexem
ption': 1, '1946,3051728,942b600105': 1, '1317,365non-food': 1, 'b610106': 1, '1011,577e
xperimental': 1, 'atemporary': 1, 'toleranceor': 1, 'exemptionexperimental': 1, 'establi
shpermanent': 1, 'toleranceexemptionno': 1, 'change': 1, 'per-manent': 1, 'b611newb612ne
w107': 1, '1211,577108': 1, '1015,918swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1,
'ingredients—continueddecisionreviewtime': 1, '1115,918epano.newcrno.actionb613new109':
1, 'convert': 1, 'apermanent': 1, 'exemp-tion': 1, '1368public': 1, '2012extension.': 1,
'ingredients—continuedepano.newcrno.actiondecisionreviewtime': 1, 'b620110': 1, '75,789e
xperimental': 1, 'destruct': 1, 'greater': 1, 'whichcase': 1, 'newinert': 1, 'thesame':
1, 'labelingwill': 1, 'theregistration': 1, 'thatfirst': 1, 'addi-tional': 1, 'informa-t
ion': 1, 'tosupport': 1, 'deficiencyscreening': 1, 'b630111': 1, 'petition1311,577b63111
2': 1, 'petition1211,577to': 1, 'estab-lished': 1, 'b640113': 1, 'petition1917,365b643ne
wb642newb644new114': 1, 'petition1011,577115': 1, '1228,942food/food': 1, 'toan': 1, '81
1,577b650117': 1, '75,789': 1, '1370public': 1, 'b652new118': 1, 'registered1311,577sour
ce': 1, 'pe-tition': 1, 'es-tablished': 1, 'ortolerance': 1, 'submis-sion': 1, 'cita-tio
n': 1, 'acceptеddata': 1, 'generated': 1, 'atgovernment': 1, 'expense': 1, 'orcitation':
1, 'scientif-ically-sound': 1, 'rationalebased': 1, 'publiclyavailable': 1, 'literatureo
r': 1, 'in-formation': 1, 'ad-dresses': 1, 're-quirement': 1, 'sub-mission': 1, 'request
for': 1, 'require-ment': 1, 'waivedsupported': 1, 'sci-entifically-sound': 1, 'ra-tional
e': 1, 'explainingwhy': 1, 'apply': 1, '1372public': 1, 'b660119': 1, 'registered41,159s
ource': 1, 'nochange': 1, 'exemption.no': 1, 'oronly': 1, 'orauthorization': 1, 'fromdat
a': 1, 'dem-onstrated': 1, 'categoryincludes': 1, 'microbialpesticides': 1, 'mustnot':
1, 're-isolated': 1, 'b670120': 1, 'registered74,631source': 1, '1374public': 1, 'b67112
1': 1, 'unregis-1711,577tered': 1, 'toamend': 1, 'establishedtolerance': 1, 'requires:
1': 1, 'ofproduct': 1, 'specific': 1, 'pre-viously': 1, 'andaccepted': 1, 'gen-erated':
1, 'governmentexpense': 1, 'scientifically-soundrationale': 1, 'based': 1, 'onpublicly':
1, 'lit-erature': 1, 'rel-evant': 1, 'informationthat': 1, 'addresses': 1, 'thedata': 1,
'requirement': 1, 'or5': 1, 're-quest': 1, 'bewaived': 1, 'supported': 1, 'bya': 1, 'b67
2122': 1, 'unregis-138,269tered': 1, 'usewith': 1, 'exemptionpreviously': 1, 'submission
of': 1, 'specificdata': 1, 'ofpreviously': 1, 'reviewedand': 1, 'or3': 1, '1376public':
1, 'products—continueddecisionreviewtime': 1, '104,63141,159epano.newcrno.actionb673newb
674newb675new123': 1, 'mup/ep': 1, 'oftechnical': 1, 'grade': 1, 'tgai': 1, 'anagency':
1, 'determina-tion': 1, 'citeddata': 1, 'supports': 1, 're-pack': 1, 'istered': 1, 'comp
letelynew': 1, 'generic': 1, 'registereduses': 1, 'reg-108,269swalcilbuphtiwdorp1nvx7thk
sdnoesuarkdverdate': 1, 'b676new126': 1, 'more138,269than': 1, 'oneactive': 1, 'isan':
1, 'besubmitted': 1, '1378public': 1, '108,000epano.newcrno.actionb677new127': 1, 'amend
mentsepano.newcrno.actiondecisionreviewtime': 1, 'b621128': 1, 'experi-74,631b622new12
9': 1, 'experi-1111,577b641130': 1, 'es-1311,577b680131': 1, 'registered54,631b681132':
1, 'unregis-75,513mental': 1, 'temporarytolerance': 1, 'toleranceexemption.mental': 1,
'anestablished': 1, 'tem-porary': 1, 'exemption.tablished': 1, 'exemption.source': 1, 'r
equiresdata': 1, 'tered': 1, '1380public': 1, 're-64,631b683newb684newquires': 1, 'revie
w/updateof': 1, 'previous': 1, 'withoutdata': 1, 'torei': 1, 'ani-88,000extension': 1,
'fees.amendmentshandledbyswalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'straight':
1, 'chain': 1, 'lepidopteranpheromones': 1, 'sclps': 1, 'epano.newcrno.actiondecisionrev
iewtime': 1, 'b690b700135': 1, 'newactive': 1, 'ornew': 1, 'use.b701137': 1, 'extend':
1, 'per-mit.7742,3161,1591,159swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1382pub
lic': 1, '—continuedepano.newcrno.actiondecisionreviewtime': 1, 'b710138': 1, 'b720139':
1, 'registered51,159source': 1, 'b721140': 1, 'unregis-72,426swalcilbuphtiwdorp1nvx7thks
dnoesuarkdverdate': 1, '1384public': 1, 'b722141': 1, 'amend-72,246ment': 1, 'quiring':
1, 'b730142': 1, 're-51,159extension': 1, 'activeingredient.fast-track': 1, '1386publi
c': 1, 'pollutionprevention': 1, 'actepano.newcrno.actiondecisionreviewtime': 1, 'b614ne
w143': 1, 'on32,294preapplication': 1, 'applicant-initiatedb615new144': 1, 'initiated32,
294swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'act—continuedepano.newcrno.actiond
ecisionreviewtime': 1, 'b682145': 1, 'appli-32,205cant': 1, 'ex-cludes': 1, 'hsrbrevie
w': 1, 'plant': 1, 'incorporated': 1, 'protectants': 1, 'pips': 1, 'b740146': 1, 'per-68
6,823mit': 1, 'nopetition': 1, 'tolerance/tolerance': 1, 'exemption.includes:1': 1, 'non
-food/feed': 1, 'pip': 1, 'food/feed': 1, 'anew': 1, 'pipwith': 1, 'pipin': 1, 'toleranc
e/toler-ance': 1, 'ex-ists': 1, 'intendeduse': 1, '1388public': 1, '—continuedepano.newc
rno.b750actiondecisionreviewtime': 1, '9115,763b77015173,644147': 1, 'witha': 1, 'establ
isha': 1, 'tolerance/tol-erance': 1, 'ingredient.includes': 1, 'petitionto': 1, 'b771fe
e': 1, 'ingredientthat': 1, 'sap': 1, 'follows.b77110115,763swalcilbuphtiwdorp1nvx7thksd
noesuarkdverdate': 1, '—continuedepano.newcrno.b772actiondecisionreviewtime': 1, '311,57

7b773528,942150': 1, 'amendor': 1, 'since': 1, 'theestablished': 1, 'exemptionfor': 1,
'unaffected.151': 1, 'ex-tend': 1, 'tol-erance/tolerance': 1, 'ingredient.152': 1, 'non-
food/feed.153': 1, 'sapreview': 1, 'permanenttolerance/toleranceexemption': 1, 'basedo
n': 1, 'existing': 1, 'exemption.b780b790b80012144,70418202,58512231,585swalcilbuphtiwdo
rp1nvx7thksdnoesuarkdverdate': 1, '1390public': 1, '—continuedepano.newcrno.b810actionde
cisionreviewtime': 1, '18289,407155': 1, 'exemption.sap': 1, 'exemptionof': 1, 'ingre-di
ent.157': 1, 'b82015289,407b84021347,288swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate':
1, 'b851158': 1, 'applica-9115,763b870934,729b880928,942tion': 1, 'event': 1, 'aprevious
ly': 1, 'registeredpip': 1, 'petitionsince': 1, 'permanent': 1, 'alreadyestablished': 1,
'theactive': 1, '.159': 1, 'additionaldata': 1, 'isalready': 1, 'establishedfor': 1, '13
92public': 1, '—continuedepano.newcrno.b881actiondecisionreviewtime': 1, '1586,823161':
1, 'withnegotiated': 1, 'acreagecap': 1, 'time-limitedregistration': 1, 'b883new9115,763
swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '—continueddecisionreviewtime': 1, '12
144,704986,823epano.newcrno.actionb884newb885new163': 1, 'breeding': 1, 'stack': 1, 'app
rovedpips': 1, 'convertsregistration': 1, 'com-mercial': 1, '.b890165': 1, 'a957,882swal
cilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1394public': 1, 'b891166': 1, 'a15115,763
seed': 1, 'exemptionalready': 1, 'asextending': 1, 'expira-tion': 1, 'modifyingan': 1,
'irm': 1, 'plan': 1, 'add-ing': 1, 'insect': 1, 'becontrolled': 1, 'b900167': 1, 'a611,5
77b901168': 1, 'a1269,458b902169': 1, 'review35,789swalcilbuphtiwdorp1nvx7thksdnoesuarkd
verdate': 1, 'b903170inert': 1, 'toler-657,882ance': 1, 'marker': 1, 'asnpt': 1, 'inbpp
d.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, '1396public': 1, '—continuedepano.ne
wcrno.b904171actionimport': 1, 'processed': 1, 'commod-ities/food': 1, 'inertor': 1, '.d
ecisionreviewtime': 1, '9115,763extension': 1, 'beenregistered.rently': 1, 'cur-': 1, 't
ransfer': 1, 'conventional': 1, 'fornew': 1, 'field': 1, 'corn': 1, 'sweet': 1, 'scienti
fic': 1, 'involved': 1, 'complex': 1, 'epaoften': 1, 'seeks': 1, 'advice': 1, 'advisor
y': 1, 'onrisks': 1, 'pesticides': 1, 'pose': 1, 'wildlife': 1, 'farm': 1, 'workers': 1,
'appli-cators': 1, 'non-target': 1, 'well': 1, 'resistance': 1, 'novelscientific': 1, 'i
ssues': 1, 'surrounding': 1, 'technologies': 1, 'scientists': 1, 'thesap': 1, 'make': 1,
'recommend': 1, 'policy': 1, 'provideadvice': 1, 'used': 1, 'isinvaluable': 1, 'strive
s': 1, 'protect': 1, 'envi-ronment': 1, 'risks': 1, 'posed': 1, 'takes': 1, 'toschedul
e': 1, 'prepare': 1, 'meetings': 1, 'timeand': 1, 'needed': 1, 'stacked': 1, 'deploymen
t': 1, 'different': 1, 'blend': 1, 'negotiated': 1, 'acreage': 1, 'cap': 1, 'depend': 1,
'deter-mination': 1, 'potential': 1, 'environmental': 1, 'exposure': 1, 'organisms': 1,
'targeted': 1, 'developing': 1, 'resist-ance': 1, 'substance': 1, 'uncertainty': 1, 'ris
ksmay': 1, 'reduce': 1, 'allowable': 1, 'quantity': 1, 'andtype': 1, 'organism': 1, 'lac
k': 1, 'insectresistance': 1, 'management': 1, 'usually': 1, 'forseed-increase': 1, 'reg
istrants': 1, 'encouraged': 1, 'consultwith': 1, 'thiscategory': 1, 'concurrently': 1,
