Name: Venkata Krishnarjun Vuppala **Semester**: 6
SRN: PES2UG19CS451 **Section**: G
Subject: Algorithms for Intelligence Web and Information Retrieval

Assignment 3

1. tCase folding

Code and Output:

PES2UG19CS451_VENKATAKRISH

https://selfcare.actcorp.in/group/blr/myaccount

Apps Bookmarks Honda Assure Insur... Find a Doctor – Tele... The Art of Living DS... Parkwest Residents... Mutual Fund NAV... HDFC Mutual Fund ACT MyAccount

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ip

Run Code

4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

In [5]:
```
df["tokenized"] = df["text"]
df = df.apply(lambda x: x.str.lower() if x.dtype == "object" else x)
df["tokenized"] = df['tokenized'].str.replace('[^\w\s]','')
```

In [6]:
```
df["tokenized"] = df["tokenized"].apply(word_tokenize)
df["tokenized"].head()
```

Out[6]:
```
0    [our, deeds, are, the, reason, of, this, earth...
1        [forest, fire, near, la, ronge, sask, canada]
2    [all, residents, asked, to, shelter, in, place...
3    [13000, people, receive, wildfires, evacuation...
4    [just, got, sent, this, photo, from, ruby, ala...
Name: tokenized, dtype: object
```

In [7]:
```
df.head()
```

Out[7]:

	id	keyword	location	text	target	tokenized
0	1	NaN	NaN	our deeds are the reason of this #earthquake m...	1	[our, deeds, are, the, reason, of, this, earth...
1	4	NaN	NaN	forest fire near la ronge sask. canada	1	[forest, fire, near, la, ronge, sask, canada]
2	5	NaN	NaN	all residents asked to 'shelter in place' are ...	1	[all, residents, asked, to, shelter, in, place...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13000, people, receive, wildfires, evacuation...
4	7	NaN	NaN	just got sent this photo from ruby #alaska as ...	1	[just, got, sent, this, photo, from, ruby, ala...

2. stemming/lemmatization

Code and Output:

PES2UG19CS451_VENKATAKRISH

http://localhost:8888/notebooks/Documents/work/AIWIR-LAB/A3/PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G.ipynb

Apps | Bookmarks | Honda Assure Insur... | Find a Doctor – Tele... | The Art of Living DS... | Parkwest Residents... | Mutual Fund NAV... | HDFC Mutual Fund | ACT MyAccount

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel

Run Code

In [26]:
```
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
df["tokenized"] = df["tokenized"].apply(lambda words: [word for word in words if word not in stop_words])
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\arjun\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [27]:
```
df.head()
```

Out[27]:

	id	keyword	location	text	target	tokenized
0	1	NaN	NaN	our deeds are the reason of this #earthquake m...	1	[deeds, reason, earthquake, may, allah, forgiv...
1	4	NaN	NaN	forest fire near la ronge sask. canada	1	[forest, fire, near, la, ronge, sask, canada]
2	5	NaN	NaN	all residents asked to 'shelter in place' are ...	1	[residents, asked, shelter, place, notified, o...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13000, people, receive, wildfires, evacuation...
4	7	NaN	NaN	just got sent this photo from ruby #alaska as ...	1	[got, sent, photo, ruby, alaska, smoke, wildfi...

In [28]:
```
stemmer = PorterStemmer()
df['stemmed'] = df['tokenized'].apply(lambda x: [stemmer.stem(y) for y in x])
```

PES2UG19CS451_VENKATAKRISH

http://localhost:8888/notebooks/Documents/work/AIWIR-LAB/A3/PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G.ipynb

Apps | Bookmarks | Honda Assure Insur... | Find a Doctor – Tele... | The Art of Living DS... | Parkwest Residents... | Mutual Fund NAV... | HDFC Mutual Fund | ACT MyAccount

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [29]:
```
df.head()
```

Out[29]:

	id	keyword	location	text	target	tokenized	stemmed
0	1	NaN	NaN	our deeds are the reason of this #earthquake m...	1	[deeds, reason, earthquake, may, allah, forgiv...	[deed, reason, earthquak, may, allah, forgiv, us]
1	4	NaN	NaN	forest fire near la ronge sask. canada	1	[forest, fire, near, la, ronge, sask, canada]	[forest, fire, near, la, rong, sask, canada]
2	5	NaN	NaN	all residents asked to 'shelter in place' are ...	1	[residents, asked, shelter, place, notified, o...	[resid, ask, shelter, place, notifi, offic, ev...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13000, people, receive, wildfires, evacuation...	[13000, peopl, receiv, wildfir, evacu, order, ...
4	7	NaN	NaN	just got sent this photo from ruby #alaska as ...	1	[got, sent, photo, ruby, alaska, smoke, wildfi...	[got, sent, photo, rubi, alaska, smoke, wildfi...

3. Standard inverted index construction

Code and Output:

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 AM (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (

Run Code

Standard inverted index construction

In [121]:
```
my_list = df["stemmed"]
my_list[:5]
```

Out[121]:
```
0    [deed, reason, earthquak, may, allah, forgiv, us]
1         [forest, fire, near, la, rong, sask, canada]
2    [resid, ask, shelter, place, notifi, offic, ev...
3    [13000, peopl, receiv, wildfir, evacu, order, ...
4    [got, sent, photo, rubi, alaska, smoke, wildfi...
Name: stemmed, dtype: object
```

In [*]:
```
words = []
for key in my_dict:
    for i in range(len(my_dict[key])):
        term = my_dict[key][i]
        words.append(term)
length = len(my_list)
inverted_index = {}
for i in range(length):
    check = my_list[i]
    for item in words:
        if item in check:
            if item not in inverted_index:
                inverted_index[item] = []
            if item in inverted_index:
                inverted_index[item].append(i+1)
for key in inverted_index:
    inverted_index[key] = list(set(inverted_index[key]))
```

In [*]:
```
for key in inverted_index:
    print(key,": ",inverted_index[key])
```

```
deed :  [1, 4986]
reason :  [4992, 1, 1921, 4998, 5001, 5002, 782, 6167, 5373, 305, 306, 7219, 6454, 2748, 6460, 4670, 318, 320, 2113, 2253, 69
92, 6233, 6243, 747, 4844, 4334, 6899, 4085, 764, 3453, 895]
earthquak :  [1, 6025, 6031, 7311, 6946, 6947, 7590, 6955, 7600, 6973, 6975, 7137, 3028, 3029, 3030, 3031, 3033, 3034, 3035,
3036, 7130, 3038, 3039, 3040, 7132, 3042, 7133, 3044, 3045, 7136, 3047, 3048, 3049, 3050, 3051, 3052, 5738, 7143, 3055, 7144,
3059, 3060, 3061, 3062, 3063, 3065, 3066]
may :  [1, 5662, 5157, 3113, 4139, 4141, 2115, 4178, 1634, 4720, 4724, 4216, 5245, 5246, 5248, 5249, 5250, 5251, 5252, 5254,
4233, 5258, 3724, 4236, 5260, 5261, 5263, 5265, 5267, 3222, 5270, 5271, 5272, 5273, 5274, 5275, 5276, 5277, 5278, 4256, 5802,
3774, 5840, 4817, 7383, 5853, 2785, 2794, 241, 5368, 5395, 5411, 2342, 808, 1833, 3394, 6468, 2375, 3400, 6485, 4955, 7004, 2
400, 5472, 6501, 895, 6533, 1935, 1936, 5011, 2463, 4517, 939, 4011, 7084, 7602, 2997, 1977, 5054, 1476, 2002, 3026, 4563, 20
15, 2016, 6647]
allah :  [4256, 1, 4291, 4680, 4011, 4300, 3726, 6448, 4313]
forgiv :  [1, 2266, 2533]
us :  [2560, 1, 514, 5122, 5123, 5125, 3078, 2056, 5130, 2059, 1036, 3084, 7178, 5139, 5142, 3096, 7193, 4636, 5150, 5157, 4
3, 5677, 3634, 7225, 7228, 4670, 7233, 5698, 5189, 1097, 2633, 1100, 4177, 85, 1627, 1116, 2141, 5214, 3680, 1633, 7268, 112
5, 4200, 2153, 1646, 628, 629, 1158, 4230, 4232, 7305, 1675, 7311, 6290, 3734, 7320, 5793, 4258, 1699, 1193, 682, 7340, 4782,
2224, 4282, 1221, 1741, 6350, 2262, 6875, 2274, 3309, 3824, 3313, 247, 3834, 251, 252, 254, 260, 2820, 267, 269, 271, 272, 38
60, 1304, 4889, 284, 6946, 6947, 2854, 2855, 3368, 2857, 2858, 4908, 6967, 6969, 2877, 830, 6461, 6975, 7486, 2373, 2885, 596
0, 2892, 2384, 338, 4436, 856, 4980, 4981, 4985, 6011, 3967, 7042, 3979, 2956, 5008, 7063, 6553, 1439, 5537, 945, 6577, 5559,
```

In [ ]:

4. Positional index construction

Code and Output:

PES2UG19CS451_VENKATAKRISH

http://localhost:8888/notebooks/Documents/work/AIWIR-LAB/A3/PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G.ipynb

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [30]:
```
pos_index = {}
document = df["stemmed"]
```

```
<class 'pandas.core.series.Series'>
```

In [31]:
```
my_dict = pd.Series(df.stemmed.values,index=df.id).to_dict()
```

In [32]:
```
my_dict[1]
```

Out[32]:
```
['deed', 'reason', 'earthquak', 'may', 'allah', 'forgiv', 'us']
```

In [33]:
```
pos_ind = {}
fileno = 0
for key in my_dict:
    for i in range(len(my_dict[key])):
        term = my_dict[key][i]
        if term in pos_ind:
            pos_ind[term][0] = pos_ind[term][0] + 1
            if fileno in pos_ind[term][1]:
                pos_ind[term][1][fileno].append(key)
            else:
                pos_ind[term][1][fileno] = [key]
        else:
            pos_ind[term] = []
            pos_ind[term].append(1)
            pos_ind[term].append({})
            pos_ind[term][1][fileno] = [key]
        fileno += 1
```

PES2UG19CS451_VENKATAKRISH

http://localhost:8888/notebooks/Documents/work/AIWIR-LAB/A3/PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G.ipynb

jupyter PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: Last Thursday at 11:03 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [36]:
```
#searching for a particular term
print(pos_ind["name"])
```

```
[30, {2835: [436], 6536: [959], 10416: [1540], 10654: [1582], 15055: [2195], 20486: [2969], 24035: [3472], 25147: [3610], 2537
6: [3639], 36564: [5179], 45566: [6398], 45570: [6398], 47080: [6603], 50572: [7112], 51004: [7169], 53574: [7555], 55241: [778
4], 58733: [8279], 58742: [8279], 59786: [8452], 60189: [8526], 60447: [8581], 60451: [8581], 62083: [8815], 62418: [8875], 647
53: [9207], 64913: [9225], 66002: [9384], 66097: [9402], 73576: [10478]}]
```