In [4]:

```
function processTweets(title)
    path = "Health-Tweets/" * title
    news = readlines(path * ".txt")

    tweets = Any[]
    for x in news
        push!(tweets,String(split(x,"|")[3]))
    end

    c = Corpus([])
    for x in tweets
        sd = StringDocument(x)
        prepare!(sd, strip_articles)
        prepare!(sd, strip_indefinite_articles)
        prepare!(sd, strip_definite_articles)
        prepare!(sd, strip_pronouns)
        prepare!(sd, strip_stopwords)
        prepare!(sd, strip_numbers)
        prepare!(sd, strip_non_letters)
        prepare!(sd, strip_frequent_terms)
        prepare!(sd, strip_html_tags)
        prepare!(sd, strip_punctuation)
        push!(c, sd)
    end

    stem!(c)
    update_lexicon!(c)

    d = DocumentTermMatrix(c)
    serialize(path * "_freq.jldata", dtm(d, :sparse))
    serialize(path * "_term.jldata", d.terms)

    freq = Serialization.deserialize(path * "_freq.jldata")
    term = Serialization.deserialize(path * "_term.jldata")

    terms_appear = 0 .< freq
    term_count = sum(terms_appear, dims = 1)

    top20term = sortperm(vec(term_count), rev = true)
    top20term_idx = top20term[1:20]
    return term[top20term_idx]
end
```

Out[4]:

```
processTweets (generic function with 1 method)
```

In [7]:

```
list = String["bbchealth", "cbchealth", "cnnhealth", "everydayhealth", "foxnewshealth",
"gdnhealthcare", "KaiserHealthNews", "latimeshealth", "msnhealthnews", "NBChealth", "nprh
ealth","nytimeshealth", "reuters_health", "usnewshealth", "wsjhealth"]
df = DataFrame()
count = 1

for x in list
    println("processing " * x)
    addCol = list[count]
    df[!, addCol] = processTweets(x)
    count += 1
end
```

```
processing bbchealth
processing cbchealth
processing cnnhealth
processing everydayhealth
processing foxnewshealth
processing gdnhealthcare
```

```
processing gdnhealthcare
processing KaiserHealthNews
processing latimeshealth
processing msnhealthnews
processing NBChealth
processing nprhealth
processing nytimeshealth
processing reuters_health
processing usnewshealth
processing wsjhealth
```

In [21]:

```
df
```

Out[21]:

20 rows × 15 columns (omitted printing of 9 columns)

	bbchealth	cbchealth	cnnhealth	everydayhealth	foxnewshealth	gdnhealthcare
	String	String	String	String	String	String
1	VIDEO	Ebola	RT	RT	studi	NHS
2	NHS	health	getfit	healthtalk	Ebola	RT
3	Ebola	RT	via	HealthTalk	cancer	GdnHealthcar
4	cancer	Canada	health	A	Newser	health
5	AUDIO	US	tip	food	help	healthcar
6	health	outbreak	cnnhealth	The	risk	patient
7	care	patient	CNN	eat	US	miss
8	death	doctor	The	EverydayHealth	health	care
9	patient	Canadian	Today	What	drug	via
10	UK	studi	What	How	patient	How
11	hospit	cancer	amp	weight	diseas	ViewsfromtheNHSfrontlin
12	risk	death	cancer	Q	brain	New
13	drug	drug	help	health	heart	Don
14	help	risk	kid	amp	vaccin	week
15	AampE	hospit	A	healthi	die	nurs
16	The	vaccin	How	Here	How	AampE
17	test	test	brain	help	flu	What
18	Hospit	Health	stori	diet	hospit	The
19	warn	WHO	Ebola	eatsmartbd	link	free
20	How	Ontario	drsanjaygupta	reason	New	staff

After looking at the top 20 frequently used words from 15 major news agencies, 11 of themajor agencies appear to be reporting on the same topics. They frequently use the words Ebola, cancer, and drug which are likely to be events that have occured in 2015 that caught national attention. The other remaining words are ominous words such as death, risk, outbreak, etc which likely written together in one tweet. However, four news agencies did not frequently mention Ebola, cancer, or drugs in their tweets. Two of the agencies (everydayhealth and usnewshealth) most frequent tweets were regarding eating, food, and diet. The two remaining agencies (gdnhealthcare and KaiserHealthNews) most frequent tweets were regarding health care (Obamacare, medicare, medicaid, healthcare).