Understanding Mental Health through The NYTimes Articles

Mohammed Khan, Purbi Kanungo, Ruchil Barya

2020-12-15

# Introduction

New York Times is bestselling newspaper in New York State and read all over the world, it is an important part of today’s democrat world by educating and addressing the important issues; But did the paper prioritize the mental health the same way they prioritize other important issues for instance politics?

In today’s world mental health is an important health related issue that needs to be educate and address, because people are virtually connected and spend less time interacting with others outside the digital world leading to depression, stress and anxiety issues.

The idea of the project was to address mental health by collecting the articles from NYTimes archives and with the help of language modeling understanding the sentiment of how the mental health related issues were addressed over the last three decades? We proposed that over the decades the issues related to mental health has been improved and correctly addressed.

In a study (Wahl, 2011), 300 articles related to mental health were collected (1998-1999) and analyzed. The results indicated that negative tone of the article reduced slightly in the subsequent year. Another study was conducted collecting 30K NYT articles to explore the topics covered in the articles (Charissa R.,2018). Results suggested that over time, family/home/work-life and sociological trends have taken up a larger share in the total number of articles mentioning ‘mental health’.

Results from a more recent study in the United states (McGinty, 2016) covering 400 random news articles (1995-2014) helped us in understanding the change in topics covered over decades. According to the study, the two most frequently mentioned topics in news coverage about mental illness were violence and suicide. These studies inspired us to investigate more about this topic.

# Research Question

As most of the knowledge is gained through news media, it becomes very important to know what news media publish about mental health. It is also helpful to understand if there is a difference in news reporting of such sensitive topic over the years. (Wahl, 2002)

The goal of this research is to understand how Mental Health has been talked over the years by analyzing ‘The NY Times’ articles. As part of this analysis, main themes and overall tone of the articles published in 20th century and 21st century is identified. The major differences between 20th century and 21st century articles are also observed.

# Method

## Data

The study includes the analysis of 10,963 articles of ‘NY Times’ from year 1987 to 2017 which mentions ‘mental health’. The data consists of the features like ‘Title of the Article’, ’Author’, ‘Date’, ’Year’, ‘Text’. As part of the analysis, ‘Year’ and ‘Text’ has been used to understand the main topics and tone of the articles over the years.

For Factor Analysis, the data was converted to a Document-Term Matrix with the help of MEH tool. Since there are approximately 10,000 articles, it becomes difficult to analyze them all together. So, the data was split into two parts by ‘Year’. For the 20th century analysis, articles from year 1987 to 1999 were considered.3088 articles were analyzed as part of it. For the 21st century analysis, articles from year 2000 to 2017 were considered. 7083 articles were analyzed as part of it. Rows with any missing values are removed.

For Sentiment Analysis, the original data was used with features ‘Year’ and ‘Text’. The pre-processing of the data was done in python. The data was split into two 20th century and 21st century similarly as Factor Analysis.

## Method: Factor Analysis

To understand the themes of the articles, Factor analysis was applied. Since our data set was quite large, it was extremely important to use a method which converts the data into smaller and understandable form. Factor Analysis meets these requirements. It decreases several variables to few latent variables and describes group attributes. It finds hidden dimensions. It is efficient in finding the interrelationship between the variables which is needed in our current analysis to find the relationship and differences between 20th and 21st century articles.To decide the number of factors, a scree plot was created with the help of adjusted eigen values, unadjusted eigen values and estimate bias.Non parametric test was used to find the differences with the group

## Method: Sentiment Analysis

For Sentiment Analysis of mental health related articles, the pre-built model from the NLTK package, VADER was used. It takes text as input and returns polarity scores for positive, negative and neutral sentiments. It also provides a compound score which ranges between -1 (extreme negative) and 1 (extreme positive). For this analysis, the compound polarity score was used for understanding tone of the articles over the years.

VADER model is more suitable for shorter texts like tweets, comments and reviews. After using it to get sentiment scores for large articles, it was observed that the scores were inaccurate and vague. To get best results with this model, each article was divided into sentences using sent\_tokenize() method. Polarity scores of these sentences were calculated and the average score was calculated for the entire article.

The compound polarity score for each year was aggregated (mean score) and label for century (20th /21st Century) was added to create the final scores.

To understand emotions, NCR lexicon from the R’s syuzhet library was used. 5 emotions (anger, fear, sadness, surprise, trust) and 2 sentiments (negative, positive) were focused in the text. The function get\_nrc\_sentiment() takes the article text and returns valence score for the sentiments and emotions. These scores are aggregated across years to understand the trend.

# Analyis

## Factor Analyis

### Libraries

Loading the required libraries

library(psych)  
library(readr)

library(FactoMineR)  
library(GPArotation)  
library(paran)

# sentiment analysis  
library(ggplot2)

library(reticulate)  
library(reshape2)  
library(viridis)

### 

### Data 20th Century

The data has been pre-processed through MEH tool.The output from the MEH tool has been loaded into the R.The data is divided into two parts based on the ‘year’ variables.

setwd("data20century/")  
  
X2020\_12\_13\_MEH\_DTM\_Verbose<- read\_csv("2020-12-13\_MEH\_DTM\_Verbose.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_double()  
## )  
## ℹ Use `spec()` for the full column specifications.

data20 <- X2020\_12\_13\_MEH\_DTM\_Verbose[,-c(2,3,4)]  
  
names(data20)[1] <- "decade"  
data20 <- data20[complete.cases(data20), ]  
data20$decade <- factor(ifelse(data20$decade<=1999, "20th\_Century", "21th\_Century"))

### Data 21st Century

The data has been pre-processed through MEH tool.The output from the MEH tool has been loaded into the R.

X2020\_11\_21\_MEH\_DTM\_Verbose <- read\_csv("verbose/2020-11-21\_MEH\_DTM\_Verbose.csv")

## Warning: Duplicated column names deduplicated: 'year' => 'year\_1' [194]

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_double(),  
## Filename = col\_character()  
## )  
## ℹ Use `spec()` for the full column specifications.

verbose <- X2020\_11\_21\_MEH\_DTM\_Verbose  
verbose <- verbose[complete.cases(verbose), ]  
verbose$century <- factor(ifelse(verbose$year>=2000, "21th\_Century", "20th\_Century"))  
table(verbose$century)

##   
## 20th\_Century 21th\_Century   
## 3880 7083

century <- verbose$century  
  
data21 <- read\_csv("newdata21century/data21.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_double()  
## )  
## ℹ Use `spec()` for the full column specifications.

data21 <- data21[,-c(2,3,4)]  
  
names(data21)[1] <- "decade"  
data21 <- data21[complete.cases(data21), ]  
  
data21$decade <- factor(ifelse(data21$decade>=2000, "21th\_Century", "20th\_Century"))

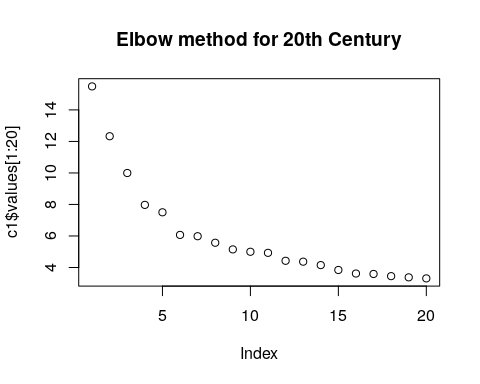
### Analysis for 20th Century

The elbow method was used to determine the number of factors .As the scree plot shows, there is a change of slope for n = 5 and 6. We have tried to find the best possibles themes based on these,n=6 showed promising results(themes were interpretable).So, we have used number of factors as 6.

paran(data20[,2:1002], iterations = 100, centile = 95)

##   
## Using eigendecomposition of correlation matrix.  
## Computing: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
##   
##   
## Results of Horn's Parallel Analysis for component retention  
## 100 iterations, using the 95 centile estimate  
##   
## --------------------------------------------------   
## Component Adjusted Unadjusted Estimated   
## Eigenvalue Eigenvalue Bias   
## --------------------------------------------------   
## 1 14.212245 15.490268 1.278023  
## 2 11.074789 12.328614 1.253824  
## 3 8.754154 9.995296 1.241141  
## 4 6.744538 7.972765 1.228227  
## 5 6.284779 7.502172 1.217393  
## 6 4.858511 6.067744 1.209232  
## 7 4.783790 5.984376 1.200585  
## 8 4.381773 5.570489 1.188715  
## 9 3.967532 5.148919 1.181387  
## 10 3.824567 4.999112 1.174545  
## 11 3.764406 4.930447 1.166040  
## 12 3.267224 4.425914 1.158689  
## 13 3.218673 4.370203 1.151530  
## 14 3.010060 4.152765 1.142704  
## 15 2.705608 3.843294 1.137686  
## 16 2.486952 3.617921 1.130968  
## 17 2.467991 3.590666 1.122674  
## 18 2.335493 3.451701 1.116207  
## 19 2.264338 3.375170 1.110831  
## 20 2.203969 3.308768 1.104799  
## 21 2.125445 3.224140 1.098695  
## 22 2.099723 3.193408 1.093685  
## 23 2.022770 3.109295 1.086525  
## 24 1.983434 3.064350 1.080915  
## 25 1.965243 3.040288 1.075045  
## 26 1.940032 3.008913 1.068880  
## 27 1.902027 2.965157 1.063129  
## 28 1.885059 2.942808 1.057749  
## 29 1.819810 2.872317 1.052506  
## 30 1.777061 2.825228 1.048166  
## 31 1.757059 2.799742 1.042682  
## 32 1.748645 2.786869 1.038224  
## 33 1.723969 2.756534 1.032565  
## 34 1.675467 2.704238 1.028770  
## 35 1.665257 2.689269 1.024011  
## 36 1.648229 2.667034 1.018805  
## 37 1.634648 2.648709 1.014060  
## 38 1.590721 2.598181 1.007459  
## 39 1.581860 2.586423 1.004562  
## 40 1.570232 2.569325 0.999093  
## 41 1.563015 2.558971 0.995955  
## 42 1.551607 2.541925 0.990318  
## 43 1.519619 2.505087 0.985467  
## 44 1.501484 2.482040 0.980555  
## 45 1.497489 2.473365 0.975875  
## 46 1.474470 2.446207 0.971736  
## 47 1.468460 2.436504 0.968043  
## 48 1.439305 2.402640 0.963334  
## 49 1.427453 2.385680 0.958226  
## 50 1.413105 2.367202 0.954097  
## 51 1.411287 2.362253 0.950965  
## 52 1.404010 2.351082 0.947072  
## 53 1.394360 2.336124 0.941763  
## 54 1.387952 2.325467 0.937515  
## 55 1.388588 2.321713 0.933125  
## 56 1.362900 2.292535 0.929634  
## 57 1.358588 2.285113 0.926525  
## 58 1.351209 2.273168 0.921959  
## 59 1.331917 2.249190 0.917273  
## 60 1.328430 2.242209 0.913778  
## 61 1.321049 2.230612 0.909562  
## 62 1.317139 2.222004 0.904864  
## 63 1.307347 2.209356 0.902008  
## 64 1.298511 2.194800 0.896288  
## 65 1.299196 2.192686 0.893490  
## 66 1.293973 2.183780 0.889807  
## 67 1.287643 2.173987 0.886344  
## 68 1.291799 2.173398 0.881599  
## 69 1.277698 2.156308 0.878610  
## 70 1.266253 2.140421 0.874168  
## 71 1.266211 2.137379 0.871168  
## 72 1.264450 2.131348 0.866898  
## 73 1.254929 2.118980 0.864051  
## 74 1.251121 2.111406 0.860284  
## 75 1.251308 2.107374 0.856065  
## 76 1.247961 2.099789 0.851827  
## 77 1.237122 2.085809 0.848687  
## 78 1.232550 2.078130 0.845579  
## 79 1.232834 2.073871 0.841037  
## 80 1.231013 2.068654 0.837641  
## 81 1.227898 2.062363 0.834465  
## 82 1.221123 2.051316 0.830193  
## 83 1.221558 2.048092 0.826534  
## 84 1.211915 2.034629 0.822714  
## 85 1.205633 2.026130 0.820497  
## 86 1.206121 2.022796 0.816674  
## 87 1.200078 2.013765 0.813687  
## 88 1.201951 2.012247 0.810295  
## 89 1.202142 2.007936 0.805794  
## 90 1.192840 1.995969 0.803129  
## 91 1.192784 1.992621 0.799836  
## 92 1.181418 1.977852 0.796433  
## 93 1.183279 1.976436 0.793156  
## 94 1.183778 1.973422 0.789644  
## 95 1.181053 1.966949 0.785895  
## 96 1.175050 1.958247 0.783197  
## 97 1.172261 1.951418 0.779157  
## 98 1.158605 1.934502 0.775897  
## 99 1.151402 1.924295 0.772892  
## 100 1.141512 1.912003 0.770491  
## 101 1.136319 1.902971 0.766652  
## 102 1.137987 1.901000 0.763012  
## 103 1.136316 1.896553 0.760237  
## 104 1.131976 1.889723 0.757746  
## 105 1.132704 1.886930 0.754225  
## 106 1.131722 1.882273 0.750550  
## 107 1.126765 1.874245 0.747479  
## 108 1.124509 1.868510 0.744001  
## 109 1.121238 1.861842 0.740604  
## 110 1.120611 1.858190 0.737578  
## 111 1.115909 1.850845 0.734935  
## 112 1.114948 1.846683 0.731735  
## 113 1.106847 1.834816 0.727969  
## 114 1.106812 1.831776 0.724964  
## 115 1.101212 1.822973 0.721760  
## 116 1.099939 1.818780 0.718840  
## 117 1.094551 1.811037 0.716486  
## 118 1.093483 1.806025 0.712541  
## 119 1.094010 1.802685 0.708674  
## 120 1.095343 1.801757 0.706413  
## 121 1.094505 1.798262 0.703757  
## 122 1.094773 1.795521 0.700748  
## 123 1.089008 1.786737 0.697729  
## 124 1.089446 1.784565 0.695119  
## 125 1.087889 1.780328 0.692438  
## 126 1.087867 1.776151 0.688283  
## 127 1.086791 1.772456 0.685665  
## 128 1.079227 1.761638 0.682411  
## 129 1.076477 1.756073 0.679595  
## 130 1.073741 1.750483 0.676742  
## 131 1.067966 1.742062 0.674096  
## 132 1.063932 1.735195 0.671262  
## 133 1.061035 1.729640 0.668604  
## 134 1.061516 1.727889 0.666372  
## 135 1.058966 1.721608 0.662641  
## 136 1.062177 1.721456 0.659279  
## 137 1.055356 1.711852 0.656495  
## 138 1.052837 1.706626 0.653788  
## 139 1.049215 1.700633 0.651418  
## 140 1.050574 1.698530 0.647956  
## 141 1.044318 1.689743 0.645424  
## 142 1.044117 1.686555 0.642438  
## 143 1.041711 1.680944 0.639233  
## 144 1.039956 1.676248 0.636291  
## 145 1.036899 1.671155 0.634256  
## 146 1.031715 1.662887 0.631172  
## 147 1.033350 1.661532 0.628182  
## 148 1.033113 1.658262 0.625148  
## 149 1.028380 1.650638 0.622257  
## 150 1.026966 1.647159 0.620192  
## 151 1.024979 1.641674 0.616695  
## 152 1.022595 1.636919 0.614323  
## 153 1.022868 1.634394 0.611525  
## 154 1.020217 1.628622 0.608405  
## 155 1.019156 1.624380 0.605223  
## 156 1.018201 1.621537 0.603335  
## 157 1.019491 1.620099 0.600607  
## 158 1.015857 1.613980 0.598123  
## 159 1.014291 1.609182 0.594891  
## 160 1.014335 1.606808 0.592472  
## 161 1.012837 1.602785 0.589948  
## 162 1.011036 1.598518 0.587482  
## 163 1.012740 1.597126 0.584386  
## 164 1.009873 1.591916 0.582042  
## 165 1.009114 1.588465 0.579350  
## 166 1.007893 1.584907 0.577013  
## 167 1.006505 1.581020 0.574515  
## 168 1.004188 1.575465 0.571277  
## --------------------------------------------------   
##   
## Adjusted eigenvalues > 1 indicate dimensions to retain.  
## (168 components retained)

c1 = principal(data20[,2:1002], nfactors=6, rotate = 'oblimin')  
plot(c1$values[1:20])#Select the n values that are reasonable solutions  
title("Elbow method for 20th Century")



### Themes based on Six factors - Best Solution for 20th Century

Theme1: Theme 1 talks about that mental illness has been emerged as a problem and solution has to be found for this.

Theme 2: Theme 2 talks about the senate, legislative, bill, election, committee, budget, tax. This theme seems off the topic which talks about bill and elections.

Theme 3: Theme 3 includes many words related to research, psychiatry, medication, treatment, therapy. This theme is trying to emphasize on the importance of medication and treatment of mentally ill patients, how the mental patients should be treated and how more research can improve their condition.

Theme 4: Theme 4 talks about the homeless, care, shelter, outpatient, community. This theme mainly talks about how the community services can be provided.

Theme 5: Theme 5 includes words like court, lawyer, guilty, murder, justice, death investigate. It is talking about the people who perform crime due to their illness.

Theme 6: Theme 6 talks about the words like medical aid, insurance, facility, finance, hospital, services, coverage which mainly talks about the services, insurance provided by the government for mental health.

#Six factors  
loads5 = as.data.frame(c1$loadings[1:1001,1:6])  
theme1.20 = subset(loads5[order(loads5$TC1, decreasing = T),], TC1 >= 0.15, select = c(TC1))  
theme2.20 = subset(loads5[order(loads5$TC2, decreasing = T),], TC2 >= 0.15, select = c(TC2))  
theme3.20 = subset(loads5[order(loads5$TC3, decreasing = T),], TC3 >= 0.15, select = c(TC3))  
theme4.20 = subset(loads5[order(loads5$TC4, decreasing = T),], TC4 >= 0.15, select = c(TC4))  
theme5.20 = subset(loads5[order(loads5$TC5, decreasing = T),], TC5 >= 0.15, select = c(TC5))  
theme6.20 = subset(loads5[order(loads5$TC6, decreasing = T),], TC6 >= 0.15, select = c(TC6))  
  
theme1.20

## TC1  
## said 0.3229875  
## people 0.3183784  
## time 0.2461815  
## problem 0.2403625  
## see 0.2373638  
## find 0.2103357  
## give 0.2063042  
## look 0.1876638  
## call 0.1794379  
## change 0.1787507  
## day 0.1769548  
## talk 0.1735151  
## turn 0.1691626  
## show 0.1628124  
## better 0.1605493  
## add 0.1587263  
## place 0.1586886  
## thought 0.1567241  
## number 0.1563122  
## happen 0.1548297  
## stop 0.1514701

theme2.20

## TC2  
## budget 0.5854845  
## republican 0.5602051  
## senate 0.5249389  
## governor 0.4961936  
## legislature 0.4958711  
## democrat 0.4941367  
## assembly 0.4767150  
## leader 0.4660157  
## spend 0.4395604  
## legislative 0.4334059  
## cut 0.4257425  
## tax 0.3962008  
## gov 0.3745962  
## bill 0.3731359  
## propose 0.3723979  
## fiscal 0.3684185  
## democratic 0.3602871  
## vote 0.3594861  
## proposal 0.3538146  
## cuomo 0.3467736  
## state 0.3260878  
## pass 0.3187573  
## legislator 0.3161321  
## legislation 0.3073022  
## senator 0.3059823  
## election 0.3056260  
## committee 0.2969880  
## political 0.2886649  
## majority 0.2769826  
## approve 0.2746376  
## money 0.2690472  
## government 0.2566987  
## albany 0.2552730  
## measure 0.2517062  
## increase 0.2370666  
## today 0.2368347  
## local 0.2268316  
## balance 0.2253936  
## year 0.2231980  
## congress 0.2210963  
## campaign 0.2139759  
## oppose 0.2113453  
## favor 0.2006211  
## eliminate 0.1985925  
## aide 0.1917193  
## politics 0.1907056  
## saving 0.1888022  
## agreement 0.1830062  
## final 0.1829701  
## close 0.1748699  
## process 0.1733838  
## president 0.1733659  
## reject 0.1719693  
## position 0.1717403  
## welfare 0.1712889  
## chairman 0.1707460  
## agree 0.1700671  
## medicaid 0.1689146  
## clinton 0.1660541  
## party 0.1656593  
## education 0.1643922  
## federal 0.1639440  
## aid 0.1596655  
## representative 0.1584215  
## council 0.1584170  
## mayor 0.1581574  
## race 0.1563151  
## raise 0.1553998  
## plan 0.1532806  
## issue 0.1524917  
## district 0.1522884  
## reduce 0.1516569  
## impose 0.1514948  
## debate 0.1503606

theme3.20

## TC3  
## study 0.5314151  
## dr 0.4123077  
## researcher 0.3890516  
## research 0.3585503  
## institute 0.3503466  
## disorder 0.3462265  
## depress 0.3282978  
## national 0.3278421  
## symptom 0.3012999  
## journal 0.3012742  
## finding 0.2989636  
## treatment 0.2970925  
## disease 0.2925876  
## psychiatry 0.2904180  
## scientist 0.2879030  
## brain 0.2832708  
## effect 0.2825677  
## illness 0.2808089  
## publish 0.2548285  
## factor 0.2525412  
## severe 0.2488806  
## american 0.2466420  
## treat 0.2434072  
## risk 0.2432343  
## suggest 0.2346910  
## suffer 0.2291006  
## psychological 0.2271784  
## effective 0.2228344  
## psychiatrist 0.2217994  
## schizophrenia 0.2188264  
## professor 0.2185096  
## behavior 0.2167589  
## therapy 0.2164745  
## problem 0.2088886  
## medical 0.2023373  
## mental 0.2004067  
## professional 0.1991341  
## medicine 0.1978847  
## editor 0.1966777  
## author 0.1936860  
## physical 0.1936660  
## clinical 0.1930507  
## affect 0.1929805  
## experience 0.1919160  
## develop 0.1902990  
## result 0.1898231  
## science 0.1871016  
## data 0.1868337  
## psychologist 0.1852139  
## difficulty 0.1818318  
## colleague 0.1796729  
## function 0.1789263  
## emotion 0.1783065  
## medication 0.1782516  
## evidence 0.1774736  
## normal 0.1769393  
## health 0.1767727  
## drug 0.1762479  
## show 0.1706369  
## produce 0.1691645  
## major 0.1685549  
## relate 0.1661531  
## understanding 0.1639979  
## report 0.1634959  
## serious 0.1633618  
## important 0.1610445  
## recognize 0.1607697  
## heart 0.1589773  
## expert 0.1585955  
## diagnose 0.1580567  
## conduct 0.1576951  
## subject 0.1574451  
## involve 0.1569116  
## stress 0.1568474  
## feeling 0.1565813  
## sex 0.1553276  
## ability 0.1550576  
## compare 0.1546621  
## influence 0.1545260  
## difference 0.1531920  
## practice 0.1518398  
## lead 0.1504659

theme4.20

## TC4  
## mental 0.5105373  
## mentally 0.4636396  
## hospital 0.4523920  
## state 0.4212765  
## psychiatric 0.4018745  
## service 0.3990116  
## patient 0.3849575  
## city 0.3787691  
## health 0.3748969  
## community 0.3741437  
## provide 0.3506143  
## homeless 0.3471517  
## care 0.3204609  
## bed 0.3112158  
## treatment 0.2903716  
## program 0.2822521  
## official 0.2791264  
## commissioner 0.2563427  
## york 0.2470762  
## emergency 0.2467282  
## advocate 0.2413302  
## shelter 0.2406789  
## system 0.2397626  
## center 0.2390244  
## director 0.2291447  
## ward 0.2264146  
## office 0.2237476  
## staff 0.2231933  
## institution 0.2213954  
## outpatient 0.2160459  
## department 0.2127017  
## agency 0.2126472  
## residence 0.2071515  
## editor 0.2054650  
## hospitalize 0.1920306  
## permanent 0.1902944  
## illness 0.1870832  
## resource 0.1867385  
## room 0.1848089  
## street 0.1716520  
## number 0.1675995  
## medical 0.1668416  
## resident 0.1635814  
## advocacy 0.1635010  
## private 0.1621737  
## release 0.1603724  
## commitment 0.1574653  
## long-term 0.1570444  
## administration 0.1565140  
## manhattan 0.1563922  
## mayor 0.1556862  
## quality 0.1540736  
## under 0.1540330  
## reduce 0.1538656  
## dangerous 0.1513178  
## executive 0.1505665  
## social 0.1501209

theme5.20

## TC5  
## court 0.5021434  
## judge 0.4529031  
## trial 0.4327485  
## lawyer 0.4264589  
## charge 0.4259348  
## attorney 0.3628996  
## case 0.3539052  
## supreme 0.3438282  
## murder 0.3434636  
## prosecutor 0.3373949  
## sentence 0.3324800  
## criminal 0.3286066  
## justice 0.3233284  
## defense 0.3158973  
## accuse 0.2990919  
## kill 0.2912545  
## guilty 0.2905446  
## order 0.2797488  
## convict 0.2747745  
## appeal 0.2728992  
## legal 0.2702825  
## release 0.2650488  
## rule 0.2647659  
## law 0.2630223  
## prison 0.2509616  
## file 0.2495499  
## decision 0.2476767  
## death 0.2434251  
## arrest 0.2412383  
## right 0.2391672  
## police 0.2353837  
## official 0.2351936  
## stand 0.2334172  
## argue 0.2311440  
## told 0.2297156  
## today 0.2283194  
## assault 0.2276560  
## commit 0.2245334  
## hear 0.2216597  
## refuse 0.2113293  
## investigation 0.2073152  
## said 0.2044025  
## suit 0.2011653  
## year-old 0.2002185  
## evidence 0.1969604  
## lawsuit 0.1960462  
## spokesman 0.1910886  
## yesterday 0.1906813  
## argument 0.1848356  
## incident 0.1839728  
## record 0.1808962  
## investigate 0.1799589  
## civil 0.1787485  
## comment 0.1785302  
## claim 0.1678094  
## statement 0.1663934  
## deny 0.1657090  
## jail 0.1648549  
## opinion 0.1640973  
## determine 0.1621367  
## protect 0.1616335  
## contend 0.1594706  
## review 0.1589410  
## condition 0.1568890  
## psychiatric 0.1566464  
## permit 0.1560346  
## district 0.1549303  
## crime 0.1531079  
## under 0.1510813

theme6.20

## TC6  
## insurance 0.4667477  
## coverage 0.4453739  
## company 0.4208724  
## care 0.4158674  
## cost 0.4127113  
## benefit 0.4100646  
## health 0.4042970  
## percent 0.3997597  
## employer 0.3979255  
## bond 0.3602728  
## pay 0.3385055  
## revenue 0.3371291  
## plan 0.3167443  
## price 0.3075081  
## rate 0.2974104  
## standard 0.2842796  
## manage 0.2823352  
## market 0.2821242  
## cover 0.2812583  
## employee 0.2727461  
## require 0.2648456  
## limit 0.2533252  
## business 0.2411854  
## offer 0.2372984  
## congress 0.2346155  
## federal 0.2334072  
## finance 0.2308723  
## security 0.2290239  
## note 0.2263685  
## provide 0.2145234  
## income 0.2116879  
## buy 0.2080766  
## payment 0.2057965  
## increase 0.2012879  
## general 0.1992827  
## medicaid 0.1900118  
## fee 0.1886897  
## share 0.1869440  
## tax 0.1864449  
## expensive 0.1848619  
## bill 0.1812308  
## poor 0.1791223  
## dollar 0.1774729  
## industry 0.1770673  
## under 0.1758466  
## 000 0.1748199  
## average 0.1741738  
## government 0.1730573  
## doctor 0.1690929  
## rise 0.1690463  
## facility 0.1676132  
## base 0.1673357  
## management 0.1654543  
## network 0.1653814  
## schedule 0.1653590  
## amount 0.1629136  
## 150 0.1611179  
## total 0.1595487  
## clinton 0.1573364  
## sale 0.1540414

### Statistics (Fit Indices)- 20th Century

The best solution was with number of factors as 6 for 20th century,the statistics are as shown below. Root mean square of residuals is 0.02. It’s lower value is considered as good.So, Root mean square of residuals is good for this model.Root mean squared error of approximation is 0.0156. Its upper confidence interval is 0.0156 and lower confidence interval is 0.0155. It’s value is quite small which is good for current model. The tucker lewis index is 0.24 which is not close to 1. So it can be improved.

#Root mean square of the residuals  
#The lower the better  
fa.stats(data20[,2:1002],f = c1)$rms # for 5 factors

## Warning in fa.stats(data20[, 2:1002], f = c1): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## [1] 0.02387591

#Root mean squared error of approximation  
#Smaller values better  
fa.stats(data20[,2:1002],f = c1)$RMSEA # for 5 factors

## Warning in fa.stats(data20[, 2:1002], f = c1): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## RMSEA lower upper confidence   
## 0.01560048 0.01554927 0.01565995 0.95000000

#tucker lewis index  
#Closer to 1 is better  
  
fa.stats(data20[,2:1002],f = c1)$TLI # for 5 factors

## Warning in fa.stats(data20[, 2:1002], f = c1): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## [1] 0.2362547

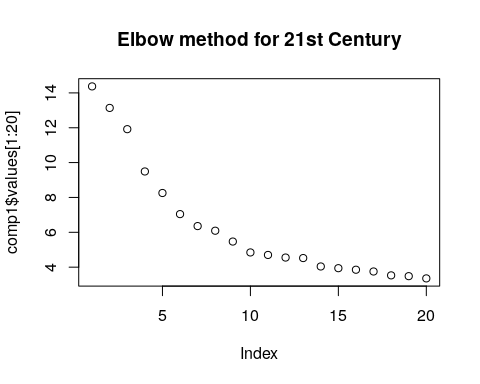
### Analysis for 21st century

The elbow method was used to determine the number of factors .As the principal plot shows, there is a change of slope for n = 7. We have tried to find the best possibles themes based on these,n=7 showed promising results(themes were more interpretable ).So, we have used number of factors as 7.

paran(data21[,2:1231], iterations = 100, centile = 95)

##   
## Using eigendecomposition of correlation matrix.  
## Computing: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
##   
##   
## Results of Horn's Parallel Analysis for component retention  
## 100 iterations, using the 95 centile estimate  
##   
## --------------------------------------------------   
## Component Adjusted Unadjusted Estimated   
## Eigenvalue Eigenvalue Bias   
## --------------------------------------------------   
## 1 13.360520 14.372136 1.011616  
## 2 12.140976 13.136951 0.995975  
## 3 10.931515 11.917646 0.986130  
## 4 8.513695 9.490696 0.977001  
## 5 7.283776 8.255351 0.971575  
## 6 6.080913 7.044089 0.963175  
## 7 5.398528 6.357002 0.958473  
## 8 5.139049 6.091266 0.952216  
## 9 4.524008 5.469783 0.945775  
## 10 3.903805 4.844696 0.940890  
## 11 3.765779 4.700781 0.935001  
## 12 3.622605 4.551734 0.929128  
## 13 3.596601 4.522294 0.925693  
## 14 3.118666 4.039347 0.920681  
## 15 3.024783 3.941419 0.916635  
## 16 2.941142 3.852963 0.911821  
## 17 2.843998 3.751524 0.907525  
## 18 2.626560 3.529133 0.902572  
## 19 2.584535 3.483641 0.899106  
## 20 2.460475 3.355232 0.894756  
## 21 2.421204 3.311859 0.890654  
## 22 2.358540 3.246099 0.887559  
## 23 2.278279 3.161380 0.883100  
## 24 2.226880 3.106117 0.879237  
## 25 2.193384 3.068979 0.875595  
## 26 2.144189 3.015757 0.871568  
## 27 2.121059 2.989177 0.868118  
## 28 2.032370 2.897173 0.864802  
## 29 2.025513 2.886138 0.860624  
## 30 1.973735 2.830195 0.856460  
## 31 1.926934 2.781182 0.854247  
## 32 1.843604 2.694880 0.851275  
## 33 1.825067 2.672422 0.847354  
## 34 1.818267 2.662126 0.843859  
## 35 1.816221 2.657050 0.840829  
## 36 1.753230 2.591561 0.838331  
## 37 1.732354 2.566313 0.833958  
## 38 1.721620 2.553095 0.831475  
## 39 1.689104 2.517212 0.828108  
## 40 1.685856 2.511189 0.825332  
## 41 1.667363 2.489176 0.821812  
## 42 1.665814 2.484504 0.818690  
## 43 1.635903 2.451443 0.815540  
## 44 1.632200 2.445034 0.812834  
## 45 1.594020 2.403889 0.809868  
## 46 1.569180 2.376431 0.807251  
## 47 1.559828 2.363785 0.803957  
## 48 1.555964 2.356957 0.800992  
## 49 1.554693 2.352908 0.798215  
## 50 1.547105 2.341866 0.794761  
## 51 1.527510 2.319921 0.792410  
## 52 1.517337 2.306916 0.789578  
## 53 1.513489 2.300307 0.786818  
## 54 1.487974 2.272945 0.784971  
## 55 1.471180 2.252688 0.781507  
## 56 1.467490 2.245481 0.777991  
## 57 1.463820 2.239670 0.775850  
## 58 1.454051 2.226776 0.772724  
## 59 1.448192 2.217909 0.769716  
## 60 1.423601 2.191171 0.767570  
## 61 1.420784 2.186015 0.765230  
## 62 1.416965 2.179111 0.762145  
## 63 1.406280 2.166450 0.760169  
## 64 1.401998 2.159357 0.757359  
## 65 1.391094 2.145125 0.754030  
## 66 1.380890 2.133244 0.752353  
## 67 1.379341 2.128559 0.749218  
## 68 1.376638 2.123352 0.746714  
## 69 1.364007 2.107560 0.743553  
## 70 1.361806 2.103609 0.741803  
## 71 1.359692 2.098410 0.738717  
## 72 1.351109 2.087543 0.736433  
## 73 1.345713 2.079308 0.733595  
## 74 1.343610 2.075263 0.731652  
## 75 1.334795 2.064538 0.729742  
## 76 1.333676 2.061134 0.727457  
## 77 1.330841 2.055162 0.724320  
## 78 1.326389 2.048499 0.722110  
## 79 1.321599 2.040852 0.719253  
## 80 1.313037 2.030526 0.717489  
## 81 1.311376 2.026524 0.715148  
## 82 1.312485 2.024958 0.712472  
## 83 1.307434 2.017232 0.709798  
## 84 1.301818 2.009739 0.707920  
## 85 1.285723 1.991229 0.705506  
## 86 1.286042 1.989042 0.702999  
## 87 1.277578 1.977443 0.699865  
## 88 1.278543 1.975994 0.697450  
## 89 1.272116 1.967933 0.695817  
## 90 1.272233 1.964968 0.692735  
## 91 1.262601 1.953519 0.690917  
## 92 1.260802 1.949835 0.689033  
## 93 1.256101 1.941875 0.685774  
## 94 1.255055 1.938866 0.683810  
## 95 1.244832 1.926710 0.681877  
## 96 1.246782 1.925855 0.679072  
## 97 1.244130 1.920824 0.676693  
## 98 1.239709 1.914782 0.675072  
## 99 1.231760 1.904718 0.672958  
## 100 1.227769 1.898344 0.670574  
## 101 1.227423 1.895803 0.668380  
## 102 1.221743 1.887925 0.666181  
## 103 1.220648 1.884969 0.664321  
## 104 1.219254 1.881189 0.661935  
## 105 1.215718 1.875475 0.659756  
## 106 1.214022 1.871584 0.657562  
## 107 1.210928 1.866133 0.655205  
## 108 1.209759 1.863205 0.653445  
## 109 1.204362 1.855186 0.650823  
## 110 1.199743 1.848343 0.648600  
## 111 1.195019 1.842332 0.647313  
## 112 1.191639 1.836329 0.644689  
## 113 1.186680 1.829376 0.642695  
## 114 1.182413 1.822658 0.640244  
## 115 1.180521 1.819131 0.638609  
## 116 1.180387 1.816485 0.636097  
## 117 1.180384 1.814442 0.634058  
## 118 1.179274 1.811749 0.632475  
## 119 1.170685 1.800936 0.630251  
## 120 1.167967 1.795815 0.627847  
## 121 1.166069 1.791968 0.625898  
## 122 1.165775 1.789271 0.623496  
## 123 1.165446 1.787228 0.621782  
## 124 1.163913 1.783929 0.620016  
## 125 1.159839 1.777950 0.618111  
## 126 1.158184 1.773390 0.615206  
## 127 1.154811 1.768342 0.613530  
## 128 1.153235 1.765042 0.611806  
## 129 1.150122 1.759986 0.609864  
## 130 1.147452 1.755281 0.607828  
## 131 1.148664 1.754364 0.605700  
## 132 1.148231 1.751544 0.603312  
## 133 1.145651 1.747200 0.601549  
## 134 1.141451 1.741478 0.600027  
## 135 1.142129 1.740271 0.598141  
## 136 1.138649 1.734768 0.596119  
## 137 1.136646 1.730970 0.594324  
## 138 1.129625 1.722294 0.592669  
## 139 1.130405 1.720693 0.590287  
## 140 1.125439 1.713670 0.588231  
## 141 1.126540 1.712307 0.585767  
## 142 1.122403 1.706888 0.584484  
## 143 1.122370 1.704118 0.581747  
## 144 1.123860 1.703880 0.580019  
## 145 1.119178 1.697562 0.578383  
## 146 1.119593 1.695848 0.576255  
## 147 1.118413 1.692869 0.574455  
## 148 1.116408 1.689951 0.573543  
## 149 1.116045 1.686158 0.570113  
## 150 1.112116 1.681313 0.569197  
## 151 1.109324 1.676071 0.566746  
## 152 1.109775 1.674522 0.564747  
## 153 1.108046 1.671080 0.563033  
## 154 1.106791 1.667940 0.561148  
## 155 1.107243 1.666037 0.558794  
## 156 1.105119 1.662150 0.557031  
## 157 1.102115 1.657865 0.555749  
## 158 1.102573 1.656130 0.553556  
## 159 1.099489 1.651213 0.551723  
## 160 1.095642 1.645393 0.549751  
## 161 1.095758 1.643727 0.547968  
## 162 1.091519 1.637839 0.546319  
## 163 1.090031 1.634323 0.544291  
## 164 1.088616 1.631356 0.542739  
## 165 1.085519 1.625877 0.540358  
## 166 1.083999 1.622517 0.538518  
## 167 1.078564 1.615519 0.536954  
## 168 1.076718 1.612039 0.535321  
## 169 1.078385 1.611908 0.533523  
## 170 1.077752 1.609343 0.531591  
## 171 1.078735 1.608776 0.530041  
## 172 1.074208 1.602142 0.527933  
## 173 1.071686 1.597994 0.526308  
## 174 1.071783 1.596537 0.524754  
## 175 1.068763 1.591554 0.522791  
## 176 1.068538 1.589342 0.520804  
## 177 1.065526 1.584469 0.518943  
## 178 1.064493 1.581410 0.516916  
## 179 1.064850 1.580256 0.515406  
## 180 1.063394 1.576578 0.513183  
## 181 1.063613 1.575116 0.511503  
## 182 1.061935 1.571961 0.510025  
## 183 1.060652 1.569420 0.508767  
## 184 1.057060 1.563686 0.506625  
## 185 1.055899 1.560563 0.504663  
## 186 1.054635 1.557779 0.503143  
## 187 1.051188 1.552421 0.501233  
## 188 1.049499 1.548835 0.499336  
## 189 1.048126 1.545768 0.497642  
## 190 1.046543 1.542419 0.495875  
## 191 1.045482 1.539555 0.494072  
## 192 1.045050 1.537470 0.492419  
## 193 1.044219 1.535138 0.490918  
## 194 1.045365 1.534340 0.488975  
## 195 1.044873 1.532371 0.487497  
## 196 1.043300 1.528821 0.485520  
## 197 1.040695 1.524726 0.484031  
## 198 1.040013 1.522213 0.482200  
## 199 1.037404 1.517933 0.480528  
## 200 1.036418 1.515299 0.478880  
## 201 1.036103 1.513970 0.477866  
## 202 1.036903 1.512354 0.475451  
## 203 1.033178 1.507406 0.474228  
## 204 1.033257 1.505494 0.472236  
## 205 1.031446 1.502143 0.470697  
## 206 1.032573 1.501442 0.468869  
## 207 1.030360 1.497218 0.466857  
## 208 1.027834 1.493329 0.465495  
## 209 1.028574 1.492480 0.463905  
## 210 1.028041 1.490238 0.462197  
## 211 1.028226 1.488878 0.460652  
## 212 1.028343 1.487452 0.459109  
## 213 1.028603 1.486432 0.457828  
## 214 1.024674 1.480182 0.455507  
## 215 1.022358 1.476961 0.454602  
## 216 1.019925 1.472747 0.452821  
## 217 1.019949 1.471185 0.451235  
## 218 1.019568 1.468693 0.449125  
## 219 1.017420 1.465162 0.447742  
## 220 1.012706 1.458976 0.446269  
## 221 1.012366 1.457230 0.444864  
## 222 1.012811 1.455461 0.442649  
## 223 1.010842 1.452017 0.441174  
## 224 1.010579 1.450013 0.439433  
## 225 1.009111 1.446805 0.437693  
## 226 1.006375 1.442592 0.436216  
## 227 1.007176 1.442020 0.434844  
## 228 1.007658 1.440751 0.433093  
## 229 1.007076 1.438554 0.431478  
## 230 1.005822 1.435708 0.429885  
## 231 1.005599 1.434418 0.428818  
## 232 1.003021 1.430183 0.427162  
## 233 1.003098 1.428556 0.425458  
## 234 1.002616 1.426142 0.423526  
## 235 1.002575 1.424688 0.422112  
## 236 1.002187 1.423060 0.420873  
## --------------------------------------------------   
##   
## Adjusted eigenvalues > 1 indicate dimensions to retain.  
## (236 components retained)

comp1 = principal(data21[,2:1231], nfactors=5, rotate = 'oblimin')  
plot(comp1$values[1:20])#Select the n values that are reasonable solutions  
title("Elbow method for 21st Century")



### Themes based on Seven factors - Best Solution for 21st Century

Although first two themes does not provide useful information, but all other themes were having most meaningful conclusions which was not possible with other as 5 or 6. We have used number of factors as 5 and 6 as well but the grouping of variables was best in 7.Themes were most interpretable.

Theme1 and Theme 2: Theme 1 and Theme 2 does not provide any useful information, it has separated the words which do not belong to any themes.

Theme 3: Theme 3 talks about the words like medical aid, insurance, facility, finance, hospital, services, coverage which mainly talks about the services, insurance provided by the government for mental health.

Theme 4: Theme 4 includes many words related to research, psychiatry, medication, treatment, therapy. This theme is trying to emphasize on the importance of medication and treatment of mentally ill patients, how the mental patients should be treated and how more research can improve their condition.

Theme 5: Theme 5 includes words like court, lawyer, guilty, murder, justice, death investigate. It is talking about the people who perform crime due to their illness.

Theme 6: Theme6 has taken most of the words related to politics, political parties, president. This theme mostly talks about how the government is taking measures to improve mental illness

Theme 7: Theme 7 includes words like army, stress, suicide, Afghanistan, Iraq, trauma. This theme talks about the veterans, army man how they may be suffering.

#seven factors  
comp3 = principal(data21[,2:1231], nfactors=7, rotate = 'oblimin')  
loadings7 = as.data.frame(comp3$loadings[1:1230,1:7])  
theme1.21 = subset(loadings7[order(loadings7$TC1, decreasing = T),], TC1 >= 0.15, select = c(TC1))  
theme2.21 = subset(loadings7[order(loadings7$TC2, decreasing = T),], TC2 >= 0.15, select = c(TC2))  
theme3.21 = subset(loadings7[order(loadings7$TC3, decreasing = T),], TC3 >= 0.15, select = c(TC3))  
theme4.21 = subset(loadings7[order(loadings7$TC4, decreasing = T),], TC4 >= 0.15, select = c(TC4))  
theme5.21 = subset(loadings7[order(loadings7$TC5, decreasing = T),], TC5 >= 0.15, select = c(TC5))  
theme6.21 = subset(loadings7[order(loadings7$TC6, decreasing = T),], TC6 >= 0.15, select = c(TC6))  
theme7.21 = subset(loadings7[order(loadings7$TC7, decreasing = T),], TC7 >= 0.15, select = c(TC7))  
  
theme1.21

## TC1  
## write 0.1647656

theme2.21

## TC2  
## people 0.2341137  
## said 0.2056887  
## time 0.1847339  
## year 0.1717340  
## look 0.1714175  
## see 0.1700926  
## percent 0.1684862  
## number 0.1592620  
## find 0.1534778  
## problem 0.1524787

theme3.21

## TC3  
## health 0.4790904  
## service 0.4615948  
## care 0.4508984  
## provide 0.3936006  
## cost 0.3893934  
## state 0.3866642  
## 000 0.3767376  
## program 0.3741193  
## plan 0.3222828  
## city 0.3127894  
## budget 0.3061938  
## pay 0.3039862  
## department 0.3008853  
## insurance 0.2988930  
## money 0.2889801  
## system 0.2853621  
## percent 0.2809455  
## federal 0.2737225  
## increase 0.2737050  
## coverage 0.2563273  
## medicaid 0.2552480  
## benefit 0.2517101  
## mental 0.2507643  
## cut 0.2477467  
## center 0.2467973  
## fund 0.2463742  
## agency 0.2448428  
## number 0.2425725  
## large 0.2416131  
## finance 0.2311862  
## year 0.2291045  
## require 0.2285484  
## house 0.2273859  
## provider 0.2271466  
## under 0.2257023  
## estimate 0.2236827  
## community 0.2200819  
## assistance 0.2182457  
## spend 0.2152550  
## hospital 0.2145571  
## expand 0.2143611  
## facility 0.2120546  
## population 0.2115254  
## emergency 0.2112659  
## inmate 0.2109543  
## executive 0.2095062  
## poor 0.2088663  
## nonprofit 0.2054701  
## cover 0.2039306  
## york 0.2035913  
## worker 0.2030854  
## official 0.2021098  
## medical 0.2013157  
## reduce 0.2007929  
## county 0.1969141  
## administration 0.1961680  
## staff 0.1941135  
## office 0.1920112  
## jail 0.1918442  
## treatment 0.1911444  
## operate 0.1893989  
## employee 0.1883006  
## income 0.1867116  
## commissioner 0.1866201  
## dollar 0.1860527  
## 500 0.1843572  
## homeless 0.1831423  
## access 0.1821703  
## resident 0.1813843  
## tax 0.1768404  
## propose 0.1765823  
## 200 0.1745596  
## additional 0.1745003  
## improve 0.1733376  
## 300 0.1701266  
## policy 0.1700038  
## government 0.1683137  
## advocacy 0.1649311  
## prison 0.1645905  
## aid 0.1634244  
## bill 0.1630346  
## proposal 0.1605479  
## mentally 0.1586529  
## gov 0.1572561  
## rate 0.1563089  
## monitor 0.1560204  
## resource 0.1549870  
## company 0.1535217  
## shelter 0.1534259  
## substance 0.1517725  
## contract 0.1515579  
## project 0.1512285  
## receive 0.1510039  
## average 0.1507165  
## correction 0.1503567  
## management 0.1502760

theme4.21

## TC4  
## study 0.4965634  
## mental 0.4345604  
## disorder 0.4201520  
## researcher 0.3974156  
## treatment 0.3957175  
## depress 0.3688347  
## psychiatry 0.3669746  
## editor 0.3648646  
## health 0.3578450  
## symptom 0.3547948  
## treat 0.3467929  
## illness 0.3452731  
## dr 0.3379936  
## research 0.3330912  
## risk 0.3327351  
## effect 0.3222769  
## writer 0.3169912  
## psychiatrist 0.3032722  
## effective 0.2967062  
## patient 0.2891349  
## professor 0.2884092  
## medication 0.2867200  
## journal 0.2851855  
## finding 0.2820744  
## drug 0.2805687  
## psychiatric 0.2776156  
## behavior 0.2768433  
## clinical 0.2727438  
## diagnose 0.2702618  
## medicine 0.2671108  
## problem 0.2627710  
## author 0.2578509  
## publish 0.2502703  
## diagnosis 0.2468655  
## factor 0.2435648  
## behavioral 0.2430500  
## therapy 0.2425679  
## adolescent 0.2425307  
## prescribe 0.2388639  
## institute 0.2370299  
## psychological 0.2340492  
## medical 0.2287279  
## disease 0.2248700  
## doctor 0.2238625  
## develop 0.2200595  
## suggest 0.2192044  
## severe 0.2142755  
## result 0.2127079  
## data 0.2078591  
## link 0.2068296  
## mood 0.2068259  
## national 0.2058419  
## physical 0.2051621  
## expert 0.2046187  
## common 0.2009652  
## schizophrenia 0.2002923  
## brain 0.2000529  
## condition 0.1983526  
## professional 0.1981830  
## experience 0.1931346  
## compare 0.1912729  
## physician 0.1895540  
## article 0.1874478  
## association 0.1864567  
## increase 0.1849899  
## lead 0.1849353  
## child 0.1835866  
## evidence 0.1818493  
## scientist 0.1790469  
## stress 0.1788603  
## chronic 0.1777552  
## university 0.1756990  
## analysis 0.1754273  
## significant 0.1734046  
## suffering 0.1727329  
## percent 0.1717675  
## anxiety 0.1705029  
## improve 0.1703019  
## american 0.1667941  
## access 0.1656090  
## survey 0.1655471  
## suffer 0.1652321  
## substance 0.1646257  
## serious 0.1639685  
## report 0.1626533  
## prevent 0.1619726  
## important 0.1604252  
## people 0.1587945  
## identify 0.1565177  
## function 0.1558934  
## psychologist 0.1554623  
## attention 0.1536063  
## emotion 0.1524328  
## approach 0.1505107  
## understanding 0.1503048

theme5.21

## TC5  
## court 0.5887182  
## lawyer 0.5383196  
## judge 0.5131249  
## case 0.4655201  
## prosecutor 0.4158400  
## trial 0.3830979  
## supreme 0.3685583  
## attorney 0.3634911  
## prison 0.3544079  
## charge 0.3484626  
## defense 0.3434594  
## justice 0.3419676  
## order 0.3386610  
## file 0.3362716  
## guilty 0.3342603  
## appeal 0.3267109  
## accuse 0.3240421  
## sentence 0.3136744  
## inmate 0.3135198  
## legal 0.3069719  
## decision 0.2981377  
## lawsuit 0.2929956  
## murder 0.2819396  
## criminal 0.2798031  
## federal 0.2771631  
## violate 0.2752491  
## state 0.2718787  
## argue 0.2696693  
## law 0.2651720  
## district 0.2635508  
## release 0.2631676  
## convict 0.2585527  
## investigation 0.2530050  
## correction 0.2458752  
## jail 0.2431313  
## request 0.2349106  
## evidence 0.2318695  
## crime 0.2287320  
## prisoner 0.2243223  
## official 0.2241828  
## civil 0.2213834  
## mentally 0.2202047  
## death 0.2184853  
## statement 0.2180320  
## commit 0.2162350  
## suit 0.2149172  
## document 0.2109102  
## record 0.2043083  
## department 0.1972555  
## rule 0.1893123  
## facility 0.1878823  
## comment 0.1844972  
## right 0.1820311  
## represent 0.1813556  
## deny 0.1806873  
## system 0.1786967  
## cell 0.1778434  
## repeatedly 0.1775812  
## hear 0.1760768  
## review 0.1750557  
## evaluation 0.1671731  
## panel 0.1670806  
## usa 0.1666836  
## claim 0.1661184  
## investigate 0.1658245  
## report 0.1623357  
## determine 0.1601644  
## refuse 0.1601425  
## oversee 0.1586469  
## opinion 0.1563666  
## argument 0.1540709  
## office 0.1522962  
## authority 0.1502552

theme6.21

## TC6  
## republican 0.6314201  
## democrat 0.6033183  
## gun 0.5626507  
## senator 0.5365416  
## senate 0.5280002  
## bill 0.4717589  
## legislation 0.4590156  
## vote 0.4330315  
## background 0.4223991  
## congress 0.4209187  
## measure 0.4192044  
## democratic 0.3970287  
## ban 0.3829953  
## weapon 0.3767487  
## president 0.3721254  
## law 0.3553402  
## obama 0.3510795  
## leader 0.3496851  
## pass 0.3482481  
## lawmaker 0.3466773  
## representative 0.3286959  
## debate 0.3224903  
## political 0.3144176  
## conservative 0.2990737  
## proposal 0.2822022  
## control 0.2818236  
## oppose 0.2816906  
## lobby 0.2676590  
## majority 0.2642692  
## campaign 0.2630297  
## party 0.2505186  
## governor 0.2459701  
## election 0.2294630  
## politics 0.2293799  
## support 0.2292865  
## shooting 0.2207692  
## sale 0.2195138  
## broad 0.2159788  
## magazine 0.2139491  
## supporter 0.2126075  
## issue 0.2110571  
## buy 0.2103247  
## gov 0.2089799  
## candidate 0.2080834  
## allow 0.2033243  
## violence 0.2013097  
## bush 0.1998619  
## limit 0.1994868  
## national 0.1988971  
## legislature 0.1960263  
## propose 0.1952893  
## coverage 0.1899097  
## prevent 0.1884378  
## right 0.1869505  
## american 0.1841845  
## committee 0.1837105  
## nation 0.1830408  
## require 0.1795273  
## virginia 0.1787881  
## insurance 0.1733966  
## state 0.1688653  
## speech 0.1688419  
## approve 0.1650197  
## house 0.1649240  
## mass 0.1634021  
## action 0.1633848  
## safety 0.1624406  
## expand 0.1620073  
## washington 0.1586598  
## politician 0.1551672  
## favor 0.1542513  
## enforcement 0.1537829  
## shoot 0.1531701  
## strong 0.1523632

theme7.21

## TC7  
## iraq 0.7075728  
## soldier 0.6193717  
## afghanistan 0.6184292  
## war 0.5990550  
## military 0.5755767  
## army 0.5449262  
## combat 0.5443210  
## post-traumatic 0.5291429  
## veteran 0.5206906  
## stress 0.4572252  
## affair 0.3805862  
## return 0.3504874  
## duty 0.2848798  
## serve 0.2493414  
## injury 0.2491441  
## tour 0.2443149  
## suicide 0.2349401  
## member 0.2296066  
## wound 0.2270027  
## base 0.2168418  
## secretary 0.1988635  
## kill 0.1904570  
## defense 0.1874750  
## force 0.1843110  
## department 0.1838492  
## trauma 0.1754679  
## problem 0.1609068  
## 2009 0.1507459  
## psychological 0.1501994

### Statistics (Fit Indices)

The best solution was with number of factors as 7 for 21st century,the statistics are as shown below. Root mean square of residuals is 0.02. It’s lower value is considered as good.So, Root mean square of residuals is good for this model.Root mean squared error of approximation is 0.0134. It’s upper confidence interval is 0.01351 and lower confidence interval is 0.01344. It’s value is quite small which is good for current model. The tucker lewis index is 0.26 which is not close to 1. So it can be improved.

#Root mean square of the residuals  
#The lower the better  
fa.stats(data21[,2:1231],f = comp3)$rms # for 5 factors

## Warning in fa.stats(data21[, 2:1231], f = comp3): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## [1] 0.01926848

#Root mean squared error of approximation  
#Smaller values better  
fa.stats(data21[,2:1231],f = comp3)$RMSEA # for 5 factors

## Warning in fa.stats(data21[, 2:1231], f = comp3): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## RMSEA lower upper confidence   
## 0.01347901 0.01344895 0.01351245 0.95000000

#tucker lewis index  
#Closer to 1 is better  
  
fa.stats(data21[,2:1231],f = comp3)$TLI # for 5 factors

## Warning in fa.stats(data21[, 2:1231], f = comp3): The estimated weights for the  
## factor scores are probably incorrect. Try a different factor score estimation  
## method.

## [1] 0.2572697

## Group Differences

Null hypothesis: Mental health has been changed in the last 30 years (i.e. There are difference in the articles for 20th and 21st century)

Alternative hypothesis: Mental health has not been changed in the last 30 years (i.e. There are no difference in the articles for 20th and 21st century)

Based on the above hypothesis, the difference between the articles of the 20th and 21st century are interpreted by the non-parametric test .If p value is less 0.05 null hypothesis is true.As per the non-parametric test, Theme 2,4,5,6 for belongs to 20th century all other themes belong to 21st century.

In the 20th century articles, they specifically mentioned about homeless, shelter, community services(theme4.20). In the 20th century articles, Theme 2 talks about the elections, legislation, political agenda but this was not mentioned in 21st century articles. In theme 5 and 6 for 20th century, there is almost same content as 21st century,but the p value is greater than 0.05, there might be minor differences but it was not visible clearly. Theme 7 for 21st century specifically belongs to it as it talks about the soldiers and the causes of their depression.

df <- full\_join(data20[-1],data21[,-1])

## Joining, by = c("dr", "robert", "william", "director", "county", "mental", "health", "clinic", "psychiatrist", "hospital", "early", "year", "die", "wednesday", "cancer", "assistant", "professor", "psychiatry", "state", "university", "york", "medical", "center", "associate", "mary", "born", "college", "school", "serve", "resident", "international", "association", "private", "practice", "park", "survive", "wife", "son", "jr", "david", "mark", "paul", "daughter", "sister", "manhattan", "family", "court", "judge", "look", "suffering", "day", "trouble", "night", "fill", "voice", "case", "lawyer", "police", "officer", "parent", "show", "time", "rule", "child", "abuse", "involve", "poor", "drug", "welfare", "program", "city", "house", "place", "emotion", "term", "emergency", "said", "group", "light", "wait", "room", "black", "building", "street", "live", "problem", "resource", "largely", "turn", "care", "system", "accord", "complaint", "200", "percent", "call", "chief", "criticize", "charge", "number", "post", "open", "public", "reporter", "view", "party", "identify", "name", "hope", "apartment", "stand", "year-old", "boy", "sentence", "foot", "guard", "month", "refer", "bronx", "give", "answer", "anger", "describe", "order", "move", "spoke", "decide", "ignore", "expert", "urge", "short", "mother", "client", "function", "fire", "force", "develop", "department", "offer", "spend", "afternoon", "work", "woman", "half", "continue", "ibe", "opportunity", "send", "job", "hear", "determine", "paper", "law", "sit", "eye", "progress", "represent", "husband", "control", "high", "treatment", "cut", "examine", "letter", "file", "stop", "attend", "under", "remember", "document", "floor", "issue", "father", "long", "history", "cover", "left", "accept", "car", "head", "late", "economic", "market", "benefit", "people", "increase", "period", "community", "service", "society", "neighborhood", "social", "agency", "organize", "advocate", "project", "approach", "conduct", "research", "information", "battle", "concern", "demand", "share", "criminal", "justice", "education", "train", "population", "disable", "love", "healthy", "easy", "disability", "face", "cope", "physical", "financial", "estimate", "000", "young", "adult", "great", "general", "single", "significant", "find", "difficult", "best", "read", "remain", "met", "staff", "receive", "assistance", "organization", "base", "income", "provide", "fund", "volunteer", "together", "strike", "alternative", "homeless", "support", "owner", "money", "capital", "development", "additional", "finance", "investigate", "district", "fight", "change", "improve", "board", "member", "campaign", "medicaid", "win", "primary", "create", "join", "rest", "stay", "american", "victim", "management", "interest", "rate", "low", "price", "counselor", "happen", "land", "local", "government", "step", "study", "past", "dead", "future", "emerge", "detail", "middle", "west", "sell", "town", "country", "seek", "lost", "experience", "stage", "range", "sense", "reflect", "century", "raise", "walk", "bank", "plan", "entire", "recall", "better", "life", "learn", "important", "today", "club", "insurance", "see", "hour", "week", "difference", "paid", "fact", "student", "measure", "dollar", "commission", "field", "spring", "22", "story", "region", "told", "eat", "start", "seeing", "grow", "rise", "cost", "decline", "summer", "fall", "mount", "balance", "sleep", "worry", "argument", "wall", "large", "replace", "depress", "admit", "kid", "friend", "alone", "failure", "correction", "right", "individual", "achieve", "goal", "set", "affect", "sure", "mean", "suffer", "personal", "fail", "total", "add", "food", "water", "aid", "reach", "ready", "hurt", "event", "establish", "free", "nurse", "decade", "youth", "sought", "address", "word", "career", "300", "choice", "survey", "drop", "compare", "children's", "teach", "write", "result", "begin", "girl", "delay", "sex", "counseling", "perform", "special", "graduate", "class", "operation", "responsible", "indicate", "bring", "condition", "unit", "long-term", "commitment", "pressure", "mentally", "leader", "person", "available", "cite", "announce", "bed", "effort", "eliminate", "current", "budget", "allow", "office", "pay", "white", "nonprofit", "corporation", "end", "action", "oppose", "man", "consider", "facility", "area", "nature", "discuss", "hire", "official", "unable", "obtain", "record", "position", "propose", "psychiatric", "acknowledge", "legal", "refuse", "meeting", "held", "vote", "express", "degree", "illness", "effect", "procedure", "request", "commissioner", "decision", "patient", "major", "disorder", "executive", "return", "understanding", "400", "sick", "shelter", "thoma", "operate", "talk", "present", "pass", "quick", "promise", "employee", "contribute", "500", "unite", "hand", "bill", "insist", "schedule", "main", "article", "play", "idea", "july", "death", "senator", "senate", "chairman", "committee", "store", "brought", "partner", "visit", "annual", "peter", "agree", "grant", "tuesday", "sign", "safe", "security", "response", "favor", "complete", "jail", "political", "attention", "convict", "release", "drive", "san", "telephone", "interview", "small", "morning", "friday", "human", "strong", "trial", "spent", "occur", "south", "longer", "age", "arrest", "subject", "psychological", "reason", "prison", "kill", "blame", "comment", "exist", "appeal", "supreme", "monday", "shot", "maintain", "real", "speak", "hard", "editor", "report", "situation", "amount", "fear", "assault", "threat", "contact", "citizen", "mind", "advance", "push", "question", "brother", "basic", "lead", "reduce", "finally", "understand", "matter", "st", "north", "island", "21", "panel", "commit", "crisis", "medication", "powerful", "save", "prevent", "expect", "california", "treat", "schizophrenia", "professional", "attack", "side", "contract", "symptom", "extend", "feeling", "deal", "nothing", "minute", "door", "form", "huge", "watch", "happy", "murder", "book", "outside", "arrive", "buy", "manage", "close", "neighbor", "challenge", "evidence", "damage", "charle", "attorney", "independent", "carry", "ground", "reject", "site", "approve", "23", "washington", "national", "alcohol", "1990", "1980", "realize", "stress", "key", "limit", "warn", "appear", "conflict", "message", "news", "deputy", "institute", "impact", "behavior", "dangerous", "usa", "publish", "nation", "suicide", "researcher", "administration", "encourage", "level", "disease", "risk", "baby", "jame", "federal", "gain", "process", "extreme", "average", "serious", "senior", "source", "photo", "incident", "suspect", "jan", "test", "institution", "body", "discover", "authority", "conference", "journal", "weekend", "author", "note", "leave", "argue", "protect", "award", "richard", "prove", "deny", "dispute", "thought", "psychologist", "heart", "protection", "division", "anxiety", "factor", "led", "focus", "require", "doctor", "series", "broad", "safety", "advocacy", "access", "violent", "crime", "investigation", "industry", "mayor", "network", "inside", "thursday", "common", "responsibility", "role", "relationship", "teacher", "blood", "sort", "shop", "avenue", "draw", "influence", "therapist", "pick", "direct", "design", "company", "east", "account", "press", "tend", "highly", "quality", "model", "team", "top", "road", "dozen", "struggle", "reality", "lose", "unusual", "simple", "guilty", "separate", "sound", "veteran", "game", "session", "statement", "notice", "environment", "success", "review", "potential", "standard", "clear", "break", "true", "prepare", "scientist", "president", "couple", "difficulty", "figure", "firm", "america", "enter", "therapy", "task", "childhood", "chance", "adopt", "marry", "avoid", "connect", "master", "meet", "recommend", "reform", "andraw", "apply", "remains", "representative", "spokesman", "directly", "relate", "specific", "handle", "respond", "spokeswoman", "data", "april", "list", "manager", "worker", "complex", "diagnose", "democratic", "explain", "big", "columbia", "jersey", "psychology", "credit", "activity", "wear", "effective", "generally", "angeles", "administrator", "michael", "vice", "war", "35", "eventually", "remove", "hit", "business", "clean", "attempt", "substance", "type", "suit", "hill", "depend", "election", "retire", "full", "severe", "science", "lack", "central", "24", "relative", "loss", "worse", "strategy", "air", "sunday", "wrong", "31", "intend", "physician", "medicine", "agreement", "march", "27", "clinical", "brown", "similar", "finding", "page", "power", "civil", "monitor", "consequence", "point", "proposal", "queen", "republican", "january", "governor", "apparent", "recognize", "john", "brooklyn", "mass", "violence", "pain", "harvard", "boston", "deep", "26", "threaten", "yesterday", "affair", "ability", "secretary", "george", "policy", "arm", "computer", "conclude", "sale", "picture", "congress", "tax", "choose", "writer", "expand", "joseph", "29", "hold", "accuse", "earn", "follow", "space", "complain", "lower", "television", "successful", "suggest", "council", "union", "newspaper", "democrat", "legislation", "increasingly", "hallway", "addition", "specialist", "phone", "produce", "28", "simply", "daily", "final", "moment", "defense", "coverage", "disturb", "brain", "appropriate", "june", "sept", "1970", "link", "race", "debate", "claim", "lie", "candidate", "normal", "danger", "block", "legislature", "prosecutor", "church", "consultant", "tough", "shift", "bear", "instance", "colleague", "wide", "lawsuit", "memory", "majority", "foundation", "opinion", "gov", "critic", "build", "trade", "politics", "track", "fine")

df$century <- century  
table(df$century)

##   
## 20th\_Century 21th\_Century   
## 3880 7083

df$Theme1.20 = apply(df[,row.names(theme1.20 )], MARGIN = 1, sum)  
df$Theme2.20 = apply(df[,row.names(theme2.20 )], MARGIN = 1, sum)  
df$Theme3.20 = apply(df[,row.names(theme3.20 )], MARGIN = 1, sum)  
df$Theme4.20 = apply(df[,row.names(theme4.20 )], MARGIN = 1, sum)  
df$Theme5.20 = apply(df[,row.names(theme5.20 )], MARGIN = 1, sum)  
df$Theme6.20 = apply(df[,row.names(theme6.20 )], MARGIN = 1, sum)  
  
df$Theme1.21 = apply(df[,row.names(theme1.21)], MARGIN = 1, sum)  
df$Theme2.21 = apply(df[,row.names(theme2.21)], MARGIN = 1, sum)  
df$Theme3.21 = apply(df[,row.names(theme3.21)], MARGIN = 1, sum)  
df$Theme4.21 = apply(df[,row.names(theme4.21)], MARGIN = 1, sum)  
df$Theme5.21 = apply(df[,row.names(theme5.21)], MARGIN = 1, sum)  
df$Theme6.21 = apply(df[,row.names(theme5.21)], MARGIN = 1, sum)  
df$Theme7.21 = apply(df[,row.names(theme5.21)], MARGIN = 1, sum)  
  
describeBy(df[,c('Theme1.20','Theme2.20','Theme3.20','Theme4.20','Theme5.20','Theme6.20','Theme1.21','Theme2.21','Theme3.21','Theme4.21','Theme5.21','Theme6.21','Theme7.21')], df$century)

##   
## Descriptive statistics by group   
## group: 20th\_Century  
## vars n mean sd median trimmed mad min max range skew  
## Theme1.20 1 3880 2.15 1.02 2.18 2.16 0.98 0.00 6.35 6.35 0.02  
## Theme2.20 2 1105 2.59 2.51 1.78 2.13 1.39 0.00 17.21 17.21 2.33  
## Theme3.20 3 3880 3.31 2.64 2.42 2.85 1.74 0.26 18.80 18.54 1.71  
## Theme4.20 4 1105 5.34 4.02 4.30 4.77 3.54 0.34 23.95 23.61 1.29  
## Theme5.20 5 1105 2.69 2.12 2.09 2.40 1.66 0.00 11.11 11.11 1.25  
## Theme6.20 6 1105 2.42 2.38 1.67 1.95 1.25 0.11 16.55 16.44 2.75  
## Theme1.21 7 3880 0.05 0.12 0.00 0.02 0.00 0.00 1.42 1.42 3.66  
## Theme2.21 8 3880 1.81 0.95 1.77 1.78 0.93 0.00 7.78 7.78 0.46  
## Theme3.21 9 2775 4.75 3.47 3.72 4.21 2.57 0.26 24.15 23.88 1.55  
## Theme4.21 10 2775 3.95 3.07 2.95 3.43 2.09 0.31 24.40 24.09 1.69  
## Theme5.21 11 2775 2.55 2.48 1.69 2.10 1.51 0.00 16.55 16.55 1.78  
## Theme6.21 12 2775 2.55 2.48 1.69 2.10 1.51 0.00 16.55 16.55 1.78  
## Theme7.21 13 2775 2.55 2.48 1.69 2.10 1.51 0.00 16.55 16.55 1.78  
## kurtosis se  
## Theme1.20 0.03 0.02  
## Theme2.20 7.04 0.08  
## Theme3.20 3.23 0.04  
## Theme4.20 1.62 0.12  
## Theme5.20 1.27 0.06  
## Theme6.20 9.53 0.07  
## Theme1.21 19.33 0.00  
## Theme2.21 0.89 0.02  
## Theme3.21 2.70 0.07  
## Theme4.21 3.28 0.06  
## Theme5.21 3.46 0.05  
## Theme6.21 3.46 0.05  
## Theme7.21 3.46 0.05  
## ------------------------------------------------------------   
## group: 21th\_Century  
## vars n mean sd median trimmed mad min max range skew  
## Theme1.20 1 7083 2.05 1.10 2.09 2.05 1.06 0.00 6.09 6.09 0.09  
## Theme2.20 2 2775 2.77 2.83 1.74 2.21 1.53 0.00 18.34 18.34 2.04  
## Theme3.20 3 7083 3.53 2.87 2.53 3.05 2.01 0.22 20.00 19.78 1.54  
## Theme4.20 4 2775 4.95 3.52 4.10 4.52 3.13 0.18 21.81 21.62 1.13  
## Theme5.20 5 2775 2.64 2.29 1.96 2.26 1.48 0.00 15.89 15.89 1.77  
## Theme6.20 6 2775 2.65 2.71 1.76 2.08 1.36 0.07 21.47 21.40 2.54  
## Theme1.21 7 7083 0.04 0.12 0.00 0.02 0.00 0.00 1.52 1.52 4.71  
## Theme2.21 8 7083 1.78 1.03 1.79 1.75 1.03 0.00 7.58 7.58 0.38  
## Theme3.21 9 4308 5.26 3.81 4.14 4.71 3.04 0.29 24.35 24.05 1.35  
## Theme4.21 10 4308 4.53 3.55 3.37 3.96 2.55 0.29 24.35 24.06 1.50  
## Theme5.21 11 4308 2.40 2.57 1.40 1.91 1.32 0.00 19.20 19.20 1.96  
## Theme6.21 12 4308 2.40 2.57 1.40 1.91 1.32 0.00 19.20 19.20 1.96  
## Theme7.21 13 4308 2.40 2.57 1.40 1.91 1.32 0.00 19.20 19.20 1.96  
## kurtosis se  
## Theme1.20 -0.17 0.01  
## Theme2.20 4.46 0.05  
## Theme3.20 2.39 0.03  
## Theme4.20 1.20 0.07  
## Theme5.20 3.54 0.04  
## Theme6.20 7.93 0.05  
## Theme1.21 32.11 0.00  
## Theme2.21 0.26 0.01  
## Theme3.21 1.82 0.06  
## Theme4.21 2.17 0.05  
## Theme5.21 4.56 0.04  
## Theme6.21 4.56 0.04  
## Theme7.21 4.56 0.04

wilcox.test(df$Theme1.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme1.20 by df$century  
## W = 14478714, p-value = 3.227e-06  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme2.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme2.20 by df$century  
## W = 1532577, p-value = 0.9845  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme3.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme3.20 by df$century  
## W = 13347010, p-value = 0.0129  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme4.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme4.20 by df$century  
## W = 1585354, p-value = 0.09762  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme5.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme5.20 by df$century  
## W = 1585682, p-value = 0.09552  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme6.20~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme6.20 by df$century  
## W = 1485544, p-value = 0.1303  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme1.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme1.21 by df$century  
## W = 14479363, p-value = 1.703e-09  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme2.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme2.21 by df$century  
## W = 13975894, p-value = 0.1383  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme3.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme3.21 by df$century  
## W = 5534240, p-value = 1.33e-07  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme4.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme4.21 by df$century  
## W = 5444153, p-value = 2.195e-10  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme5.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme5.21 by df$century  
## W = 6438591, p-value = 4.007e-08  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme6.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme6.21 by df$century  
## W = 6438591, p-value = 4.007e-08  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(df$Theme7.21~df$century)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$Theme7.21 by df$century  
## W = 6438591, p-value = 4.007e-08  
## alternative hypothesis: true location shift is not equal to 0

### Sentiment Analysis

To start our analysis, the original data file in loaded in the R chunk below:

df = read.csv("NYTMentalHealthArticles.csv")  
str(df)

## 'data.frame': 11935 obs. of 6 variables:  
## $ Filename: chr "1000970;katherine-oettinger-94-expert-on-retardation.html.txt" "1001627;news-summary-716251.html.txt" "1001756;wider-mental-health-benefits-threatened-by-exemption-plan.html.txt" "1001857;experts-denounce-children-s-agency.html.txt" ...  
## $ Title : chr "Katherine Oettinger, 94, Expert on Retardation" "NEWS SUMMARY" "Wider Mental Health Benefits Threatened by Exemption Plan" "EXPERTS DENOUNCE CHILDREN'S AGENCY" ...  
## $ Date : chr "OCT. 19, 1997" "OCT. 21, 1997" "OCT. 21, 1997" "OCT. 22, 1997" ...  
## $ By : chr "By HOLCOMB B. NOBLE" "" "By ROBERT PEAR" "By RACHEL L. SWARNS" ...  
## $ Year : chr "1997" "1997" "1997" "1997" ...  
## $ Text : chr "\"Katherine B. Oettinger an authority on the care and upbringing of children who was head of the Federal agen"| \_\_truncated\_\_ "\"INTERNATIONAL A3-13 Clinton Policy on Tibet A Sore Point With China The Clinton Administration's pledge to ap"| \_\_truncated\_\_ "\"New rules being drafted by the Clinton Administration could allow many employers to exempt themselves from a "| \_\_truncated\_\_ "\"Offering a bleak portrayal of New York City's efforts to protect its most vulnerable children a court-appoin"| \_\_truncated\_\_ ...

#head(df)  
#nrow(df)

Load Python Libraries necessary for the analysis and read in the python chunk.

#nltk.download('vader\_lexicon')  
#nltk.download('stopwords', 'punkt', 'wordnet')  
  
import nltk  
import warnings  
warnings.filterwarnings('ignore')  
from nltk.sentiment.vader import SentimentIntensityAnalyzer  
  
import re  
import pandas as pd  
from nltk.corpus import stopwords  
from wordcloud import WordCloud, STOPWORDS  
import dateutil.parser as dparser  
import spacy  
  
from nltk import tokenize  
from datetime import datetime, timedelta  
import time  
import pprint  
from nltk import word\_tokenize  
from statistics import mean  
  
df = r.df  
df.head()

## Filename ... Text  
## 0 1000970;katherine-oettinger-94-expert-on-retar... ... "Katherine B. Oettinger an authority on the c...  
## 1 1001627;news-summary-716251.html.txt ... "INTERNATIONAL A3-13 Clinton Policy on Tibet A...  
## 2 1001756;wider-mental-health-benefits-threatene... ... "New rules being drafted by the Clinton Admini...  
## 3 1001857;experts-denounce-children-s-agency.htm... ... "Offering a bleak portrayal of New York City's...  
## 4 1001909;l-medical-research-saves-money-too-doc... ... "To the Editor: The news that 'Analysts Expect...  
##   
## [5 rows x 6 columns]

The date in the original dataset is stored as characters. Before analysis we converted dates to a standard formatted date type. Removed rows where date fields had characters and string values.

df['NewDate'] = pd.to\_datetime(df['Date'], errors='coerce')  
df = df.dropna(axis=0)   
df[pd.isnull(pd.to\_datetime(df.NewDate))]

## Empty DataFrame  
## Columns: [Filename, Title, Date, By, Year, Text, NewDate]  
## Index: []

analyzer = SentimentIntensityAnalyzer()  
  
posts = df['Text']  
date\_sentiments = {}  
  
# Used VADER with entire article text  
  
#posts = df['Text'].head()  
#for post in posts:  
# date = df.loc[df['Text'] == post, 'NewDate'].iloc[0]  
# for sentence in sentence\_list:  
# sentiment = analyzer.polarity\_scores(post)['compound']  
# date\_sentiments.setdefault(date, []).append(sentiment)  
  
for post in posts:  
 sentence\_list = tokenize.sent\_tokenize(post)  
 date = df.loc[df['Text'] == post, 'NewDate'].iloc[0]  
   
 for sentence in sentence\_list:  
 sentiment = analyzer.polarity\_scores(sentence)['compound']  
 date\_sentiments.setdefault(date, []).append(sentiment)

After calculating the polarity scores for every sentence in the article, we store the scores in a dictionary with date as the key. For every date we calculate the mean score and store it.

date\_sentiment = {}  
  
for k,v in date\_sentiments.items():  
 date\_sentiment[k.date()] = round(sum(v)/float(len(v)),3)  
  
sentiments\_df = pd.DataFrame(list(date\_sentiment.items()),columns = ['Date','Score'])   
sentiments\_df.to\_csv('sentiment\_scores.csv',index=False)

Next we aggregate these scores over the years by taking average of all the scores dated in a year.

sentiment\_scores = read\_csv("sentiment\_scores.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## Date = col\_date(format = ""),  
## Score = col\_double()  
## )

sentiment\_scores$Year = as.numeric(format(sentiment\_scores$Date,'%Y'))  
sentiment\_scores\_year = sentiment\_scores %>%  
 group\_by(Year) %>%  
 summarise(mean\_score = round(mean(Score),3))

## `summarise()` ungrouping output (override with `.groups` argument)

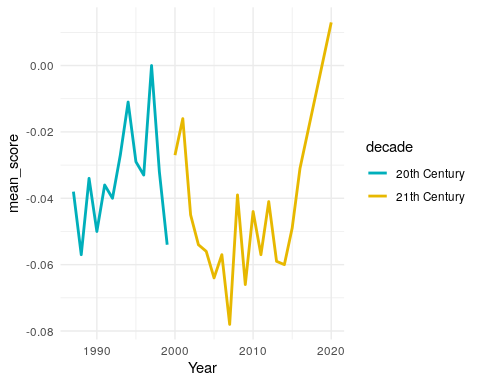
sentiment\_scores\_year$decade <- factor(ifelse(sentiment\_scores\_year$Year<=1999, "20th Century", "21th Century"))  
  
# normalized, weighted composite score  
head(sentiment\_scores\_year)

## # A tibble: 6 x 3  
## Year mean\_score decade   
## <dbl> <dbl> <fct>   
## 1 1987 -0.038 20th Century  
## 2 1988 -0.057 20th Century  
## 3 1989 -0.034 20th Century  
## 4 1990 -0.05 20th Century  
## 5 1991 -0.036 20th Century  
## 6 1992 -0.04 20th Century

Plotting the average compound polarity score for each year, we can observe that in the 20th Century, mental health was discussed in the news articles with a more negative sentiment. We can observe some peaks during the middle of the century, however by year 2000, most articles had an extreme negative sentiment.

The score went extreme low around year 2008. On the contrary, we see a distinctive rise in the positive sentiment in the latter half of the 21st century. Awareness around the subject of mental health has increased and newspaper articles are talking about it with a more positive sentiment since the last 5 years.

# Multiple line plot  
ggplot(sentiment\_scores\_year, aes(x = Year, y = mean\_score)) +   
 geom\_line(aes(color = decade), size = 1) +  
 scale\_color\_manual(values = c("#00AFBB", "#E7B800")) +  
 theme\_minimal()



df.to\_csv('updated\_dataframe.csv',index=False)

To understand emotions, NCR lexicon from the R’s syuzhet library was used. 5 emotions (anger, fear, sadness, surprise, trust) and 2 sentiments (negative, positive) were focused in the text. The function get\_nrc\_sentiment() takes the article text and returns valence score for the sentiments and emotions. These scores are aggregated across years to understand the trend.

library(syuzhet)

##   
## Attaching package: 'syuzhet'

## The following object is masked from 'package:psych':  
##   
## rescale

df = read\_csv("updated\_dataframe.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## Filename = col\_character(),  
## Title = col\_character(),  
## Date = col\_character(),  
## By = col\_character(),  
## Year = col\_double(),  
## Text = col\_character(),  
## NewDate = col\_datetime(format = "")  
## )

nrc = get\_nrc\_sentiment(df$Text, language = "english")  
  
nrcdf = cbind(df,nrc)

nrcdf$Year = as.numeric(format(nrcdf$NewDate,'%Y'))  
  
# emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative  
  
nrc\_scores\_year = nrcdf %>%  
 group\_by(Year) %>%  
 summarise(mean\_anger = round(mean(anger),3),  
 #mean\_anticipation = round(mean(anticipation),3),  
 #mean\_disgust = round(mean(disgust),3),  
 mean\_fear = round(mean(fear),3),  
 #mean\_joy = round(mean(joy),3),  
 mean\_sadness = round(mean(sadness),3),  
 mean\_surprise = round(mean(surprise),3),  
 mean\_trust = round(mean(trust),3),  
 mean\_negative = round(mean(negative),3),  
 mean\_positive = round(mean(positive),3))

## `summarise()` ungrouping output (override with `.groups` argument)

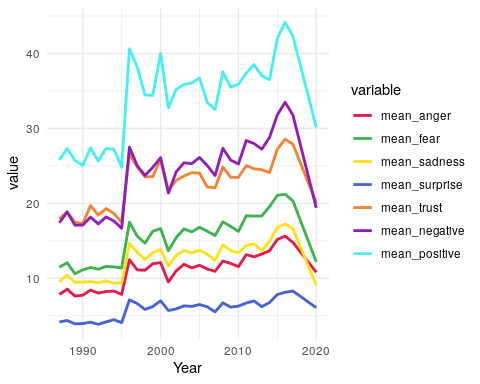
nrc\_scores\_year$decade <- factor(ifelse(nrc\_scores\_year$Year<=1999, "20th Century", "21th Century"))  
  
head(nrc\_scores\_year)

## # A tibble: 6 x 9  
## Year mean\_anger mean\_fear mean\_sadness mean\_surprise mean\_trust mean\_negative  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1987 7.84 11.5 9.57 4.17 17.9 17.4  
## 2 1988 8.54 12.1 10.4 4.36 18.9 18.8  
## 3 1989 7.63 10.6 9.45 3.92 17.5 17.1  
## 4 1990 7.74 11.1 9.5 3.94 17.3 17.1  
## 5 1991 8.42 11.4 9.55 4.14 19.7 18.2  
## 6 1992 8.02 11.2 9.41 3.85 18.4 17.2  
## # … with 2 more variables: mean\_positive <dbl>, decade <fct>

## NRC Emotion Lexicon

The NRC Scores represents the average valence scores of articles aggregated over years. The colors represent the emotions. We can observe that the emotions fear and sadness were high in the 20th century and declined towards the end of 21st century.

melted\_df=melt(nrc\_scores\_year,id.vars=c("Year","decade"))  
  
# Multiple line plot  
ggplot(melted\_df, aes(x = Year, y = value)) +   
 geom\_line(aes(color = variable), size = 1) +  
 scale\_color\_manual(values = c('#e6194b', '#3cb44b', '#ffe119', '#4363d8', '#f58231', '#911eb4', '#46f0f0')) +  
 theme\_minimal()



# Discussion

From the above analysis, we see there are differences in the articles of 20th and 21st century but very few variations are visible. In 20th century articles were mostly based on the medical aid, treatment of mental patients, building hospitals, further research areas, government initiatives to improve the medical services. It also talks about the judiciary system courts, laws, murders, death which specify that convicts suffer depression. Mental illness can be the reason for their misdeeds. These terms are similar to the 21st century articles.These topics are also used in the 21st century articles. The main differences with 20th century articles are, they talk more about politics, community services for homeless, shelter, change in thinking where these topics were not the part of the 21st century articles. The 21st century articles speak about the depression faced by the military soldiers which was not there in 20th century articles. In conclusion, major topics covered in both the articles were same over 30 years with only minor variations.

After looking at the results from the Vader sentiment analysis, we also observed changes in the tone used in the articles in the two centuries. The sentiment is more positive in the latter century and the emotions like fear, sadness and negative sentiment also declined in 21st century as compared to the 20th. Although sentiment analysis with Vader doesnot necesarily cover the entire tone of the article, it still helps us gather an idea around the changes in emotion and sentiment over the years.

# Credit Guidelines

* Developed Research Question - Ruchil , Purbi , Mohammed
* Acquired Data - Purbi , Ruchil , Mohammed
* Analyzed Data - Ruchil , Purbi , Mohammed
* Researched Related Work - Purbi
* Wrote Introduction - Mohammed
* Wrote Method - Ruchil, Purbi , Mohammed
* Wrote Analysis/Results - Ruchil , Purbi
* Wrote Discussion - Mohammed, Ruchil , Purbi
* Revisions/Edits - Ruchil

# References

* The Meaning Extraction Helper by Ryan l. Boyd, Ph.D. – Boyd, R.L. (2018). MEH: Meaning Extraction Helper (version 2.1.07) [Software]. Available from <https://meh.ryanb.cc>
* Otto E Wahl, Amy Wood & Renee Richards (2002) Newspaper Coverage of Mental Illness: Is It Changing?, Psychiatric Rehabilitation Skills, 6:1, 9-31, DOI: 10.1080/10973430208408417
* McGinty, E. E., RC, K., PS, W., RG, F., PJ, C., MF, H., . . . Pescosolido, B. A. (2016, June 01). Trends In News Media Coverage Of Mental Illness In The United States: 1995–2014: Health Affairs Journal. Retrieved from <https://www.healthaffairs.org/doi/10.1377/hlthaff.2016.0011>
* Mohammad, S.M. and Turney, P.D. (2013), CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. Computational Intelligence, 29: 436-465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
* Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 2098–2110. DOI:https://doi.org/10.1145/2858036.2858207