'anapplication': 1, 'commercial': 1, 'example': 1, 'modifications': 1, 'applicant-ini-ti
ated': 1, 'fees.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'ingredients': 1, 'ext
ernal': 1, 'reviewand': 1, 'miscellaneous': 1, 'actionsdecisionreviewtime': 1, '1218,000
105,0003,00010,000888epano.newcrno.actioni001i002newi003newi004newi005newi006new172': 1,
'ap-proved': 1, 'proved': 1, 'useinert': 1, 'withnew': 1, 'pattern': 1, 'newdata': 1, 'n
onew': 1, 'ap-5,000177': 1, 'ap-63,000swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1,
'1398public': 1, 'actions—continueddecisionreviewtime': 1, '41,500543,4002,80097,200epan
o.newcrno.actioni007newi008newi009newi010newm001newm002new178': 1, 'substan-tially': 1,
'ingre-dients': 1, 'originalinert': 1, 'iscompositionally': 1, 'usepattern': 1, 'poly-me
r': 1, 'non': 1, 'erance': 1, 'exemptiondescriptor': 1, 'oneor': 1, 'casrns': 1, 'requir
-ing': 1, 'studiesreview': 1, 'board': 1, 'reviewas': 1, 'defined': 1, 'cfr26': 1, 'anac
tive': 1, 'stud-ies': 1, '40cfr': 1, 'ofan': 1, 're-97,200181': 1, 'tol-61,500swalcilbup
htiwdorp1nvx7thksdnoesuarkdverdate': 1, 'actions—continuedepano.newcrno.actionm003newdec
isionreviewtime': 1, '1258,000184': 1, 'peerreview': 1, 'consultation': 1, 'withfifra':
1, 'ad-visory': 1, 'anaction': 1, 'decisiontimeframe': 1, 'lessthan': 1, 'ad-ministrato
r': 1, 'de-fined': 1, 'anovel': 1, 'unique': 1, 'applicationtechnology': 1, 'excludespi
p': 1, '1400public': 1, 'actions—continuedepano.newcrno.actionm004newdecisionreviewtim
e': 1, '1858,000185': 1, 'greaterthan': 1, 'actions—continuedepano.newcrno.actionm005new
decisionreviewtime': 1, '920,000186': 1, 'combina-tion': 1, 'contains': 1, 'unreg-istere
d': 1, 'con-ventional': 1, 'bio-pesticide': 1, 'requirescoordination': 1, 'withother':
1, 'regulatory': 1, 'divi-sions': 1, 'conduct': 1, 'labeland/or': 1, 'verify': 1, 'va-li
dity': 1, 'dataas': 1, 'exist-ing': 1, 'thecombination': 1, 'request': 1, 'let-1250m006n
ewm007newters': 1, 'certification': 1, 'gold': 1, 'seal': 1, 'oneactively': 1, 'register
edproduct.clusive': 1, 'asprovided': 1, 'fifrasection': 1, 'ex-125,000swalcilbuphtiwdorp
1nvx7thksdnoesuarkdverdate': 1, '1402public': 1, 'm008new189': 1, 'grant': 1, 'exclu-10
1,500sive': 1, 'minor': 1, 'sec-tion': 1, 'depend-ent': 1, 'eachapplication': 1, 'respec
tive': 1, 'fee.the': 1, 'match': 1, 'pria': 1, 'pendinginert': 1, 'unless': 1, 'as-socia
ted': 1, 'sub-ject': 1, 'mul-tiple': 1, 'grouped': 1, 'chemical': 1, 'class': 1, 'single

```
registration': 1, 'ofsuch': 1, 'rule': 1, 'de-pendent': 1, 'hsrb': 1, 'times': 1, 'assoc
iatedactions': 1, 'run': 1, 'latest': 1, 'actionwill': 1, '1403labeling.deadlines.notifi
cation': 1, 'agency.': 1, 'october': 1, 'atthe': 1, 'administratorsubsectionapplicatio
n': 1, 'undertherejected': 1, 'fund.—section': 1, '7u.s.c': 1, 'grants': 1, '500,000.':
1, '1404public': 1, 'reforms': 1, 'periods.—section': 1, 'thefederal': 1, 'u.s.c.136w-
8': 1, 'registrationimprovement': 1, 'renewal': 1, 'publishin': 1, 'administratorshall':
1, 'publicly': 1, 'appearing': 1, 'thecongressional': 1, 'record': 1, 'pages': 1, 's1040
9': 1, 'followsthrough': 1, 'completeness': 1, 'later': 1, 'inclause': 1, 'initial': 1,
'content': 1, 'preliminary': 1, 'technicalscreenings.—': 1, 'screenings.—': 1, 'content.
—not': 1, 'designated': 1, 'addingat': 1, 'screening.—after': 1, 'conducting': 1, 'scree
ningdescribed': 1, 'accordance': 1, 'withclause': 1, 'apreliminary': 1, 'screening—': 1,
'aa': 1, 'dateon': 1, 'periodbegins': 1, 'timereview': 1, 'periods': 1, 'thefollowing':
1, 'rejection.—': 1, 'general.—if': 1, 'deter-mines': 1, 'com-pletes': 1, 'underclause':
1, 'failed': 1, 'initialcontent': 1, 'theapplicant': 1, 'correct': 1, 'failure': 1, 'the
date': 1, 'applicantdeadlines.determination.swalcilbuphtiwdorp1nvx7thksdnoesuarkdverdat
e': 1, '1405time': 1, 'period.receives': 1, 'notification': 1, 'reject': 1, 'notificatio
n.—the': 1, 'every': 1, 'effort': 1, 'writtennotification': 1, 'rejection': 1, '10-day':
1, 'begins': 1, 'datethe': 1, 'completes': 1, 'tech-nical': 1, 'screening.': 1, 'headin
g': 1, 'cc': 1, 'con-tains': 1, 'contain': 1, 'requirements': 1, 'technicalscreening.—i
n': 1, 'technicalscreening': 1, 'shalldetermine': 1, 'if—determination.': 1, 'accuratean
d': 1, 'complete': 1, 'areconsistent': 1, 'pro-posal': 1, 'therequirement': 1, '21u.s.
c': 1, 'fullreview': 1, 'standards': 1, 'could': 1, 'resultin': 1, 'granting': 1, 'repor
ts.—section': 1, 'march': 1, 'atend': 1, 'viii': 1, 'extensions': 1, 'agreed': 1, 'along
with': 1, 'description': 1, 'reason': 1, 'administratorwas': 1, 'unable': 1, 'deci-sio
n': 1, 'theend': 1, 'periodand': 1, 'progress': 1, 'toward—': 1, 'carrying': 1, 'amounts
from': 1, 'processingfund': 1, '1406public': 1, '2012publicinformation.deadline.assessme
nt.evaluation.deadline.notification.review.assessment.': 1, 'systems': 1, 'electronictra
cking': 1, 'december': 1, '31,2013': 1, 'system': 1, 'statusof': 1, 'making': 1, 'noncon
-fidential': 1, 'related': 1, 'endan-gered': 1, 'knowledge': 1, 'database': 1, 'makingno
nconfidential': 1, 'databasepublicly': 1, 'electronicallysubmit': 1, 'registrationaction
s': 1, 'actionsby': 1, 'facilitating': 1, 'participation': 1, 'processby': 1, 'providin
g': 1, 'interested': 1, 'partiesof': 1, 'additions': 1, 'docket': 1, 'rejected': 1, 'the
administrator': 1, 'preliminarytechnical': 1, 'conducted': 1, 'updating': 1, 'thepestici
de': 1, 'incident': 1, 'towardmaking': 1, 'availableto': 1, 'appro-priate': 1, 'availabi
lity': 1, 'sum-mary': 1, 'usage': 1, 'data.': 1, 'report.—': 1, 'scope.—in': 1, 'repor
t': 1, 'describedin': 1, 'submit': 1, 'committee': 1, 'agri-culture': 1, 'committeeon':
1, 'agriculture': 1, 'nutrition': 1, 'forestry': 1, 'areport': 1, 'analysis': 1, 'impac
t': 1, 'mainte-nance': 1, 'businesses': 1, 'have—': 1, 'notexceed': 1, '2,000,000.': 1,
'required.—in': 1, 'anal-ysis': 1, 'shallcollect': 1, 'on—': 1, 'insubparagraph': 1, 'pa
ying': 1, 'companyholds.': 1, 'termination': 1, 'effectiveness.—section': 1, 'amended—sw
alcilbuphtiwdorp1nvx7thksdnoesuarkdverdate': 1, 'andand': 1, 'date.—this': 1, 'madeby':
1, 'take': 1, 'effect': 1, 'relationship': 1, 'law.—in': 1, 'conflictbetween': 1, 'join
t': 1, 'resolution': 1, 'continuing': 1, 'appropriations': 1, 'forfiscal': 1, 'year': 1,
'jointresolution': 1, 'sectionshall': 1, 'control.7': 1, '136a-1note.7': 1, '136a-1note.
approved': 1, '2012.legislative': 1, 'history—s': 1, 'congressional': 1, 'vol': 1, 'sep
t.': 1, 'passed': 1, 'senate.sept': 1, 'house.æswalcilbuphtiwdorp1nvx7thksdnoesuarkdverd
ate': 1}
```

In [13]:

```
print(type(nb_env_input))
print(len(nb_env_input))
```

```
<class 'list'>
12
```

In [14]:

```
legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0] + nb_not
#legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0] + nb_no
#legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0] + nb_no
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-14-8bf9ed7d4988> in <module>
----> 1 legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0]
```

```
+ nb_notenv_input[0][0])
      2 #legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0]
 + nb_notenv_input[0][0])
      3 #legislative_subject_classifier = NaiveBayesClassifier.train(nb_env_input[0][0]
 + nb_notenv_input[0][0])

~/anaconda3/envs/metis/lib/python3.8/site-packages/nltk/classify/naivebayes.py in train
(cls, labeled_featuresets, estimator)
    204         # Count up how many times each feature value occurred, given
    205         # the label and featurename.
--> 206         for featureset, label in labeled_featuresets:
    207             label_freqdist[label] += 1
    208             for fname, fval in featureset.items():

ValueError: not enough values to unpack (expected 2, got 1)
```

In [ ]:
```
nltk.classify.util.accuracy(mr_op_classifier, nb_mr_train_input + nb_op_train_input)*10
```

Test the model

- essentially just run the accuracy but pass in the testing data instead of the training data

Using TFIDF vectorizer

In [615…
```
from sklearn.model_selection import train_test_split, StratifiedKFold, RepeatedStratifi
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, fb
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics import confusion_matrix, classification_report
from xgboost import XGBClassifier
import seaborn as sns

def vectorize(X_text_tr, X_text_val, vectorizer_type):
    '''
    Vectorizes text into machine readable and model ready input format.

    Input:
        X_text_tr (Training set texts)
        X_text_val (Validation set texts)
        vectorizer_type ['tfidf', 'counter']
    Output:
        X_tr_vector: sparse matrix; vectorized training text features
        X_val_vector: sparse matrix; vectorized validation text features
        train_df: doc-term term frequencies dataframe (training)
        val_df:
        vectorizer: vectorizer object; fitted on X_text_tr to be used to transform test
    '''
    if vectorizer_type == 'tfidf':
        vectorizer = TfidfVectorizer(max_df=0.8, min_df=0.1, lowercase=True, analyzer='
                                     stop_words= 'english',ngram_range=(1,1))
    elif vectorizer_type == 'counter':
        vectorizer = TfidfVectorizer(max_df=0.8, min_df=0.1, lowercase=True, analyzer='
                                     stop_words= 'english',ngram_range=(1,1))

    X_tr_vector = vectorizer.fit_transform(X_text_tr.ravel())
    X_val_vector = vectorizer.transform(X_text_val.ravel())

    feature_names = vectorizer.get_feature_names()
```

```
    dense1 = X_tr_vector.todense().tolist()
    dense2 = X_val_vector.todense().tolist()
    train_df = pd.DataFrame(dense1, columns=feature_names)
    val_df = pd.DataFrame(dense2, columns=feature_names)

    return X_tr_vector, X_val_vector, train_df, val_df, vectorizer
```

In [683…

```
text = doc_df['text_cleaned']
y = (doc_df['env_label'] == 'Environmental')

#test random_state
text_train, text_test, y_train, y_test = train_test_split(text, y, test_size=0.2)
text_train, text_val, y_train, y_val = train_test_split(text_train, y_train, test_size=

kf = RepeatedStratifiedKFold(n_splits=5, n_repeats=25)
#model using TFIDF vectorizer
X_train, X_val, df_train, df_val, vectorizer = vectorize(text_train, text_val, 'tfidf')
X, y = X_train.toarray(), np.array(y_train)

train_accuracies, train_precisions, train_recalls, train_f1s, train_fbetas = [], [], []
val_accuracies, val_precisions, val_recalls, val_f1s, val_fbetas = [], [], [], [], []

for train_ind, val_ind in kf.split(X,y):
    X_train, y_train = X[train_ind], y[train_ind]
    X_val, y_val = X[val_ind], y[val_ind]

    model = BernoulliNB()
    model.fit(X_train, y_train)
    ytrain_preds = model.predict(X_train)
    ytrain_preds_probs = model.predict_proba(X_train)[:,1]
    train_preds = np.where(ytrain_preds_probs > 0.5, 1, 0)
    yval_preds = model.predict(X_val)
    yval_preds_probs = model.predict_proba(X_val)[:,1]
    val_trains = np.where(yval_preds_probs > 0.5, 1, 0)

#     train_acc = model.score(X_train, y_train)
#     val_acc = model.score(X_val, y_val)
    train_accuracies.append(accuracy_score(y_train, ytrain_preds))
    train_precisions.append(precision_score(y_train, train_preds))
    train_recalls.append(recall_score(y_train, ytrain_preds))
    train_f1s.append(f1_score(y_train, ytrain_preds))
    train_fbetas.append(fbeta_score(y_train, ytrain_preds, beta=0))

    val_accuracies.append(accuracy_score(y_val, yval_preds))
    val_precisions.append(precision_score(y_val, yval_preds))
    val_recalls.append(recall_score(y_val, yval_preds))
    val_f1s.append(f1_score(y_val, yval_preds))
    val_fbetas.append(fbeta_score(y_val, yval_preds, beta=0))

train_scores = [train_accuracies, train_precisions, train_recalls, train_f1s, train_fbe
val_scores = [val_accuracies, val_precisions, val_recalls, val_f1s, val_fbetas]
scores = ['Accuracy', 'Precision', 'Recall', 'F1', 'FBeta']
```

In [686…

```
for score, tr, val in zip(scores, train_scores, val_scores):
    print(score)
    print(f'Train: {np.mean(tr):.3f} +- {np.std(tr):.3f}')
    print(f'Val: {np.mean(val):.3f} +- {np.std(val):.3f}')
    print('-------------')
```

```
Accuracy
Train: 0.810 +- 0.048
Val: 0.731 +- 0.112
-------------
Precision
Train: 0.855 +- 0.062
Val: 0.713 +- 0.210
-------------
Recall
Train: 0.609 +- 0.096
Val: 0.540 +- 0.198
-------------
F1
Train: 0.709 +- 0.078
Val: 0.597 +- 0.180
-------------
FBeta
Train: 0.855 +- 0.062
Val: 0.713 +- 0.210
-------------
```

