Homework 2 - Text Processing and Unsupervised Learning - Due March 23rd

1. In this question, we will create a data set of Donald Trump's press releases and study the topics Donald Trump's press releases talk about.

(a) Download all of the press releases on Donald Trump's website (https://www.donaldjtrump.com/press-releases). Save each of these press releases as an html file (for example, one press release is https://www.donaldjtrump.com/press-releases/crooked-hillary-question-of-the-day27). There are approximately 900 press releases in total.

(b) Create a data set where you extract the text of the press release and the date of the press release.

(c) Apply LDA to the data set coming up with a reasonable labeling of the topics that the press releases relate to.

(d) Study the evolution of what Donald Trump's press releases talk about over time? Which topics is he talking more about towards the end of the campaign? Which topics is he talking less about.

Some hints: to complete (a), you will have to come up with a way to get the links for all the press releases. See the example I posted for scraping the congressional record for some ideas. Using regular expressions, you can use ".+" to match anything except endlines. To match anything including endlines, you can use "(.|\n)+". For viewing patterns over time, you can aggregate to the monthly level. Though probably not necessary, you can use the following code to strip html code from the text you extract:

```
doc <- htmlParse(text, asText=TRUE)
res <- try(xpathSApply(doc, "//p", xmlValue))
```

```
    if(!inherits(res, "try-error")) {
            plaintext <- paste(res,collapse="\n")
    } else {
            plaintext <- "ERROR"
    }
```

which uses the packages RCurl and XML. When converting html to text, there will be some garbage characters in the text, but these will probably not cause problems once you tokenize the text. Finally, so you can work on parts (b) and (c) if you are stuck on the previous parts, I posted the html files I downloaded as well as the text I extracted on google drive (pressrel.dat in the zip file).