# abc_segments

March 6, 2022

[1]:
```
import docx
from simplify_docx import simplify
```

[2]:
```
filename = "Analysis_12_30_21_Colorado_Fire_segments.docx"
```

[12]:
```
def docx_to_clean_dict(docx_as_json, first_table_index=1):
    """Takes docx_as_json and cleans it up
    return: list of dicts
      {
        "time": ___,
        "location": ____,
        "station": ____,
        "text": _____________
      }
    """
    clean_data = []

    for blob in docx_as_json['VALUE'][0]['VALUE'][first_table_index:]:
        text_end = False

        if blob['TYPE'] == 'table':
            time = blob['VALUE'][0]['VALUE'][0]['VALUE'][0]['VALUE'][0]['VALUE']
            location =␣
 ↪blob['VALUE'][0]['VALUE'][1]['VALUE'][0]['VALUE'][0]['VALUE']
            station =␣
 ↪blob['VALUE'][0]['VALUE'][2]['VALUE'][0]['VALUE'][0]['VALUE']

        if blob['TYPE'] == 'paragraph':
            text = blob['VALUE'][0]['VALUE']
            text_end = True

        if text_end:
            clean_data.append({
                "time": time,
                "location": location,
                "station": station,
                "text": text
```

```
          })

  return clean_data



def read_docx_to_dict(filename):
    """Reads in docx file and converts it to a list of dicts"""
  # read in a document
  doc = docx.Document(filename)

  # coerce to JSON using the standard options
  docx_as_json = simplify(doc)

  blob_types = [blob['TYPE'] for blob in docx_as_json['VALUE'][0]['VALUE']]

  first_table_index = blob_types.index('table')

  return docx_to_clean_dict(docx_as_json, first_table_index)
```

[14]:

```
data = read_docx_to_dict(filename)
```

[15]:

```
import pandas as pd
pd.options.display.max_rows = 500

# create dataframe
df = pd.DataFrame.from_dict(data)
```

[16]:

```
df['text'] = df['text'].str.lower()
```

[18]:

```
df.head()
```

[18]:

```
                 time                      location station  \
0  2021-12-30 6:14 PM              Salt Lake City    KTVX
1  2021-12-30 6:12 PM          Phoenix (Prescott)    KNXV
2  2021-12-30 6:03 PM          Phoenix (Prescott)    KNXV
3  2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4  2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS

                                                text
0  u.s. many of these travel troubles will likely…
1  500, 80 homes and businesses destroyed. it is …
2  reporting live from downtown flagstaff. luzdel…
3  injured. the sheriff did not rule out the poss…
4  need this time. it's outside gambling groups a…
```

[19]:

```
import time
from thefuzz import fuzz

def check_text_likeness(df, text, ratio=85, row_name='text'):
    """For a given dataframe (df), loop through the column (row_name),
    calculate the partial ratio between given text (text) and the text in each␣
 ↪row,
    and return the indexes where the partial ratio is greater than or equal to␣
 ↪the ratio
    """
    matches = df.apply(lambda row: (fuzz.partial_ratio(row[row_name], text) >=␣
 ↪ratio), axis=1)
    return [i for i, x in enumerate(matches) if x]

# def extract_similar_texts(df, text, ratio=85, row_name='text'):
#     start = time.time()
#     start_time = time.strftime("%a, %d %b %Y %H:%M:%S", time.localtime())
#     print(f'start time: {start_time}')

#     check_text_likeness(df, text, ratio=85, row_name='text')

#     end = time.time()
#     minutes = (end - start)/60.0
#     end_time = time.strftime("%a, %d %b %Y %H:%M:%S", time.localtime())
#     print(f'end time: {end_time} -- took {minutes} minutes')
```

[26]:

```
df['matches'] = df.apply(lambda row: check_text_likeness(df, row['text']),␣
 ↪axis=1)
```

[29]:

```
def fetch_biggest_text(idx_list):
    """For the rows with similar text, fetch the biggest text's index"""
    biggest_length = 0
    idx = None

    if len(idx_list) == 1:
        return idx_list[0]

    for i in idx_list:
        current_length = len(df['text'][i])
        if current_length > biggest_length:
            biggest_length = current_length
            idx = i
    return idx
```

[30]:

```
df['row_to_use'] = df.apply(lambda row: fetch_biggest_text(row['matches']),␣
 ↪axis=1)
```

[32]:
```
def mark_use_row():
    # mark rows to use
    idxs = list(df['row_to_use'].unique())

    for index, row in df.iterrows():
        df.at[index,'use_row'] = index in idxs

    return 'done'
```

[34]:
```
mark_use_row()
```

[34]:
```
'done'
```

[64]:
```
df.head(20)
```

[64]:
```
                  time                    location station  \
0   2021-12-30 6:14 PM              Salt Lake City    KTVX
1   2021-12-30 6:12 PM          Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM          Phoenix (Prescott)    KNXV
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                Jacksonville    WJXX
6   2021-12-30 5:54 PM                Jacksonville    WJXX
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                    New York    WABC
9   2021-12-30 5:38 PM                 Los Angeles    KABC
10  2021-12-30 5:37 PM                 Los Angeles    KABC
11  2021-12-30 5:35 PM            Dallas-Ft. Worth    WFAA
12  2021-12-30 5:35 PM              Seattle-Tacoma    KOMO
13  2021-12-30 5:35 PM              Seattle-Tacoma    KOMO
14  2021-12-30 5:35 PM               Oklahoma City    KOCO
15  2021-12-30 5:35 PM                   Milwaukee    WISN
16  2021-12-30 5:35 PM                     Houston    KTRK
17  2021-12-30 5:35 PM                   Las Vegas    KTNV
18  2021-12-30 5:35 PM                   Las Vegas    KTNV
19  2021-12-30 5:35 PM                 Kansas City    KMBC

                                                 text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
```

```
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
11  senior meteorologist rob marciano, and rob, yo…
12                     attle-tacoma2021-12-30 5:35 pm
13  concern. will, thank you. let's get right to a…
14  >>> let's get right to abc's senior meteorolog…
15  senior meteorologist rob marciano, and rob, yo…
16  senior meteorologist rob marciano, and rob, yo…
17                         as vegas2021-12-30 5:35 pm
18  >> until then, still so much concern. will, th…
19  senior meteorologist rob marciano, and rob, yo…

                                              matches  row_to_use use_row
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
10                                               [10]          10    True
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
12                                       [12, 25, 49]          12    True
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
17                                       [17, 25, 49]          17    True
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
```

[68]:

```
df['words'] = df['text'].str.lower().str.replace(',', '').str.replace('>', '').
 ↪str.replace('.', '').str.replace('\n', '').str.replace('’', "'").str.replace(
   '!', '').str.replace('?', '').str.replace('%', '').str.replace(')', '').str.
 ↪replace('(', '').str.replace('_', '').str.replace(':', '').str.strip().str.
 ↪split(' ')
```

```
/Users/loren/.pyenv/versions/3.7.4/lib/python3.7/site-
packages/ipykernel_launcher.py:1: FutureWarning: The default value of regex will
change from True to False in a future version. In addition, single character
regular expressions will *not* be treated as literal strings when regex=True.
  """Entry point for launching an IPython kernel.
/Users/loren/.pyenv/versions/3.7.4/lib/python3.7/site-
packages/ipykernel_launcher.py:2: FutureWarning: The default value of regex will
change from True to False in a future version. In addition, single character
```

```
regular expressions will *not* be treated as literal strings when regex=True.
```

[69]:
```
df.head(20)
```

[69]:
```
                   time                     location station  \
0   2021-12-30 6:14 PM               Salt Lake City    KTVX
1   2021-12-30 6:12 PM           Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM           Phoenix (Prescott)    KNXV
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                 Jacksonville    WJXX
6   2021-12-30 5:54 PM                 Jacksonville    WJXX
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                     New York    WABC
9   2021-12-30 5:38 PM                  Los Angeles    KABC
10  2021-12-30 5:37 PM                  Los Angeles    KABC
11  2021-12-30 5:35 PM             Dallas-Ft. Worth    WFAA
12  2021-12-30 5:35 PM               Seattle-Tacoma    KOMO
13  2021-12-30 5:35 PM               Seattle-Tacoma    KOMO
14  2021-12-30 5:35 PM                Oklahoma City    KOCO
15  2021-12-30 5:35 PM                    Milwaukee    WISN
16  2021-12-30 5:35 PM                      Houston    KTRK
17  2021-12-30 5:35 PM                    Las Vegas    KTNV
18  2021-12-30 5:35 PM                    Las Vegas    KTNV
19  2021-12-30 5:35 PM                  Kansas City    KMBC

                                                 text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
11  senior meteorologist rob marciano, and rob, yo…
12                    attle-tacoma2021-12-30 5:35 pm
13  concern. will, thank you. let's get right to a…
14  >>> let's get right to abc's senior meteorolog…
15  senior meteorologist rob marciano, and rob, yo…
16  senior meteorologist rob marciano, and rob, yo…
17                        as vegas2021-12-30 5:35 pm
18  >> until then, still so much concern. will, th…
```

```
19  senior meteorologist rob marciano, and rob, yo…

                                              matches  row_to_use use_row  \
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
10                                               [10]          10    True
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
12                                       [12, 25, 49]          12    True
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
17                                       [17, 25, 49]          17    True
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False

                                                words
0   [us, many, of, these, travel, troubles, will, …
1   [500, 80, homes, and, businesses, destroyed, i…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, , po…
9   [exploding, in, size, , burping, more, than, t…
10  [12, story, apartment, building, collapsing, o…
11  [senior, meteorologist, rob, marciano, and, ro…
12                  [attle-tacoma2021-12-30, 535, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                     [as, vegas2021-12-30, 535, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…
```

[ ]:

```
import sys
sys.path.append('../')
```

[71]:

```
from helpers.utils import parse_words
df['clean_words'] = df.apply(lambda row: parse_words(row['words']), axis=1)
```

[72]:

```
df.head(20)
```

[72]:

```
                   time                      location station  \
0   2021-12-30 6:14 PM                Salt Lake City    KTVX
1   2021-12-30 6:12 PM            Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM            Phoenix (Prescott)    KNXV
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                  Jacksonville    WJXX
6   2021-12-30 5:54 PM                  Jacksonville    WJXX
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                      New York    WABC
9   2021-12-30 5:38 PM                   Los Angeles    KABC
10  2021-12-30 5:37 PM                   Los Angeles    KABC
11  2021-12-30 5:35 PM              Dallas-Ft. Worth    WFAA
12  2021-12-30 5:35 PM                Seattle-Tacoma    KOMO
13  2021-12-30 5:35 PM                Seattle-Tacoma    KOMO
14  2021-12-30 5:35 PM                 Oklahoma City    KOCO
15  2021-12-30 5:35 PM                     Milwaukee    WISN
16  2021-12-30 5:35 PM                       Houston    KTRK
17  2021-12-30 5:35 PM                     Las Vegas    KTNV
18  2021-12-30 5:35 PM                     Las Vegas    KTNV
19  2021-12-30 5:35 PM                   Kansas City    KMBC

                                                 text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
11  senior meteorologist rob marciano, and rob, yo…
12                     attle-tacoma2021-12-30 5:35 pm
13  concern. will, thank you. let's get right to a…
14  >>> let's get right to abc's senior meteorolog…
15  senior meteorologist rob marciano, and rob, yo…
```

```
16  senior meteorologist rob marciano, and rob, yo…
17                         as vegas2021-12-30 5:35 pm
18  >> until then, still so much concern. will, th…
19  senior meteorologist rob marciano, and rob, yo…

                                              matches  row_to_use use_row  \
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
10                                               [10]          10    True
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
12                                       [12, 25, 49]          12    True
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
17                                       [17, 25, 49]          17    True
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False

                                                words  \
0   [us, many, of, these, travel, troubles, will, …
1   [500, 80, homes, and, businesses, destroyed, i…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, , po…
9   [exploding, in, size, , burping, more, than, t…
10  [12, story, apartment, building, collapsing, o…
11  [senior, meteorologist, rob, marciano, and, ro…
12                  [attle-tacoma2021-12-30, 535, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                     [as, vegas2021-12-30, 535, pm]
18  [until, then, still, so, much, concern, will, …
```

```
19  [senior, meteorologist, rob, marciano, and, ro…

                                          clean_words
0   [us, many, of, these, travel, troubles, will, …
1   [homes, and, businesses, destroyed, it, is, un…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, poli…
9   [exploding, in, size, burping, more, than, twi…
10  [story, apartment, building, collapsing, outsi…
11  [senior, meteorologist, rob, marciano, and, ro…
12                    [attle, tacoma2021, 12, 30, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                        [as, vegas2021, 12, 30, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…
```

[56]:

```

```

[60]:

```
from helpers.utils import fetch_climate_words_in_words,␣
 ↪fetch_climate_phrases_in_text
```

[62]:

```
fetch_climate_words_in_words(["adapt","for", "climate", "change"])
# segment_df['climate_words_found'] = segment_df.apply(lambda row:␣
 ↪fetch_climate_words_in_text(row['clean_words']), axis=1)
```

[62]:

```
['adapt', 'climate']
```

[63]:

```
fetch_climate_phrases_in_text("adapt for climate change")
```

[63]:

```
['climate change']
```

[73]:

```
df['climate_words_found'] = df.apply(lambda row:␣
 ↪fetch_climate_words_in_words(row['clean_words']), axis=1)
```

[74]:

```
df.head(20)
```

[74]:

```
                   time                   location station  \
0   2021-12-30 6:14 PM             Salt Lake City    KTVX
1   2021-12-30 6:12 PM         Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM         Phoenix (Prescott)    KNXV
```

```
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                Jacksonville    WJXX
6   2021-12-30 5:54 PM                Jacksonville    WJXX
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                    New York    WABC
9   2021-12-30 5:38 PM                 Los Angeles    KABC
10  2021-12-30 5:37 PM                 Los Angeles    KABC
11  2021-12-30 5:35 PM            Dallas-Ft. Worth    WFAA
12  2021-12-30 5:35 PM              Seattle-Tacoma    KOMO
13  2021-12-30 5:35 PM              Seattle-Tacoma    KOMO
14  2021-12-30 5:35 PM               Oklahoma City    KOCO
15  2021-12-30 5:35 PM                   Milwaukee    WISN
16  2021-12-30 5:35 PM                     Houston    KTRK
17  2021-12-30 5:35 PM                   Las Vegas    KTNV
18  2021-12-30 5:35 PM                   Las Vegas    KTNV
19  2021-12-30 5:35 PM                 Kansas City    KMBC

                                                 text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
11  senior meteorologist rob marciano, and rob, yo…
12                     attle-tacoma2021-12-30 5:35 pm
13  concern. will, thank you. let's get right to a…
14  >>> let's get right to abc's senior meteorolog…
15  senior meteorologist rob marciano, and rob, yo…
16  senior meteorologist rob marciano, and rob, yo…
17                         as vegas2021-12-30 5:35 pm
18  >> until then, still so much concern. will, th…
19  senior meteorologist rob marciano, and rob, yo…

                                              matches  row_to_use use_row  \
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
```

```
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
10                                               [10]          10    True
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
12                                       [12, 25, 49]          12    True
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
17                                       [17, 25, 49]          17    True
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False

                                                words  \
0   [us, many, of, these, travel, troubles, will, …
1   [500, 80, homes, and, businesses, destroyed, i…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, , po…
9   [exploding, in, size, , burping, more, than, t…
10  [12, story, apartment, building, collapsing, o…
11  [senior, meteorologist, rob, marciano, and, ro…
12                  [attle-tacoma2021-12-30, 535, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                     [as, vegas2021-12-30, 535, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…

                                          clean_words  \
0   [us, many, of, these, travel, troubles, will, …
1   [homes, and, businesses, destroyed, it, is, un…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, poli…
```

```
9   [exploding, in, size, burping, more, than, twi…
10  [story, apartment, building, collapsing, outsi…
11  [senior, meteorologist, rob, marciano, and, ro…
12                     [attle, tacoma2021, 12, 30, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                         [as, vegas2021, 12, 30, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…

                                 climate_words_found
0                          [life-threatening, record]
1                                    [effect, severe]
2                                                  []
3                                                  []
4                                              [high]
5   [climate, history, hottest, record, unpreceden…
6   [climate, drought, extreme, historic, policy, …
7                                              [high]
8                                               [gas]
9                                            [record]
10  [arctic, climate, crisis, drought, extreme, gl…
11           [high, historic, meteorologist, severe]
12                                                 []
13           [high, historic, meteorologist, severe]
14           [high, historic, meteorologist, severe]
15           [high, historic, meteorologist, severe]
16           [high, historic, meteorologist, severe]
17                                                 []
18           [high, historic, meteorologist, severe]
19           [high, historic, meteorologist, severe]
```

[75]:
```
df['climate_phrases_found'] = df.apply(lambda row:␣
 ↪fetch_climate_phrases_in_text(row['text']), axis=1)
```

[76]:
```
df.head(20)
```

[76]:
```
                  time                        location station  \
0   2021-12-30 6:14 PM                  Salt Lake City    KTVX
1   2021-12-30 6:12 PM              Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM              Phoenix (Prescott)    KNXV
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                    Jacksonville    WJXX
6   2021-12-30 5:54 PM                    Jacksonville    WJXX
```

```
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                    New York    WABC
9   2021-12-30 5:38 PM                 Los Angeles    KABC
10  2021-12-30 5:37 PM                 Los Angeles    KABC
11  2021-12-30 5:35 PM           Dallas-Ft. Worth    WFAA
12  2021-12-30 5:35 PM             Seattle-Tacoma    KOMO
13  2021-12-30 5:35 PM             Seattle-Tacoma    KOMO
14  2021-12-30 5:35 PM              Oklahoma City    KOCO
15  2021-12-30 5:35 PM                  Milwaukee    WISN
16  2021-12-30 5:35 PM                    Houston    KTRK
17  2021-12-30 5:35 PM                  Las Vegas    KTNV
18  2021-12-30 5:35 PM                  Las Vegas    KTNV
19  2021-12-30 5:35 PM                Kansas City    KMBC

                                                 text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
11  senior meteorologist rob marciano, and rob, yo…
12                     attle-tacoma2021-12-30 5:35 pm
13  concern. will, thank you. let's get right to a…
14  >>> let's get right to abc's senior meteorolog…
15  senior meteorologist rob marciano, and rob, yo…
16  senior meteorologist rob marciano, and rob, yo…
17                         as vegas2021-12-30 5:35 pm
18  >> until then, still so much concern. will, th…
19  senior meteorologist rob marciano, and rob, yo…

                                              matches  row_to_use use_row  \
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
```

```
10                                               [10]          10    True
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
12                                       [12, 25, 49]          12    True
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
17                                       [17, 25, 49]          17    True
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28   False

                                                words  \
0   [us, many, of, these, travel, troubles, will, …
1   [500, 80, homes, and, businesses, destroyed, i…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, , po…
9   [exploding, in, size, , burping, more, than, t…
10  [12, story, apartment, building, collapsing, o…
11  [senior, meteorologist, rob, marciano, and, ro…
12                  [attle-tacoma2021-12-30, 535, pm]
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                     [as, vegas2021-12-30, 535, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…

                                          clean_words  \
0   [us, many, of, these, travel, troubles, will, …
1   [homes, and, businesses, destroyed, it, is, un…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, poli…
9   [exploding, in, size, burping, more, than, twi…
10  [story, apartment, building, collapsing, outsi…
11  [senior, meteorologist, rob, marciano, and, ro…
12                    [attle, tacoma2021, 12, 30, pm]
```

```
13  [concern, will, thank, you, let's, get, right,…
14  [let's, get, right, to, abc's, senior, meteoro…
15  [senior, meteorologist, rob, marciano, and, ro…
16  [senior, meteorologist, rob, marciano, and, ro…
17                        [as, vegas2021, 12, 30, pm]
18  [until, then, still, so, much, concern, will, …
19  [senior, meteorologist, rob, marciano, and, ro…

                                  climate_words_found  \
0                          [life-threatening, record]
1                                    [effect, severe]
2                                                  []
3                                                  []
4                                              [high]
5   [climate, history, hottest, record, unpreceden…
6   [climate, drought, extreme, historic, policy, …
7                                              [high]
8                                               [gas]
9                                            [record]
10  [arctic, climate, crisis, drought, extreme, gl…
11            [high, historic, meteorologist, severe]
12                                                 []
13            [high, historic, meteorologist, severe]
14            [high, historic, meteorologist, severe]
15            [high, historic, meteorologist, severe]
16            [high, historic, meteorologist, severe]
17                                                 []
18            [high, historic, meteorologist, severe]
19            [high, historic, meteorologist, severe]

               climate_phrases_found
0                  [record breaking]
1                                 []
2                                 []
3                                 []
4                                 []
5                   [climate change]
6                   [climate change]
7                                 []
8                                 []
9                                 []
10  [climate crisis, global climate]
11                                []
12                                []
13                                []
14                                []
15                                []
```

```
16                                  []
17                                  []
18                                  []
19                                  []
```

```
[112]: # save data to csv
       df.to_csv('reports/abc_all.csv', encoding='utf-8')
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[78]: unique_df = df[df['use_row']]
```

```
[79]: unique_df
```

```
[79]:
                  time                    location station  \
0   2021-12-30 6:14 PM              Salt Lake City    KTVX
1   2021-12-30 6:12 PM          Phoenix (Prescott)    KNXV
2   2021-12-30 6:03 PM          Phoenix (Prescott)    KNXV
3   2021-12-30 6:00 PM  San Francisco-Oak-San Jose     KGO
4   2021-12-30 5:54 PM   Tampa-St. Pete (Sarasota)    WFTS
5   2021-12-30 5:54 PM                Jacksonville    WJXX
6   2021-12-30 5:54 PM                Jacksonville    WJXX
7   2021-12-30 5:48 PM   Washington, DC (Hagrstwn)    WJLA
8   2021-12-30 5:42 PM                    New York    WABC
9   2021-12-30 5:38 PM                 Los Angeles    KABC
10  2021-12-30 5:37 PM                 Los Angeles    KABC
12  2021-12-30 5:35 PM              Seattle-Tacoma    KOMO
17  2021-12-30 5:35 PM                   Las Vegas    KTNV
20  2021-12-30 5:35 PM                       Tulsa    KTUL
23  2021-12-30 5:35 PM        Minneapolis-St. Paul    KSTP
28  2021-12-30 5:35 PM                Portland, OR    KATU
38  2021-12-30 5:34 PM                   Las Vegas    KTNV
42  2021-12-30 5:34 PM               Oklahoma City    KOCO
55  2021-12-30 5:34 PM  Raleigh-Durham (Fayetvlle)    WTVD
67  2021-12-30 5:32 PM                   Milwaukee    WISN
68  2021-12-30 5:32 PM                     Houston    KTRK

                                               text  \
0   u.s. many of these travel troubles will likely…
1   500, 80 homes and businesses destroyed. it is …
2   reporting live from downtown flagstaff. luzdel…
3   injured. the sheriff did not rule out the poss…
4   need this time. it's outside gambling groups a…
```

```
5   northwest. it was sporadic weather patterns co…
6   be evacuated because there are several grass-f…
7   living with a risk factor they may not know ab…
8   but jeff says too little too late. >>> police …
9   exploding in size. >> burping more than twice …
10  >> 12 story apartment building collapsing outs…
12                     attle-tacoma2021-12-30 5:35 pm
17                         as vegas2021-12-30 5:35 pm
20                           ltulsa2021-12-30 5:35 pm
23                   polis-st. paul2021-12-30 5:35 pm
28  it is supposed to start snowing in the morning…
38  to heed the warning and get out. >> i have a p…
42                    oklahoma city2021-12-30 5:34 pm
55  the rescheduled game. >> i don't like it. disr…
67  murder in texas. the teenager considered armed…
68                             ston2021-12-30 5:32 pm

                                              matches  row_to_use use_row  \
0                                                 [0]           0    True
1                                                 [1]           1    True
2                                                 [2]           2    True
3                                                 [3]           3    True
4                                                 [4]           4    True
5                                                 [5]           5    True
6                                                 [6]           6    True
7                                                 [7]           7    True
8                                                 [8]           8    True
9                                                 [9]           9    True
10                                               [10]          10    True
12                                       [12, 25, 49]          12    True
17                                       [17, 25, 49]          17    True
20                                       [20, 25, 49]          20    True
23                                       [23, 25, 49]          23    True
28  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…          28    True
38  [37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 5…          38    True
42                                       [25, 42, 49]          42    True
55                                               [55]          55    True
67  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…          67    True
68                                       [25, 49, 68]          68    True

                                                words  \
0   [us, many, of, these, travel, troubles, will, …
1   [500, 80, homes, and, businesses, destroyed, i…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
```

```
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, , po…
9   [exploding, in, size, , burping, more, than, t…
10  [12, story, apartment, building, collapsing, o…
12                  [attle-tacoma2021-12-30, 535, pm]
17                     [as, vegas2021-12-30, 535, pm]
20                        [ltulsa2021-12-30, 535, pm]
23                [polis-st, paul2021-12-30, 535, pm]
28  [it, is, supposed, to, start, snowing, in, the…
38  [to, heed, the, warning, and, get, out, , i, h…
42                [oklahoma, city2021-12-30, 534, pm]
55  [the, rescheduled, game, , i, don't, like, it,…
67  [murder, in, texas, the, teenager, considered,…
68                           [ston2021-12-30, 532, pm]

                                          clean_words  \
0   [us, many, of, these, travel, troubles, will, …
1   [homes, and, businesses, destroyed, it, is, un…
2   [reporting, live, from, downtown, flagstaff, l…
3   [injured, the, sheriff, did, not, rule, out, t…
4   [need, this, time, it's, outside, gambling, gr…
5   [northwest, it, was, sporadic, weather, patter…
6   [be, evacuated, because, there, are, several, …
7   [living, with, a, risk, factor, they, may, not…
8   [but, jeff, says, too, little, too, late, poli…
9   [exploding, in, size, burping, more, than, twi…
10  [story, apartment, building, collapsing, outsi…
12                    [attle, tacoma2021, 12, 30, pm]
17                        [as, vegas2021, 12, 30, pm]
20                           [ltulsa2021, 12, 30, pm]
23                  [polis, st, paul2021, 12, 30, pm]
28  [it, is, supposed, to, start, snowing, in, the…
38  [to, heed, the, warning, and, get, out, i, hav…
42                   [oklahoma, city2021, 12, 30, pm]
55  [the, rescheduled, game, i, don't, like, it, d…
67  [murder, in, texas, the, teenager, considered,…
68                             [ston2021, 12, 30, pm]

                                  climate_words_found  \
0                          [life-threatening, record]
1                                    [effect, severe]
2                                                  []
3                                                  []
4                                              [high]
5   [climate, history, hottest, record, unpreceden…
6   [climate, drought, extreme, historic, policy, …
```

```
7                                              [high]
8                                               [gas]
9                                            [record]
10  [arctic, climate, crisis, drought, extreme, gl…
12                                                 []
17                                                 []
20                                                 []
23                                                 []
28                  [historic, meteorologist, severe]
38                                   [record-setting]
42                                                 []
55                                             [high]
67                                 [life-threatening]
68                                                 []

               climate_phrases_found
0                  [record breaking]
1                                 []
2                                 []
3                                 []
4                                 []
5                   [climate change]
6                   [climate change]
7                                 []
8                                 []
9                                 []
10  [climate crisis, global climate]
12                                []
17                                []
20                                []
23                                []
28                                []
38                                []
42                                []
55                                []
67                                []
68                                []
```

[80]:

```
total_words = unique_df['clean_words'].str.len().sum()
total_words
```

[80]:

```
2427
```

[81]:

```
def words_found_master_list(df_clean_words):
    """Given a column of words, aggregate master list"""
    words_found = list()
    for chunk in df_clean_words:
        words_found += chunk

    return words_found
```

[86]:

```
words_found = words_found_master_list(unique_df['clean_words'])
len(words_found)
```

[86]: 2427

[96]:

```
from nltk.corpus import stopwords
from nltk import pos_tag
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from wordcloud import STOPWORDS

def master_stopwords_list():
    """Creates a master list of stopwords from pre-existing stopwords found in nltk and wordcloud"""
    stop_words = set(stopwords.words("english"))
    final_stopwords = list(STOPWORDS) + list(stop_words)
    return [i.lower() for i in set(final_stopwords)]

def lemmatize_words(words):
    """Given a list of words, distill to root words"""
    lem = WordNetLemmatizer()

    lemma_list = []
    for word, tag in pos_tag(words):
        wntag = tag[0].lower()
        wntag = wntag if wntag in ['a', 'r', 'n', 'v'] else None
        if not wntag:
            lemma = word
        else:
            lemma = lem.lemmatize(word, pos=wntag)
        lemma_list.append(lemma)
    return lemma_list

def clean_lemmatized_words(lemma_words):
    """Removes stop words from the lemma list"""
    nonstop_lemma_words = []
    final_stopwords = master_stopwords_list()

    for word in lemma_words:
        if word not in final_stopwords:
            nonstop_lemma_words.append(word)

    return list(filter(None, nonstop_lemma_words))
```

[100]:

```
clean_lemma_words = clean_lemmatized_words(lemmatize_words(words_found))
```

[101]:

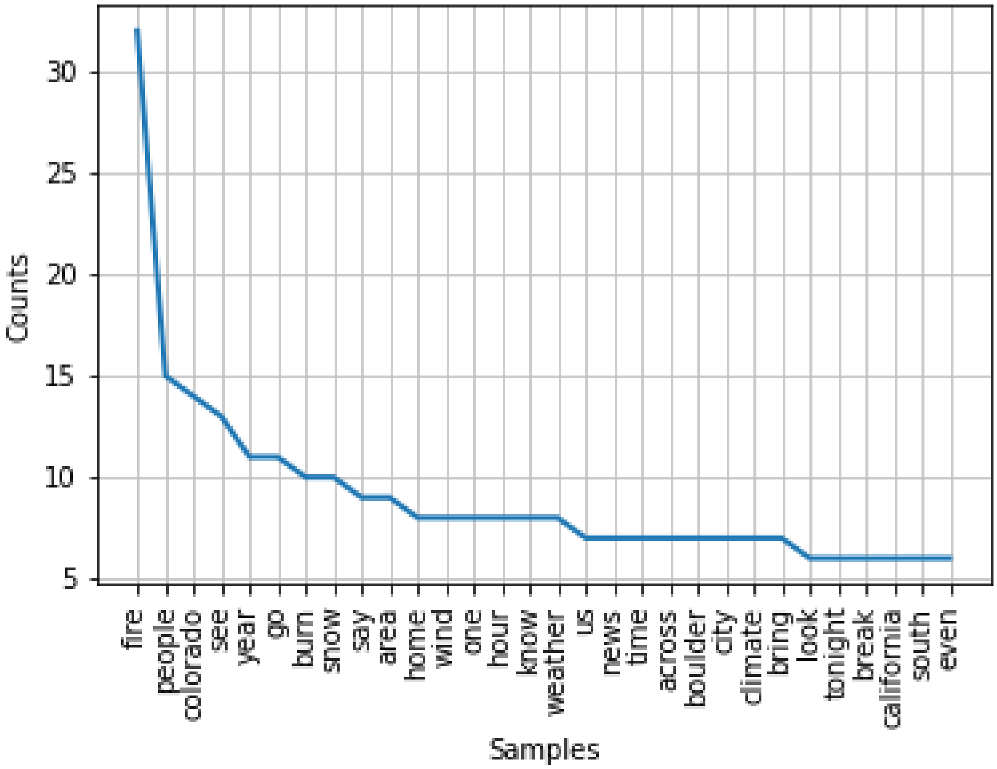
```
from nltk.probability import FreqDist

lfdist = FreqDist(clean_lemma_words)
lfdist
```

[101]:

```
FreqDist({'fire': 32, 'people': 15, 'colorado': 14, 'see': 13, 'year': 11, 'go':
11, 'burn': 10, 'snow': 10, 'say': 9, 'area': 9, …})
```

[102]:

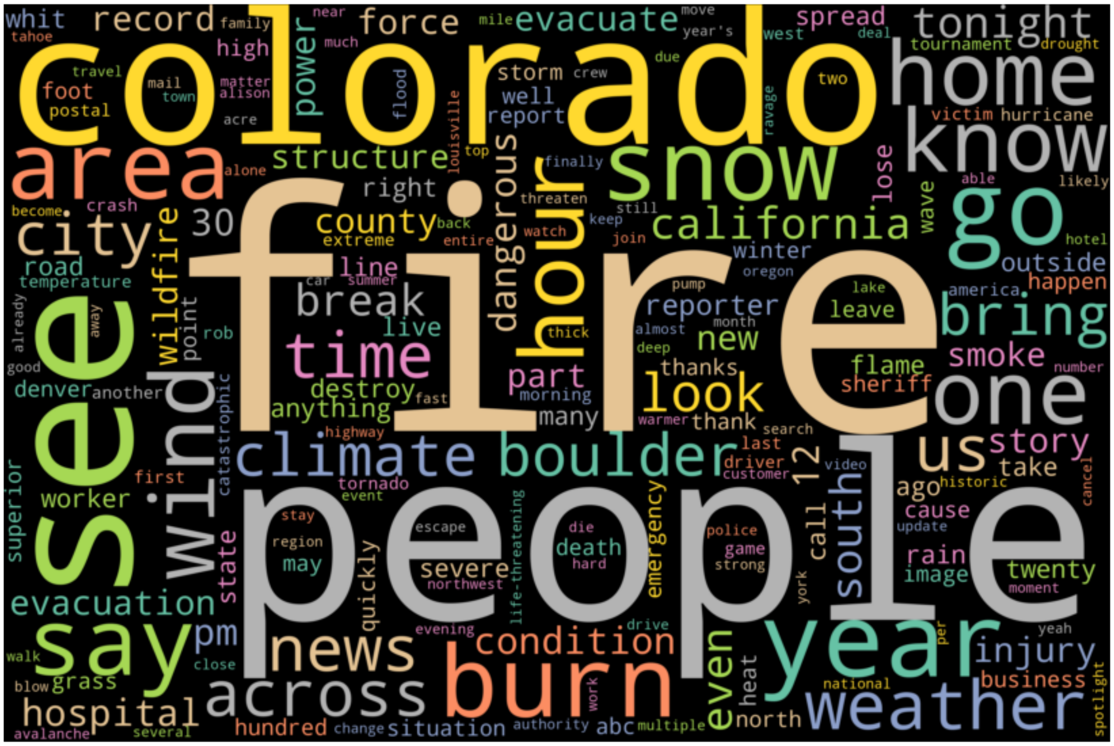
```
import matplotlib.pyplot as plt
lfdist.plot(30,cumulative=False)
plt.show()
```



[103]:

```
from wordcloud import WordCloud
from wordcloud import ImageColorGenerator
from wordcloud import STOPWORDS
import matplotlib.pyplot as plt

wordcloud = WordCloud(width = 3000, height = 2000, random_state=1,␣
 ↪background_color='black', colormap='Set2', collocations=False, stopwords =␣
 ↪master_stopwords_list()).generate_from_frequencies(lfdist)
```

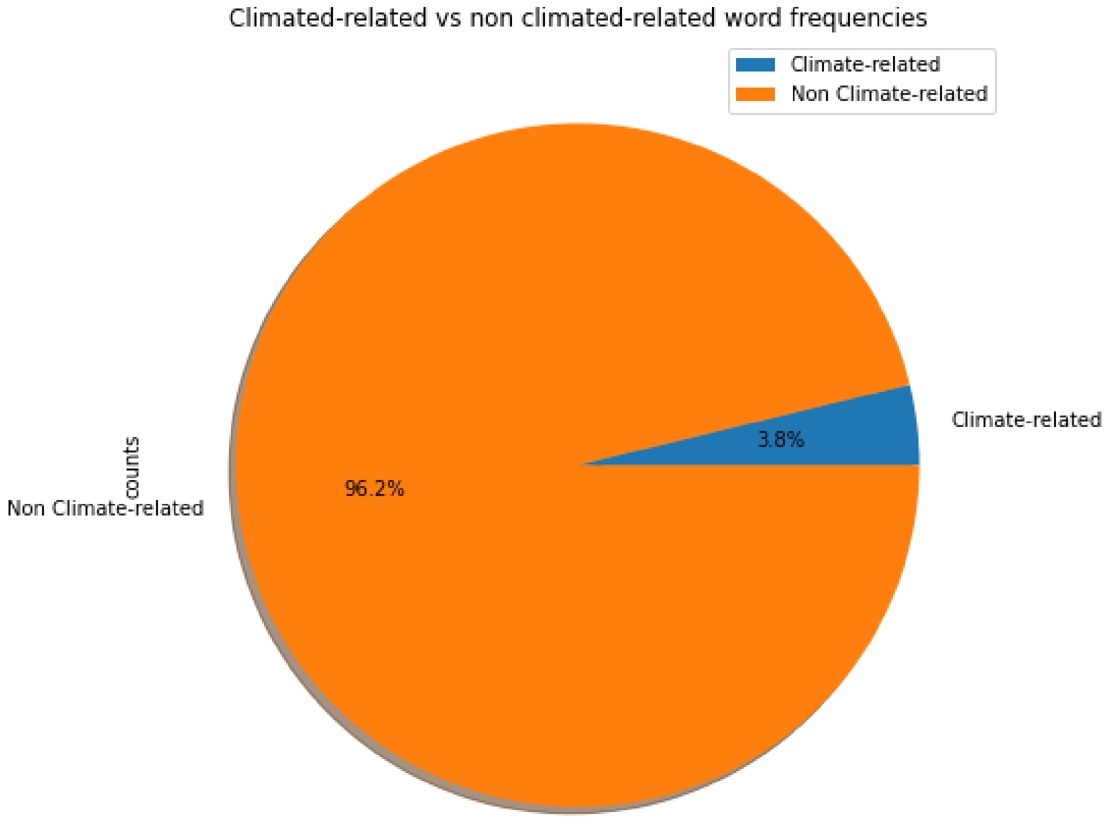
```
# Plot
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()

#plt.savefig('word_cloud.png')
```



[104]:
```
import pandas as pd
pd.options.display.max_rows = 500
words_df = pd.DataFrame(lfdist.items(), columns=['Word', 'Count'])

words_df.sort_values(by=['Count'], ascending=False, inplace=True)
len(words_df)
# 1374 total words

words_df['Count'].sum()

# create data
climate_change_words_df = words_df.loc[words_df['Word'].
 ↪isin(CLIMATE_CHANGE_RELATED_WORDS)]
```

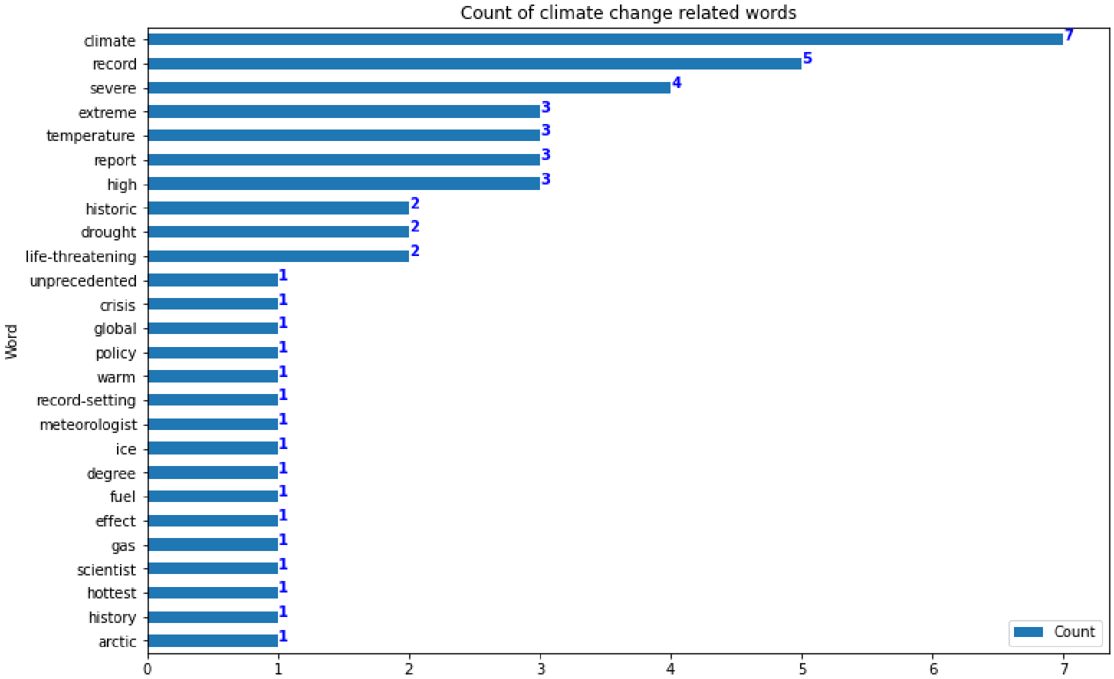
```
climate_words_count = climate_change_words_df['Count'].sum()
non_climate_words_count = words_df['Count'].sum() - climate_words_count

comparison_df = pd.DataFrame({'Words': ['Climate-related', 'Non␣
 ↪Climate-related'],
                              'counts': [climate_words_count,␣
 ↪non_climate_words_count]})
comparison_df.set_index('Words', inplace=True)
print(comparison_df)

plot = comparison_df.plot.pie(y='counts', title="Climated-related vs non␣
 ↪climated-related word frequencies", legend=True, autopct='%1.1f%%',␣
 ↪shadow=True, figsize=(8, 8))

fig = plot.get_figure()
#fig.savefig("comparison.png")
```

```
                     counts
Words
Climate-related          50
Non Climate-related    1270
```



[105]:

```
# find climate related word frequencies

# set figure size
fig, ax = plt.subplots(figsize=(12, 8))
# plot horizontal bar plot
climate_change_words_df.sort_values(by='Count').plot.barh(x="Word", y="Count",␣
 ↪ax=ax)
# set the title
plt.title("Count of climate change related words")

for i, v in enumerate(climate_change_words_df['Count'].sort_values()):
    ax.text(v, i , str(v),
            color = 'blue', fontweight = 'bold')

plt.show()
# plt.savefig('climate-related-words-breakdown.png', transparent=False)
```



[106]:

```
# find segments
climate_change_words_found = list(climate_change_words_df['Word'].unique())
climate_change_words_found
```

```
[106]: ['climate',
        'record',
        'severe',
        'extreme',
        'temperature',
        'report',
        'high',
        'historic',
        'drought',
        'life-threatening',
        'crisis',
        'global',
        'arctic',
        'warm',
        'record-setting',
        'meteorologist',
        'ice',
        'degree',
        'fuel',
        'effect',
        'gas',
        'scientist',
        'hottest',
        'history',
        'unprecedented',
        'policy']
```

```
[110]: unique_df[unique_df["climate_words_found"].str.len() != 0].to_csv('reports/
        ↪abc_final.csv', encoding='utf-8')
```

```
[ ]:
```

```
[ ]:
```