Summary of the Analysis

This document explains the step by step analysis of how the analysis was done.

Data was scraped from The Hindu for the years 2000-2019. These were named as hindu_data/yx.csv for all years x.

The first step in the analysis was to score the articles based on a list of words and associated words returned by the Word2Vec model. The list had 30 words which had the minimum cosine distance from the word “corruption”.
The word corruption was added to the list and given a weight of 1. Also the weight of the word corrupt was changed to 1.

First the list created for tokenized articles without Lemmatization or Stemming:

```
corruption 1
nepotism 0.9195553064346313
favouritism 0.8965622186660767
indiscipline 0.8548580408096313
maladministration 0.8346401453018188
graft 0.8276455998420715
wrongdoing 0.7929514646530151
mismanagement 0.7868432998657227
illegality 0.7804960012435913
criminality 0.7802691459655762
horse trading 0.7802132368087769
misuse 1
corrupt 0.7622466087341309
omission 0.7567644119262695
impropriety 0.7549698352813721
dishonest 0.752644419670105
inefficiency 0.7506660223007202
scandal 0.7474774122238159
scam 0.7340414524078369
malice 0.7330329418182373
tactic 0.727906346321106
vendetta 0.7277158498764038
allegation 0.725343644618988
cash for vote 0.7236220240592957
coercion 0.7228101491928101
bribe 0.7221366763114929
conspiracy 0.7185566425323486
undue 0.7079461812973022
accusation 0.7077040672302246
irregularity 0.7007368206977844
malafide 0.7007308602333069
```

Then I used nltk.tokenize to tokenize the articles and check for the presence of these words in the articles in lowercase. The score of an article = ∑**(frequency[i]*weight[i])**

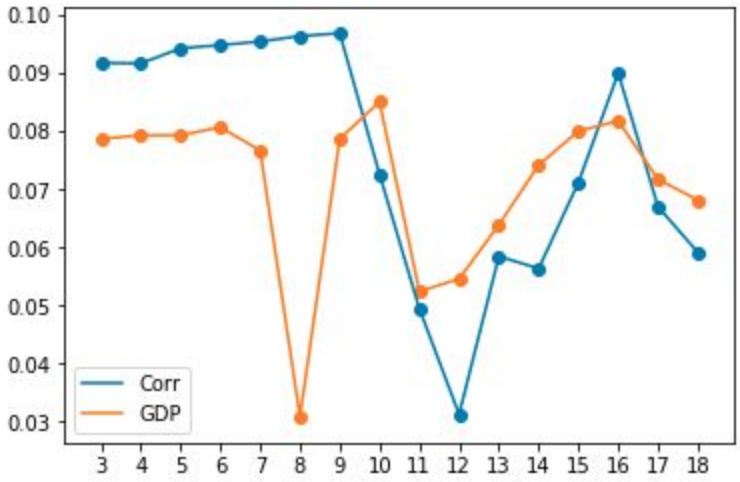
Then I redid the analysis for Lemmatized and Stemmed words from the gensim module. All articles were lemmatized using WordNetLemmatizer and Stemmed using SnowballStemmer.

Once the articles were scored, yearly data was saved as hindu_data_Scored_yx.csv for all years x.

Then for each year I calculated the ratio of articles crossing the cutoff of (1.5 for unlemmatized and 1.13 for lemmatized) which was the sum of two least weighted associated words in the list.

The corruption index for a year was calculated as 0.1 - ratio for that year. This was done to incorporate the fact that higher value of ration corresponds to lower GDP growth value for that year. I call this the Corruption Perception Index Inverse for a year. Ths was plotted against the GDP value for the next year. For example, the CPII for 2007 was plotted against the GDP value for 2008. This 1 year impact window was chosen to incorporate the fact that the corruption related events in this year would affect the GDP growth values released in March of Next Year.

The plot obtained is shown below:



The x axis is for Year post 2000. And the Y-Axis shows the value of CPII and GDP growth rate. The correlation fails for the year 2007 due to the Financial Distress caused due to the burst of the Housing Bubble.

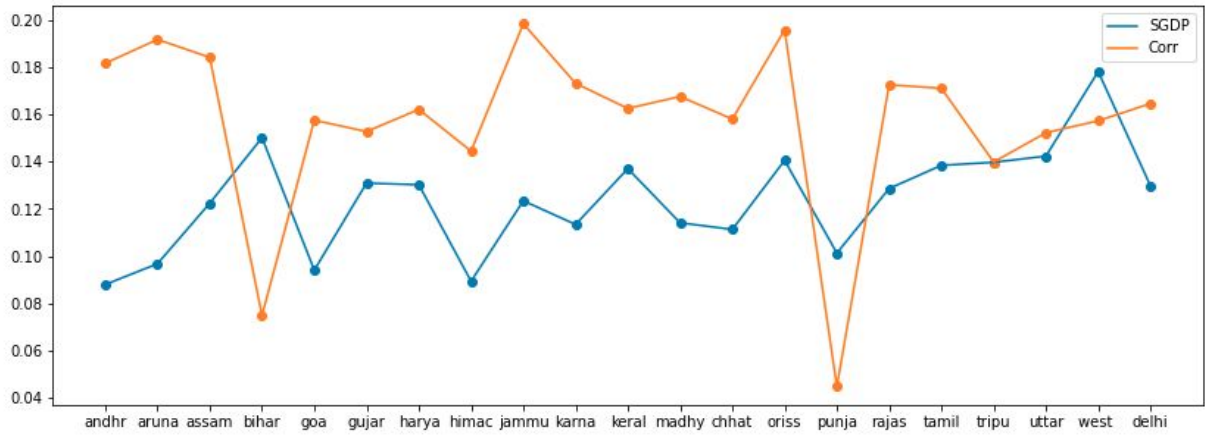
To summarize, this can **predict the GDP of the country for the next year using the newspaper articles of this year!!!**

Then I proceeded to do the State wise analysis. The first step was ofcourse to know which article belongs to which state or in other words tag states to articles. The Hindu has a Category column in their data which doesnt always tags articles to states. It does have values like “Andhra Pradesh”, “Telangana”, “Karnataka” etc.
To tag states to articles for each article I used a dictionary with state names as keys. Using locations.csv containing Places, Districts and States I incremented the corresponding state whenever an occurrence of any of its Places or Districts appeared in the article. Then the state with the highest value was chosen as the tagged_state for that article.

To incorporate the tagging of some articles that The Hindu does, the final tag of an article was the same as the one appearing in its Category column if available, or else was the tagged_state created by my code.

Now with the States tagged to each article I repeated a similar analysis as the GDP one to get the SGDP growth rate plotted against the number of articles from that state.

For 2017, States with the CPII and SGDP growth rates are shown below:



After this, I got the state wise time series plots as well for the years 2012-2018.

The correlation was decent considering the lesser amounts of data present and better plots can be expected when multiple newspapers are used in the analysis.

Plots for Jharkhand and Kerala follow: