Technical Part

Introduction

In order to gain more understanding, and deepen economic analysis of the correlation of Internet penetration, and level of development, We use *Times of India* as the main unstructured data source. *Times of India* is an Indian-English based newspaper. With track records over 150 years, this newspaper has been acknowledged as reliable, actual, and factual source of information. Our group has newspaper data for *Times of India* dated back from 2006 until 2012. Over the span of years, there are thousands of articles which some of them have some correlations with the Internet penetration in India.

Objective

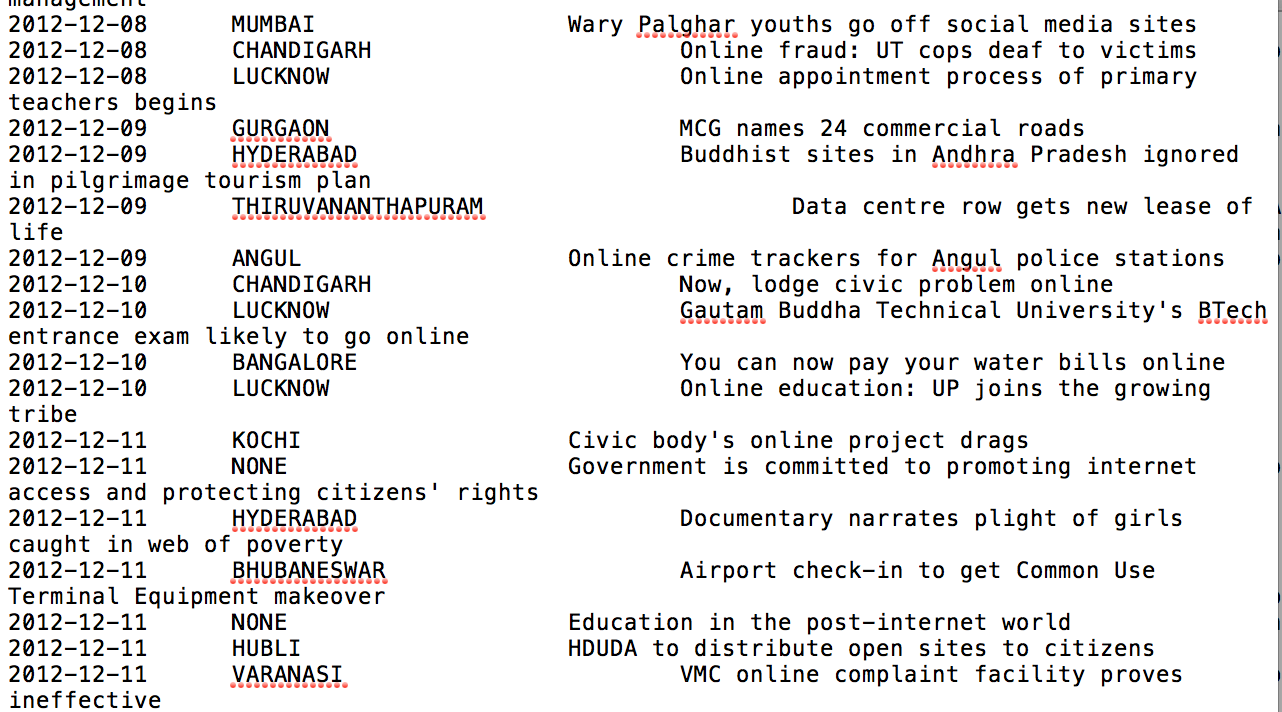
The objective of using unstructured newspaper data is to calculate or predict the presence of Internet or mobile within India from 2006 until 2012. Our hypothesis is that there is a causal correlation between the presence keywords that are correlated with newspaper data, with the Internet penetration. Consequently, these causal correlations might be an alternative economic indicator of India’s development. The purpose of this new unstructured data is to give new alternative of economic indicator aside from GDP, level of health, and education. From these unstructured data, our group creates a financial model which compares the analysis from the unstructured data over the course of 8 years with the structured data of internet penetration rate, and other development indicator such as GDP per year, and standard of living (Gini) number.

Based on regression model analysis, our group want to prove our hypothesis that there is a correlation between newspaper article trending, and internet penetration rate, which might lead to increase in economic growth of India. Moreover, a good regression between those factors will be really useful to create another financial projection in the near future. Consequently, predicting future economic growth based on the regression is a really useful economic tool which may lead to increase in investment, and bonds due to investor confidence in India’s market. Hence, our project is quite pertinent since the idea can be implemented in a large scale to increase the economic development in India.

Technical specification

The times of India data is taken from <http://timesofindia.indiatimes.com/archive.cms> as JSON (Javascript Object Notation) nodes. The information is stored in the JSON such as city, code, title, and text. In order to find the correct articles, we have dictionary that contains various keywords that have some correlations with the presence or development of Internet such as ADDRESS, INTERNET, 3G, 4G, MOBILE, BANDWITCH, PROVIDER, and so on (reference:analysis.py).

Our group created a script to read every title of all the articles, and filtered all articles that have some correlation with keywords inside the dictionary. Aside from the titles of the articles, the script also keeps track of the date of the articles, and the location of events, which might be useful for Graph theory analysis. Moreover, another source[[1]](#footnote-1) is used to provide the coordinate geo-location of cities in India. From that particular source, the locations of all the cities in India can be identified. Some of the Graph theory concepts such as UDG, NNG, and Voronoi concepts can be applied to strengthen the argument, or make a prediction about the graph pattern of the economic development in India.



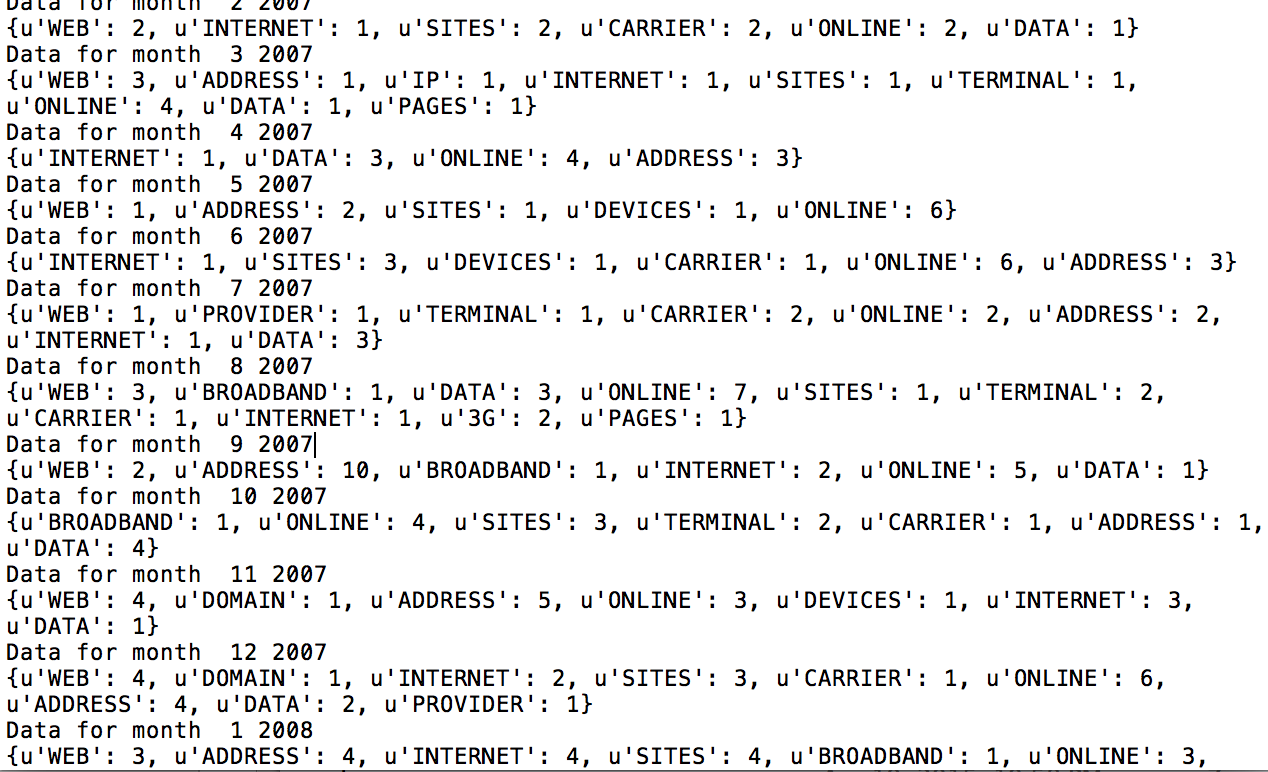
*Figure 1*

The most important part of my script is how our group decided to use the tokenizing technique after the script read the JSON line. According to Wikipedia, in lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens[[2]](#footnote-2). Our group use a ‘regular expression’ to tokenize a sentence, and to remove comma, semicolon, space, and so on.

After tokenizing title of the each article is done, the script will check every tokenized word from the title with the keywords dictionary. Based on the time complexity analysis, the time taken to check each word in keyword dictionary is constant time complexity O(1). Hence, the complexity time taken to check n-tokenized words is linear time O(n).

Classification of the unstructured result

The purpose of tokenizing words is to create a frequency table per month consisting of occurrence of each of the keywords in the dictionary. Moreover, the cumulative sum of those keywords in the dictionary that present in the title of the articles will be tracked in order to create a time-series regression model.



*Figure 2*

Based on figure 2, on each month, the frequencies of keywords that are related to internet are keep tracked in order to build a regression model. According to the data gathered, our group found that there is an increasing trend of internet-related discussion in the newspapers from 2006-2012. Our group created a financial model that correlate the presence of Internet penetration in the newspaper with other economic development within the same timeframe period. Their correlation might prove that these unstructured data can be used as a new economic indicator of development. Consequently, through several financial models, our team explores the possibility to use newspaper or media as a tool that can replace GDP, level of education as an economic indicator. The selling point of our project is to minimize cost of creating an economic indicator which can gauge the level of development in India, that can increase the confidence of government and investors about the opportunity to invest in India.

Challenge

The challenge of the python script is detecting the false positive keyword in the article that appears in the dictionary but has different meaning, or so called homonyms. For example, the word ‘address’ might be interpreted as the location of web address, or the location of someone, or synonymy’s of ‘talk’.

The alternative solution is to create a word tagger which detect whether a word within a sentence can be considered as subject, predicate, or object. NLTK library has a word tagger function that is pretty useful to classify words. The ‘address’ keyword in the dictionary has higher tendency to be regarded as a noun, instead of the verb. Hence, aside from checking the word, we also need to check the position of the word within a sentence. However, classification based on the NLTK tagger might cause some relevant words might be considered as false negative since those words didn’t fulfill the criteria.

Moreover, another challenge is to determine whether a word in a substring of another word in the article, and vice versa. Some of the common example is between plural and singular words that might exempt the use of ‘S’.

1. http://download.geonames.org/export/dump/ [↑](#footnote-ref-1)
2. http://en.wikipedia.org/wiki/Tokenization\_(lexical\_analysis) [↑](#footnote-ref-2)