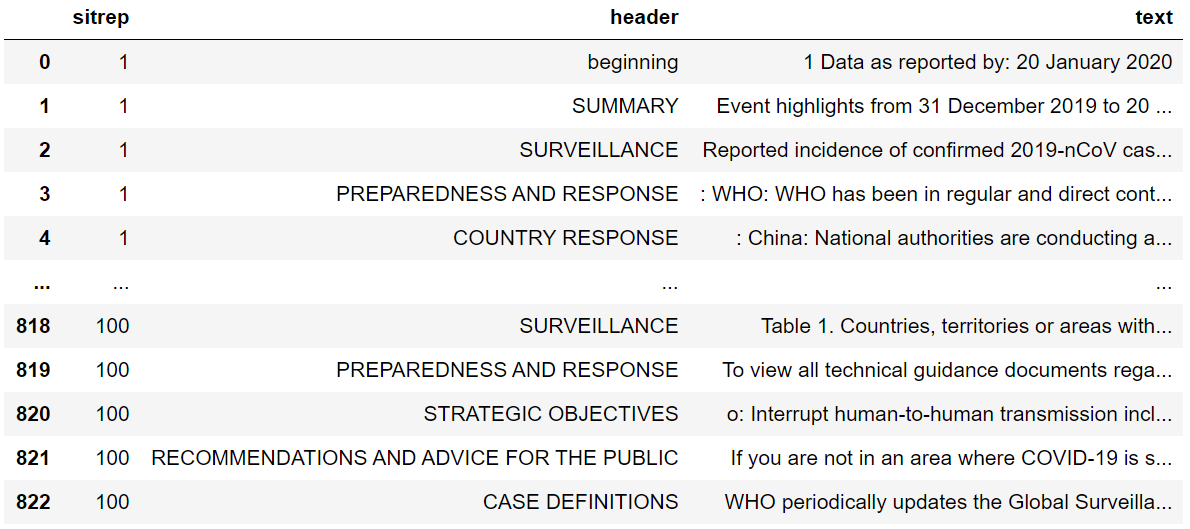
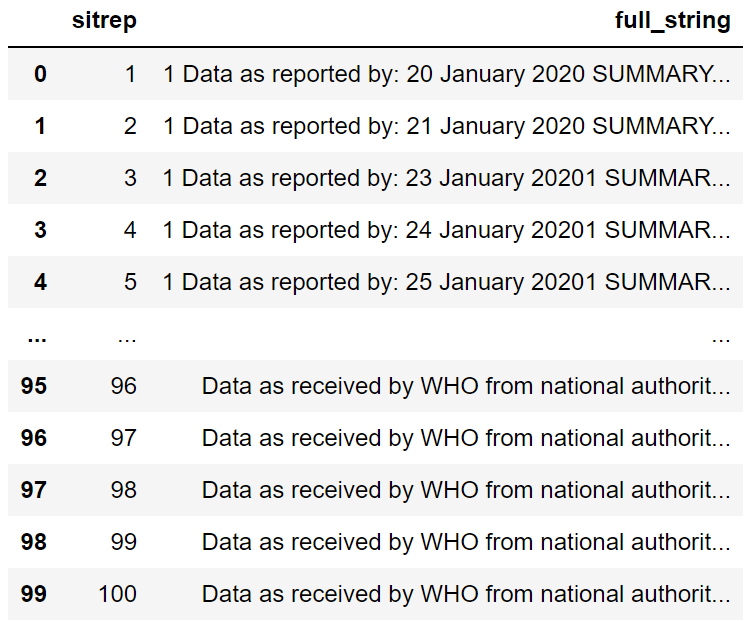
Text Analysis of WHO COVID Situational Reports

1. Download 1st 100 Reports (Jan 21st – April 29th)
   1. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
   2. 
2. GetPDFText Function in Python
   1. Extracts text using PyPDF2 package
   2. Formats text
      1. 
   3. Extracts Header titles using regex
   4. Separates text by header
   5. 
   6. Export dataframes as XLSX
3. Import XLSX files into SAS
4. Manipulate datasets in SAS
   1. Create header aliases
   2. Create data structure for each unit of observation
      1. Sentences
      2. Every 100 words
      3. Individual words
   3. Add variables that provide more info about each sitrep
      1. Date
      2. URL
      3. # territories w/ covid cases
      4. # cases worldwide
      5. # deaths worldwide
   4. Clean individual word dataset
      1. Remove 100 most common English words
      2. Remove all punctuation and numbers
      3. Lowercase all letters
5. Import datasets to Visual Analytics
   1. Used individual word dataset to explore contents
   2. Filter by headers, date, and word
   3. Text Topics
      1. Derived topic scores
      2. Compared topic relevance to COVID stats

Things to Consider:

* Unit of observation
* Removing certain words
* # topics derived