In [687…

```
print('Validation Results')
print(classification_report(y_val, yval_preds))
print('')
```

```
Validation Results
              precision    recall  f1-score   support

       False       0.67      0.80      0.73        10
        True       0.60      0.43      0.50         7

    accuracy                           0.65        17
   macro avg       0.63      0.61      0.61        17
weighted avg       0.64      0.65      0.63        17
```

Testing Model on Hold-out/Test Set

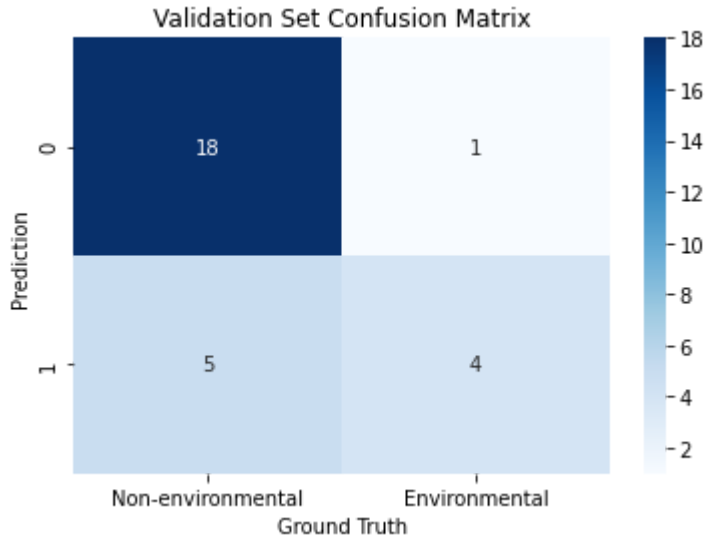
In [688…

```
X_test = vectorizer.transform(text_test)
preds = model.predict(X_test)
print(f'Test Accuracy: ', model.score(X_test, y_test))
```

```
Test Accuracy:  0.6428571428571429
```

In [661…

```
cm = confusion_matrix(y_test, preds)
sns.heatmap(cm, annot=True, cmap='Blues', xticklabels=['Non-environmental', 'Environmen
plt.title('Validation Set Confusion Matrix')
plt.xlabel('Ground Truth')
plt.ylabel('Prediction');
```



In [ ]:
```
# text = doc_df['text_cleaned']
# y = (doc_df['env_label'] == 'Environmental')

# # text_train, text_test, y_train, y_test = train_test_split(text, y, test_size=0.2)
# text_train, text_val, y_train, y_val = train_test_split(text_train, y_train, test_siz

# X_train, X_val, df_train, df_val, vectorizer = vectorize(text_train, text_val, 'tfidf
# X, y = X_train.toarray(), np.array(y_train)
# X_test = vectorizer.transform(text_test)
```

In [ ]:
```
# X_train.shape, X_test.shape
```

1-run of XGBoost

In [641…
```
xgb = XGBClassifier(n_estimators=100)
xgb.fit(X_train, y_train)

train_preds = xgb.predict(X_train)
val_preds = xgb.predict(X_val)
test_preds = xgb.predict(X_test)

train_acc_xgb = (train_preds == y_train).sum().astype(float) / len(train_preds)*100
val_acc_xgb = (val_preds == y_val).sum().astype(float) / len(val_preds)*100
test_acc_xgb = (test_preds == y_test).sum().astype(float) / len(test_preds)*100

print("XGBoost's train prediction accuracy is: %3.2f" % (train_acc_xgb))
print("XGBoost's val prediction accuracy is: %3.2f" % (val_acc_xgb))
print("XGBoost's test prediction accuracy is: %3.2f" % (test_acc_xgb))
```

```
XGBoost's train prediction accuracy is: 100.00
XGBoost's val prediction accuracy is: 82.61
XGBoost's test prediction accuracy is: 89.29
```

In [642…
```
print('Validation Results')
print(classification_report(y_val, val_preds))
print('')
```

```
Validation Results
              precision    recall  f1-score   support

       False       0.92      0.79      0.85        14
        True       0.73      0.89      0.80         9

    accuracy                           0.83        23
   macro avg       0.82      0.84      0.82        23
weighted avg       0.84      0.83      0.83        23
```

XGBoost with Repeated Kfold Cross Validation

In [689…

```
# text = doc_df['text_cleaned']
# y = (doc_df['env_label'] == 'Environmental')

# #test random_state
# text_train, text_test, y_train, y_test = train_test_split(text, y, test_size=0.2)
# text_train, text_val, y_train, y_val = train_test_split(text_train, y_train, test_siz

# kf = RepeatedStratifiedKFold(n_splits=5, n_repeats=25)
# #model using TFIDF vectorizer
# X_train, X_val, df_train, df_val, vectorizer = vectorize(text_train, text_val, 'tfidf
# X, y = X_train.toarray(), np.array(y_train)

train_accuracies, train_precisions, train_recalls, train_f1s, train_fbetas = [], [], []
val_accuracies, val_precisions, val_recalls, val_f1s, val_fbetas = [], [], [], [], []

for train_ind, val_ind in kf.split(X,y):
    X_train, y_train = X[train_ind], y[train_ind]
    X_val, y_val = X[val_ind], y[val_ind]

    xgb = XGBClassifier(n_estimators=100)
    xgb.fit(X_train, y_train)
    ytrain_preds = xgb.predict(X_train)
    ytrain_preds_probs = xgb.predict_proba(X_train)[:,1]
    train_preds = np.where(ytrain_preds_probs > 0.5, 1, 0)
    yval_preds = xgb.predict(X_val)
    yval_preds_probs = xgb.predict_proba(X_val)[:,1]
    val_trains = np.where(yval_preds_probs > 0.5, 1, 0)

#     train_acc = model.score(X_train, y_train)
#     val_acc = model.score(X_val, y_val)
    train_accuracies.append(accuracy_score(y_train, ytrain_preds))
    train_precisions.append(precision_score(y_train, train_preds))
    train_recalls.append(recall_score(y_train, ytrain_preds))
    train_f1s.append(f1_score(y_train, ytrain_preds))
    train_fbetas.append(fbeta_score(y_train, ytrain_preds, beta=0))

    val_accuracies.append(accuracy_score(y_val, yval_preds))
    val_precisions.append(precision_score(y_val, yval_preds))
    val_recalls.append(recall_score(y_val, yval_preds))
    val_f1s.append(f1_score(y_val, yval_preds))
    val_fbetas.append(fbeta_score(y_val, yval_preds, beta=0))

train_scores = [train_accuracies, train_precisions, train_recalls, train_f1s, train_fbe
val_scores = [val_accuracies, val_precisions, val_recalls, val_f1s, val_fbetas]
scores = ['Accuracy', 'Precision', 'Recall', 'F1', 'FBeta']
```

for score, tr, val in zip(scores, train_scores, val_scores): print(score) print(f'Train: {np.mean(tr):.3f} +- {np.std(tr):.3f}') print(f'Val: {np.mean(val):.3f} +- {np.std(val):.3f}') print('------------')

In [690…

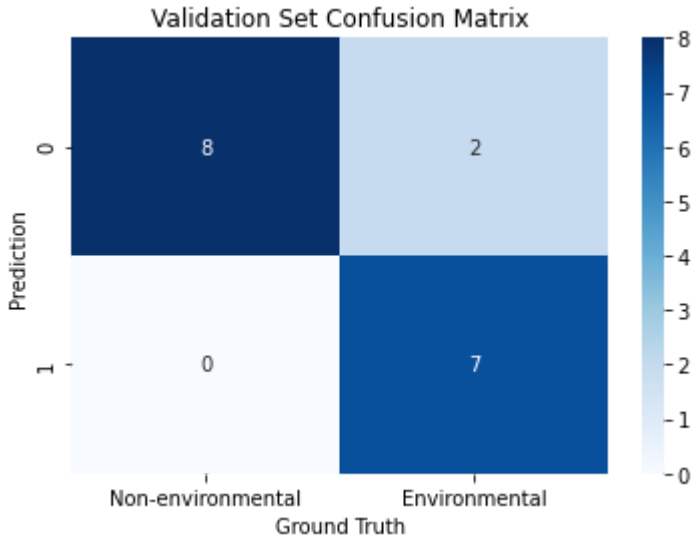
```
print('Training Results')
print(classification_report(y_train, ytrain_preds))
print("ROC AUC Score Training = " + str(roc_auc_score(y_train, ytrain_preds_probs)))
print("Training Accuracy = " + str(xgb.score(X_train, y_train)))
print('\n-------------------------\n')
print('Validation Results')
print(classification_report(y_val, yval_preds))
print("ROC AUC Score Validation = " + str(roc_auc_score(y_val, yval_preds_probs)))
print("Validation Accuracy = " + str(xgb.score(X_val, y_val)))
```

```
Training Results
              precision    recall  f1-score   support

       False       1.00      1.00      1.00        44
        True       1.00      1.00      1.00        27

    accuracy                           1.00        71
   macro avg       1.00      1.00      1.00        71
weighted avg       1.00      1.00      1.00        71

ROC AUC Score Training = 1.0
Training Accuracy = 1.0

-------------------------

Validation Results
              precision    recall  f1-score   support

       False       1.00      0.80      0.89        10
        True       0.78      1.00      0.88         7

    accuracy                           0.88        17
   macro avg       0.89      0.90      0.88        17
weighted avg       0.91      0.88      0.88        17

ROC AUC Score Validation = 1.0
Validation Accuracy = 0.8823529411764706
```

In [691…

```
cm = confusion_matrix(y_val, yval_preds)
sns.heatmap(cm, annot=True, cmap='Blues', xticklabels=['Non-environmental', 'Environmen
plt.title('Validation Set Confusion Matrix')
plt.xlabel('Ground Truth')
plt.ylabel('Prediction');
print('Recall:', recall_score(y_val, yval_preds))
print('Precision:', precision_score(y_val, yval_preds))
```

```
Recall: 1.0
Precision: 0.7777777777777778
```



In [692…
```
X_train.shape, X_test.shape
```

Out[692… ((71, 610), (28, 610))

In [693…
```
test_preds = xgb.predict(X_test)
test_acc_xgb = (test_preds == y_test).sum().astype(float) / len(test_preds)*100
print("XGBoost's test prediction accuracy is: %3.2f" % (test_acc_xgb))
```

```
XGBoost's test prediction accuracy is: 92.86
```

In [694…
```
pd.to_pickle(xgb, 'xgbclassifier.pkl')
```

Extras

In [262…
```
#No Kfold cross validation yet
def run_model_n_times(n_times):
    train_scores, test_scores = [], []

    runs = 0
    while runs < n_times:
        print(f'{runs+1} Run(s)')
        text_train, text_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
        #first model using TFIDF vectorizer
        X_train, X_test, df_train, df_test, vectorizer = vectorize(text_train, text_tes
        X_train.shape, X_test.shape

        model = BernoulliNB()
        model.fit(X_train, y_train)
        train_score = model.score(X_train, y_train)
        test_score = model.score(X_test, y_test)

        train_scores.append(train_score)
        test_scores.append(test_score)
        print('Train Score:', train_score)
        print('Test Score:', test_score)
        print('--------------------------')
        runs+=1
```

```
    print(f'Average Train Score ({n_times} Runs): {np.mean(train_scores)}')
    print(f'Average Test Score ({n_times} Runs): {np.mean(test_scores)}')

run_model_n_times(10)
```

```
1 Run(s)
Train Score: 0.8557692307692307
Test Score: 0.7714285714285715
--------------------------
2 Run(s)
Train Score: 0.8846153846153846
Test Score: 0.8
--------------------------
3 Run(s)
Train Score: 0.875
Test Score: 0.9428571428571428
--------------------------
4 Run(s)
Train Score: 0.75
Test Score: 0.6285714285714286
--------------------------
5 Run(s)
Train Score: 0.9326923076923077
Test Score: 0.9142857142857143
--------------------------
6 Run(s)
Train Score: 0.75
Test Score: 0.7142857142857143
--------------------------
7 Run(s)
Train Score: 0.8076923076923077
Test Score: 0.7428571428571429
--------------------------
8 Run(s)
Train Score: 0.7403846153846154
Test Score: 0.5714285714285714
--------------------------
9 Run(s)
Train Score: 0.8557692307692307
Test Score: 0.7428571428571429
--------------------------
10 Run(s)
Train Score: 0.8653846153846154
Test Score: 0.8857142857142857
--------------------------
Average Train Score (10 Runs): 0.8317307692307694
Average Test Score (10 Runs): 0.7714285714285716
```

Using CounterVectorizer

In [226…

```
X_train, X_test, df_train, df_test, vectorizer = vectorize(text_train, text_test, 'coun
X_train.shape, X_test.shape
```

Out[226…

```
((104, 475), (35, 475))
```

In [227…

```
model = BernoulliNB()
model.fit(X_train, y_train)
train_score = model.score(X_train, y_train)
test_score = model.score(X_test, y_test)
print('Train Score:', train_score)
print('Test Score:', test_score)
```

```
Train Score: 0.8173076923076923
Test Score: 0.7142857142857143
```

After trying out both TFIDF and Counter vectorizers, neither really have any discernable difference in performance.

In [ ]:

In [493…

```
from preprocessing import get_text_from_pdf, clean_text
file = 'Test Data/BILLS-116hr8915ih.pdf'
doc_str = get_text_from_pdf(file)
txt_cln = clean_text(doc_str)
txt_cln
```

Out[493…

```
'ith congressd session  h r to  amend  the  comprehensive  environmental  response  comp
ensation  andliability  act  of    to  provide  for  the  consideration  of  climate  ch
angeand for other purposesin  the  house  of  representativesdecember  mr  cleaver for
himself  and  ms  bass  introduced  the  following  bill  whichwas referred to the commi
ttee on energy and commerce and in additionto  the  committee  on  transportation  and
infrastructure  for  a  period  tobe  subsequently  determined  by  the  speaker  in  ea
ch  case  for  consideration  of  such  provisions  as  fall  within  the  jurisdiction
of  the  committeeconcerneda  billto amend the comprehensive environmental response comp
ensation  and  liability  act  of    to  provide  for  theconsideration  of  climate  ch
ange  and  for  other  purposesbe  it  enacted  by  the  senate  and  house  of  represe
ntatives of the united states of america in congress assembledsection  short titlethis
act  may  be  cited  as  the  preparing  superfundfor climate change act of sllibhtiwdor
pclksdnonosnhokjverdate sep     dec   jkt  po  frm  fmt  sfmt  ebillshih hsec  climate c
hange mitigationsection  of the comprehensive environmental response  compensation  and
liability  act  of    usc  is amended  in  subsection  b  in  the  fifth  sentencea  in
the  matter  preceding  subparagrapha  by  striking  account  and  inserting  accountand
at the endbin  subparagraph  f  by  strikingc  in  subparagraph  g  by  striking  theper
iod at the end and inserting  and andd  by  inserting  after  subparagraph  gthe followi
ngh  the  potential  threat  to  human  healthand  the  environment  associated  with  l
ocal  natural  disasters  and  extreme  weather  hazards  including  any  projected  exa
cerbation  or  change  inthose  disasters  and  hazards  due  to  climatechange and in s
ubsection c by inserting after the firstsentence  the  following  the  president  shall
includein  the  review  an  assessment  of  whether  the  selectedremedial  action  rema
ins  protective  after  taking  intoaccount  local  natural  disasters  and  extreme  we
atherhazards  including  any  projected  exacerbation  orsllibhtiwdorpclksdnonosnhokjver
date sep     dec   jkt  po  frm  fmt  sfmt  ebillshih hhr    ihchange in those disasters
and hazards due to climatechangesllibhtiwdorpclksdnonosnhokjverdate sep     dec   jkt  p
o  frm  fmt  sfmt  ebillshih hhr    ih'
```

In [494…

```
test_tfidf = vectorizer.transform([txt_cln])
model.predict(test_tfidf)[0]
```

Out[494…

```
True
```

In [ ]: