File = WNIT3rdRnd\_WomensNIT\_UNCvUCLA\_12.out

I created this document to understand the tweet structure which drives the cleaning

process. This is a JSON format that can be read into a data frame quite nicely using the

jsonlite package. Further processing to get the word list includes removing; URL,

punctuation, numbers and stopwords. There still exists the ellipsis before or after some

words in the word list, e.g. "quar…". Further evaluation shows the ellipsis to be the Unicode

character “U+0085”. There are also word concatentations,

e.g. "usmgoldeneagles", and abbreviations, e.g. "miss" for Mississippi.

The data contains the BOM or Byte Order Mark found at the beginning of the file and

signifies the file is unicode. This is great but is a headache when processing text as it appear as "U+FEFF" in the string as what I believe is an HTML tag as it is enclosed in carrot

brackets. I found a quick solution that requires opening the file in NotePad++ and resaving as UTF-8 without BOM. The file without the BOM is called "WomensNIT noBOM.txt". We

need a more robust solution to process very large file sizes. Based on my research, the BOM plagues all programming languages; R, Python, Linux, JavaScript, etc. The BOM is really "0xEF,0xBB,0xBF". There are various regex solutions on the internet to try in the near future.

A later project may be to see if the tweets can be extracted from Twitter in such a way the data are already or mostly cleaned and the BOM removed. The ndjson package has functionality to stream in from an http .gz file. This package requires compiling before use.

{

"contributors":null,

"coordinates":null,

"created\_at":"Fri Mar 27 02:05:40 +0000 2015",

"entities":

{

"hashtags":

[

{

"indices":[126,133],

"text":"Temple"

},

{

"indices":[134,139],

"text":"WNIT"

}

],

"symbols":[],

"urls":[],

"user\_mentions":

[

{

"id":201739006,

"id\_str":"201739006",

"indices":[8,19],

"name":"Temple WBB",

"screen\_name":"TUOWLS\_WBB"

},

{

"id":861745004,

"id\_str":"861745004",

"indices":[72,82],

"name":"WNIT",

"screen\_name":"WomensNIT"

}

]

},

"favorite\_count":0,

"favorited":false,

"geo":null,

"id":581275866082185217,

"id\_str":"581275866082185217",

"in\_reply\_to\_screen\_name":null,

"in\_reply\_to\_status\_id":null,

"in\_reply\_to\_status\_id\_str":null,

"in\_reply\_to\_user\_id":null,

"in\_reply\_to\_user\_id\_str":null,

"lang":"en",

"metadata":

{

"iso\_language\_code":"en",

"result\_type":"recent"

},

"place":null,

"retweet\_count":0,

"retweeted":false,

"source":"<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",

"text":"Update: @TUOWLS\_WBB takes on Middle Tennessee State in Quarterfinals of @WomensNIT on Sunday at 4PM at Middle Tennessee State #Temple #WNIT",

"truncated":false,

"user":

{

"contributors\_enabled":false,

"created\_at":"Tue Oct 30 23:51:42 +0000 2012",

"default\_profile":true,

"default\_profile\_image":false,

"description":"Philadelphia Sports Writer for @PhillySportLive , Philadelphia Eagles Writer for @Phillysoc NCAA Action, Boxing & MMA & events in Philly, 26 years old",

"entities":

{

"description":{"urls":[]},

"url":

{

"urls":

[

{

"display\_url":"phillysportslive.com/author/davidm/",

"expanded\_url":"http://phillysportslive.com/author/davidm/",

"indices":[0,22],

"url":<http://t.co/YNAISMp2Oy>

}

]

}

},

"favourites\_count":397,

"follow\_request\_sent":null,

"followers\_count":792,

"following":null,

"friends\_count":1762,

"geo\_enabled":false,

"id":915754364,

"id\_str":"915754364",

"is\_translation\_enabled":false,

"is\_translator":false,

"lang":"en",

"listed\_count":13,

"location":"Havertown PA",

"name":"David Malandra Jr",

"notifications":null,

"profile\_background\_color":"C0DEED",

"profile\_background\_image\_url":"http://abs.twimg.com/images/themes/theme1/bg.png",

"profile\_background\_image\_url\_https":"https://abs.twimg.com/images/themes/theme1/bg.png",

"profile\_background\_tile":false,

"profile\_banner\_url":"https://pbs.twimg.com/profile\_banners/915754364/1353790586",

"profile\_image\_url":"http://pbs.twimg.com/profile\_images/2786715393/2ec3a5103e459313ee0ac1e6f69a2846\_normal.jpeg",

"profile\_image\_url\_https":"https://pbs.twimg.com/profile\_images/2786715393/2ec3a5103e459313ee0ac1e6f69a2846\_normal.jpeg",

"profile\_link\_color":"0084B4",

"profile\_location":null,

"profile\_sidebar\_border\_color":"C0DEED",

"profile\_sidebar\_fill\_color":"DDEEF6",

"profile\_text\_color":"333333",

"profile\_use\_background\_image":true,

"protected":false,

"screen\_name":"DaveM\_PSL",

"statuses\_count":24743,

"time\_zone":null,

"url":"http://t.co/YNAISMp2Oy",

"utc\_offset":null,

"verified":false

}

}