analysis_02

August 18, 2019

In [1]:
```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
import numpy as np
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
import seaborn as sns
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
```

In [2]:
```
psy_df = pd.read_csv('depression.csv')
environ_df = pd.read_csv('climate_change.csv')
soci_df = pd.read_csv('institution.csv')
```

In [3]:
```
pd.set_option('display.max_colwidth', 120)
```

In [4]:
```
df = psy_df.append(environ_df,ignore_index=True).append(soci_df,ignore_index=True)
df = df.sample(frac=1).drop(columns=['Unnamed: 0'])
print('Shape before dropping duplicates: ' + str(df.shape))
df = df.drop_duplicates()
print('Shape after dropping duplicates: ' + str(df.shape))
df.reset_index(drop=True,inplace=True)
df
```

```
Shape before dropping duplicates: (1151, 8)
Shape after dropping duplicates: (1112, 8)
```

Out[4]:

```
                                                                            url  \
0      https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.60.110707.163655
1                https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.24.1.345
2             https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.24.1.461
3             https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.24.1.513
4     https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.32.080106.133554
5         https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010416-043958
6     https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.31.102505.133552
7         https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015104
8                https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.52.1.1
9              https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.25.1.441
```

10	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-071112-054629
11	https://www.annualreviews.org/doi/abs/10.1146/annurev.environ.33.050707.085733
12	https://www.annualreviews.org/doi/abs/10.1146/annurev.environ.33.022007.145142
13	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.21.1.69
14	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.29.010202.100030
15	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-070308-120029
16	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-122216-011732
17	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.25.1.441
18	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.51.1.599
19	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071811-145449
20	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.55.090902.141528
21	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-081309-150008
22	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071913-043215
23	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-060116-053252
24	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.187
25	https://www.annualreviews.org/doi/abs/10.1146/annurev-environ-102014-021358
26	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.58.110405.085535
27	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.33.040406.131647
28	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.25.1.245
29	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010416-044138
...	...
1082	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.012809.102629
1083	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-060116-053403
1084	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010213-115154
1085	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.30.012703.110644
1086	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.34.040507.134740
1087	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.54.101601.145240
1088	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.28.011503.163459
1089	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.26.1.49
1090	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010213-115100
1091	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.60.110707.163639
1092	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.31.041304.122312
1093	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.31.020105.100307
1094	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.33.040406.131651
1095	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.24.1.395
1096	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015038
1097	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-073018-022351
1098	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.29.010202.100113
1099	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-060116-053411
1100	https://www.annualreviews.org/doi/abs/10.1146/annurev-environ-102016-061128
1101	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.31.020105.100157
1102	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.59.103006.093643
1103	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.23.1.63
1104	https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.283
1105	https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-120710-100328
1106	https://www.annualreviews.org/doi/abs/10.1146/annurev.energy.21.1.31
1107	https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.57.102904.190205
1108	https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-081309-150219

1109 https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.55.090902.142054
1110 https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.28.110601.141117
1111 https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015240

0 Personality: The Universal and the
1 Measurin
2 HIGH-LEVEL NUCLEAR WASTE: The Sta
3 A REVIEW OF TECHNICAL CHANGE IN ASSESSM
4 Renewable Energy Futures: Targets, Sc
5 Interactions With Robots: The Truths We R
6 Sustainability Values, Attitudes, and Behaviors: A Review of Multinatio
7 Critical Periods in Speech Perce
8 Social Cognitive Theory: A
9 POLITICS AND INSTITUTIONALISM: Explaining
10 A Cultural Neuroscience Approach to the Biosocial Natu
11 Global Environmental Governance: Taking
12 Sanitation for Unserved Populations: Technologies, Implementation Challeng
13 ON THE CONCEPT
14 Skills Mismatc
15 The Political Consequence
16 Developmental Adaptation to Stress: An Evo
17 Water Vapor Feedba
18 Memory
19 Immigrants
20 Devel
21 Research on Adolescence in the
22 The Political Mobilization of
23 Social Structure, Adversity, Toxic Stress, and Intergenerational Poverty: An
24
25 State of t
26 The Elaboration of Personal
27 Embeddedness and the Intellectual Projects
28 APHORISMS AND CLICHÉS: The Generation and Dissipation o
29 Ea
...
1082 A World of Standards but not a Standard World: Toward a Sociology of Standard
1083 Stress-Related Biosocial Mechanisms of Discrimination and African Ameri
1084 The Cognitive Ne
1085 Reflections on a Half-Century of Org
1086
1087 Biology, Context, and
1088 GREEN CHEMISTRY AND ENGINEERING: Drivers, Metrics, and
1089 INDICATORS OF ENERGY USE AND CARBON EMISSIONS: Explaining th
1090 Gene-En
1091 An Odor is Not Worth a Thousand Words: From Multidimensional Odors to Unidim
1092 Reading and the Reading Class in the
1093 Soils: A Con

1094
1095 Intermarriage and Homogamy: Caus
1096 Childhood Antecedents and Risk for Ac
1097 Fascism and Populism: Are They Useful Categories for Comparative Sc
1098 Narrative Explanation: An Alternative to Variable-C
1099 Poverty in America: New Di
1100 Linking Urbanization and the Environment: Conceptual ar
1101 Energy Efficiency Policies: A Retro
1102 Health Psychology: The Search for Pathways betweer
1103 Growing Up American: The Challenge Confronting Immigrant Children and Cl
1104 Collective Identity
1105 Consequences of Age-Relate
1106 THE EVOLUTION OF
1107 Stressful Experience and Learning
1108 The Integration Imperative: The Children of Low-Status Immigrants in the Schools
1109 Motivational Influences
1110 From Factors to Actors: Computational Sociology and
1111 The Neuroendocrinology

0 Steven J. Heine a
1
2
3 Christian Azar ar
4 Eric Martinot, Carmen Dienst, Liu Weilia
5 I
6 Anthony A. Leiserowitz, Robert W. Kates, ar
7 Janet F. Werker a
8
9 Elisabeth S. Clemens and
10 Shihui Han, Georg Northoff, Kai Vogeley, Bruce E. Wexler, Shinobu Kitayama, and I
11 Frank Biermann ar
12 Kara L. Nelsor
13
14
15 Edwin Amenta, Neal Caren, Elizabeth Chia
16 Bruce J. Ellis and
17 Isaac M. Held
18
19 Mary C. Waters, Philip Kasinitz
20
21 Robert Crosnoe and Monica I
22 Edward T. Walker and
23 Craig A. McEwen a
24 Robert E. Washingt
25 Al
26 Beverly M. Walker a
27 Greta R. Krippner and

```
28
29                                                                                    I
...
1082                                                       Stefan Timmermans
1083                                     Bridget J. Goosby, Jacob E. Cheadle, a
1084                                                              John Kouni
1085
1086                                                        Philip S. Gorski
1087
1088                           Anne E. Marteel, Julian A. Davies, Walter W. Olson
1089                                   Lee Schipper, Fridtjof Unander, Scott Mu
1090                                                     Stephen B. Manuck and J
1091                                                             Yaara Yeshu
1092                                          Wendy Griswold, Terry McDonnell
1093                                     Cheryl Palm, Pedro Sanchez, Sonya Ahan
1094                                        Pamela Paxton, Sheri Kunovich, and
1095
1096                                                              Daniel S. Pine
1097
1098
1099                                                             Matthew Desmond
1100  Xuemei Bai, Timon McPhearson, Helen Cleugh, Harini Nagendra, Xin Tong, Tong Zhu
1101                                       Kenneth Gillingham, Richard Newell
1102                     Howard Leventhal, John Weinman, Elaine A. Leventhal, and
1103
1104                                                       Francesca Polletta a
1105
1106
1107
1108                                          Richard Alba, Jennifer Sloan, an
1109                                       Timothy B. Baker , Thomas H. Brandon ,
1110                                                       Michael W. Macy and
1111                     John T. Cacioppo, Stephanie Cacioppo, John P. Capitanio,

      year  \
0     2009
1     1998
2     1999
3     1999
4     2007
5     2017
6     2006
7     2015
8     2001
9     1999
10    2013
11    2008
12    2008
```

13	1996
14	2003
15	2010
16	2019
17	2000
18	2000
19	2014
20	2004
21	2011
22	2014
23	2017
24	2001
25	2015
26	2007
27	2007
28	1999
29	2017
...	...
1082	2010
1083	2018
1084	2014
1085	2004
1086	2008
1087	2003
1088	2003
1089	2001
1090	2014
1091	2010
1092	2005
1093	2007
1094	2007
1095	1998
1096	2015
1097	2019
1098	2004
1099	2018
1100	2017
1101	2006
1102	2008
1103	1997
1104	2001
1105	2012
1106	1996
1107	2006
1108	2011
1109	2004
1110	2002
1111	2015

	html_url
0	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.60.110707.163655
1	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.24.1.345
2	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.24.1.461
3	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.24.1.513
4	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.32.080106.133554
5	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010416-043958
6	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.31.102505.133552
7	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010814-015104
8	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.52.1.1
9	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.25.1.441
10	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-071112-054629
11	https://www.annualreviews.org/doi/full/10.1146/annurev.environ.33.050707.085733
12	https://www.annualreviews.org/doi/full/10.1146/annurev.environ.33.022007.145142
13	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.21.1.69
14	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.29.010202.100030
15	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-070308-120029
16	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-122216-011732
17	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.25.1.441
18	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.51.1.599
19	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-071811-145449
20	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.55.090902.141528
21	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-081309-150008
22	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-071913-043215
23	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-060116-053252
24	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.27.1.187
25	https://www.annualreviews.org/doi/full/10.1146/annurev-environ-102014-021358
26	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.58.110405.085535
27	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.33.040406.131647
28	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.25.1.245
29	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010416-044138
...	...
1082	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.012809.102629
1083	https://www.annualreviews.org/doi/full/10.1146/annurev-soc-060116-053403
1084	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010213-115154
1085	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.30.012703.110644
1086	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.34.040507.134740
1087	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.54.101601.145240
1088	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.28.011503.163459
1089	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.26.1.49
1090	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010213-115100
1091	https://www.annualreviews.org/doi/full/10.1146/annurev.psych.60.110707.163639
1092	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.31.041304.122312
1093	https://www.annualreviews.org/doi/full/10.1146/annurev.energy.31.020105.100307
1094	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.33.040406.131651
1095	https://www.annualreviews.org/doi/full/10.1146/annurev.soc.24.1.395
1096	https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010814-015038

```
1097          https://www.annualreviews.org/doi/full/10.1146/annurev-soc-073018-022351
1098       https://www.annualreviews.org/doi/full/10.1146/annurev.soc.29.010202.100113
1099          https://www.annualreviews.org/doi/full/10.1146/annurev-soc-060116-053411
1100      https://www.annualreviews.org/doi/full/10.1146/annurev-environ-102016-061128
1101    https://www.annualreviews.org/doi/full/10.1146/annurev.energy.31.020105.100157
1102     https://www.annualreviews.org/doi/full/10.1146/annurev.psych.59.103006.093643
1103                https://www.annualreviews.org/doi/full/10.1146/annurev.soc.23.1.63
1104               https://www.annualreviews.org/doi/full/10.1146/annurev.soc.27.1.283
1105        https://www.annualreviews.org/doi/full/10.1146/annurev-psych-120710-100328
1106             https://www.annualreviews.org/doi/full/10.1146/annurev.energy.21.1.31
1107     https://www.annualreviews.org/doi/full/10.1146/annurev.psych.57.102904.190205
1108          https://www.annualreviews.org/doi/full/10.1146/annurev-soc-081309-150219
1109     https://www.annualreviews.org/doi/full/10.1146/annurev.psych.55.090902.142054
1110       https://www.annualreviews.org/doi/full/10.1146/annurev.soc.28.110601.141117
1111        https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010814-015240
```

```
0      Abstract There appears to be a universal desire to understand individual differen
1      Abstract The recent cultural turn in American sociology has inspired a number of
2      Abstract  AbstractYucca Mountain, NV, is being characterized for disposal of U.S
3      Abstract  AbstractClimate policy is often discussed as a lever with which to brin
4      Abstract Scenarios for the future of renewable energy through 2050 are reviewed t
5      Abstract In movies, robots are often extremely humanlike. Although these robots a
6      Abstract  AbstractThis review surveys five major efforts to identify and declare
7      Abstract A continuing debate in language acquisition research is whether there a
8      Abstract  AbstractThe capacity to exercise control over the nature and quality o
9      Abstract  AbstractFrom the complex literatures on institutionalisms in political
10     Abstract Cultural neuroscience (CN) is an interdisciplinary field that investigat
11     Abstract This article provides a focused review of the current literature on glo
12     Abstract The global population without complete sanitation services is enormous;
13     Abstract  AbstractThe term industrial ecology was conceived to suggest that indus
14     Abstract  AbstractResearchers across a wide range of fields, policy makers, and l
15     Abstract Research on the political consequences of social movements has recently
16     Abstract The assumption that early stress leads to dysregulation and impairment i
17     Abstract  AbstractWater vapor is the dominant greenhouse gas, the most important
18     Abstract The operation of different brain systems involved in different types of
19     Abstract We examine how recent immigration to the United States has affected Afri
20     Abstract In this chapter we review theoretical conceptual and empirical advances
21     Abstract Recent methodological advances have allowed empirical research on adoles
22     Abstract Corporate political activity is both a long-standing preoccupation and a
23     Abstract Why are children of poor parents more likely to be poor as adults than o
24     Abstract  AbstractDespite its economic and cultural centrality, sport is a relati
25     Abstract The Anthropocene is characterized by a widespread biodiversity crisis th
26     Abstract AbstractăMore than half a century has passed since the publication of Ge
27     Abstract AbstractIn this review, we explore how the concept of embeddedness has s
28     Abstract  AbstractThe motivating engines of intellectual life are not true ideas
29     Abstract For more than four decades, I have been studying human memory. My resear
...
```

```
1082  Abstract Standards and standardization aim to render the world equivalent across
1083  Abstract This review describes stress-related biological mechanisms linking inter
1084  Abstract Insight occurs when a person suddenly reinterprets a stimulus, situation
1085  Abstract For the past half-century, the study of organizations has been an active
1086  Abstract The study of secularization appears to be entering a new phase. Supply-s
1087  Abstract This chapter summarizes some of the conceptual changes in developmental
1088  Abstract  AbstractGreen chemistry and engineering is the design of chemical manu
1089  Abstract  AbstractThis article reviews energy indicators, which are developed to
1090  Abstract With the advent of increasingly accessible technologies for typing gene
1091  Abstract Olfaction is often referred to as a multidimensional sense. It is multic
1092  Abstract Sociological research on reading, which formerly focused on literacy, no
1093  Abstract Soils are viewed in the context of ecosystem services, soil processes an
1094  Abstract AbstractWomen's political participation and representation vary dramatic
1095  Abstract People have a tendency to marry within their social group or to marry a
1096  Abstract Progress in treating and preventing mental disorders may follow from res
1097  Abstract Political developments in the United States and Europe have generated a
1098  Abstract The nature of narrative explanations is explored as an alternative to th
1099  Abstract Reviewing recent research on poverty in the United States, we derive a
1100  Abstract Urbanization is one of the biggest social transformations of modern time
1101  Abstract  AbstractWe review literature on several types of energy efficiency poli
1102  Abstract This review of the current status of theoretically based behavioral rese
1103  Abstract Since the 1980s, immigrant children and children of immigrant parentage
1104  Abstract Sociologists have turned to collective identity to fill gaps in resource
1105  Abstract Adult age differences in a variety of cognitive abilities are well docum
1106  Abstract  AbstractAn analysis of the forces that have shaped energy and energy-re
1107  Abstract It is usually assumed that stressful life events interfere with our abil
1108  Abstract Because demographic shifts will affect their labor forces in the immedia
1109  Abstract Cigarette smoking is a leading cause of mortality and morbidity and a pa
1110  Abstract  AbstractSociologists often model social processes as interactions among
1111  Abstract Social isolation has been recognized as a major risk factor for morbidit
```

	ack_idx	key_word
0	73804	depression
1	66814	institution
2	56279	climate change
3	72888	climate change
4	86968	climate change
5	70822	depression
6	74846	climate change
7	58440	depression
8	71421	depression
9	60960	institution
10	80236	depression
11	48252	climate change
12	93831	climate change
13	59764	climate change
14	78444	institution
15	60247	institution

```
16      85340      depression
17      76441  climate change
18      67280      depression
19      59070     institution
20      88591      depression
21      61855     institution
22      60814     institution
23      69679     institution
24      65353     institution
25      68565  climate change
26      60088      depression
27      56472     institution
28      58650     institution
29      54305      depression
...       ...             ...
1082    66235     institution
1083    56735     institution
1084    65473      depression
1085    50089     institution
1086    91548     institution
1087    62180      depression
1088    62799  climate change
1089    59567  climate change
1090    88239      depression
1091    52403      depression
1092    40145     institution
1093    68108  climate change
1094    50838     institution
1095    69968     institution
1096    77297      depression
1097    47212     institution
1098    60059     institution
1099    37316     institution
1100    63288  climate change
1101    69674  climate change
1102    79863      depression
1103    86018     institution
1104    57081     institution
1105    69822      depression
1106    97703  climate change
1107    55796      depression
1108    60330     institution
1109    66173      depression
1110    60700     institution
1111    62834      depression

[1112 rows x 8 columns]
```

0.1 Checking the most frequent words in each topic

```
In [8]: vectorizer = CountVectorizer(lowercase = True,
                                      stop_words= 'english',
                                      max_df = .80,
                                      min_df = .02)
```

```
In [9]: def get_fre_word(text_list):
            vectorizer.fit(text_list)
            frequency_array = vectorizer.transform(text_list)
            word_frequen_df = pd.DataFrame(frequency_array.toarray(),
                                           columns = vectorizer.get_feature_names())
            no_feature_names = len(vectorizer.get_feature_names())
            return word_frequen_df.sum()
```

Finding the most frequent words in "depression" articles

```
In [13]: print("the number of feature keywords is: " + str(get_fre_word(df[df['key_word'] == 'd
         get_fre_word(df[df['key_word'] == 'depression']['article_text']).sort_values(ascending
```

```
the number of feature keywords is: 11765
```

```
Out[13]: children       6380
         health         3601
         memory         3079
         treatment      2978
         stress         2963
         personality    2920
         brain          2770
         2000           2669
         risk           2648
         1999           2583
         family         2581
         outcomes       2498
         age            2483
         2002           2446
         2001           2416
         cultural       2392
         child          2338
         patients       2337
         learning       2278
         2003           2273
         dtype: int64
```

Finding the most frequent words in "climate change" articles

```
In [14]: print("the number of feature keywords is: " + str(get_fre_word(df[df['key_word'] == 'c
         get_fre_word(df[df['key_word'] == 'climate change']['article_text']).sort_values(ascen
```

```
the number of feature keywords is: 11457
```

```
Out[14]: emissions       6546
         carbon          5571
         social          4583
         species         4014
         fuel            3927
         co2             3900
         costs           3889
         cost            3790
         power           3733
         technologies    3547
         technology      3481
         efficiency      3159
         air             2836
         soil            2747
         public          2723
         gas             2671
         governance      2655
         urban           2617
         forest          2617
         policies        2567
         dtype: int64
```

Finding the most frequent words in "institution" articles

```
In [15]: print("the number of feature keywords is: " + str(get_fre_word(df[df['key_word'] == 'i
         get_fre_word(df[df['key_word'] == 'institution']['article_text']).sort_values(ascendi
```

```
the number of feature keywords is: 12077
```

```
Out[15]: women            7439
         children         5195
         family           4763
         countries        4360
         gender           4359
         labor            4261
         inequality       3867
         market           3864
         education        3779
         organizations    3719
         health           3686
         2001             3615
         racial           3439
         2003             3356
         2005             3356
         school           3339
```

```
policy          3325
2006            3311
men             3218
race            3210
dtype: int64
```

1 Classify the articles into the "key_word" categories

```
In [32]: nb_classifier = MultinomialNB()

In [33]: vectorizer = CountVectorizer(lowercase = True,
                                      ngram_range = (1,2),
                                      stop_words= 'english',
                                      max_df = .70,
                                      min_df = 5,
                                      max_features = None)

In [34]: # fit the model
         vectorizer.fit(df['article_text'])
         print(len(vectorizer.get_feature_names()))

149335


In [35]: df['key_word'].value_counts()

Out[35]: institution       479
         climate change    343
         depression        290
         Name: key_word, dtype: int64

In [36]: # create the data based on the model
         review_word_counts = vectorizer.transform(df['article_text'])

In [37]: ## model.fit(X,Y)
         nb_classifier.fit(review_word_counts, df['key_word'])

Out[37]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)

In [38]: review_word_counts

Out[38]: <1112x149335 sparse matrix of type '<class 'numpy.int64'>'
                 with 3009872 stored elements in Compressed Sparse Row format>

In [39]: nb_classifier.coef_[0]

Out[39]: array([-11.25143124,  -7.55292089, -11.21968254, ..., -11.0218568 ,
               -11.79504669, -10.62497544])

In [40]: np.shape(nb_classifier.coef_[0])
```

```
Out[40]: (149335,)
```

```
In [41]: coefficients = pd.Series(nb_classifier.coef_[0],
                                  index = vectorizer.get_feature_names())
```

```
In [42]: coefficients.sort_values()[:20]
```

```
Out[42]: intramural              -14.685418
         homology                -14.685418
         homophilous             -14.685418
         homophilous networks    -14.685418
         homophily occurs        -14.685418
         homophily social        -14.685418
         homophily tendency      -14.685418
         homophobia              -14.685418
         homophobic              -14.685418
         homosexual              -14.685418
         homosexual identity     -14.685418
         homosexuality           -14.685418
         homogenizes             -14.685418
         homosexuals             -14.685418
         hondagneu               -14.685418
         hondagneu sotelo        -14.685418
         honduras el             -14.685418
         honesty                 -14.685418
         hong et                 -14.685418
         hooghe                  -14.685418
         dtype: float64
```

```
In [43]: coefficients.sort_values(ascending=False)[:20]
```

```
Out[43]: energy           -5.105033
         environmental    -5.322357
         climate          -5.591724
         water            -5.608609
         global           -5.804555
         emissions        -5.898656
         carbon           -6.059909
         land             -6.147443
         countries        -6.164831
         policy           -6.213223
         production       -6.218677
         impacts          -6.354314
         figure           -6.374512
         management       -6.381914
         species          -6.387626
         fuel             -6.409533
         co2              -6.416430
         costs            -6.419254
```

```
climate change   -6.426219
cost             -6.445033
dtype: float64
```

```
In [44]: # create the predicted data
         nb_classifier.predict(review_word_counts)
```

```
Out[44]: array(['depression', 'depression', 'depression', ..., 'institution',
                'climate change', 'climate change'], dtype='<U14')
```

```
In [45]: df['prediction'] = nb_classifier.predict(review_word_counts)
```

```
In [46]: pd.crosstab(df['key_word'], df['prediction'])
```

```
Out[46]: prediction      climate change  depression  institution
         key_word
         climate change             343           0            0
         depression                   0         285            5
         institution                  0           9          470
```

```
In [47]: accuracy_score(df['key_word'], df['prediction'])
```

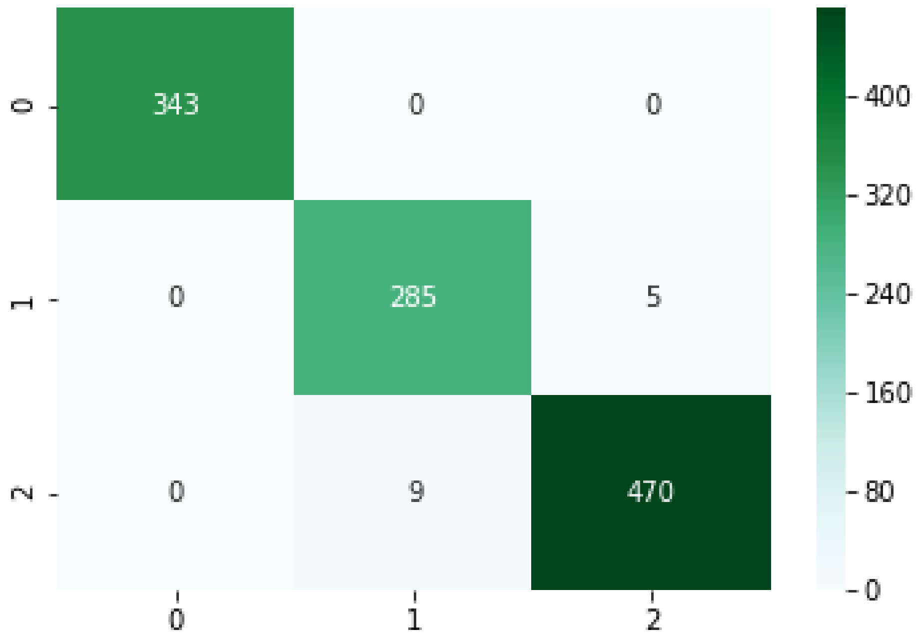
```
Out[47]: 0.987410071942446
```

```
In [48]: print(classification_report(df['key_word'], df['prediction']))
```

```
              precision    recall  f1-score   support

climate change       1.00      1.00      1.00       343
    depression       0.97      0.98      0.98       290
   institution       0.99      0.98      0.99       479

     micro avg       0.99      0.99      0.99      1112
     macro avg       0.99      0.99      0.99      1112
  weighted avg       0.99      0.99      0.99      1112
```

```
In [49]: cm = confusion_matrix(df['key_word'], df['prediction'])
         sns.heatmap(cm, annot = True, cmap = "BuGn", fmt = 'g')
```

```
Out[49]: <matplotlib.axes._subplots.AxesSubplot at 0x1781a6c6cc0>
```



In [50]:
```
## predict the probability of the keywords
nb_classifier.predict_proba(review_word_counts)
```

Out[50]:
```
array([[0., 1., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       ...,
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.]])
```

In [51]:
```
predict_df = pd.DataFrame(nb_classifier.predict_proba(review_word_counts),
                          columns = nb_classifier.classes_ )
predict_df.head()
```

Out[51]:

	climate change	depression	institution
0	0.0	1.0	0.0
1	0.0	1.0	0.0
2	0.0	1.0	0.0
3	0.0	1.0	0.0
4	0.0	1.0	0.0

In [52]:
```
## Testing for overfitting
train, test  = train_test_split(df, test_size = 0.6)
vectorizer = CountVectorizer(lowercase = True,
                             ngram_range = (1,2),
```

```
                                    stop_words= 'english',
                                    max_df = .70,
                                    min_df = 5,
                                    max_features = None)
        vectorizer.fit(df['article_text'])
```

```
Out[52]: CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                 dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                 lowercase=True, max_df=0.7, max_features=None, min_df=5,
                 ngram_range=(1, 2), preprocessor=None, stop_words='english',
                 strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                 tokenizer=None, vocabulary=None)
```

```
In [54]: X_train = vectorizer.transform(train['article_text'])
         nb_classifier.fit(X_train, train['key_word'])
         print(accuracy_score(train['key_word'],
                              nb_classifier.predict(X_train)))
```

```
0.9954954954954955
```

```
In [55]: test_wf = vectorizer.transform(test['article_text'])
         test_prediction = nb_classifier.predict(test_wf)
         print(accuracy_score(test['key_word'], test_prediction))
```

```
0.9535928143712575
```

2 Sentiment Analysis

```
In [5]: from afinn import Afinn
        afinn = Afinn(language = 'en')
```

```
In [6]: def word_count(text_string):

           "'Calculate the number of words in a string'"
           return len(text_string.split())
```

```
In [7]: df['afinn_score'] = df['article_text'].apply(afinn.score)
```

```
In [61]: df.head()
```

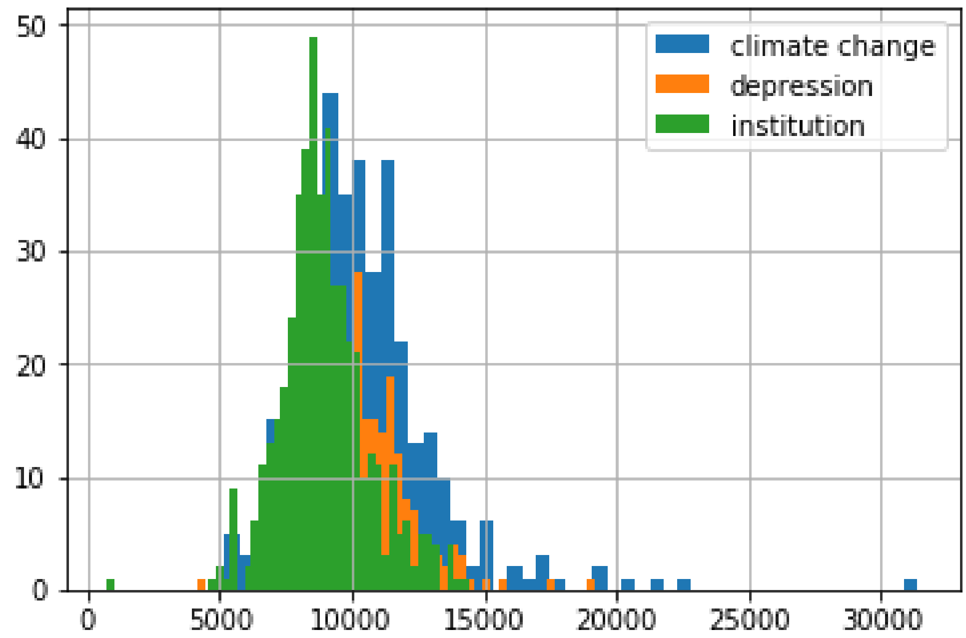
```
Out[61]:                                                                       url  \
         0  https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.58.110405.085516
         1          https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.50.1.165
         2     https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-120710-100356
         3           https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.47.1.33
         4          https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.50.1.599
```

```
0             Research on Attention Networks as a Model for the Integration of Psychol
1                                                                          INTERVENTI
2  Child Development in the Context of Disaster, War, and Terrorism: Pathways of Risk
3                      THEORETICAL FOUNDATIONS OF COGNITIVE-BEHAVIOR THERAPY FOR ANXIETY
4                                                               SINGLE-GENE INFLUENCES ON BRA

                                  author  year  \
0  Michael I. Posner and Mary K. Rothbart  2007
1         A. Christensen and C. L. Heavey  1999
2     Ann S. Masten and Angela J. Narayan  2012
3                        Chris R. Brewin  1996
4                            D. Wahlsten  1999

                                                                        html_url  \
0  https://www.annualreviews.org/doi/full/10.1146/annurev.psych.58.110405.085516
1          https://www.annualreviews.org/doi/full/10.1146/annurev.psych.50.1.165
2     https://www.annualreviews.org/doi/full/10.1146/annurev-psych-120710-100356
3           https://www.annualreviews.org/doi/full/10.1146/annurev.psych.47.1.33
4          https://www.annualreviews.org/doi/full/10.1146/annurev.psych.50.1.599


0  Abstract AbstractAs Titchener pointed out more than one hundred years ago, attentic
1  Abstract  AbstractA substantial body of empirical research has documented both the
2  Abstract This review highlights progress over the past decade in research on the e
3  Abstract  AbstractCognitive-behavior therapy (CBT) involves a highly diverse set o
4  Abstract  AbstractAs traditional behavioral genetics analysis merges with neurogene

   ack_idx    key_word  afinn_score  word_count  afinn_adjusted  prediction
0    55133  depression        164.0        8465        0.019374  depression
1    61144  depression        173.0        9168        0.018870  depression
2    89520  depression       -772.0       13178       -0.058582  depression
3    59781  depression       -245.0        8686       -0.028206  depression
4    51930  depression        129.0        8159        0.015811  depression
```

In [9]:
```
df['word_count'] = df['article_text'].apply(word_count)
```

In [70]:
```
df['word_count'][df['key_word']=='climate change'].hist(bins=50,label='climate change
df['word_count'][df['key_word']=='depression'].hist(bins=50,label='depression')
df['word_count'][df['key_word']=='institution'].hist(bins=50,label='institution')
plt.legend()
```

Out[70]:
```
<matplotlib.legend.Legend at 0x17821a7b630>
```



In [20]:
```
df.groupby('key_word')['afinn_score'].describe()
```

Out[20]:
```
                count        mean         std     min     25%    50%    75%  \
key_word
climate change  343.0  217.396501  239.080757  -914.0  103.00  221.0  339.0
depression      290.0   13.875862  423.535080 -1933.0 -119.25  105.0  244.5
institution     479.0  103.659708  321.893065 -1903.0   39.00  158.0  269.0

                   max
key_word
climate change  1323.0
depression      1196.0
institution     1342.0
```

In [10]:
```
df['afinn_adjusted'] = df['afinn_score']/df['word_count']
```

Calculate the average afinn score for each word in the articles

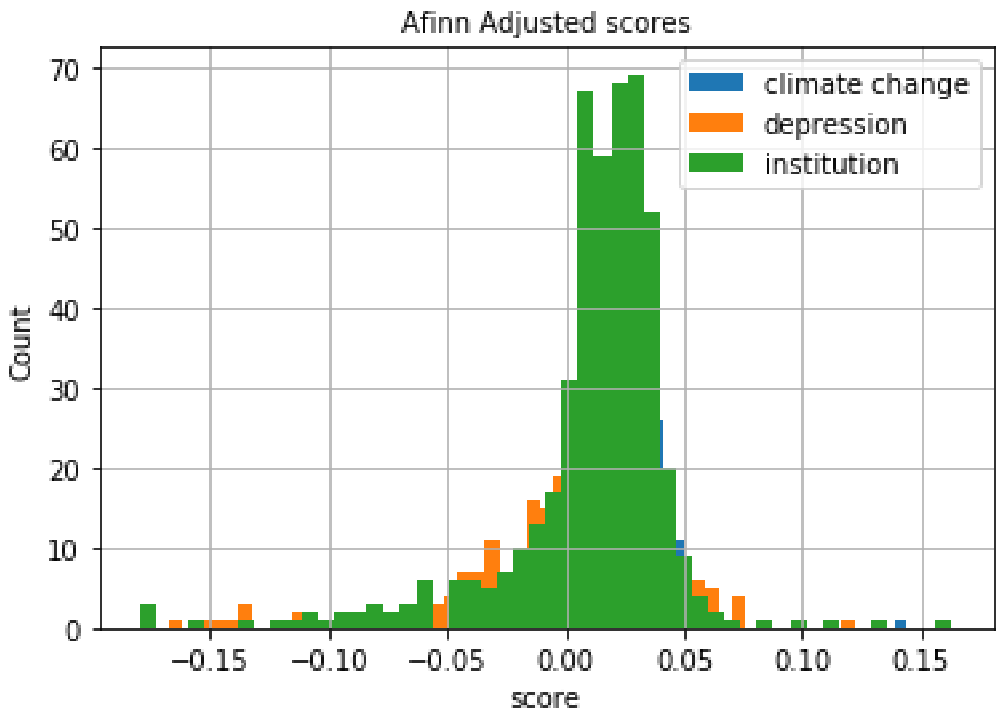
In [25]:
```
df.groupby('key_word')['afinn_adjusted'].describe()
```

Out[25]:
```
                count      mean       std       min       25%       50%  \
key_word
climate change  343.0  0.019958  0.022281 -0.090540  0.010821  0.021235
depression      290.0  0.001213  0.040806 -0.167200 -0.012342  0.010475
institution     479.0  0.011286  0.035840 -0.179608  0.005379  0.018123
```

```
                     75%       max
key_word
climate change  0.031715  0.143804
depression      0.024499  0.122642
institution     0.030202  0.163102
```

In [11]:
```
df['afinn_adjusted'][df['key_word']=='climate change'].hist(bins=50,label='climate cha
df['afinn_adjusted'][df['key_word']=='depression'].hist(bins=50,label='depression')
df['afinn_adjusted'][df['key_word']=='institution'].hist(bins=50,label='institution')
plt.legend()
plt.xlabel('score')
plt.ylabel('Count')
plt.title('Afinn Adjusted scores',fontsize=10)
```

Out[11]:
```
Text(0.5, 1.0, 'Afinn Adjusted scores')
```



In [29]:
```
# top ten negative scores 'depression' articles
columns_to_display = ['title', 'year', 'afinn_score']
df[df['key_word'] == 'depression'].sort_values(by = 'afinn_score')[columns_to_display
```

Out[29]:
```
679                                    The Effects of Family and Community Violence o
716  KEY ISSUES IN THE DEVELOPMENT OF AGGRESSION AND VIOLENCE FROM CHILDHOOD TO EARLY
25                                                              Moral Emotions and Mora
```

```
252                                                                   Famil
417                                      The Psychology and Neurobiology of Suicida
104                                                           Child Maltreatment 
849                         Bullying in Schools: The Power of Bullies and the Plight 
540                                         COMORBIDITY OF ANXIETY AND UNIPOLAR MOOD
997                           Workplace Victimization: Aggression from the Target's P
264                                       Childhood Antecedents and Risk for Adult Mental

     year  afinn_score
679  2000      -1933.0
716  1997      -1604.0
25   2007      -1539.0
252  2006      -1473.0
417  2005      -1369.0
104  2010      -1279.0
849  2014      -1268.0
540  1998      -1230.0
997  2009      -1186.0
264  2015      -1137.0
```

In [30]:
```
# top ten negative scores 'climate change' articles
df[df['key_word'] == 'climate change'].sort_values(by = 'afinn_score')[columns_to_disp
```

Out[30]:
```
147                                                              Disaster Governance: Soc
612   HOW ENVIRONMENTAL HEALTH RISKS CHANGE WITH DEVELOPMENT: The Epidemiologic and E
357                                                           Pyrogeography and the Globa
304                                                                                     
552                                                                Assessing the Vulner
1038          HARMFUL ALGAL BLOOMS: An Emerging Public Health Problem with Possible L
348
410                                                                  Water Security an
813
151                                                             Emerging Threats to Human 

      year  afinn_score
147   2012       -914.0
612   2005       -871.0
357   2013       -545.0
304   2016       -485.0
552   2006       -475.0
1038  1999       -424.0
348   2010       -406.0
410   2014       -355.0
813   2006       -299.0
151   2009       -277.0
```

In [31]:
```
# top ten negative scores 'institution' articles
df[df['key_word'] == 'institution'].sort_values(by = 'afinn_score')[columns_to_displa
```

Out[31]:

```
475
1070                                                                           Hate Cri
888                                     Silence, Power, and Inequality: An Intersectio
491              Violence and the Life Course: The Consequences of Victimization for 
401                                                                             Mass Me
633   Macrostructural Analyses of Race, Ethnicity, and Violent Crime: Recent Lessons a
842                                                              Gender and Crime: Toward a Gend
936                                                                                   
865                 White-Collar Crime: A Review of Recent Developments and Promising
629

      year  afinn_score
475   2002      -1903.0
1070  2001      -1633.0
888   2018      -1444.0
491   2001      -1290.0
401   1996      -1200.0
633   2005      -1070.0
842   1996      -1058.0
936   1998      -1039.0
865   2013      -1033.0
629   1999       -882.0
```

Sentiment analysis using Vader

In [5]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

In [6]:
```
def vaderize(df, textfield):

        "'Compute the Vader polarity scores for a textfield'"
        analyzer = SentimentIntensityAnalyzer()

        print('Estimating polarity score for %d cases.' % len(df))
        sentiment = df[textfield].apply(analyzer.polarity_scores)

        # convert to data frame
        sdf = pd.DataFrame(sentiment.tolist()).add_prefix('vader_')

        # merge data frames
        df_combined = pd.concat([df,sdf], axis =1)
        return df_combined
```

In [8]:
```
df_vaderized = vaderize(df,'article_text')
```

```
Estimating polarity score for 1112 cases.
```

In [9]:
```
df_vaderized.head()
```

Out[9]:

```
                                                                              url  \
0       https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.47.1.143
1  https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015221
2    https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-081309-150129
3 https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.28.110601.140936
4    https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-073117-041131

                                                                                   title  \
0                    VERBAL LEARNING AND MEMORY: Does the Modal Model Still Work?
1 School Readiness and Self-Regulation: A Developmental Psychobiological Approach
2                                                           The Sociology of Finance
3                                                            Violence in Social Life
4  From Chicago to China and India: Studying the City in the Twenty-First Century

                                   author  year  \
0 Alice F. Healy and and Danielle S. McNamara  1996
1            Clancy Blair and C. Cybele Raver  2015
2       Bruce G. Carruthers and Jeong-Chul Kim  2011
3                          Mary R. Jackman  2002
4                               Xuefei Ren  2018

                                                                          html_url  \
0       https://www.annualreviews.org/doi/full/10.1146/annurev.psych.47.1.143
1  https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010814-015221
2    https://www.annualreviews.org/doi/full/10.1146/annurev-soc-081309-150129
3 https://www.annualreviews.org/doi/full/10.1146/annurev.soc.28.110601.140936
4    https://www.annualreviews.org/doi/full/10.1146/annurev-soc-073117-041131


0 Abstract  AbstractThis chapter focuses on recent research concerning verbal learning
1 Abstract Research on the development of self-regulation in young children provides a
2 Abstract The economic crisis of 20082010 stimulated an already growing sociological
3 Abstract  AbstractTwo features have marked the sociological analysis of violence: (a
4 Abstract Since the last quarter of the twentieth century, cities in the Global South

  ack_idx     key_word  vader_compound  vader_neg  vader_neu  vader_pos
0   72224   depression          0.9999      0.035      0.894      0.071
1   60398   depression          1.0000      0.051      0.790      0.159
2   60675  institution          1.0000      0.056      0.824      0.120
3   72294  institution         -1.0000      0.244      0.683      0.074
4   49502  institution          0.9997      0.031      0.905      0.064
```

In [14]:

```
#df_vaderized.to_csv('df_vaderized.csv') (save the vaderized for this will take half
df_vaderized = pd.read_csv('df_vaderized.csv')
```

In [18]:

```
df_vaderized.head(2)
```

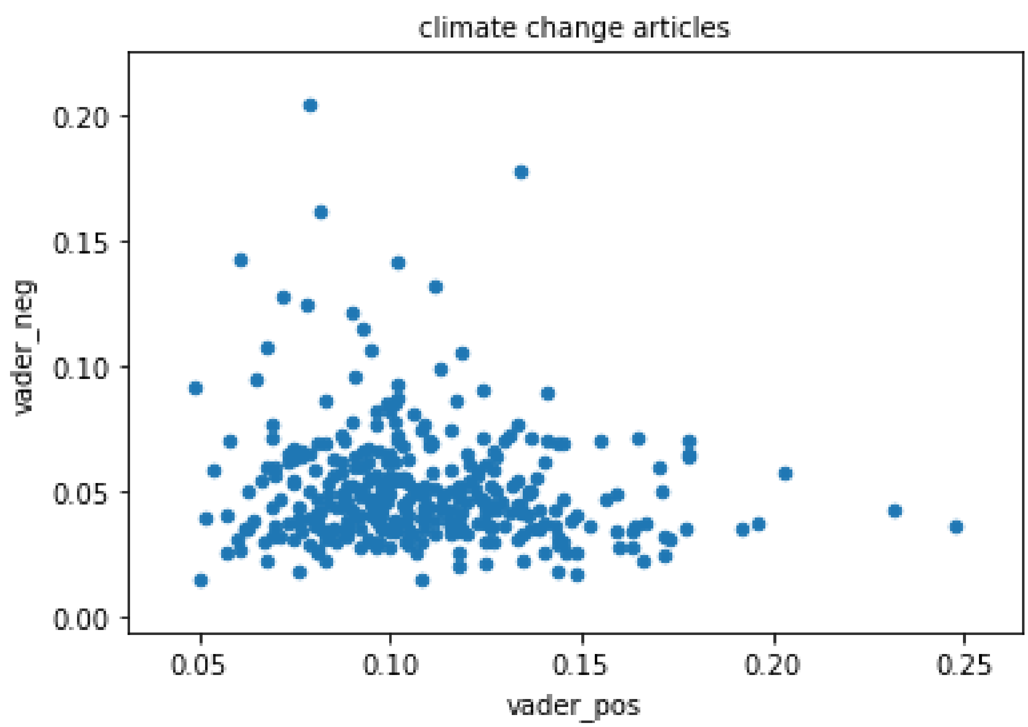
Out[18]:

```
   Unnamed: 0  \
0           0
```

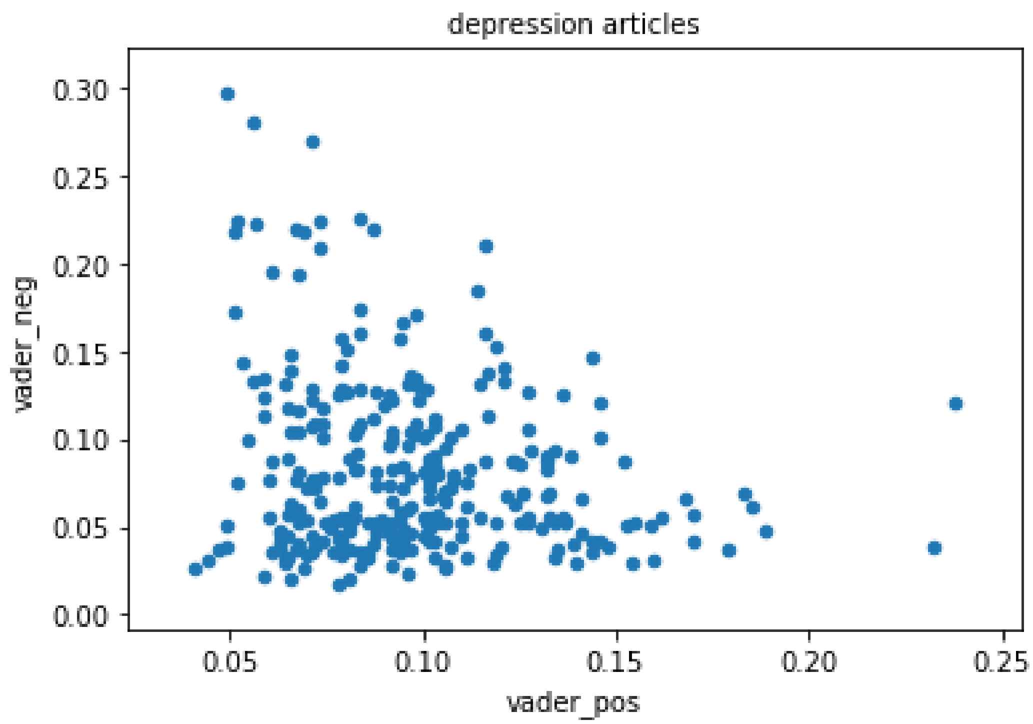
```
1           1

                                                                         url  \
0      https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.47.1.143
1  https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015221

                                                                            title  \
0                      VERBAL LEARNING AND MEMORY: Does the Modal Model Still Work?
1  School Readiness and Self-Regulation: A Developmental Psychobiological Approach

                                           author  year  \
0  Alice F. Healy and and Danielle S. McNamara  1996
1             Clancy Blair and C. Cybele Raver  2015

                                                                          html_url  \
0      https://www.annualreviews.org/doi/full/10.1146/annurev.psych.47.1.143
1  https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010814-015221


0  Abstract  AbstractThis chapter focuses on recent research concerning verbal learni
1  Abstract Research on the development of self-regulation in young children provides

   ack_idx    key_word  vader_compound  vader_neg  vader_neu  vader_pos
0    72224  depression          0.9999      0.035      0.894      0.071
1    60398  depression          1.0000      0.051      0.790      0.159
```

In [19]:
```
df_vaderized[df_vaderized['key_word']=='institution'].plot.scatter(x='vader_pos', y =
plt.title('institution articles',fontsize=10)
plt.savefig('02_vader_institution.png')
```


```
In [20]: df_vaderized[df_vaderized['key_word']=='climate change'].plot.scatter(x='vader_pos', y
         plt.title('climate change articles',fontsize=10)
         plt.savefig('02_vader_climate_change.png')
```



In [21]:
```
df_vaderized[df_vaderized['key_word']=='depression'].plot.scatter(x='vader_pos', y =
plt.title('depression articles',fontsize=10)
plt.savefig('02_vader_depression.png')
```



In [16]:
```
sentiment_variables = [ 'vader_neg', 'vader_neu', 'vader_pos', 'vader_compound']
df_vaderized.groupby('key_word')[sentiment_variables].mean()
```

Out[16]:

key_word	vader_neg	vader_neu	vader_pos	vader_compound
climate change	0.051367	0.841023	0.107592	0.879506
depression	0.082797	0.820376	0.096828	0.314858
institution	0.065058	0.840188	0.094810	0.624717

3 NLP

In [ ]:

In [5]:
```
import spacy
from spacy import displacy
```

In [6]:
```
nlp = spacy.load("en_core_web_sm")
```

```
In [7]: def extract_adjectives(text):
            adjectives = []
            doc = nlp(text)
            for token in doc:
                if token.pos_ == 'ADJ':
                    adjectives.append(token.text)
            adjectives = ', '.join(adjectives)
            return adjectives
```

```
In [92]: extract_adjectives('today is beautiful')
```

```
Out[92]: 'beautiful'
```

```
In [8]: df['adjectives'] = df['article_text'].apply(extract_adjectives)
```

```
In [9]: df['adjectives'].head()
```

```
Out[9]: 0    past, quantitative, genetic, identical, fraternal, human, heritable, more, contem
        1    common, scientific, popular, unconscious, prevalent, potent, modern, theoretical,
        2    cultural, important, past, current, substantive, Conceptual, specific, cultural,
        3    recent, psychological, physical, first, recent, selfish, psychological, physical,
        4    sociological, disparate, various, that, urgent, social, overwhelming, that, devia
        Name: adjectives, dtype: object
```

```
In [13]: df.to_csv('nlp_df.csv')
```

```
In [12]: vectorizer = CountVectorizer(lowercase = True,
                                      stop_words = 'english',
                                      max_df = 1.0,
                                      min_df = 0.0)
         vectorizer.fit(df['adjectives'])
```

```
Out[12]: CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                 dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                 lowercase=True, max_df=1.0, max_features=None, min_df=0.0,
                 ngram_range=(1, 1), preprocessor=None, stop_words='english',
                 strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                 tokenizer=None, vocabulary=None)
```

```
In [16]: wf_psy_array = vectorizer.transform(df[df['key_word']=='depression']['adjectives'])
         wf_psy_df = pd.DataFrame(wf_psy_array.todense(),
                                  columns = vectorizer.get_feature_names())
         wf_psy_df.sum().sort_values(ascending=False)[:20]
```

```
Out[16]: social          7847
         different       3827
         cognitive       3693
         important       3231
         negative        2946
```

```
        high              2940
        positive          2718
        psychological     2666
        specific          2654
        likely            2581
        behavioral        2536
        individual        2442
        cultural          2343
        early             2225
        emotional         2120
        recent            2030
        low               1986
        greater           1972
        higher            1903
        new               1874
        dtype: int64
```

```
In [17]: wf_environ_array = vectorizer.transform(df[df['key_word']=='climate change']['adjectiv
         wf_environ_df = pd.DataFrame(wf_environ_array.todense(),
                                 columns = vectorizer.get_feature_names())
         wf_environ_df.sum().sort_values(ascending=False)[:20]
```

```
Out[17]: environmental    11037
         global            6594
         social            4408
         new               4378
         economic          4344
         different         4288
         large             4250
         human             4064
         high              3972
         important         3826
         natural           3460
         local             3206
         low               2698
         urban             2542
         significant       2377
         public            2293
         recent            2292
         major             2196
         future            2195
         long              2138
         dtype: int64
```

```
In [18]: wf_soci_array = vectorizer.transform(df[df['key_word']=='institution']['adjectives'])
         wf_soci_df = pd.DataFrame(wf_soci_array.todense(),
                                 columns = vectorizer.get_feature_names())
         wf_soci_df.sum().sort_values(ascending=False)[:20]
```

Out[18]:
```
social         21107
political       7457
new             5847
economic        5611
different       4708
cultural        4570
important       3987
public          3760
racial          3375
likely          3362
high            3350
recent          3125
american        2993
national        2945
black           2851
ethnic          2728
religious       2726
large           2649
individual      2636
educational     2578
dtype: int64
```

In [ ]: