0. Load data

In [71]:
```
import pandas as pd
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, LancasterStemmer, WordNetLemmatizer
import string
from collections import Counter
import nltk as nltk
from nltk.text import Text
import re
import sys
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import jaccard_score
from itertools import combinations
import matplotlib.pyplot as plt
from nltk import ngrams
```

In [72]:
```
# import data
import csv
df = pd.read_csv('Food_Inspections_20240108.csv')
df.head(10)
```

Out[72]:

	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	
0	2587633	TACOS MARIO'S LIMITED	TACOS MARIO'S LIMITED	1447643.0	Restaurant	Risk 1 (High)	4540 W 63RD ST	CHICAGO	IL	60629.0	01/04/2024	Non-Inspection	No Entry	NaN	41.778606	-87.736592	(41.7786061 -87.7365916
1	2587625	TAQUERIA EL ARCO #3 EL POLLO FELIZ, INC.	EL POLLO CRIS CRIS #3	1964458.0	Restaurant	Risk 1 (High)	7023-7025 S PULASKI RD	CHICAGO	IL	60629.0	01/04/2024	Canvass	Out of Business	NaN	41.765126	-87.722279	(41.7651264 -87.7222789
2	2587598	Tel Aviv Slices Pizza LLC	Tel Aviv Slices Pizza LLC	2948183.0	NaN	Risk 1 (High)	3517 W DEVON AVE	CHICAGO	IL	60659.0	01/04/2024	License	Not Ready	NaN	41.997177	-87.717388	(41.997176 -87.7173876
3	2587596	KENSINGTON SCHOOL OF LINCOLN PARK	KENSINGTON SCHOOL OF LINCOLN PARK	2944036.0	Children's Services Facility	Risk 1 (High)	2745 N LINCOLN AVE	CHICAGO	IL	60614.0	01/04/2024	License	Pass	NaN	41.931779	-87.657407	(41.9317793 -87.6574069
4	2587599	Q'S AUTHENTIC COOKING	Q'S AUTHENTIC COOKING	2757521.0	Restaurant	Risk 1 (High)	306 E 75TH ST	CHICAGO	IL	60619.0	01/04/2024	Canvass	Pass	55. PHYSICAL FACILITIES INSTALLED, MAINTAINED ...	41.758467	-87.617974	(41.758467 -87.6179738
5	2587579	SUBWAY	SUBWAY	1719251.0	Restaurant	Risk 1 (High)	10354 S HALSTED ST	CHICAGO	IL	60628.0	01/04/2024	Canvass	Out of Business	NaN	41.705296	-87.642931	(41.70529649 -87.6429309
6	2587551	WE ARE THE FUTURE PRESIDENTS	WE ARE THE FUTURE PRESIDENTS	2578250.0	Daycare Above and Under 2 Years	Risk 1 (High)	843-847 W 103RD ST	CHICAGO	IL	60643.0	12/29/2023	License	No Entry	NaN	41.706792	-87.644524	(41.70679217 -87.6445236
7	2587527	ST. ANTHONY HOSPITAL	PATIENT KITCHEN/PREP AREA	2204412.0	Hospital	Risk 1 (High)	2875 W 19TH ST	CHICAGO	IL	60623.0	12/29/2023	Canvass Re-Inspection	Pass	51. PLUMBING INSTALLED; PROPER BACKFLOW DEVICE...	41.855403	-87.698708	(41.8554028 -87.6987075
8	2587465	PROVARE CHICAGO	PROVARE CHICAGO	2796734.0	Restaurant	Risk 1 (High)	1421 W CHICAGO AVE	CHICAGO	IL	60642.0	12/28/2023	Canvass	Out of Business	NaN	41.896025	-87.663272	(41.8960250 -87.6632722
9	2587467	ICOLDTEA	ICOLDTEA	2886816.0	Restaurant	Risk 2 (Medium)	1122 W WILSON AVE	CHICAGO	IL	60640.0	12/28/2023	Canvass	Out of Business	NaN	41.965497	-87.658403	(41.96549703 -87.6584028

In [73]:
```
df.shape
```

Out[73]:
```
(265653, 17)
```

Leverage the results of your homework from Week-1 and Week-2 to extract free-form text comments from inspectors

In [74]:
```
# Define a regular expression pattern to extract the comments
comment_pattern = r'Comments:\s*(.*?)(?=\s*\||$)'

# Apply the regular expression and create a new column with only the regulation descriptions
df['comments'] = df['Violations'].str.findall(comment_pattern)

# Display the DataFrame with the new column
df.head(60)
```

Out[74]:

ne	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location	comments
)'S ED	1447643.0	Restaurant	Risk 1 (High)	4540 W 63RD ST	CHICAGO	IL	60629.0	01/04/2024	Non-Inspection	No Entry	NaN	41.778606	-87.736592	(41.77860614662516, -87.73659163763277)	NaN
RIS #3	1964458.0	Restaurant	Risk 1 (High)	7023-7025 S PULASKI RD	CHICAGO	IL	60629.0	01/04/2024	Canvass	Out of Business	NaN	41.765126	-87.722279	(41.765126421774134, -87.72227898862089)	NaN
za LC	2948183.0	NaN	Risk 1 (High)	3517 W DEVON AVE	CHICAGO	IL	60659.0	01/04/2024	License	Not Ready	NaN	41.997177	-87.717388	(41.99717696905116, -87.71738760229222)	NaN
ON OF RK	2944036.0	Children's Services Facility	Risk 1 (High)	2745 N LINCOLN AVE	CHICAGO	IL	60614.0	01/04/2024	License	Pass	NaN	41.931779	-87.657407	(41.93177936843524, -87.65740695216547)	NaN
IC NG	2757521.0	Restaurant	Risk 1 (High)	306 E 75TH ST	CHICAGO	IL	60619.0	01/04/2024	Canvass	Pass	55. PHYSICAL FACILITIES INSTALLED, MAINTAINED ...	41.758467	-87.617974	(41.75846743102775, -87.61797380173462)	[OBSERVED THE FLOORS BEHIND THE COOKING EQUIPM...
AY	1719251.0	Restaurant	Risk 1 (High)	10354 S HALSTED ST	CHICAGO	IL	60628.0	01/04/2024	Canvass	Out of Business	NaN	41.705296	-87.642931	(41.705296492547895, -87.64293095094614)	NaN
HE RE TS	2578250.0	Daycare Above and Under 2 Years	Risk 1 (High)	843-847 W 103RD ST	CHICAGO	IL	60643.0	12/29/2023	License	No Entry	NaN	41.706792	-87.644524	(41.706792177425825, -87.64452365930406)	NaN
NT EP EA	2204412.0	Hospital	Risk 1 (High)	2875 W 19TH ST	CHICAGO	IL	60623.0	12/29/2023	Canvass Re-Inspection	Pass	51. PLUMBING INSTALLED; PROPER BACKFLOW DEVICE...	41.855403	-87.698708	(41.85540289350027, -87.69870754487314)	[EYE WASH STATION INSTALLED ON FAUCET OF HANDW...
GO	2796734.0	Restaurant	Risk 1 (High)	1421 W CHICAGO AVE	CHICAGO	IL	60642.0	12/28/2023	Canvass	Out of Business	NaN	41.896025	-87.663272	(41.89602502708334, -87.66327229348566)	NaN
EA	2886816.0	Restaurant	Risk 2 (Medium)	1122 W WILSON AVE	CHICAGO	IL	60640.0	12/28/2023	Canvass	Out of Business	NaN	41.965497	-87.658403	(41.965497036127346, -87.65840280954839)	NaN
) & ELI	2511672.0	Grocery Store	Risk 2 (Medium)	5601 W MADISON ST	CHICAGO	IL	60644.0	12/27/2023	Canvass Re-Inspection	Pass w/ Conditions	2. CITY OF CHICAGO FOOD SERVICE SANITATION CER...	41.880135	-87.764942	(41.880135062255434, -87.7649424467103)	[OBSERVED NO CITY OF CHICAGO CERTIFIED FOOD SE...
KS	38056.0	Grocery Store	Risk 2 (Medium)	2656 N ELSTON AVE	CHICAGO	IL	60647.0	12/27/2023	Canvass Re-Inspection	Pass	5. PROCEDURES FOR RESPONDING TO VOMITING AND D...	41.929849	-87.684644	(41.929848996596895, -87.684643783527)	[OBSERVED NO CLEAN UP SUPPLIES ON THE PREMISES...
FE	2665159.0	Restaurant	Risk 1 (High)	600 W CHICAGO AVE	CHICAGO	IL	60654.0	12/26/2023	Canvass Re-Inspection	Pass	NaN	41.896585	-87.642996	(41.896585191199556, -87.64299618172501)	NaN
LL	2609690.0	Restaurant	Risk 1 (High)	2246 N MILWAUKEE AVE	CHICAGO	IL	60647.0	12/26/2023	Complaint Re-Inspection	Pass	NaN	41.921614	-87.695481	(41.92161436223205, -87.6954814291355)	NaN

ne	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location	comments
AN NE	2948072.0	NaN	All	3422 N BROADWAY	CHICAGO	IL	60657.0	12/22/2023	License	Not Ready	NaN	41.943797	-87.645266	(41.94379677558918, -87.64526607444076)	NaN
RK NT	2713658.0	Restaurant	Risk 1 (High)	1400 E 47TH DR	CHICAGO	IL	60653.0	12/22/2023	Canvass Re-Inspection	Pass	NaN	41.809795	-87.592374	(41.80979548271692, -87.59237386663331)	NaN
ıut	2202944.0	Restaurant	Risk 2 (Medium)	2111 W DIVISION ST	CHICAGO	IL	60622.0	12/22/2023	Canvass	Pass	NaN	41.903029	-87.680197	(41.903028811141404, -87.68019723624181)	NaN
:'S 36	1276342.0	Restaurant	Risk 2 (Medium)	1701 W DIVISION ST	CHICAGO	IL	60622.0	12/22/2023	Canvass	Pass	NaN	41.903196	-87.669965	(41.90319559301502, -87.66996531168556)	NaN
AN NC	21500.0	Restaurant	Risk 2 (Medium)	1068 W TAYLOR ST	CHICAGO	IL	60607.0	12/22/2023	Non-Inspection	No Entry	NaN	41.869575	-87.653759	(41.86957530510424, -87.65375913135858)	NaN
TO	2757761.0	Restaurant	Risk 1 (High)	4811 N ROCKWELL ST	CHICAGO	IL	60625.0	12/22/2023	Non-Inspection	No Entry	NaN	41.968920	-87.693661	(41.96891957958689, -87.69366142212719)	NaN
RN PS	2125326.0	Restaurant	Risk 2 (Medium)	27 W JACKSON BLVD	CHICAGO	IL	60604.0	12/21/2023	Canvass Re-Inspection	Pass	NaN	41.878064	-87.628696	(41.878064434112346, -87.62869594583626)	NaN
FK	63936.0	Restaurant	Risk 1 (High)	432 W DIVERSEY PKWY	CHICAGO	IL	60614.0	12/21/2023	Canvass	Pass	NaN	41.932995	-87.640640	(41.93299476122033, -87.64064047605578)	NaN
RT IC.	2943732.0	Grocery Store	Risk 2 (Medium)	547 N KEDZIE AVE	CHICAGO	IL	60612.0	12/21/2023	License	Pass	NaN	41.891760	-87.706295	(41.891759531797, -87.7062947815695)	NaN
GA	2157363.0	Restaurant	Risk 1 (High)	2619 W LAWRENCE AVE	CHICAGO	IL	60625.0	12/20/2023	Non-Inspection	No Entry	NaN	41.968454	-87.694620	(41.96845444863656, -87.69461991509863)	NaN
ZA	1245673.0	Restaurant	Risk 1 (High)	2625 W LAWRENCE AVE	CHICAGO	IL	60625.0	12/20/2023	Non-Inspection	No Entry	NaN	41.968452	-87.694875	(41.96845247365508, -87.69487467681566)	NaN
LY ER	2938915.0	Children's Services Facility	Risk 1 (High)	6922-6924 W NORTH AVE	CHICAGO	IL	60707.0	12/21/2023	License Re-Inspection	Pass	NaN	41.909091	-87.798164	(41.90909072583769, -87.79816372279038)	NaN
TY SE	2753635.0	Restaurant	Risk 2 (Medium)	1811 W HARRISON ST	CHICAGO	IL	60612.0	12/20/2023	Canvass	Out of Business	NaN	41.874003	-87.672010	(41.874003274286764, -87.6720101808251)	NaN
EN SS	2943715.0	Restaurant	Risk 1 (High)	1811-1835 W HARRISON ST	CHICAGO	IL	60612.0	12/20/2023	License	Pass	NaN	41.874003	-87.672010	(41.874003274286764, -87.6720101808251)	NaN
TO	2882672.0	Restaurant	Risk 1 (High)	325 E PERSHING RD	CHICAGO	IL	60653.0	12/20/2023	License	Not Ready	NaN	41.823707	-87.618229	(41.82370709190454, -87.61822880259395)	NaN
LY IC.	2911865.0	Restaurant	Risk 1 (High)	554 W PERSHING RD	CHICAGO	IL	60609.0	12/20/2023	Complaint	Pass	NaN	41.823636	-87.640821	(41.82363599819552, -87.64082111787692)	NaN

ne	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location	comments
ɔT #3	2923949.0	Restaurant	Risk 1 (High)	3207 S HALSTED ST	CHICAGO	IL	60608.0	12/19/2023	License	Pass	55. PHYSICAL FACILITIES INSTALLED, MAINTAINED ...	41.835988	-87.646082	(41.83598843774932, -87.64608160187807)	[FOUND FOOD STORED IN BASEMENT. INSTRUCED TO P...
TH NT	1676129.0	Restaurant	Risk 1 (High)	327 S PLYMOUTH CT	CHICAGO	IL	60604.0	12/19/2023	Non-Inspection	No Entry	NaN	41.877438	-87.628582	(41.877437739687544, -87.62858234022409)	NaN
LA JR	2917549.0	Restaurant	Risk 3 (Low)	2237 N WESTERN AVE	CHICAGO	IL	60647.0	12/19/2023	License	Pass	NaN	41.922767	-87.687429	(41.92276684658204, -87.68742931955062)	NaN
ıL, IC.	2831485.0	Restaurant	Risk 1 (High)	4126 W 26TH ST	CHICAGO	IL	60623.0	12/19/2023	Non-Inspection	No Entry	NaN	41.844315	-87.728180	(41.84431470188272, -87.72817991896609)	NaN
ɔA	2943940.0	Restaurant	Risk 3 (Low)	1140 W 18TH ST	CHICAGO	IL	60608.0	12/19/2023	License	Pass	NaN	41.858087	-87.655009	(41.858087069556625, -87.65500878752039)	NaN
SS	1937825.0	Restaurant(protein shake bar)	Risk 2 (Medium)	819 S STATE ST	CHICAGO	IL	60605.0	12/18/2023	Canvass	Out of Business	NaN	41.871535	-87.627371	(41.87153497463235, -87.62737081568564)	NaN
FE	2245656.0	Restaurant	Risk 1 (High)	1559 S LAKE SHORE DR	CHICAGO	IL	60605.0	12/18/2023	Non-Inspection	No Entry	NaN	41.860450	-87.617488	(41.860449933483245, -87.6174881929214)	NaN
NT	2881898.0	Restaurant	Risk 3 (Low)	5405-5409 W MADISON ST	CHICAGO	IL	60644.0	12/18/2023	License	Pass	NaN	41.880195	-87.760339	(41.880195424513424, -87.7603388113785)	NaN
AY RT	40631.0	Daycare (2 - 6 Years)	Risk 1 (High)	6422 S KEDZIE AVE	CHICAGO	IL	60629.0	12/04/2012	Canvass	Pass	34. FLOORS: CONSTRUCTED PER CODE, CLEANED, GOO...	41.776343	-87.703224	(41.7763429708591, -87.70322370755467)	[COVING NEEDED IN STAFF AREA. MUST INSTALL AN...
ı & RT	2647720.0	Grocery Store	Risk 2 (Medium)	3925 W CHICAGO AVE	CHICAGO	IL	60651.0	12/18/2023	Canvass	Out of Business	NaN	41.895268	-87.724800	(41.895268077120654, -87.72480023226892)	NaN
NT	2881896.0	Restaurant	Risk 1 (High)	5405-5409 W MADISON ST	CHICAGO	IL	60644.0	12/18/2023	License	Not Ready	NaN	41.880195	-87.760339	(41.880195424513424, -87.7603388113785)	NaN
ET I7)	2944236.0	Grocery Store	Risk 3 (Low)	11601 W TOUHY AVE	CHICAGO	IL	60666.0	12/18/2023	License	Pass	NaN	42.008536	-87.914428	(42.008536400868735, -87.91442843927047)	NaN
ɾS'	40864.0	Restaurant	Risk 1 (High)	3855 N LINCOLN AVE	CHICAGO	IL	60613.0	12/18/2023	Non-Inspection	No Entry	NaN	41.952030	-87.677110	(41.952030017774646, -87.67710952822847)	NaN
IA, LC	2569202.0	Restaurant	Risk 1 (High)	3508 W 26TH ST	CHICAGO	IL	60623.0	12/15/2023	Canvass Re-Inspection	Fail	59. PREVIOUS PRIORITY FOUNDATION VIOLATION COR...	41.844516	-87.712776	(41.84451619427469, -87.71277572197552)	[PREVIOUS PRIORITY FOUNDATION NOT CORRECTED FR...
FE	2915944.0	Restaurant	Risk 2 (Medium)	7364 N CLARK ST	CHICAGO	IL	60626.0	12/15/2023	License	Pass	NaN	42.015583	-87.675279	(42.015583313516835, -87.67527911063985)	NaN

ne	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location	comments
RS TS	2929766.0	Bakery	Risk 2 (Medium)	2308 S ARCHER AVE	CHICAGO	IL	60616.0	12/15/2023	License Re-Inspection	Pass	10. ADEQUATE HANDWASHING SINKS PROPERLY SUPPLI...	41.850884	-87.639190	(41.85088411291624, -87.6391895226703)	[]
ET	2923803.0	Restaurant	Risk 1 (High)	2019 S LAFLIN ST	CHICAGOO	IL	60608.0	12/15/2023	License Re-Inspection	Pass w/ Conditions	5. PROCEDURES FOR RESPONDING TO VOMITING AND D...	41.854493	-87.663530	(41.854493434030644, -87.66352975090798)	[FOUND SAMPLE VOMIT AND DIARRHEA PROCEDURE ON ...
50	2590022.0	Restaurant	Risk 1 (High)	1850 W ROOSEVELT RD	CHICAGO	IL	60608.0	12/14/2023	Canvass	Fail	10. ADEQUATE HANDWASHING SINKS PROPERLY SUPPLI...	41.866906	-87.673316	(41.86690583756181, -87.6733159162735)	[FOUND CUSTOMER AND EMPLOYEE WOMEN'S AND MEN'S...
IO	2298245.0	Restaurant	Risk 1 (High)	1700 S HALSTED ST	CHICAGO	IL	60608.0	12/14/2023	Non-Inspection	No Entry	NaN	41.859104	-87.646828	(41.85910428354797, -87.64682757730905)	NaN
NA SE	2203561.0	Restaurant	Risk 1 (High)	2300 S DAMEN AVE	CHICAGO	IL	60608.0	12/14/2023	Non-Inspection	No Entry	NaN	41.850316	-87.675936	(41.850315691945504, -87.67593597054196)	NaN
RK	1122751.0	Restaurant	Risk 1 (High)	2149 S HALSTED ST	CHICAGO	IL	60608.0	12/14/2023	Non-Inspection	No Entry	NaN	41.852750	-87.646351	(41.85274953958152, -87.64635124584164)	NaN
TE	2929629.0	Restaurant	Risk 1 (High)	3622 S MORGAN ST	CHICAGO	IL	60609.0	12/14/2023	License	Pass	NaN	41.828141	-87.651052	(41.8281414018716, -87.65105153620306)	NaN
OR RE	2215996.0	Children's Services Facility	Risk 1 (High)	7463 N RIDGE BLVD	CHICAGO	IL	60645.0	12/14/2023	License	Pass	NaN	42.017637	-87.684279	(42.017637330007524, -87.6842790144969)	NaN
KA	68171.0	Restaurant	Risk 2 (Medium)	4785 S ARCHER AVE	CHICAGO	IL	60632.0	12/13/2023	Non-Inspection	No Entry	NaN	41.806547	-87.716305	(41.80654740106753, -87.71630466982813)	NaN
LO ET	2796693.0	Restaurant	Risk 1 (High)	2342 N CLARK ST	CHICAGO	IL	60614.0	12/13/2023	Canvass	Pass	NaN	41.924828	-87.640165	(41.924827819740216, -87.64016496112795)	NaN
RE	2786131.0	Restaurant	Risk 1 (High)	1001 N CALIFORNIA AVE	CHICAGO	IL	60622.0	12/13/2023	Canvass	Pass	NaN	41.899331	-87.696645	(41.899330918472295, -87.69664492473737)	NaN
EN	2845744.0	Restaurant	Risk 1 (High)	2131 N ELSTON AVE	CHICAGO	IL	60614.0	12/13/2023	Canvass	Pass	58. ALLERGEN TRAINING AS REQUIRED - Comments: ...	41.920355	-87.670463	(41.920354885900885, -87.67046254370356)	[OBSERVED NO ALLERGEN TRAINING CERTIFICATE ON ...
OS NG MY	2384881.0	Children's Services Facility	Risk 1 (High)	504-516 E 75TH ST	CHICAGO	IL	60619.0	12/12/2023	Canvass	Fail	10. ADEQUATE HANDWASHING SINKS PROPERLY SUPPLI...	41.758567	-87.612451	(41.758566800153645, -87.61245073862713)	[OBSERVED HOT RUNNING WATER TEMPERATURE AT HAN...

ne	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location	comments
’’S	2938354.0	Restaurant	Risk 3 (Low)	1633-1635 N MILWAUKEE AVE	CHICAGO	IL	60647.0	12/12/2023	License	Pass	NaN	41.911153	-87.678440	(41.91115346303384, -87.67843965041517)	NaN
ED ES	2924591.0	TAVERN	Risk 3 (Low)	7401 S SOUTH CHICAGO AVE	CHICAGO	IL	60619.0	12/12/2023	License Re-Inspection	Pass	NaN	41.760530	-87.597924	(41.76053016467859, -87.59792396776683)	NaN

In [75]:
```
df.loc[47, 'comments']
```

Out[75]:
```
["FOUND CUSTOMER AND EMPLOYEE WOMEN'S AND MEN'S TOILET ROOM WITHOUT HOT WATER. FOUND WATER TEMPERATURES RANGING FROM 80F - 85F. MUST PROVIDE HOT WATER TEMPERATURE OF 100F AT
ALL TIMES. MUST REPAIR AND MAINTAIN. PRIORITY VIOLATION. 7-38-030(C).  CITATION ISSUED",
 'FOUND CITY OF CHICAGO FOOD SERVICE MANAGER WITHOUT PROOF OF ALLERGEN TRAINING ON SITE AT THE TIME OF INSPECTION. INSTRUCTED TO PROVIDE PROOF OF ALLERGEN TRAINING FOR ALL CI
TY OF CHICAGO FOOD SERVICE MANAGERS.']
```

text pre-processing technique

In [76]:
```
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
import numpy as np
stopwords_set = set(stopwords.words('english'))
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

# Function to clean and tokenize text
def process_text(text_list):
    if text_list is None or not isinstance(text_list, list):
        return np.nan

    # Combine the list elements into a single string
    text = ' '.join(str(element) for element in text_list if isinstance(element, str))

    # Convert to lowercase
    text = text.lower()

    # Remove punctuation
    text = ''.join([char for char in text if char not in string.punctuation])

    # Tokenize words
    tokens = word_tokenize(text)

    # Remove stopwords
    stop_words = set(stopwords.words("english"))
    tokens = [word for word in tokens if word not in stop_words]

    # Perform lemmatization
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in tokens]

    # Join the processed tokens back into a string
    processed_text = " ".join(tokens)
    return processed_text
```

In [77]:
```
# Apply text processing to 'comments' column
df['cleaned comments'] = df['comments'].apply(process_text)
df.head()
```

Out[77]:

	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Locat
0	2587633	TACOS MARIO'S LIMITED	TACOS MARIO'S LIMITED	1447643.0	Restaurant	Risk 1 (High)	4540 W 63RD ST	CHICAGO	IL	60629.0	01/04/2024	Non-Inspection	No Entry	NaN	41.778606	-87.736592	(41.778606146625 -87.736591637632
1	2587625	TAQUERIA EL ARCO #3 EL POLLO FELIZ, INC.	EL POLLO CRIS CRIS #3	1964458.0	Restaurant	Risk 1 (High)	7023-7025 S PULASKI RD	CHICAGO	IL	60629.0	01/04/2024	Canvass	Out of Business	NaN	41.765126	-87.722279	(41.7651264217741 -87.722278988620
2	2587598	Tel Aviv Slices Pizza LLC	Tel Aviv Slices Pizza LLC	2948183.0	NaN	Risk 1 (High)	3517 W DEVON AVE	CHICAGO	IL	60659.0	01/04/2024	License	Not Ready	NaN	41.997177	-87.717388	(41.99717696905 -87.717387602292
3	2587596	KENSINGTON SCHOOL OF LINCOLN PARK	KENSINGTON SCHOOL OF LINCOLN PARK	2944036.0	Children's Services Facility	Risk 1 (High)	2745 N LINCOLN AVE	CHICAGO	IL	60614.0	01/04/2024	License	Pass	NaN	41.931779	-87.657407	(41.931779368435 -87.657406952165
4	2587599	Q'S AUTHENTIC COOKING	Q'S AUTHENTIC COOKING	2757521.0	Restaurant	Risk 1 (High)	306 E 75TH ST	CHICAGO	IL	60619.0	01/04/2024	Canvass	Pass	55. PHYSICAL FACILITIES INSTALLED, MAINTAINED ...	41.758467	-87.617974	(41.758467431027 -87.617973801734

In [78]:
```
df.loc[43, 'comments']
```

Out[78]:
```
['PREVIOUS PRIORITY FOUNDATION NOT CORRECTED FROM 12/8/23 REPORT 2586583 NOT CORRECTED. 14- NO WRITTEN AGREEMENT OR STATEMENT FROM THE SUPPLIER STATING THE FISH/SEAFOOD SUPPL
IED FOR SUSHI IS FROZEN TO THE REQUIRED TIME AND TEMPERATURE. INSTD TO PROVIDE THE SAME AND MAINTAIN VERIFICATION ON-SITE. CITATION ISSUED 7-42-090 PRIORITY']
```

In [79]:
```
df.loc[43, 'cleaned comments']
```

Out[79]:
```
'previous priority foundation corrected 12823 report 2586583 corrected 14 written agreement statement supplier stating fishseafood supplied sushi frozen required time tempera
ture instd provide maintain verification onsite citation issued 742090 priority'
```

Explain why you selected a particular text pre-processing technique

Before delving into model building, a series of text pre-processing techniques were executed to optimize the input data for enhanced model performance. The sequential steps involved in this process are as follows:

First of all, the entire text corpus was converted to lowercase. This standardization facilitates case-insensitive matching, promoting consistency across the dataset. And then we decided to eliminate punctuation characters from the text because punctuation typically does not carry much information and can be safely removed to simplify the text without losing important content. For analysis purpose, we also split the text into individual words (tokens), enabling a granular examination of the text. After tokenizing the text, we further removed common English stopwords from the tokenized list, because stopwords are words that are frequently used in a language but often do not contribute much to the meaning of a text. Removing them can reduce noise in the data and improve the efficiency of downstream analyses. Finally, we applied lemmatization to the tokens. This process reduced words to their base or root form, fostering standardization. The resultant effect is an improvement in model performance by curbing the dimensionality of the feature space.

Build a classification model, predicting the outcome of inspection – comments are predictors, target variable is “Results” column

In [80]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df['comments'], df['Results'], test_size=0.2, random_state=42)
```

Naive Bayes

In [81]:
```
# Drop rows with missing values in the relevant columns
df = df.dropna(subset=['comments', 'Results'])

# Flatten the lists within 'comments' column
df['comments'] = df['comments'].apply(lambda x: ' '.join(map(str, x)) if isinstance(x, list) else x)

# Convert comments to TF-IDF features
if df['comments'].dtype != 'object':
    df['comments'] = df['comments'].astype(str)

vectorizer = TfidfVectorizer(max_features=5000)
X_tfidf = vectorizer.fit_transform(df['comments'])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_tfidf, df['Results'], test_size=0.2, random_state=42)

# Train a classifier (Naive Bayes as an example)
classifier = MultinomialNB()
classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = classifier.predict(X_test)

# Evaluate the model
classification_rep = classification_report(y_test, y_pred)
print("Classification Report:\n", classification_rep)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")
```

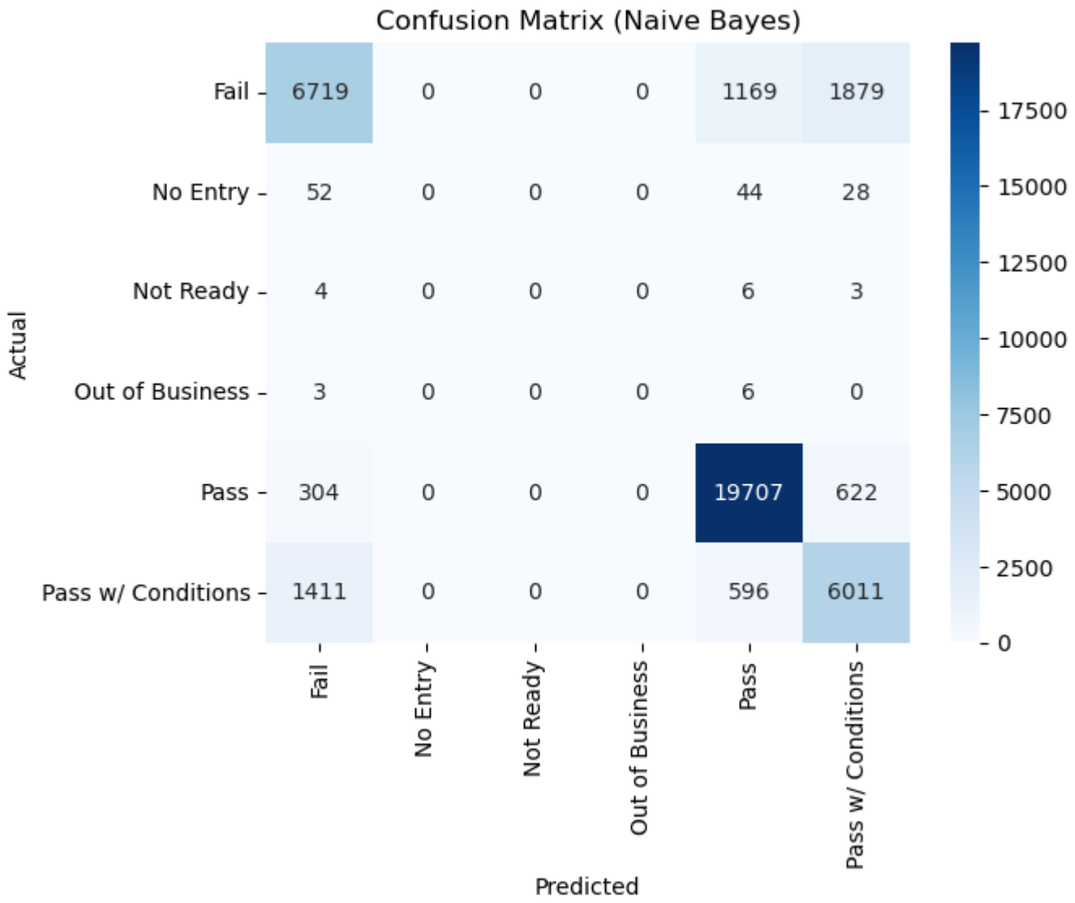
```
/Users/edwinhsu/anaconda3/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1469: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/edwinhsu/anaconda3/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1469: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

```
Classification Report:
                     precision    recall  f1-score   support

               Fail       0.79      0.69      0.74      9767
           No Entry       0.00      0.00      0.00       124
          Not Ready       0.00      0.00      0.00        13
   Out of Business       0.00      0.00      0.00         9
               Pass       0.92      0.96      0.93     20633
Pass w/ Conditions       0.70      0.75      0.73      8018

           accuracy                           0.84     38564
          macro avg       0.40      0.40      0.40     38564
       weighted avg       0.84      0.84      0.84     38564

Accuracy: 0.84
```

```
/Users/edwinhsu/anaconda3/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1469: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [82]:

```python
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Confusion Matrix for Naive Bayes
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=classifier.classes_, yticklabels=classifier.classes_)
plt.title('Confusion Matrix (Naive Bayes)')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

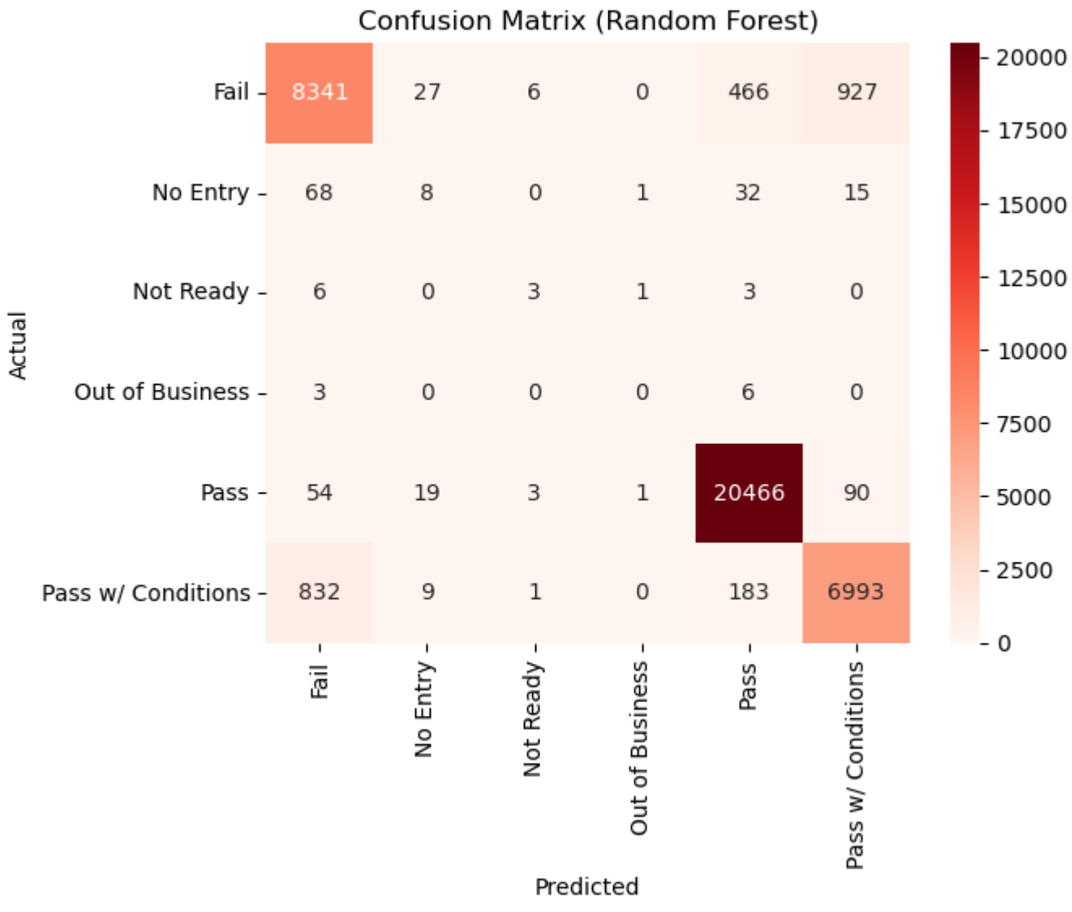


Random Forest

In [83]:
```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Train a Random Forest classifier
classifier = RandomForestClassifier(n_estimators=100, random_state=42)
classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = classifier.predict(X_test)

# Evaluate the model
classification_rep = classification_report(y_test, y_pred)
print("Classification Report:\n", classification_rep)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")
```

```
Classification Report:
                    precision    recall  f1-score   support

              Fail       0.90      0.85      0.87      9767
          No Entry       0.13      0.06      0.09       124
         Not Ready       0.23      0.23      0.23        13
   Out of Business       0.00      0.00      0.00         9
              Pass       0.97      0.99      0.98     20633
Pass w/ Conditions       0.87      0.87      0.87      8018

          accuracy                           0.93     38564
         macro avg       0.52      0.50      0.51     38564
      weighted avg       0.93      0.93      0.93     38564

Accuracy: 0.93
```

In [84]:
```python
# Confusion Matrix for Random Forest
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Reds', xticklabels=classifier.classes_, yticklabels=classifier.classes_)
plt.title('Confusion Matrix (Random Forest)')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```



Visualize results of at least two text classifiers and select the most robust one

Upon examining the classification reports and Confusion Matrices for both the Naive Bayes and Random Forest models, we observed notable distinctions. The Random Forest model demonstrated a superior performance with an accuracy of 93%, surpassing the 84% accuracy achieved by the Naive Bayes model.

While both models exhibited impressive accuracy in predicting the majority of data labeled as "Fail," "Pass," and "Pass w/ Conditions," the Random Forest model excelled further by demonstrating capability in handling minority cases such as "No Entry" and "Not Ready."

Considering the comprehensive evaluation of accuracy and effectiveness across various classifiers, we concluded that the Random Forest model stands out as the most robust choice for this specific scenario.