segments

March 5, 2022

[3]:
```
import docx
from simplify_docx import simplify

# read in a document
my_doc = docx.Document("Analysis_12_30_21_Colorado_Fire_segments.docx")

# coerce to JSON using the standard options
my_doc_as_json = simplify(my_doc)
```

[5]:
```
# my_doc_as_json

import json
with open('segments.json', 'w') as fp:
    json.dump(my_doc_as_json, fp)
```

[20]:
```
# create a list of dicts(time, location, station, text)
clean_data = []

for blob in my_doc_as_json['VALUE'][0]['VALUE'][1:]:
    text_end = False
    count = 0
    key = f'text{count}'

    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']

        # print(time, location, station)

    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
        })
        count += 1


print('done')
```

```
done
```

[23]:
```
import pandas as pd
pd.options.display.max_rows = 500

# create dataframe
segment_df = pd.DataFrame.from_dict(clean_data)
```

[25]:
```
segment_df['text'] = segment_df['text'].str.lower()
```

[27]:
```
segment_df.head()
```

[27]:

	time	location	station	text
0	2021-12-30 6:14 PM	Salt Lake City	KTVX	u.s. many of these travel troubles will likely…
1	2021-12-30 6:12 PM	Phoenix (Prescott)	KNXV	500, 80 homes and businesses destroyed. it is …
2	2021-12-30 6:03 PM	Phoenix (Prescott)	KNXV	reporting live from downtown flagstaff. luzdel…
3	2021-12-30 6:00 PM	San Francisco-Oak-San Jose	KGO	injured. the sheriff did not rule out the poss…
4	2021-12-30 5:54 PM	Tampa-St. Pete (Sarasota)	WFTS	need this time. it's outside gambling groups a…

[28]:
```
from words import CLIMATE_CHANGE_RELATED_WORDS

import nltk
from nltk.corpus import stopwords
stop_words=set(stopwords.words("english"))

from wordcloud import STOPWORDS

final_stopwords = list(STOPWORDS) + list(stop_words)
final_stopwords = set(final_stopwords)
```

[29]:
```
# remove duplicates
from thefuzz import fuzz
```

```
def duplicate_check(text):
    matches = segment_df.apply(lambda row: (fuzz.partial_ratio(row['text'],␣
 ↪text) >= 85), axis=1)
    return [i for i, x in enumerate(matches) if x]


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

[33]:
```
segment_df
```

[33]:

	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	
20	2021-12-30 5:35 PM	Tulsa	KTUL	
21	2021-12-30 5:35 PM	Tulsa	KTUL	
22	2021-12-30 5:35 PM	Nashville	WKRN	
23	2021-12-30 5:35 PM	Minneapolis-St. Paul	KSTP	
24	2021-12-30 5:35 PM	Minneapolis-St. Paul	KSTP	
25	2021-12-30 5:35 PM	St. Louis	KDNL	
26	2021-12-30 5:35 PM	St. Louis	KDNL	
27	2021-12-30 5:35 PM	San Antonio	KSAT	
28	2021-12-30 5:35 PM	Portland, OR	KATU	
29	2021-12-30 5:35 PM	San Francisco-Oak-San Jose	KGO	
30	2021-12-30 5:35 PM	Memphis	WATN	
31	2021-12-30 5:35 PM	Albuquerque-Santa Fe	KOAT	
32	2021-12-30 5:35 PM	Salt Lake City	KTVX	
33	2021-12-30 5:35 PM	Austin	KVUE	
34	2021-12-30 5:35 PM	New Orleans	WGNO	

35	2021-12-30 5:35 PM	Phoenix (Prescott)	KNXV
36	2021-12-30 5:35 PM	Sacramnto-Stkton-Modesto	KXTV
37	2021-12-30 5:34 PM	Seattle-Tacoma	KOMO
38	2021-12-30 5:34 PM	Las Vegas	KTNV
39	2021-12-30 5:34 PM	Portland, OR	KATU
40	2021-12-30 5:34 PM	Sacramnto-Stkton-Modesto	KXTV
41	2021-12-30 5:34 PM	Dallas-Ft. Worth	WFAA
42	2021-12-30 5:34 PM	Oklahoma City	KOCO
43	2021-12-30 5:34 PM	Oklahoma City	KOCO
44	2021-12-30 5:34 PM	Dallas-Ft. Worth	WFAA
45	2021-12-30 5:34 PM	Oklahoma City	KOCO
46	2021-12-30 5:34 PM	Milwaukee	WISN
47	2021-12-30 5:34 PM	Houston	KTRK
48	2021-12-30 5:34 PM	Kansas City	KMBC
49	2021-12-30 5:34 PM	Tulsa	KTUL
50	2021-12-30 5:34 PM	Tulsa	KTUL
51	2021-12-30 5:34 PM	Nashville	WKRN
52	2021-12-30 5:34 PM	Minneapolis-St. Paul	KSTP
53	2021-12-30 5:34 PM	St. Louis	KDNL
54	2021-12-30 5:34 PM	San Antonio	KSAT
55	2021-12-30 5:34 PM	Raleigh-Durham (Fayetvlle)	WTVD
56	2021-12-30 5:34 PM	Memphis	WATN
57	2021-12-30 5:34 PM	Albuquerque-Santa Fe	KOAT
58	2021-12-30 5:34 PM	Salt Lake City	KTVX
59	2021-12-30 5:34 PM	Austin	KVUE
60	2021-12-30 5:34 PM	New Orleans	WGNO
61	2021-12-30 5:34 PM	New Orleans	WGNO
62	2021-12-30 5:34 PM	Phoenix (Prescott)	KNXV
63	2021-12-30 5:34 PM	Phoenix (Prescott)	KNXV
64	2021-12-30 5:32 PM	Dallas-Ft. Worth	WFAA
65	2021-12-30 5:32 PM	Seattle-Tacoma	KOMO
66	2021-12-30 5:32 PM	Oklahoma City	KOCO
67	2021-12-30 5:32 PM	Milwaukee	WISN
68	2021-12-30 5:32 PM	Houston	KTRK
69	2021-12-30 5:32 PM	Houston	KTRK
70	2021-12-30 5:32 PM	Las Vegas	KTNV
71	2021-12-30 5:32 PM	Kansas City	KMBC
72	2021-12-30 5:32 PM	Tulsa	KTUL
73	2021-12-30 5:32 PM	Nashville	WKRN
74	2021-12-30 5:32 PM	Minneapolis-St. Paul	KSTP
75	2021-12-30 5:32 PM	St. Louis	KDNL
76	2021-12-30 5:32 PM	San Antonio	KSAT
77	2021-12-30 5:32 PM	Portland, OR	KATU
78	2021-12-30 5:32 PM	San Francisco-Oak-San Jose	KGO
79	2021-12-30 5:32 PM	Memphis	WATN
80	2021-12-30 5:32 PM	Albuquerque-Santa Fe	KOAT
81	2021-12-30 5:32 PM	Salt Lake City	KTVX

82	2021-12-30 5:32 PM	Austin	KVUE
83	2021-12-30 5:32 PM	New Orleans	WGNO
84	2021-12-30 5:32 PM	Phoenix (Prescott)	KNXV
85	2021-12-30 5:32 PM	Sacramnto-Stkton-Modesto	KXTV

	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…
20	ltulsa2021-12-30 5:35 pm
21	senior meteorologist rob marciano, and rob, yo…
22	senior meteorologist rob marciano, and rob, yo…
23	polis-st. paul2021-12-30 5:35 pm
24	senior meteorologist rob marciano, and rob, yo…
25	021-12-30 5:35 pm
26	concern. will, thank you. >>> let's get right …
27	senior meteorologist rob marciano, and rob, yo…
28	it is supposed to start snowing in the morning…
29	concern. will, thank you. let's get right to a…
30	senior meteorologist rob marciano, and rob, yo…
31	marciano, and rob, you're tracking those dange…
32	marciano, and rob, you're tracking those dange…
33	senior meteorologist rob marciano, and rob, yo…
34	marciano, and rob, you're tracking those dange…
35	marciano, and rob, you're tracking those dange…
36	marciano, and rob, you're tracking those dange…
37	>> i have a party near the element hotel that'…
38	to heed the warning and get out. >> i have a p…
39	and so, at this point, you need to heed the wa…
40	>> i have a party near the element hotel that'…

```
41  towards you. they are actively running from-fi…
42                     oklahoma city2021-12-30 5:34 pm
43  trying to evacuate on foot towards you. they a…
44  towards you. they are actively running from-fi…
45  trying to evacuate on foot towards you. they a…
46  trying to evacuate on foot towards you. they a…
47  towards you. they are actively running from-fi…
48  trying to evacuate on foot towards you. they a…
49                                   21-12-30 5:34 pm
50  trying to evacuate on foot towards you. they a…
51  towards you. they are actively running from-fi…
52  towards you. they are actively running from-fi…
53  trying to evacuate on foot towards you. they a…
54  towards you. they are actively running from-fi…
55  the rescheduled game. >> i don't like it. disr…
56  trying to evacuate on foot towards you. they a…
57  trying to evacuate on foot towards you. they a…
58  trying to evacuate on foot towards you. they a…
59  trying to evacuate on foot towards you. they a…
60  trying to evacuate on foot towards you. they a…
61  towards you. they are actively running from-fi…
62  towards you. they are actively running from-fi…
63  towards you. they are actively running from-fi…
64  the teenager considered armed and dangerous. >…
65  and dangerous. >>> and america strong tonight…
66  murder in texas. the teenager considered armed…
67  murder in texas. the teenager considered armed…
68                               ston2021-12-30 5:32 pm
69  the teenager considered armed and dangerous. >…
70  the teenager considered armed and dangerous. >…
71  murder in texas. the teenager considered armed…
72  murder in texas. the teenager considered armed…
73  the teenager considered armed and dangerous. >…
74  and dangerous. >>> and america strong tonight…
75  the prime suspect in a triple murder in texas…
76  the teenager considered armed and dangerous. >…
77  >>> news tonight on the urgent search for a 14…
78  and dangerous. >>> and america strong tonight…
79  murder in texas. the teenager considered armed…
80  the teenager considered armed and dangerous. >…
81  the teenager considered armed and dangerous. >…
82  murder in texas. the teenager considered armed…
83  and dangerous. >>> and america strong tonight…
84  the teenager considered armed and dangerous. >…
85  and dangerous. >>> and america strong tonight…

                                                  matches
```

```
0                                                          [0]
1                                                          [1]
2                                                          [2]
3                                                          [3]
4                                                          [4]
5                                                          [5]
6                                                          [6]
7                                                          [7]
8                                                          [8]
9                                                          [9]
10                                                        [10]
11  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
12                                          [12, 25, 49]
13  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
14  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
15  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
16  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
17                                          [17, 25, 49]
18  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
19  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
20                                          [20, 25, 49]
21  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
22  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
23                                          [23, 25, 49]
24  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
25                  [12, 17, 20, 23, 25, 42, 49, 68]
26  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
27  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
28  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
29  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
30  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
31  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
32  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
33  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
34  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
35  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
36  [11, 13, 14, 15, 16, 18, 19, 21, 22, 24, 26, 2…
37  [37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 5…
38  [37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 5…
39                                    [37, 38, 39, 40]
40  [37, 38, 39, 40, 41, 43, 44, 45, 46, 47, 48, 5…
41  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
42                                          [25, 42, 49]
43  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
44  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
45  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
46  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
```

```
47  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
48  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
49                    [12, 17, 20, 23, 25, 42, 49, 68]
50  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
51  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
52  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
53  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
54  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
55                                                [55]
56  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
57  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
58  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
59  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
60  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
61  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
62  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
63  [37, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 5…
64  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
65  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
66  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
67  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
68                                      [25, 49, 68]
69  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
70  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
71  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
72  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
73  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
74  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
75  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
76  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
77  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
78  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
79  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
80  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
81  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
82  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
83  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
84  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
85  [64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 7…
```

[36]:
```python
def fetch_biggest_text(idx_list):
    biggest_length = 0
    idx = None

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

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

fetch_biggest_text(segment_df['matches'][12])
```

[36]:
```
12
```

[ ]:
```

```

[37]:
```
segment_df['row_to_use'] = segment_df.apply(lambda row:␣
 ↪fetch_biggest_text(row['matches']), axis=1)
```

[38]:
```
idxs = list(segment_df['row_to_use'].unique())

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

# getting unique:
# get the first value in the matches list if len is 1
# else for duplicates, get the longest text, grab that index
```

[39]:
```
no_dup_segment_df = segment_df[segment_df['use_row']]
no_dup_segment_df
```

[39]:

	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
```

[40]:
```
# breakdown words - unique words only
no_dup_segment_df['words'] = no_dup_segment_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(' ')

def clean_text(words):
    # text.replace(',', '').replace('>','').replace('.','').replace('\n','').
 ↪replace('’',"'").replace('!','').replace('?','').replace('%','').
 ↪replace(')','').replace('(','').replace('_','').replace(':','')
    # words = text.split(' ')

    for word in words:
        if word.isdigit():
            words.pop(words.index(word))

    return words


no_dup_segment_df.apply(lambda row: clean_text(row['words']), axis=1)
```

```
/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.

/Users/loren/.pyenv/versions/3.7.4/lib/python3.7/site-
packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

[40]:
```
0     [us, many, of, these, travel, troubles, will, …
1     [80, homes, and, businesses, destroyed, it, is…
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    [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, h…
42                         [oklahoma, city2021-12-30, pm]
55    [the, rescheduled, game, , i, don't, like, it,…
67    [murder, in, texas, the, teenager, considered,…
68                                    [ston2021-12-30, pm]
dtype: object
```

[ ]:

[41]:
```
# compile the list of unique words, get that count
total_words = no_dup_segment_df['words'].str.len().sum()
total_words
```

[41]: 2421

[42]:
```
unique_words_found = set()
for chunk in no_dup_segment_df['words']:
    for word in chunk:
        unique_words_found.add(word)

total_unique_words = len(unique_words_found)
total_unique_words
```

[42]: 914

[ ]:

[43]:
```
# run lemmatization clean
# tokenize filtered word list for frequency distribution
from nltk.tokenize import word_tokenize
nltk.download('punkt')

#Lexicon Normalization
# Lemmatization -- distill to root words

nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('averaged_perceptron_tagger')
from nltk.stem import WordNetLemmatizer
```

```
lem = WordNetLemmatizer()

lemma_list = []
for word, tag in nltk.pos_tag(unique_words_found):
    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)
len(lemma_list)
```

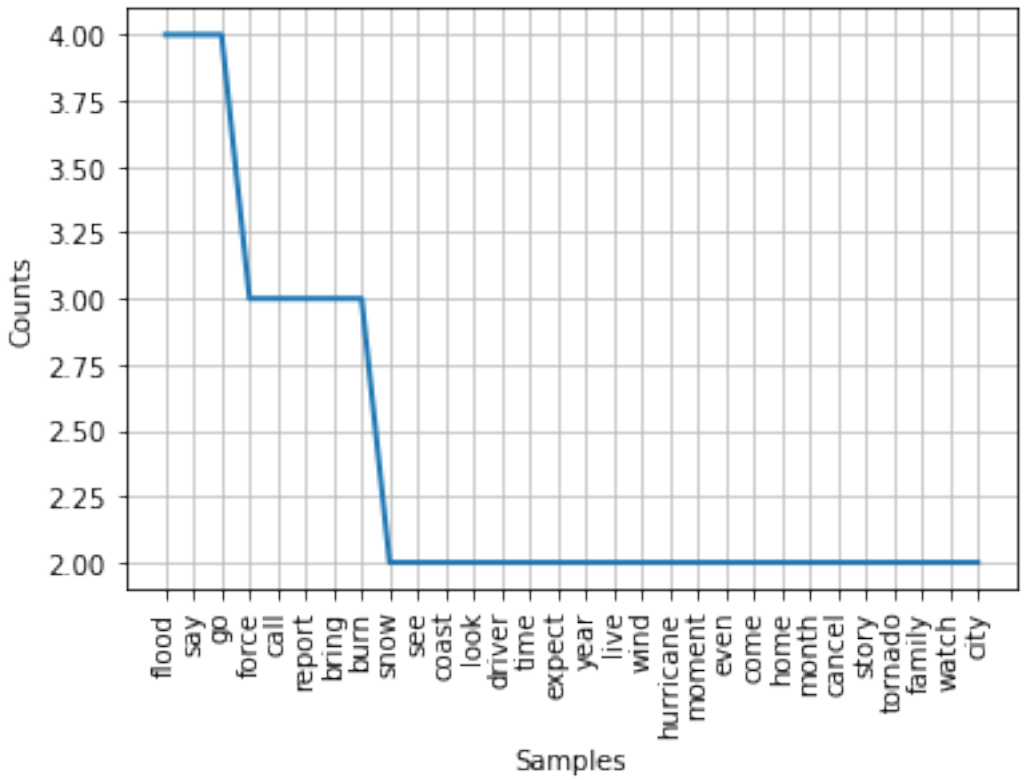
```
[nltk_data] Downloading package punkt to /Users/loren/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /Users/loren/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /Users/loren/nltk_data…
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/loren/nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

[43]:
```
914
```

[ ]:
```

```

[44]:
```
from nltk.probability import FreqDist
# no stopwords
clean_lemma_list = []

for word in lemma_list:
    if word not in final_stopwords:
        clean_lemma_list.append(word)
len(clean_lemma_list)


# graph frequence distributions of lemma words
lfdist = FreqDist(clean_lemma_list)
print(lfdist)

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

```
<FreqDist with 732 samples and 797 outcomes>
```



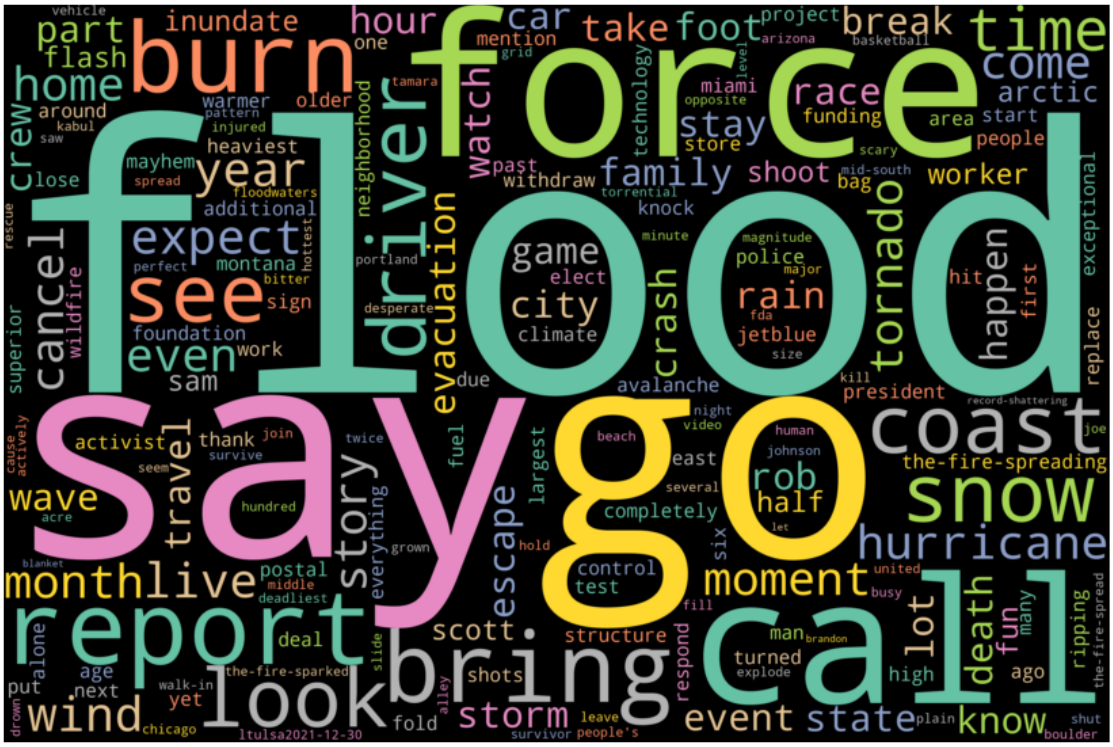
[46]:
```
lfdist
# len(clean_lemma_list)
```

[46]:
```
FreqDist({'flood': 4, 'say': 4, 'go': 4, 'force': 3, 'call': 3, 'report': 3,
'bring': 3, 'burn': 3, 'snow': 2, 'see': 2, …})
```

[ ]:
```

```

[47]:
```
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 =␣
 ↪final_stopwords).generate_from_frequencies(lfdist)

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

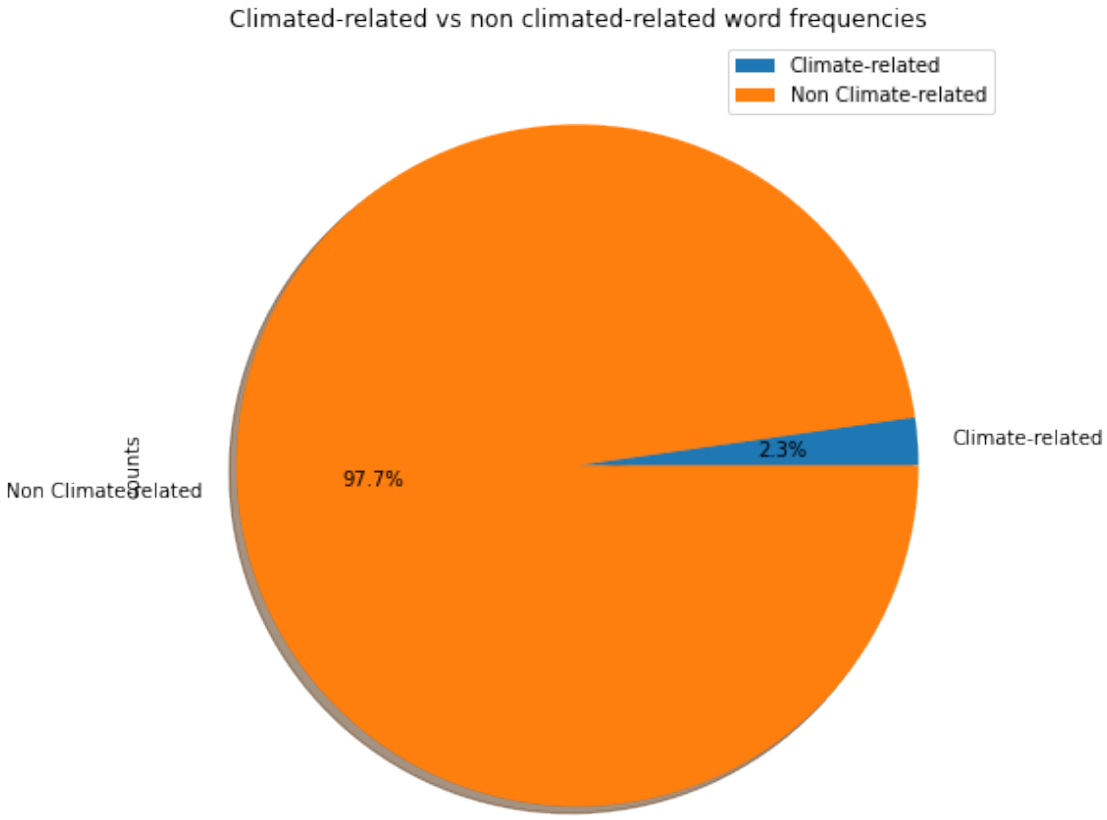
```
#plt.savefig('word_cloud.png')
```



[48]:
```
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'],
```

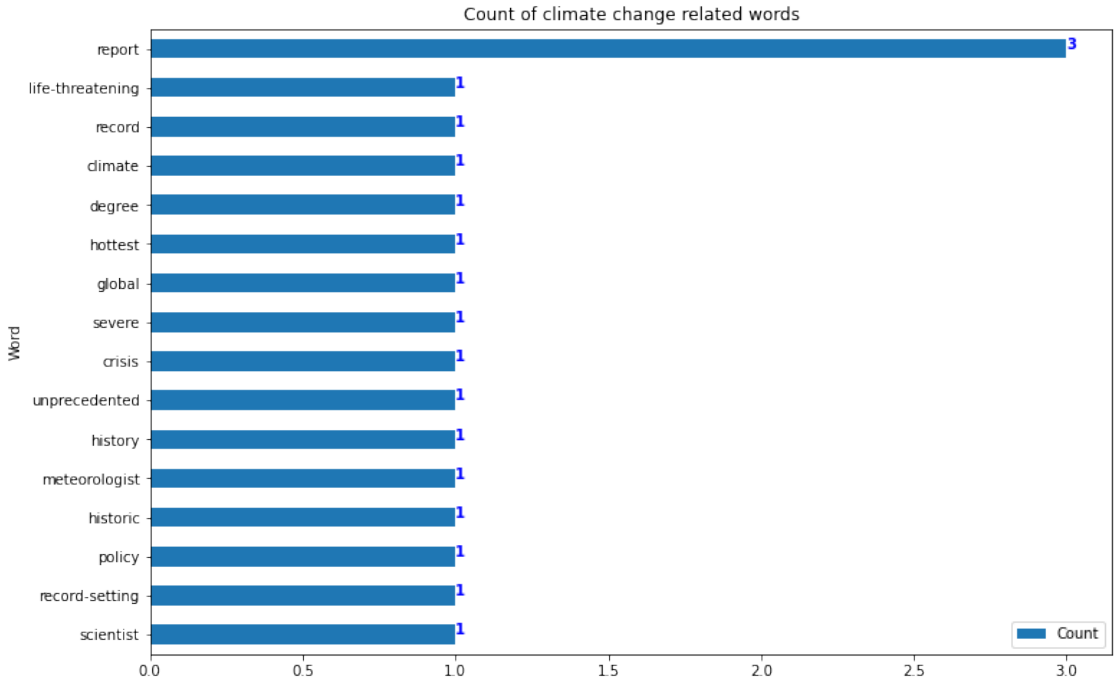
```
                            'counts': [climate_words_count,␣
 ↪non_climate_words_count]})
comparison_df.set_index('Words', inplace=True)
print(comparison_df)

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          18
Non Climate-related     779
```



[ ]:

[49]:
```
# 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)
```



[52]:
```
# find segments

climate_change_words_found = list(climate_change_words_df['Word'].unique())
climate_change_words_found

climate_change_words_found_str = '(?i)|'.join(climate_change_words_found)
climate_change_words_found_str = '(?i)' + climate_change_words_found_str
```

```
climate_change_words_found_str
```

[52]:
```
'(?i)report(?i)|scientist(?i)|record-setting(?i)|policy(?i)|historic(?i)|meteorologist(?i)|history(?i)|unprecedented(?i)|crisis(?i)|severe(?i)|global(?i)|hottest(?i)|degree(?i)|climate(?i)|record(?i)|life-threatening'
```

[ ]:
```

```

[53]:
```
pd.set_option("display.max_colwidth", -1)
```

```
/Users/loren/.pyenv/versions/3.7.4/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Passing a negative integer is deprecated in version 1.0 and will not be supported in future version. Instead, use None to not limit the column width.
  """Entry point for launching an IPython kernel.
```

[ ]:
```

```

[55]:
```
climate_change_related_df =  no_dup_segment_df[no_dup_segment_df['text'].str.lower().str.contains(climate_change_words_found_str)]
climate_change_related_df.loc[:, 'time':'text']
```

[55]:

	time	location	station	text
0	2021-12-30 6:14 PM	Salt Lake City	KTVX	u.s. many of these travel troubles will likely continue into the new year 3 in an alley, abc news, new york. >> brandon, thank you. well, time now for the mount in in this snapshot of the stories making headlines across the region in boulder county, colorado. look at these images. crews are responding to what they're calling a life-threatening situation. several small grass-fire-spreading fast burning hundreds of homes and forcing 2 entire towns to evacuate. authorities say the-fire-sparked from downed power lines. and tonight strong winds are fueling the flames. in montana. 2 snowmobilers were killed in an avalanche near cooke city. authorities say the time of the slide, 4 people were on the to were able to escape. but the other 2 were buried. the avalanche broke 5 feet deep in roughly 300 feet wide. and finally in california, south lake
1	2021-12-30 6:12 PM	Phoenix (Prescott)	KNXV	
2	2021-12-30 6:03 PM	Phoenix (Prescott)	KNXV	
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	
9	2021-12-30 5:38 PM	Los Angeles	KABC	
10	2021-12-30 5:37 PM	Los Angeles	KABC	
28	2021-12-30 5:35 PM	Portland, OR	KATU	
38	2021-12-30 5:34 PM	Las Vegas	KTNV	
55	2021-12-30 5:34 PM	Raleigh-Durham (Fayetvlle)	WTVD	
67	2021-12-30 5:32 PM	Milwaukee	WISN	

tahoe saying a record breaking amount of snow. take a look at this time lapse video showing the heaviest snow blanketing one backyard just this month alone. lake tahoe
1 500, 80 homes and businesses destroyed. it is unreal to even think of yeah. this is an area south of boulder. it's north of denver. the-fire-spread so quickly there. 1600 acres just within hours. >> we're hearing a hotel and a target has both burned and evacuations are now in effect at a hospital in the area. and 6 people already in the hospital. but the county sheriff, they're saying it is likely that there will be severe injuries and that there will people will be people who were not able to survive this. >> we know so many of you have friends and family here. so we are staying on this story committed to keeping you updated throughout the evening.
2 reporting live from downtown flagstaff. luzdelia abc, 15, arizona thanks. new year's celebrations canceled but still plenty of other winter fun to be had called snow play hotline. >> the best spots for sledding playing in the snow and so much more. that number for 4 to 5, 6, snow. >> now to the breaking news out of colorado. the national guard being deployed tonight to battle those dangerous-fires in colorado. this isn't real because this this is a live look. this is going on right now. the-fires outside boulder, threatening the towns of lewisville and superior. the sheriff there says entire neighborhoods are burning hundreds of homes and businesses. as we've reported to you this evening, already destroyed. now on top of all that the power is out because the-fires took out all the pie or power lines there. we know one-fire-has grown to 1600 acres in a matter of hours. again, these are live pictures. those are people's homes.
5 northwest. it was sporadic weather patterns connected to climate change that created a once in a millenia heat wave resulting in some of the highest temperatures ever recorded in the region. temperatures in portland, seattle and parts of canada soaring well above 100 degrees adlibs 228 people dying in washington state and oregon alone. >> electricity went off and so it's quickly getting warmer and warmer kicking off an unprecedented summer of heat. twenty twenty one will go down in history as the hottest summer on record for the united states. mega-fires in the west burning for months. the bootleg-fire-in oregon becoming the state's biggest this year burning more than 400 430000. the dixie farm becoming the second largest to ever scorched parts of california. >> i didn't know where i was, whose house is what and it was a wasteland. we're just grateful to be alive . >> we've got each other the 20
6 be evacuated because there are several grass-fires burning out there. this is happening just north of denver about a half hour out of denver. you can see the smoke from the-fires blowing in the wind. >> a hospital in the area now working to evacuate the patients there and this-fire-wrapping up a year of extreme catastrophic weather events. scientists say climate change contributes to the worsening conditions this year also brought historic levels of funding for climate related projects around the nation. nbc's al roker walks us through the severe weather and climate policy efforts of twenty twenty one, 2021. another blistering year of climate and weather extremes from wildfires and drought to catastrophic flooding hurricanes almost no state escaping unscathed this year seem to pick up where 2012 left off with one major exception. four

years after withdrawing from the paris climate agreement, newly elected president joe biden bringing the u.s. back into the fold.
7 living with a risk factor they may not know about. i am josie sterman for spotlight on america. >> this is really important information. you can batch more of her in- depth reporting on wjla.com. look under the spotlight on america section. carl? >>> well, alison, a breaking update on news we brought you a few minutes ago. a wildfire-is forcing evacuations in colorado right now. we just got this new video in from the city of superior. you can see how hard it is to see anything on that road there. driving through that dense smoke, vaccinations are under way there and the nearby city of louisville, colorado. high winds are fanning the flames and causing the-fire-to spread out quickly. alison, when you first brought us this, you saw the thick plume of smoke out there. it is getting even worse and getting closer to keeping it there. >> yeah. so scary. like you said, from afar it was such a different scene. to see up close what the people are dealing with, bill, it is
9 exploding in size. >> burping more than twice the acreage of last year's record-fire-season. >> they be -- >> catastrophic floods ravage middle tennessee. >> knocked off foundations. just folt gone. >> hurricane ida ripping through louisiana. >> the rain -- >> the remnants of hurricane ida hammered the northeast. >> never seen anything like this. >> the national weather service issuing a flash flood emergency for new york city. >> it was frightening. >> many of the victims dying in flooded basements. >> the nation's deadliest tornado outbreak in more than a decade. >> 88 people lost their lives. >> we lost everything. we basically have nothing. >> the f.d.a. authorizing the pfizer vaccine for 12 to 15-year-olds. >> the delta variant is spreading and the hospitalizations are going up. >> i'm putting people that are my age and my parents' age in body bags. >> i will not get that vaccine. >> and a busy -- >> we must stop denying that the
10 >> 12 story apartment building collapsing outside miami beach. >> desperate to find survivors. >> time to end america's longest war. >> mayhem in kabul. the airport overrun. >> the deadliest attack on u.s. forces in afghanistan in a decade. marking the end of america's longest war. >> thousands of diplomats, activists and world leaders bringing the global climate crisis to the center stage. >> we are drowning. >> the world has warmed more in the last 29 years than in the past 110 years. >> arctic air, power grids failing, and millions left to cope with the bitter cold. texas getting pummeled with snow. >> like a walk-in freezer. >> people freezing to death inside their homes. >> dangerous heat waves hitting both coasts. >> it feels like an armageddon. disblmplets record-shattering temperatures in the pacific northwest now blamed for dozens of deaths. >> it's like you're on-fire. >> nearly two thirds of the west exceptional or extreme drought. >> wildfires across the west
28 it is supposed to start snowing in the morning in colorado, but at that point, the damage will have been done. whit? >> until then, still so much concern. will, thank you. let's get right to abc's senior meteorologist rob marciano, and rob, you're tracking those dangerous-fire-conditions in colorado. when are the winds expected to finally let up? >> reporter: we expect the strongest winds, whit, to retreat into the foothills over the next couple of

hours, but this has turned out to be a historic what we call a mountain wave event, and it's going to flip the switch to snow, as will mentioned. we have advisories that are up in the colorado area and now winter storm watches that are posted for kansas city and chicago, as this low that was in california emerges into the plains. it will bring a severe weather threat again to the mid-south tomorrow. but icing conditions in kansas city by saturday morning. lincoln will see some snow. that presses into chicago and detroit by saturday. heavy rain across the area. they got the tornadoes two weeks ago. and an expansive severe weather threat across the deep south and mid-south where tornadoes will be possible in big cities like
38 to heed the warning and get out. >> i have a party near the element hotel that's going to be trying to evacuate on foot towards you. they are actively running from-fire-behind them. >> reporter: across the area, at least six people rushed to the hospital. >> due to the magnitude of this-fire, intensity of this-fire-and its presence in such a heavily populated area, we would not be surprised if there are injuries or fatalities. >> reporter: millions across the country facing dangerous weather conditions. in california, a slow-moving storm dumping torrential rain. floodwaters inundating this campground in malibu before dawn. >> everyone okay? >> reporter: first responders racing to rescue families and their pets. some 50 people in all brought to safety. to the south, a tree crashing down on the pacific coast highway, bringing utility lines with it, shutting down the roadway. while it's raining here in california, it's been a record-setting dry season in colorado. add in those winds, up to 100 miles per hour, and you're looking at perfect wildfire-conditions. it is supposed to start snowing in the morning in colorado, but at that point, the damage will have been done. whit?
55 the rescheduled game. >> i don't like it. disrespectful to everybody else. and it's bad because people can get hurt that are not even in the fight. it's dangerous and folks are cutting up. i don't know what they've been doing before they got here. >> reporter: security will be increased here for the rest of the day, but the games should go on as scheduled. tamara scott, abc11 eyewitness news. >>> police are searching for the person who opened-fire-at a high school basketball tournament in catawba county. two juveniles were hurt when shots rang out inside that gym lobby last night. their conditions aren't known. we know one of the victims was flown to the hospital. more than 400 people were attending the game as part of the sam moore basketball tournament. the rest of the tournament has now been cancelled. >>> jetblue that has a number of flights in and out of rdu is cutting 1,300 flights through january 13 because of a shortage of crew members, a lot of people
67 murder in texas. the teenager considered armed and dangerous. >>> and america strong tonight. a graduation story to remember. >>> good evening. thanks for joining us on a busy thursday night. i'm whit johnson, in for david. and we begin tonight with that breaking news. life-threatening-fires breaking out in boulder county, colorado, at this hour. the governor declaring a state of emergency just a short time ago. residents forced to evacuate with almost no notice. the images just coming in. drivers blinded by thick smoke blowing across the roads. powerful winds downing power lines, igniting fast-moving grass-fires.

homes and structures going up in flames. the-fire-spreading in just a matter of hours. reports of burn victims rushed to the hospital. customers in stores suddenly told to held for the exits. you see there. smoke fillinging the air outside. it's all part of a series of

[ ]:

[ ]:

[ ]: