# Using Topic Modeling to Interpret Discussion of Property Market Trend on Social Media

*Hao Rong*

## 0 Abstract

Housing bubble is one of the most urgent social problems to address in China. To guide investment behavior and to make effective regulatory policy, it is essential to understand how market discussion shift corresponding to the change in socio-economic background. Since most of the market prediction and evaluation from individual or agency are posted online in the form of text, text mining could be a potent tool in extracting information. This research focuses on obtaining valuable information from text data and making statistical convincing inference on estimating market expectation.

## 1 Introduction

Spreading real estate risk, especially residential property, has concerned every family in China. The property price in most of the big cities is so high that ordinary citizen who does not own a house cannot afford to live in the cities. At the meantime, population growth is not high enough to fill in the extra houses built in the new town. With soaring real estate price absorbing most of the investment, manufactory and other industries have been suffering from low investment and low interest in entrepreneurship. However, local government has been relying on finance from the land sale since the beginning of 21st century. Without smoothly transiting the taxing policy from indirect tax like land sell to direct tax, the Chinese government cannot easily restrain the real estate price which is positively related to land price.

To prevent from bubble burst causing a potential financial crisis, the Chinese government has been putting a long-term effort in renewing the regulatory policy to control the real estate price from drastic soar but also to hold the price within a reasonable small increasing rate. The property market has been experiencing more than four rounds of up and down due to recurrent loosen and tighten policy since 2005. The most recent deflate period happened in 2015. In 2016, real estate market in China experienced an astonishing price boost in recent history. Price of real estate had almost doubled in some of the major cities. On October 18, 2017, president Xi said "Houses are built to be inhabited, not for speculation (trade for profit)" at the 19th Party Congress in Beijing. "At such an important occasion, it is quite unusual for top leaders to be so straightforward" commented by Larry Hu, who think this is a signal that Chinese government will build up a long-term mechanism to cool down the housing market.

Wechat Subscription Account platform was developed in Aug 2012. It is an affiliate platform attached to Wechat, one of the largest standalone messaging apps by 2017. The reason of choosing this platform as the data source because it is currently the most widespread social media platform for any entity to register and promote ideas, as described in the official website: 'WeChat Subscription account is typically the most basic choice of the official accounts. It allows you to push frequent content to your followers. Account manager can broadcast one message per day. The account followers will see the update information in the subscription area.' Unlike other media application, Wechat Subscription Account platform provides direct access to let the article to be directly forwarded and shared to the group chat and Wechat moment. Links from other platforms need to be viewed in a different browser which creates a barrier for users to browse outside of Wechat. Hence, Articles on the platform has the characteristics of being widespread and create the effect of echo chamber.

## 2 Objectives

The main objectives of this research is to understand how market expectation shift corresponding to the change in socio-economic background. With the result of text mining, I want to analyze how each information related to the regulatory policy in a statistical sense. How does different regulatory policy influence the market? And how is the effect different from each other? The market expectation could be measured by approaches below: What the prediction towards general property price? Towards short-term and long-term price? (Increase/Decrease) What the emotional reaction in regards to the general market? What the terms and facts related to the market prediction? (Increase/Decrease) How all these topics changes corresponding to the time?

## 3 Data Source

The data source is scraped from the Wechat Subscription Account platform in forms of separate articles. The metadata fields include title, author, date, and content. I also use Sogou search engine for Wechat article to filter out the desired articles. The key word I use to search is “房价” (property price). For each month from June 2013 to October 2017, I scrape the first 200 articles from the search result, except for some month in 2013, when there were less than 200 related articles published. Duplicated articles in the same month are removed. No data are scraped before June 2013 also because of the insufficient number of articles. In those months with not sufficient related articles, the search engine starts to provide less related articles after certain numbers of results. If topic modeling algorithm could not differentiate these less related articles, I might choose to remove data from those months.

Since all the articles are in Chinese language, I need to apply tokenization to create space to segment words in Chinese. The package for segmentation in Chinese is LTP-Cloud developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology. This package also conducts sentence separating on data. All the special characters, English characters, and numbers are removed after tokenization. The result of the tokenization is not perfectly accurate. It might create some degree of bias in the final result.

When viewed from web instead of from Wechat application, the number of views, the number of likes, the originality, and the comments of the articles are omitted. So I cannot measure the effectiveness of each article. Also due to undisclosed the search algorithm from Sogou, I cannot know how the articles are ranked in the search result. If we assume search engine provided a random result, based on 99% of the confidence level, the population of average 1000 articles, we need a sample size of 143 to get an estimate with a confidence interval of 10. Since my sample size is 200, I think my samples can give a fairly good estimate of the entire data. Although the duplicated articles are removed, there still exist some degree of repetition in the data. Repetition cases are usually highly similar content with small change in titles. It could exist within the same month or among different month. Examples are: “Will housing price rise again? All people in [Certain Location] should see this”, “Real estate market collapse in [Certain year]? Let experts tell you why this is impossible.”, Where content in square bracket could be replaced in different articles. Besides repetition, there is also a significant amount of property listing advertisement in the dataset. I decide to keep these data unless there is a strong reason to remove them.

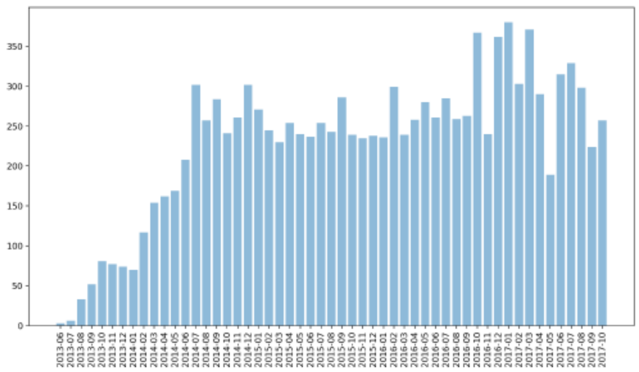

*Figure 3.1: Number of Articles per Month*

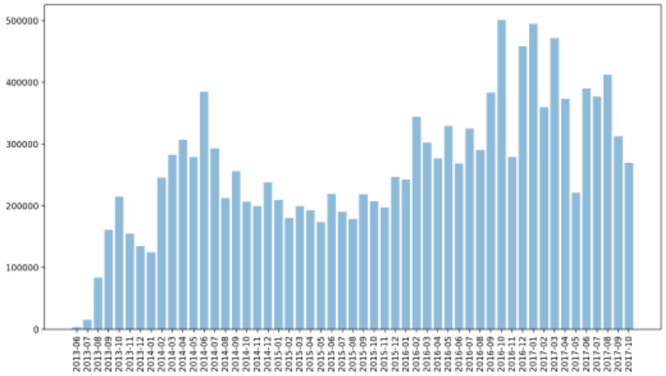

*Figure 3.2: Number of Total Tokens per Month*

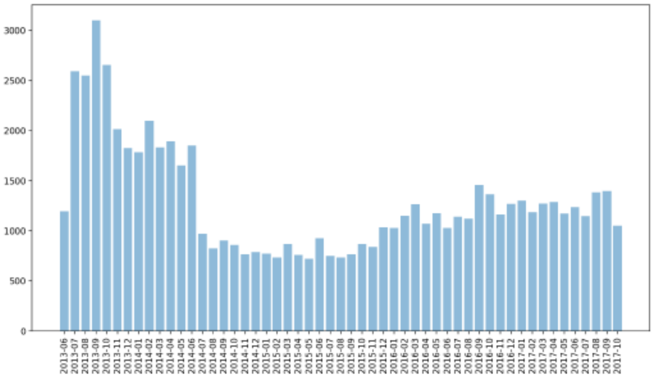

*Figure 3.3: Average Length (Token Numbers) of Articles per Month*

My scraping method setup is to collect as many articles as it can until it hit a verification request. Due to this setup, the number of articles I collect for each month is totally unstable. From Figure 3.1 and Figure 3.2 we can see the distribution of the number of articles and total text length(tokens) per month. In analogy to literature text data, it shares the similar characteristic that each document has different length and number of unit pages. So I decide to keep all the data for analysis but will normalize each quantitative result in the analysis phase. Another factor should be taken into consideration is there is not enough amount of data before 2014-06, and the search engine starts to provide totally unrelated categories of articles at the end of the search result, which might lead to the oddly long average length of articles in those months. So I will truncate the result in visualization even I still use them in training set of topic modeling.

## 4 Methods

The analytical tool of text mining associated with this research purpose involves topics modeling. Topic modeling is used to categorize the data into different topics that related to the main theme of housing price. Then I define human labels on each of the topics. Unrelated content can be identified and removed during this period. Topics will be analyzed compared with market data which includes investment data from National Bureau of Statistics of China and Hundred Cities Price Index from China Real Estate Index System. January data of each year are lost in the investment data, so I use the average of adjacent values.

Additional bar in Figure 4.1 denotes the major interference policy carried out during this time period. Two stimulus policy is released each in October 2014 and October 2015. A regulation policy is released in

December 2016. The reason that total investment and residential investment are similar because real estate in China mainly yields profit from selling residential building during the process of urbanization.

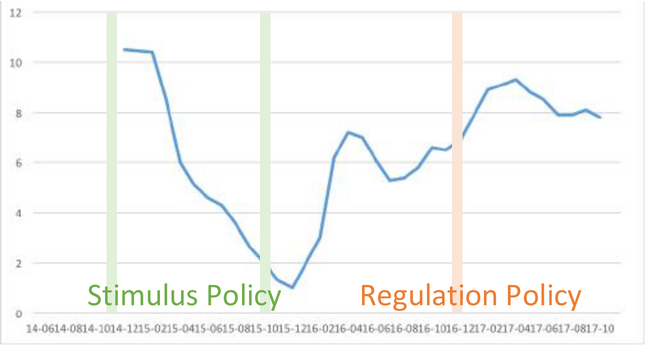

*Figure 4.1:Investment of Real Estate Accumulated Growth Rate (%)*

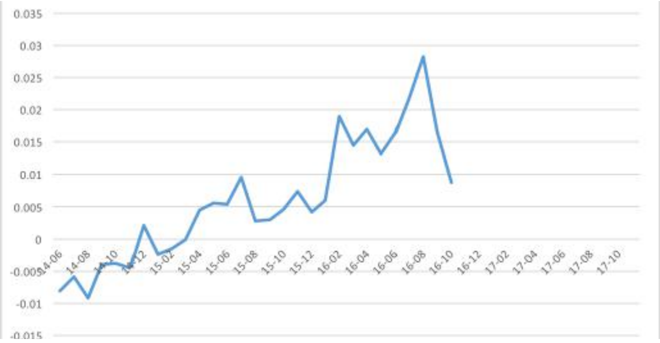

*Figure 4.2: Hundred Cities Price Index Growth Rate (%)*

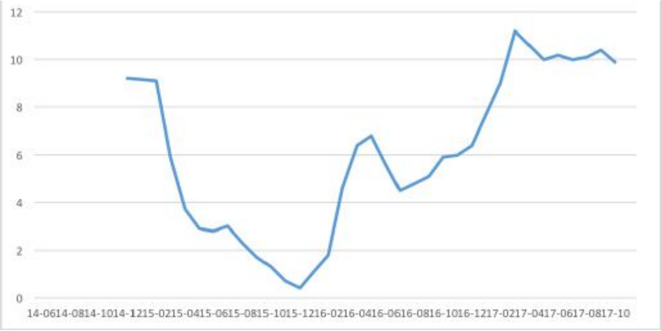

*Figure 4.3: Total Investment in Residential Buildings Growth Rate (%)*

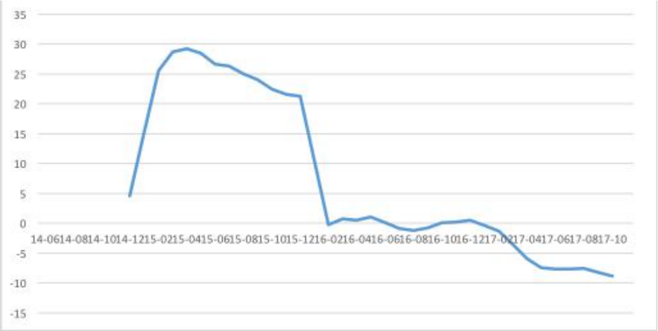

*Figure 4.4: Total Investment in Residential Buildings Below 90 Square Meters Growth Rate (%)*

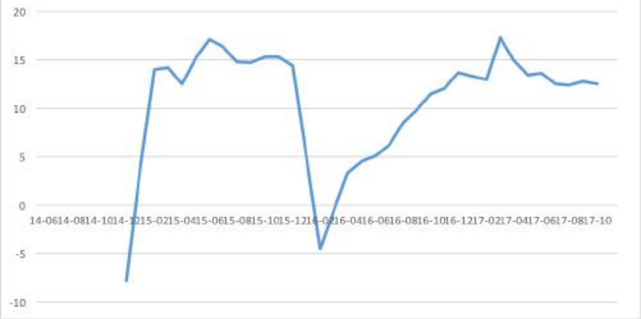

*Figure 4.5: Total Investment in Residential Buildings Above 144 Square Meters Growth Rate (%)*

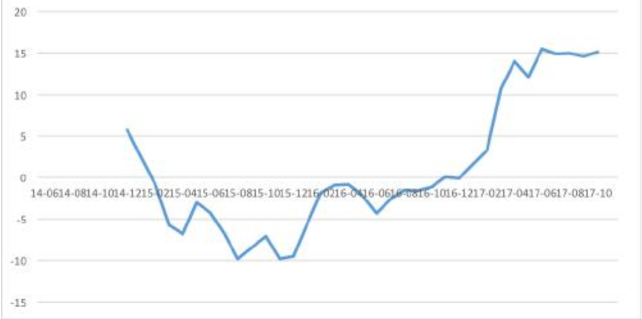

*Figure 4.6: Total Investment in Residential Buildings Villas and High-grade Apartments Growth Rate (%)*

## 5 Proportion Trends of Topics

I use MALLET to apply topic modeling. I have changed the configuration to adapt to the data. First I keep the sequence within the data. Second, since my training set is tokenized space-separated Chinese text and one-character or two-character words are common in Chinese, I change the token-regex to "\p{L}+". I also use a Chinese stop words list when importing the data. When training the topics, I have tried and finally set the topic number to 40 and the number of top words to 40 because the data are complex enough to contain such amount of topics. Since I believe each article is long enough, I use article as the unit of document to train the topics.

After I train the topics, I manually label each of the topic to a main class and sub class. The main class includes ["Oversea", "Price", "Analysis", "Ads", "Society","Stories","Purchase","Development", "Policy", "Unrelated"]. Sub class contains detailed labels which could be the same for some topics. The topic with the same class name will be aggregated together during the topic proportion trend analysis. The proportion of each topics in an article is given in the 'composition' file of the training result. The proportion of a topic T in month M is calculated through aggregation of proportion value of each article $M_i$and normalization of dividing by number of articles N in that month. The proportion of a main *Topic* is the sum of the proportion of each subclass $T_j$.

$$P_{T,M} = \frac{1}{n}\sum_{1}^{n} P_{T,M_i} \qquad\qquad P_{Topic,M} = \sum_{1}^{m} P_{T_j,M}$$

| ***Topic_ID*** | ***Manual Label*** | ***Topic_ID*** | ***Manual Label*** |
|---|---|---|---|
| 0 | Oversea: Positive | 20 | Policy: Regulation |
| 1 | Price: 1st Tier Cities | 21 | Society: Education |
| 2 | Analysis: Investment | 22 | /: Unrelated |
| 3 | Ads: Listing | 23 | Ads: Decoration |
| 4 | Price: 2nd Tier Cities | 24 | Ads: Listing |
| 5 | Analysis: Wait | 25 | /: Unrelated |
| 6 | Society: Young's Stress | 26 | Development: Pearl-Delta |
| 7 | Price: North China | 27 | Policy: System |
| 8 | Society: Mentality | 28 | Price: East China |
| 9 | Society: Travel | 29 | Purchase: Dispute |
| 10 | Analysis: Supply Demands | 30 | Ads: Financial |
| 11 | Stories: Analogy | 31 | Policy: Economy |
| 12 | Analysis: Rise | 32 | /: Unrelated |
| 13 | Purchase: Procedure | 33 | Stories: Bride Price |
| 14 | Development: Beijing Vice Capital | 34 | Oversea: Negative |
| 15 | Ads: Listing | 35 | Purchase: Mortgage |
| 16 | Ads: Estate Company | 36 | /: Unrelated |
| 17 | Analysis: Statistics | 37 | Analysis: Monetary |
| 18 | Society: Region Culture | 38 | Ads: Listing |
| 19 | Analysis: Monetary | 39 | Policy: Economy |

***Table 5.1: Manual Label for each Topic***

From the proportion trend of main topic class, we can observe a clear pattern of time series data from which we can deduct that there might exist a periodical pattern within these topics. The underlying reason might be its relationship to the market data which also has time series. Or it might because of the recurring topic effect on social media. Around three major policy action, we can observe some relatively big change in the proportion of topics. Namely, the proportion of ‘Analysis’ will first drop off then rise within a month. Another pattern we can observe is the negative correlation of ‘Analysis’ and ‘Ads’ and of ‘Analysis’ and ‘Policy’ when one topic proportion decreased, the other increased.

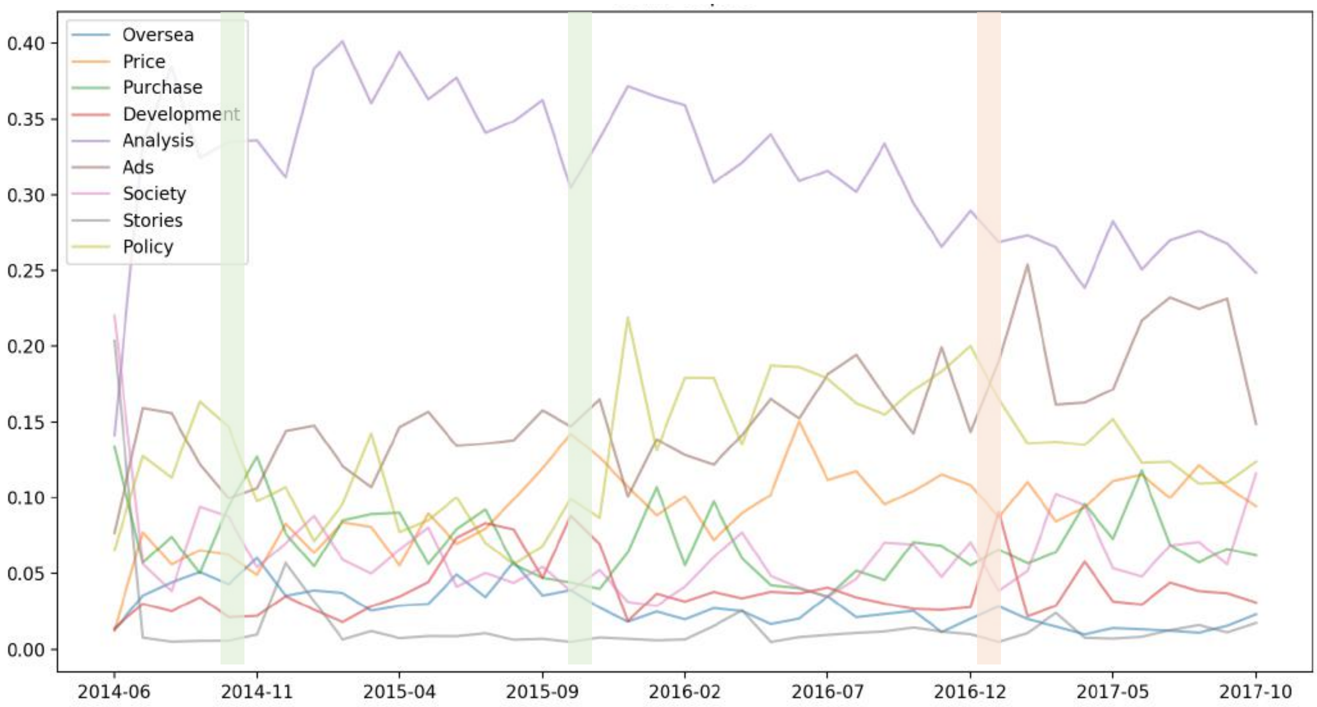

***Figure 5.1: Trend of Main Topics Class (Aggregated Proportion)***

The subclass of ‘Ads’ topic includes ‘Listing’ which mainly list the name and the price of a collection of properties, ‘Estate Company’ which describe the development, market capital, image PR of certain real estate company, ‘Decoration’ advertisement and ‘Financial’ ads that promote the financial product. Listing ads increase periodically while other three remain relatively stable. The subclass of ‘Oversea’ topic is not distinctly separated. Generally, ‘Positive’ promotes investment and immigration to US, Japan, and European countries. ‘Negative’ warns the historical bubble burst or financial crisis in US, Japan, and Hong Kong area. ‘Negative’ topic decreases while ‘positive’ topic remains stable.

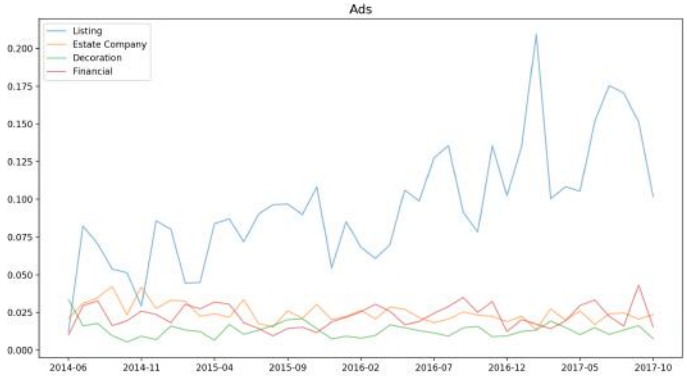

***Figure 5.2: Trend of Subclass in Advertisement Topic***

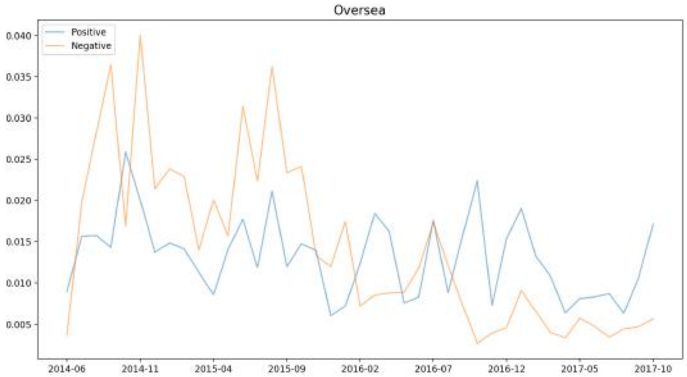

***Figure 5.3: Trend of Subclass in Oversea Topic***

The main class of ‘Analysis’ and ‘Price’ is similar while the former emphasizes on prediction and reasoning for the future trend, the latter directly provides price information for a region. Within ‘Analysis’, ‘Supply Demands’ and ‘Monetary’ topics have a high proportion in 2014 and 2015 and a low proportion in 2016 and 2017. This pattern could correspond to the dramatic drop in February 2016 in the total investment of residential buildings below 90 Square meters which could indicate that the rigid demand of residency had been largely satisfied by the beginning of 2016 and the main reason that drive people in buying property changed to investment need. The trend of ‘wait’ topic has been steadily increasing which could imply that people become hesitater in buying. Within ‘Price’ class, ‘1st Tier Cities’ topic busted at the end of 2015 which correspond to the price bust of 1st Tier Cities in 2016, following by the price bust of 2nd Tier Cities.

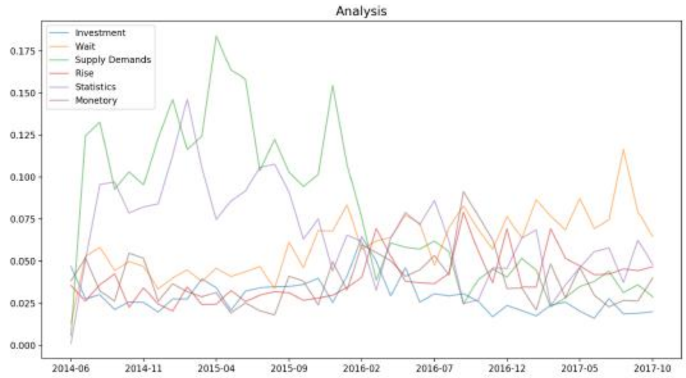

*Figure 5.4: Trend of Subclass in Analysis Topic*

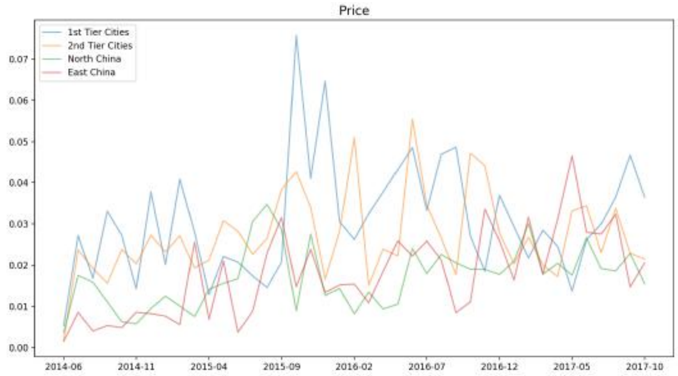

*Figure 5.5: Trend of Subclass in Price Topic*

The main class of ‘Policy’ focus on planning and adjusting the society from an academic and governmental view while ‘Purchase’ topic focus on the logistics during the process of buying. Within ‘Policy’, we can see a high period of ‘Regulation’ topics before and during the execution of the governmental regulation. During the stimulus policy period, ‘Regulation’ and ‘System’ topic which regarding the government structure are low. Three of the policy topics are active during 2016 which is the period of incredible high growth in housing price. Within ‘Purchase’ topic, ‘Procedure’ topic which regarding the purchase and taxing process peaked at the beginning of 2016. ‘Mortgage’ topic dropped down dramatically for reason I have not discovered yet.

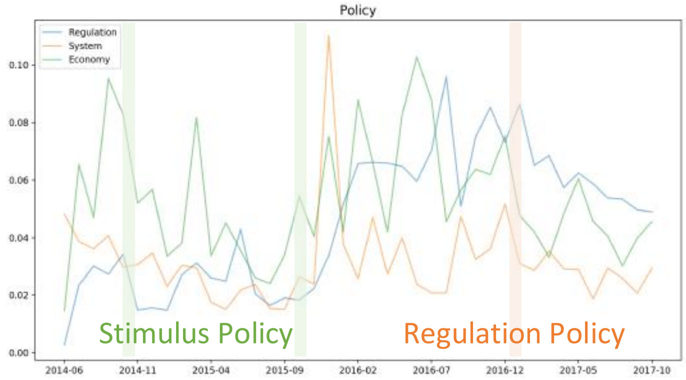

*Figure 5.6: Trend of Subclass in Policy Topic*

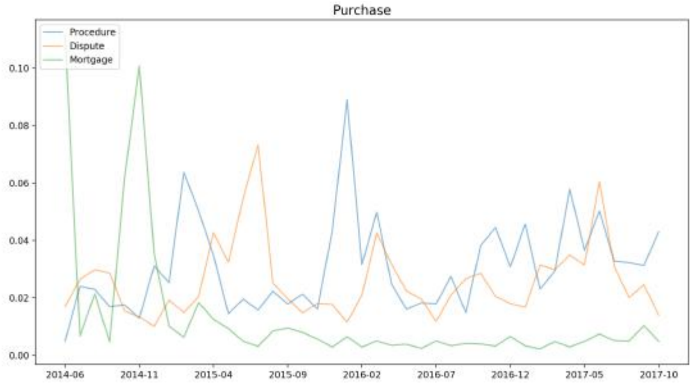

*Figure 5.7: Trend of Subclass in Purchase Topic*

Since the main class of ‘Stories’ shares the similar pattern and concept to the ‘Mentality’ class under ‘Society’ that both of them use stories to promote a living attitude, I will aggregate them together in the follow-up research. Under ‘Development’ class, each topic describes the plan that will be carried out in the future. Under ‘Society’ class, the ‘Education’ topic decrease dramatically after June 2015 which confuses me because the idea of detaching education right from residency was brought up on June 20. This could result from the possible satisfaction of rigid need of residency by the end of 2015. The topic of ‘Young’s Stress’ has been thriving since the beginning of 2017. This might relate to the worry of the stabilization of social structure after the stabilization of the property price and the unaffordability of housing for young people.

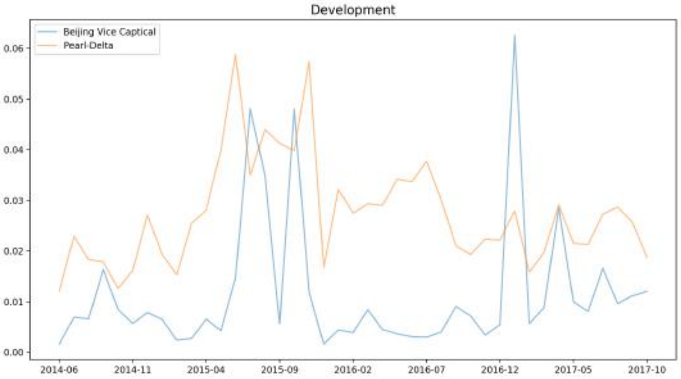

***Figure 5.8: Trend of Subclass in Development Topic***

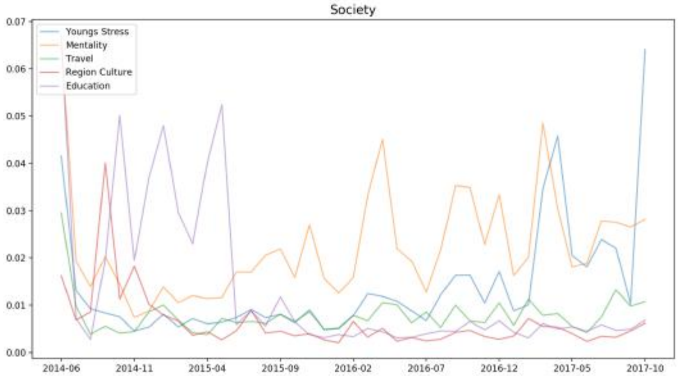

***Figure 5.9: Trend of Subclass in Society Topic***

## 6 Diagnosis Index of Topics

MALLET toolkit provides some useful diagnosis measure that could help us understand how each class of the topic is presented in the articles. Based on the standard of finding attributes that could visually separate the main classes in 2-D plot, I choose four of the diagnosis and plot them in scatter plot.

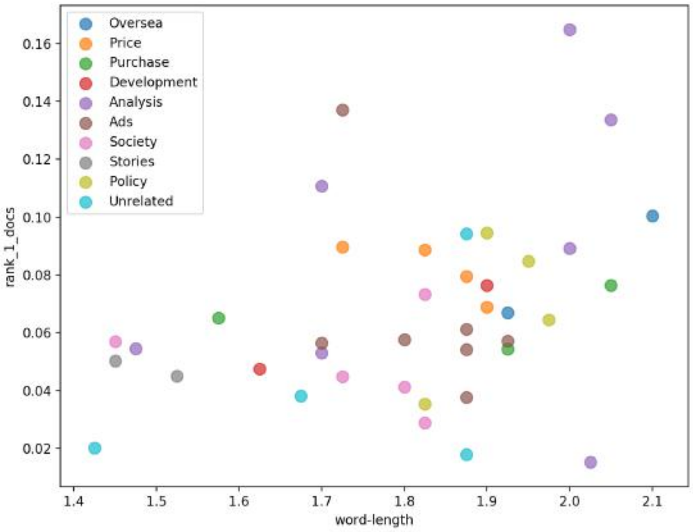

***Figure 6.1: Rank1 vs. Word Length plot***

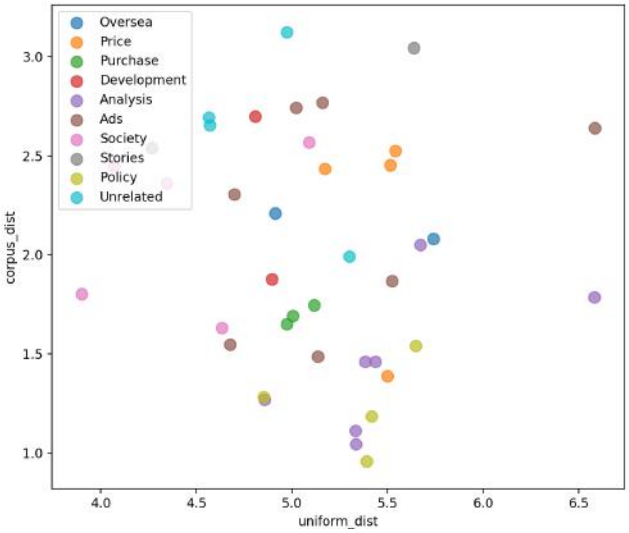

***Figure 6.2: Corpus Distance vs. Uniform Distance plot***

From Figure 6.1 we can see ‘Analysis’, ‘Price’, ‘Policy’ have relative high rank 1 documents value than others which mean are more specified and concentrated in the local document. ‘Price’, ‘Ads’, ‘Policy’, ‘Society’ have the most similar word length which might possibly indicate that these four topics may use a similar collection of specified words within each class since we already remove the words in stop lists. From Figure 6.2, uniform distance indicates a similar conclusion to rank 1 documents. Corpus distance measures how far a topic is from the overall distribution of words in the corpus. This means ‘Ads’, ‘Price’, ‘Unrelated’ topics are more different from the general corpus.

## 7 Discussion

From this project, I found that the topic trends on social media reflect amazingly similar trends in the market data especially in terms of event correlation and time series pattern. To better utilize the topic trend in helping understand the real world condition, I might need to dig into the time series analysis and social media study the find the reason behind recurring pattern in the topic trends. I will also use sentiment analysis to get a rough image on positive and negative market expectation from social media and compared with market data and topics to discover what classes of topics are more likely to increase during the positive period or the negative period.