Islamic, Asian and Latin Worlds in the New York Times, 1987-2006

Yanning Cui & Jing Zhang & Rodrigo Valdes

We study the changes in rhetoric, perceptions, and topics about three different concepts, broadly defined as Asian, Islamic, and Latin American worlds. We are particularly interested in economy, security, politics, and immigration. To do this, we get the full text of New York Times from 1987 to 2006, and we filter the full text for each region, based on keywords (see Appendix) defined for each one.

Filter the text by each area give us three different corpora of data, each one based on all the articles in the period, and extracted with a standard methodology. The keywords are selected by two criteria. First, a list of nouns including the name of all the countries in that region, and the name of the ethnic group (e.g. Iraq, Asian, Latino). Second, adjectives used to describe people who live there (e.g. Chinese, Korean, Mexican) and the ethnic group (e.g. Islamic).

We employ word frequency, part-of-speech(POS) tagging, topic modelling and word embedding to analyze our corpus. What kind of issues does New York Times focus about those regions? What kind of change happened in New York Times' rhetoric of those regions? What kind of socio-economic changes about those three societies are reflected by the text? The methods listed above help us to answer those questions.

The rest of the report is structure as follows. In section 1we discuss word frequency method. Then, we switch to section 2 and discuss of POS-tagging. In section 3, we talk about topid modelling and in section 4, we discuss about word embedding.

1.Frequency Method

Part 1: Motivation

The most immediate way to detect sequences of discrete events is through their change in frequency.
Although rough, word frequency can help us capture the shift of focus and perceptions of NYT York Times toward certain regions. For instance, increase in certain words can show the spotlight of the U.S. society on certain affairs, and vice versa. Meanwhile, change in certain words can even track shift in perceptions of the U.S. media (NYT in our case) towards other societies.

To capture the most salient change in words' usage, which might be related to changing situations, we adopt Kulkarni et al.'s (2014) method for detecting statistically significant change of word frequency.

Part 2: Change Point Detection Method

The change point detecting method by Kulkarni et al. (2014) [1]analyzes the times series data based on the Mean Shift model. The model detects a shift in the mean of the time series data by using a variant of mean shift algorithms. Basically, the algorithm calculates the mean and variance across all words, transforms the times series of words to the time series of Z-scores, use bootstrapping to estimate the statistical significance of mean shift at time point *j*, and finally estimate the change point by considering the time point *j* with the minimum *p*-value score (Kulkarni et al., 2014).

Procedures:

We get the final visualization of time series change through following step:

1. Filter corpus for articles related to the specific topic (Asian, Islamic, or Latin American) with a certain **occurrence of keywords** (no less than 1 here).
2. Tokenize the times series for each word.
3. Specify a set of words that we want to detect. We built a set of 96 words that related to economic, political, or safety issues, such as growth, democratization, and terrorism.
4. Fed the time series of tokenized words and word to be detected to the change point detection algorithm.
5. Calculated conditional probabilities of words (to be plotted on the y-axis) with statistically significant change detected and group by year (to be plotted on the x-axis).

Conditional probability of word A in 1997 = sum of frequency of word A in 1997 / total number of all words in filtered sample in 1997 (this is why I call it conditional probability; it is conditional on the sample).

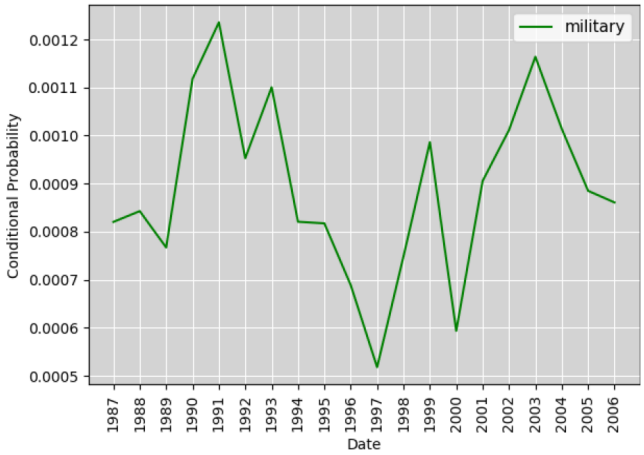
Discussion of results:

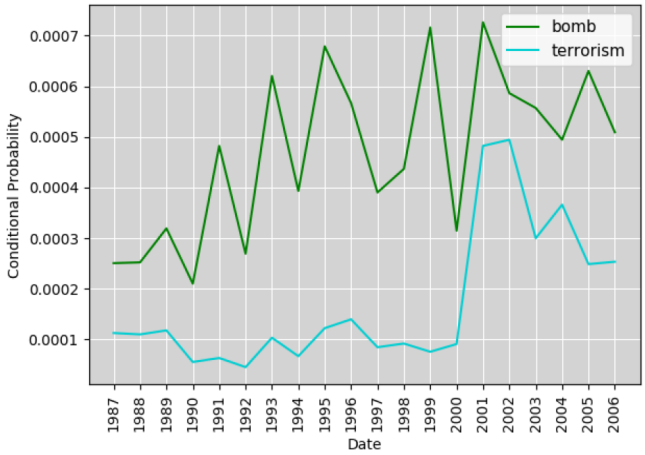
No significant change was detected for most of the words. Also, different words became significant in different samples, which tell some interesting story. The plot is shown as below:

First, the word frequency reflects the occurrence of certain events. As show in the figure below, the word frequency of "military" correspondents to wars happened in the Islamic world. The first peak of the left figure reflects the Gulf War that involving Kuwait, U.S., UK, France, Saudi Arabia, and Iraq in 1990 and 1991. The second big peak of the left figure around 2003 indicating a sharp increase mentioning of "military" during the Iraq War. The blue line in the right graph clearly shows a spotlight of terrorism in the Islamic corpus after 9·11 in 2001.

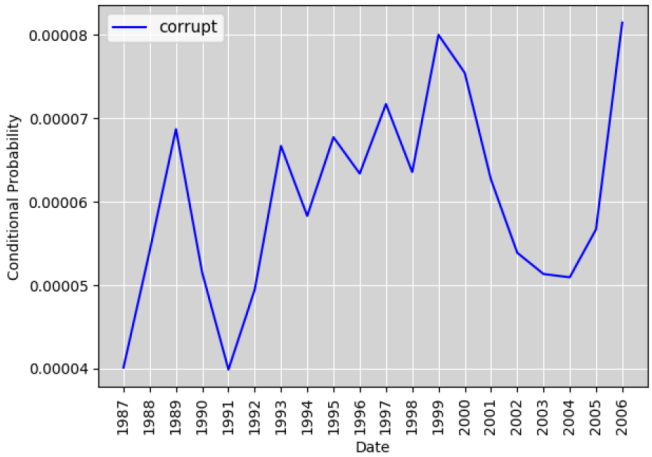
[1] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. CoRR, abs/1411.3315, 2014.

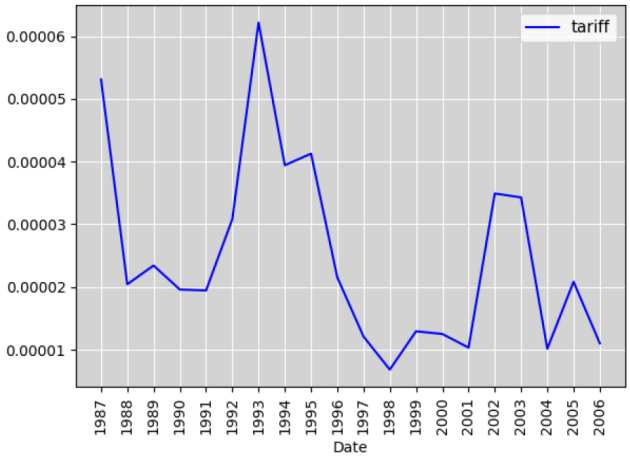
Significant Word Frequency Change in Islamic Corpus





Significant Word Frequency Change in Asian Corpus

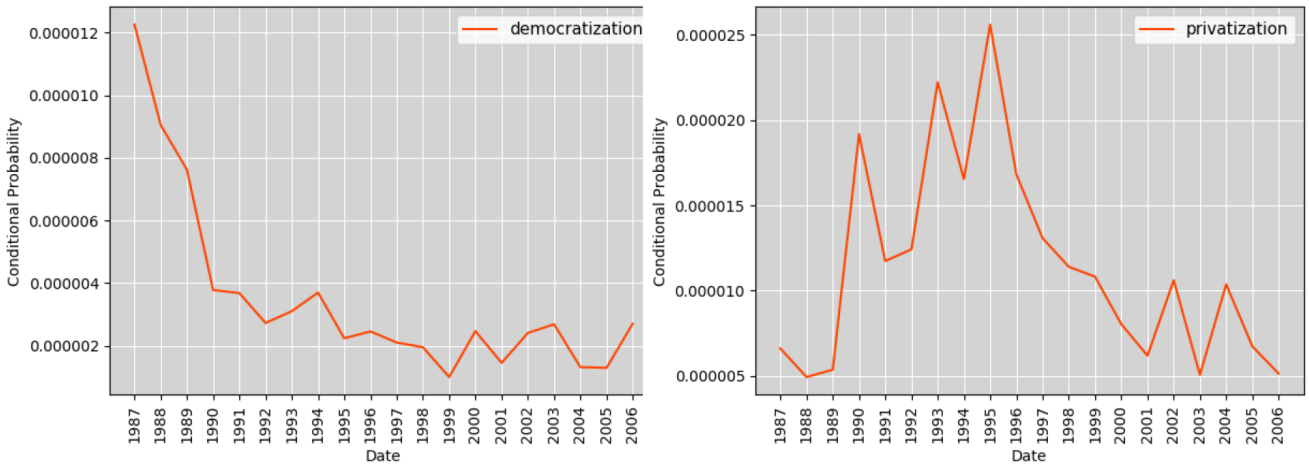




Statistically significant change in word frequency in Asian corpus also reflects events there. For instance, the small peak of word "corrupt" in 1999 and 2000 reflects a series of corrupt cases in China and Korea (those cases are not so astonishing, maybe this is why the peak is not so distinguishable). Going back to the corpus, we found that in1999, New York Times reported a series violent protests in China due to corruption of the local Communist Party officials in China. And in 2000, it reports the embezzlement and bribery of Ji Shengde, a former major-general in charge of military intelligence in the People's Liberation Army of China. The rumor that the wife of the Beijing's top Communist Party official at that time, Jia Qinglin has involved in a huge smuggling scheme in Fujian. About the right plot, the peak of the conditional probability of "tariff" in 1993 correlated to the Asia-Pacific Economic Cooperation (APEC) in Seattle, Washington. The smaller peak of "tariff" in 2002 and 2003 is correlated to China's entry of WTO at the end of 2001.

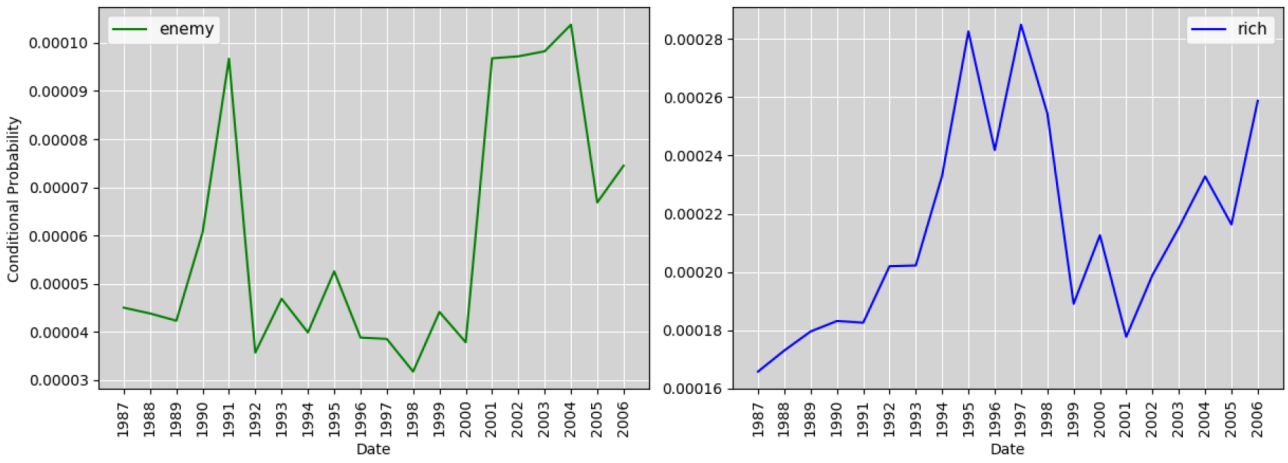
Statistically significant change of words frequency in Latin American's corpus also reflects the political-economic change in that continent.

Significant Word Frequency Change in Latin Corpus



The left plot above shows the time series of the word "democratization". Since the democratization process of the Latin America countries has finished the democratization process by the end of 1980s (e.g. Brazil in 1985-1988, Argentina in 1983), "democratization" was mentioned much less in articles related to Latin American Countries in 1990s and 2000s. The right plot of "privatization" with a peak in 1990s corresponds to the process of privatization in South America.

Besides tracking important historical issues, word frequency also reflects the change of perceptions of the U.S. society toward different countries and groups.



For instance, one interesting finding is that during the Gulf War and after 9·11, the frequency of word "enemy" increases sharply. It may indicate a kind of enmity between U.S. and the Islamic world. Also, we can find that the use of "enemy" decreased sharply after the end of Gulf war, but was still used widely after the 9·11. Maybe it reflects some difference between enmity brought by overseas military conflicts and terroristic attacks that cause death and injuries on American soil. The change of frequency of the word "rich" in the Asian corpus is also very interesting. We

can observe a sudden decrease in the frequency of the word after 1997, when the Asian Financial Crisis happened at the late 1997 and 1998.

2. POS Tagging and Named Entities

To classify tokens from the text, and understand some of their claims, we utilizes POS tagging and named entities recognition based on the software developed by the Stanford's NLP groups. The analysis comprises two parts. The first task is an identification of nouns, verbs, and adjectives related to the relevant group. We believe that this methodology helps to understand the social and economic relationships in the corpus due to the most frequent nouns, verbs, and adjectives provide useful insights of the jobs, events, and actions related with each of the groups. The second one task is to name organizations, locations, and persons connected with each of the groups. With those approaches, we can identify relevant futures linked with the recent political and economic history of the groups.

Part one: selection of the data

Due to computational limitations, we restrict our sample to four years during the analysed period, 1987, 1993, 1999, and 2006. For each year, we filtered the corpus in three different samples, one by each group. We selected the samples by considering only the articles with at least 25 mentions of words related with each of the tree groups, using the same list of keywords of the other sections (e.g., in the case of Latinos: Mexican, Mexican, Cuban, Dominican, Hispanic, among others). Additionally, from that selection, we used a random sample of 80 articles by year. Then, for each group, the number of articles is 320 (80 by year), and the total for the sections is 960 (three groups).

Part two: tagging words by their part of speech (POS)

Procedures:

1. We compute the number of occurrences of each noun inside the sample for each group, by year.
2. This number was divided by the total number of words, by sample and year. This step generates the conditional probability (conditional to the sample) of occurrence of each noun by year.
3. Then, we identify the most popular nouns during the fourth years, adding the conditional probabilities by each year and sort them by relevance.
4. From the first fifty nouns with the highest accumulated probability, we select to graph those that can help to understand the corpus. We avoid common nouns, such as today, day, month, place, and use those which talk about politics, economics, security, or immigration.
5. Finally, we repeat this process for verbs, and adjectives.

Part three: tagging words as named entities (NER)

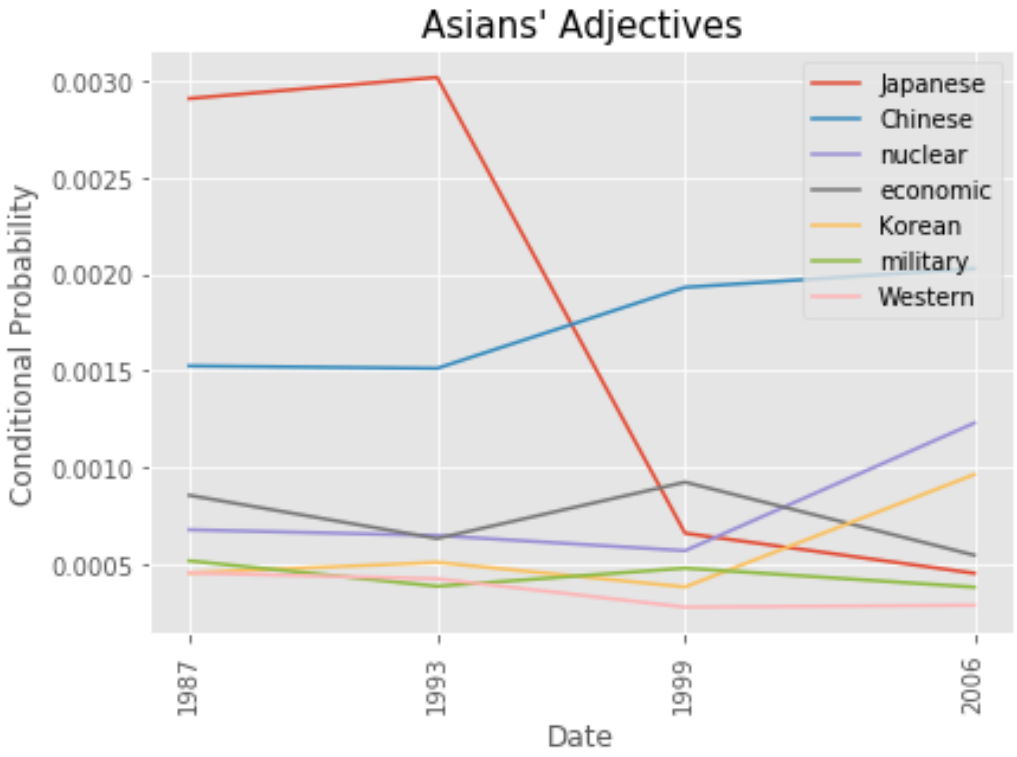
The process for analysing organizations, locations and persons follow a similar process as the described above for POS tagging.

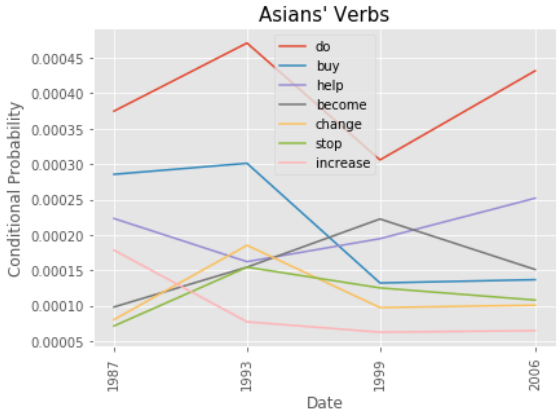
Procedures:

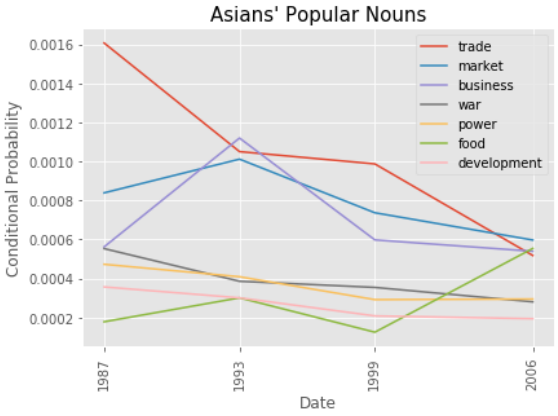
1. Run a program for named entities recognition (NER) which gives tags for each word: organization, location, person, or other.
2. Detect the most common items adding the number of times they occur.
3. Obtain the conditional probability of them, by dividing the occurrences by the total number of words in each year sample.
4. After this process, we selected the fifty more relevant entities by kind (organization, location, or person) according to their accumulative conditional probability. From those, we chose those entities which represent the general trends in the corpus. For instance, if there are many words related to trade, such as exports, imports, and commerce. We used for the graphs at least one word that can depict this trend. We avoid trivial topics, as Asian food, and focus on politics, economics, security, and immigration.
5. In the case of persons, due to the conditional probability of relevant individuals can emerge quickly we use bar plots by group and year. For instance, individuals not in the 1987 sample are relevant in 2006.

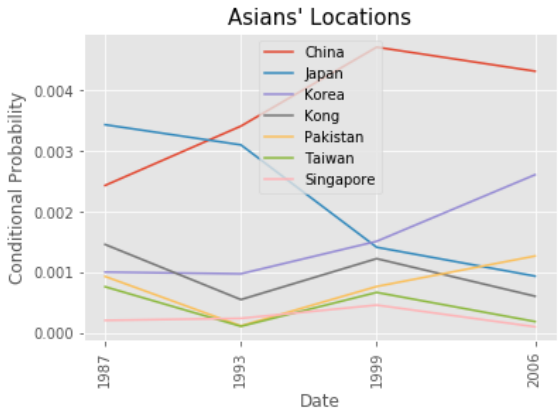
Discussion of results:

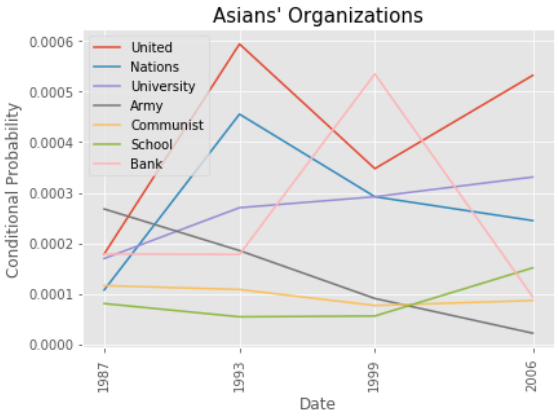
First, we will discuss some of the most relevant findings related with Asia. The most striking view is the relative importance of China growing steadily during the nineties. For locations and adjectives, China became more relevant that Japan during the nineties, a trend that continues until 2006. It is interesting to see that also the organizations school and university increases its relevance during the nineties and the first years of the twenty-one century, suggesting that there were other transformative changes in Asia besides the growing of China.


Asians' Adjectives
Conditional Probability
Date
Japanese
Chinese
nuclear
economic
Korean
military
Western
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
1987
1993
1999
2006

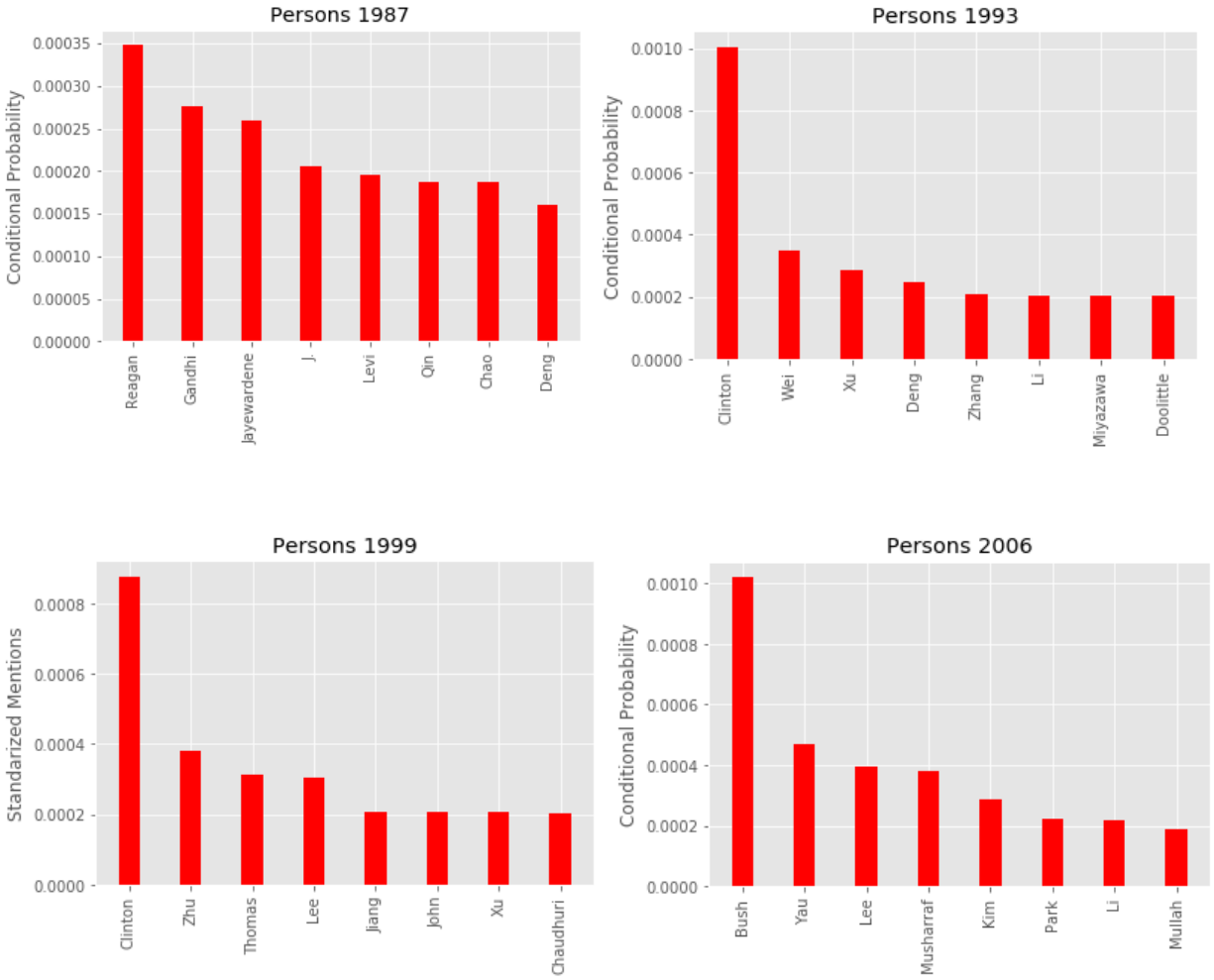

Asians' Verbs
Conditional Probability
Date
do
buy
help
become
change
stop
increase
1987
1993
1999
2006


Asians' Popular Nouns
Conditional Probability
Date
trade
market
business
war
power
food
development
1987
1993
1999
2006

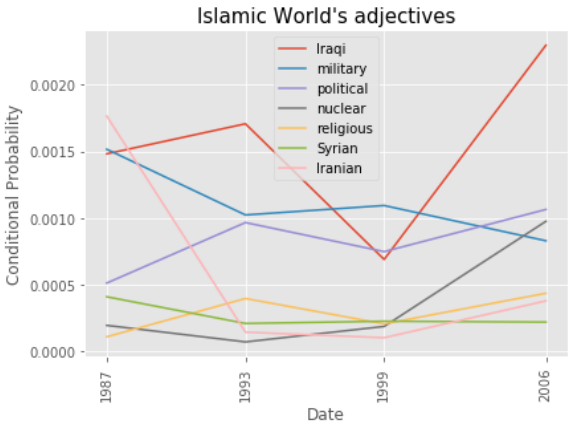

Asians' Locations
Conditional Probability
Date
China
Japan
Korea
Kong
Pakistan
Taiwan
Singapore
1987
1993
1999
2006

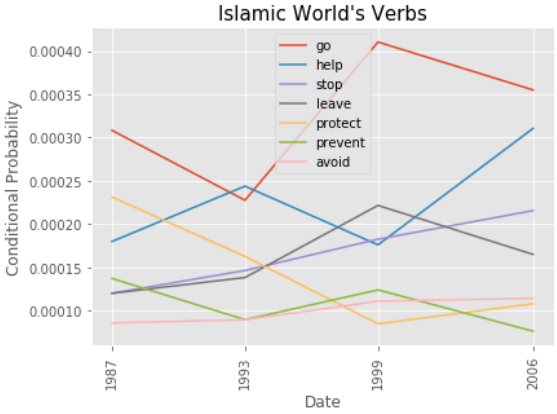

Asians' Organizations
Conditional Probability
Date
United
Nations
University
Army
Communist
School
Bank
1987
1993
1999
2006

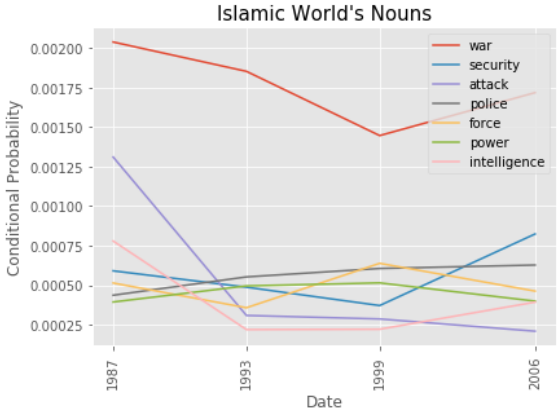
Regarding the persons, it is also possible to detect the augment in the conditional probability of Chinese leaders over Japanse ones. For instance, in 1993 the leader of the Communist Party of China, Deng Xiaoping has a similar relevance as the Prime Minister of Japan, Kiichi Miyazawa. However, for 1999, the Premier of China, Zhu Rongji, was the second in importance, just after the president of the United States, while the Japanese Primer Minister is not even in the most twenty mentioned. In the same year, Jiang Zemin, the General Secretary of the Communist Party of China had a higher conditional probability that any Japanese leader.

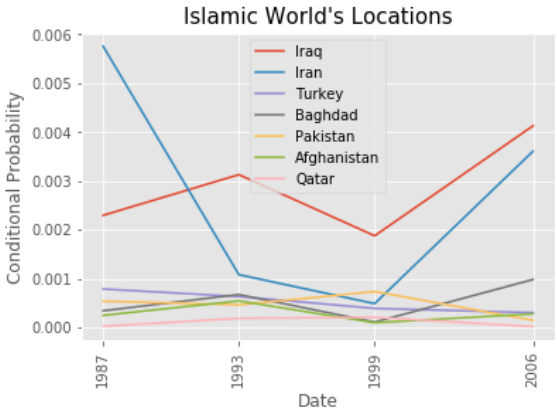


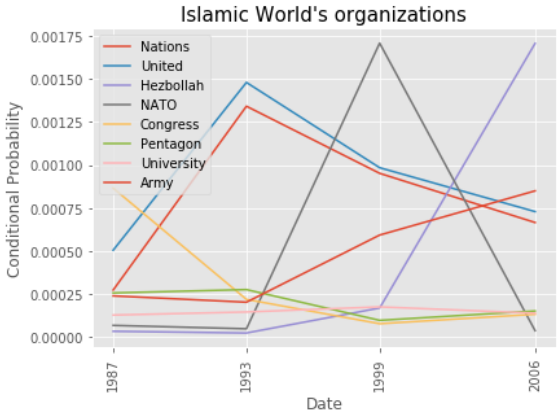
Second, for the Islamic World, the graphs regarding nouns, adjectives, and verbs show the evolution of the relative importance of Iraq through the years. In 1993, Iraqi had higher relative importance than in 1987, most probably as a result of the Gulf War. Moreover, it had the greatest conditional probability in 2006, after the invasion of Iraq in 2003. Another feature depicted by the graphs is the relevance of the nuclear program of Iran during the first year of the twenty-one century.



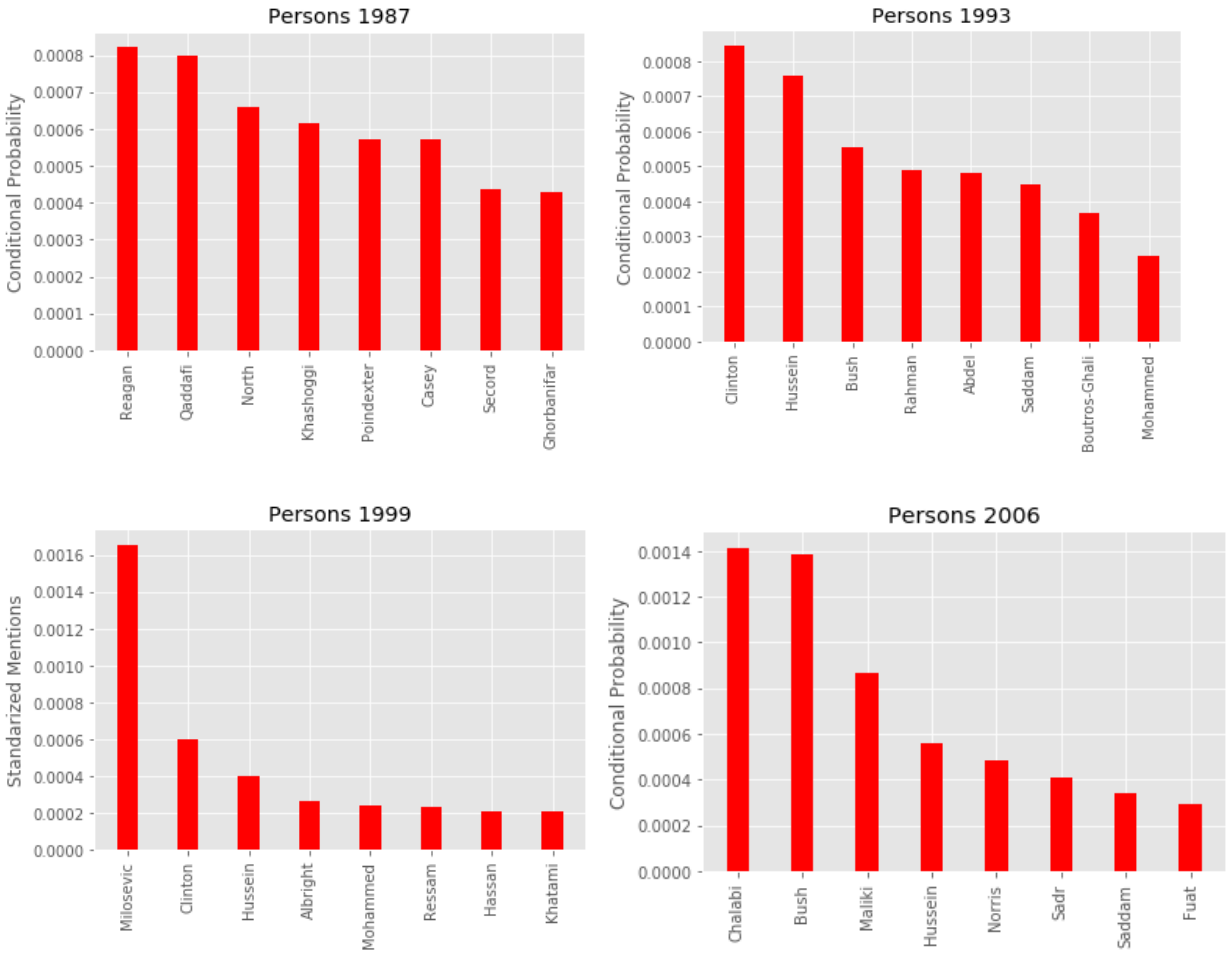






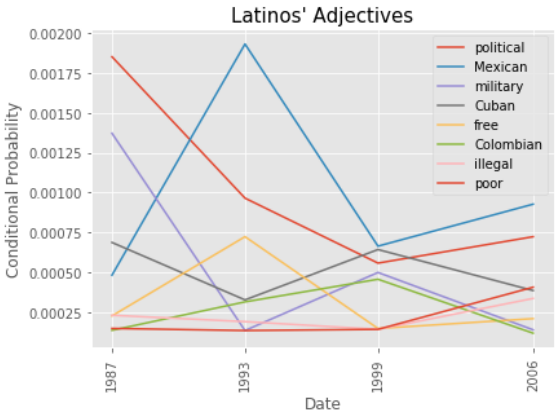


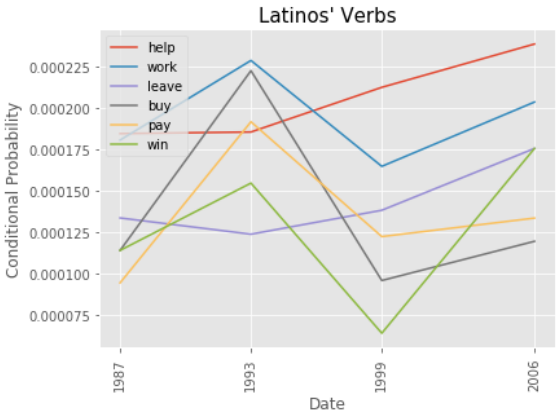
The most relevant persons through the years are the Irani and Iraqi political leaders, such as Ghorbanifar (a middleman in the Iran–Contra Affair during the Ronald Reagan presidency) in 1987, Sadam Hussein in 1993 and 1999, and Chalabi (Iraqi politician) in 2006. In addition, in 1987, the most notable names were from other regions, the most relevant one was Gaddafi, a Lybian leader. Meanwhile, other prominent names were Khashoggi, a Saudi Arabian businessman.

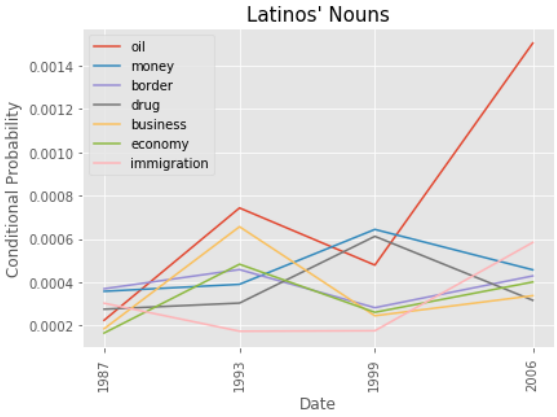


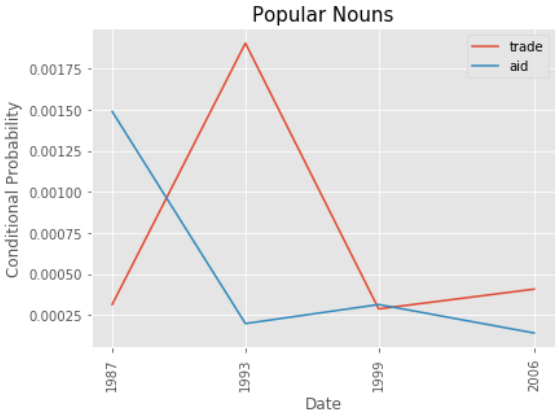
Finally, for the Latin American World, it is interesting to see that there is a climb in the graphs for nouns, verbs, and adjectives in the year 1993. Some of the prominent words are business, money, economy, work, free, and Mexican. That year was the negotiations of the North American Free Trade Agreement (NAFTA), which causes controversy. Also, the relative importance of Cuba decreases during that time.

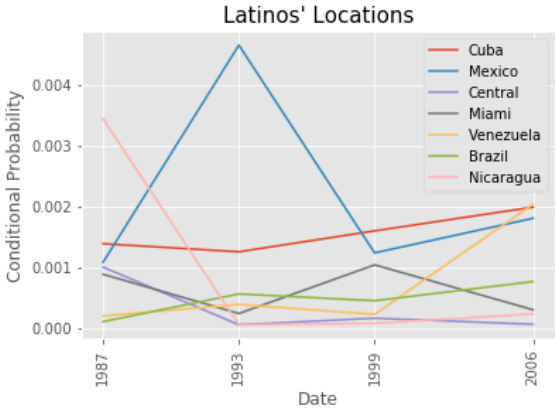
Furthermore, for 2006, there is a dramatic increase in the word oil and the verb help, which suggest the success of Hugo Chavez in Venezuela, supported by oil revenues. Moreover, the graph on locations indicates that Venezuela increases its relative importance for 2006, such as Mexico, which has a highly controverted presidential election that year, which also explain the word PRI (Institutional Revolutionary Party), one of the biggest political parties in Mexico.

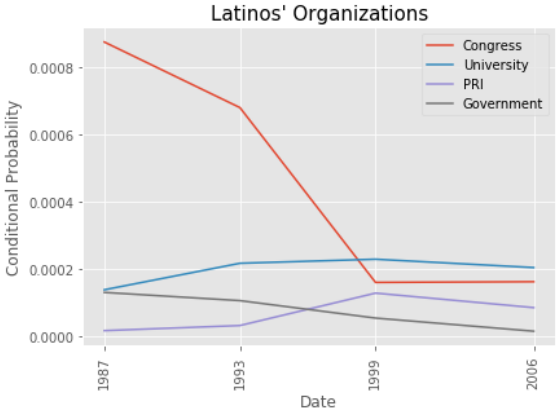




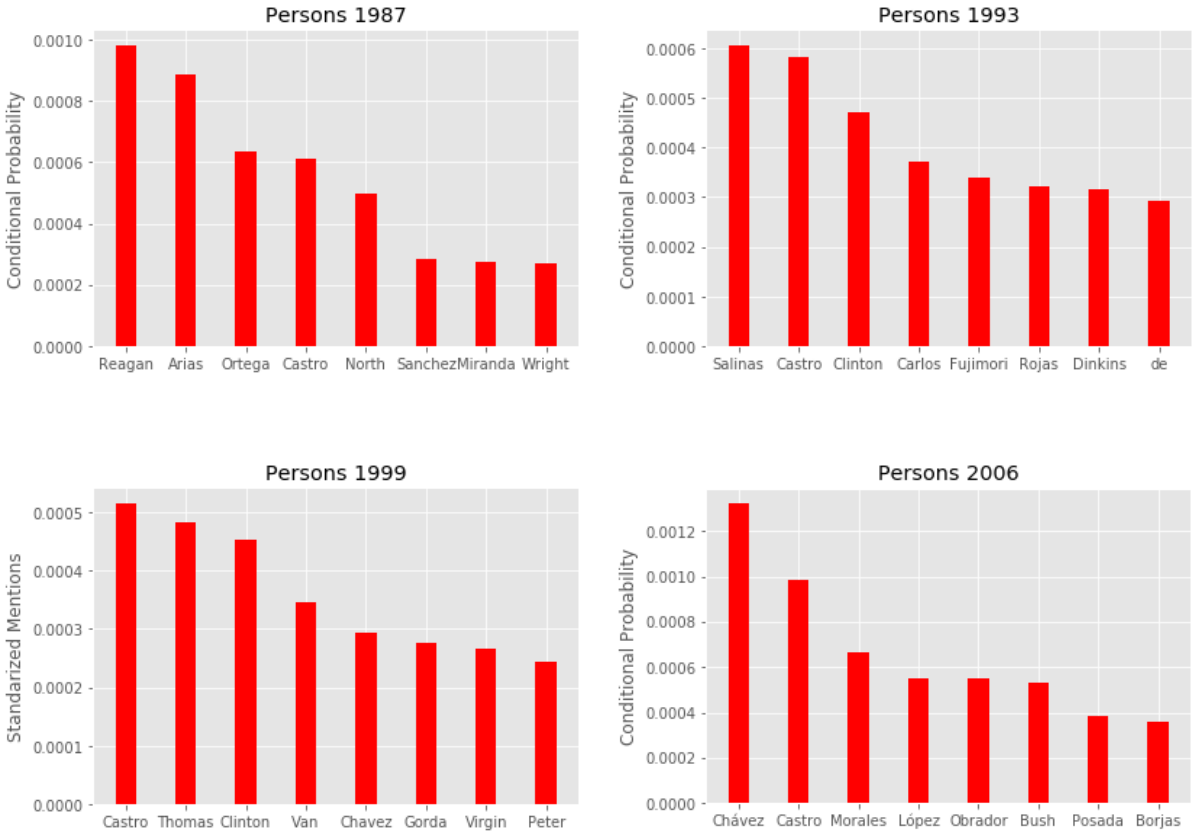








Regarding the persons, there are similar trends. In 1987, the most notable names for the region where Reagan and Arias, which received the Nobel Peace Prize that year. Then, in 1993, it turned to Salinas, the president of Mexico during the negotiations of the NAFTA; Castro for 1999; and finally, Chávez in 2006, followed by Lopez Obrador, one presidential candidate in Mexico during 2006.



Discussion and Limitations

As discussed in the topic modelling section, the trends that we were able to track in Asia are, in general, more positive than those from the Islamic and Latin American worlds. For instance, the relevant adjectives in Asia are related to the economy and trade, war and peace for the Islamic world, and drugs, immigration, and economy for Latin America.

The perception of the regions that we are analysing must be related to the information that we find here. For instance, one of the extensions for this analysis is doing a survey in Amazon Mechanical Turk about the nouns, verbs, and adjectives related with each of our worlds. Then, we can understand if the trends that we depict with the NYT's corpus are representative of the population that read this newspaper, or more broadly, the American population.

3. Topic Modeling

For topic modeling, we experiment with two different models: Non-negative Matrix Factorization(NMF) and Latent Dirichlet Allocation(LDA).

Part 1: NMF topic modeling

Procedures:
For NMF topic modeling, we get the final visualization of time series change through the following steps:

1. Filter corpus for articles related to the specific topic (Asian, Islamic, or Hispanic) with different **occurrence of keywords**
2. Set the number of topics, the number of features and start extracting topics
3. Test with different number of topics and features, until all topics can be clearly decoded
4. Assign names to every topic based on the list of top terms in that topic and our experience (we need to validate if topic names are appropriate by conducting some research on the history of a region and underlying context of its culture, economy, politics, etc.)
5. Calculate values of every topic (to be plotted on y-axis) and group by year (to be plotted on x-axis), for example:
   method 1: theoretically it is more accurate but the results don't make sense

value of topic A in 1997 = sum of weights of topic A in 1997 / sum of weights of all topics in 1997

   method 2: theoretically it cannot fully represent the actual topic loadings but the results make much sense

value of topic A in 1997 = number of articles that mention topic A in 1997

We shared results using both methods. Our discussion is based on the second method because it makes more sense of our data.

6. Put related topics (i.e. all topics related to China) in a grand topic (i.e. China) for visualization purpose because it will be too crowded to plot 20 or 30 lines on a single graph
7. Visualize topic changes by year (garbage topics will not be plotted)
8. Interpret and discuss some interesting results (significant change, peak of a curve)

At first, we set the occurrence of keywords to be 2-3, and it returns around 70,000 articles related to Asia from 1987-2005. However, there are overlaps in topics that we cannot fully interpret. Then we set the occurrence of keywords to be 10, and it returned around 20,000 articles. The topics become more distinct and easy to decode.

To extract the topics of articles, we use tf-idf (term frequency-inverse document frequency) to transform each article into a word vector and fit NMF model. Tf-idf is a statistic for each word in

an article that increases with the number of times a word appears in that article while being offset by how frequently that word appears across the entire set of documents. Using tfidfvectorizer from Scikit-learn to transform the corpus of documents[2], we extract additive models of the topic structure of the corpus. The output is a list of topics, each represented as a list of terms with weights. Tf-idf is useful when all your documents are about the same broad subject (i.e. Asian or Islamic) and likely to have a lot of words in common.

We also experiment with different number of topics and find that 30 is the best fit for Asian corpus, 20 to be the best fit for Islamic and Hispanic corpus. However, there are always mistakes when interpreting topics based on our limited experience and knowledge. Since topics don't come with labels, we have to decode them according to the top terms in every topic. And sometimes we find that there are multiple topics that might be about fundamentally the same thing (overlaps in top terms). Also, not all topics can be fully decoded (garbage topics include top terms like *"mr said lee years man minister told did wife later kim president family interview"*). And we will exclude topics related to Arts, Sports, or Food that we don't care.

For calculating values that can represent topic loadings, we try 3 different methods. First, we use the number of articles in Topic A to represent the topic loading, but we think it overlooks the actual weight of each topic extracted from the model. Then we try adding up weights of Topic A in each document to represent overall weight of Topic A over years. But it ignores the fact that topics are not discussed averagely every year. Then, we use:

value of topic A in 1997 = sum of weights of topic A in 1997 / sum of weights of all topics in 1997

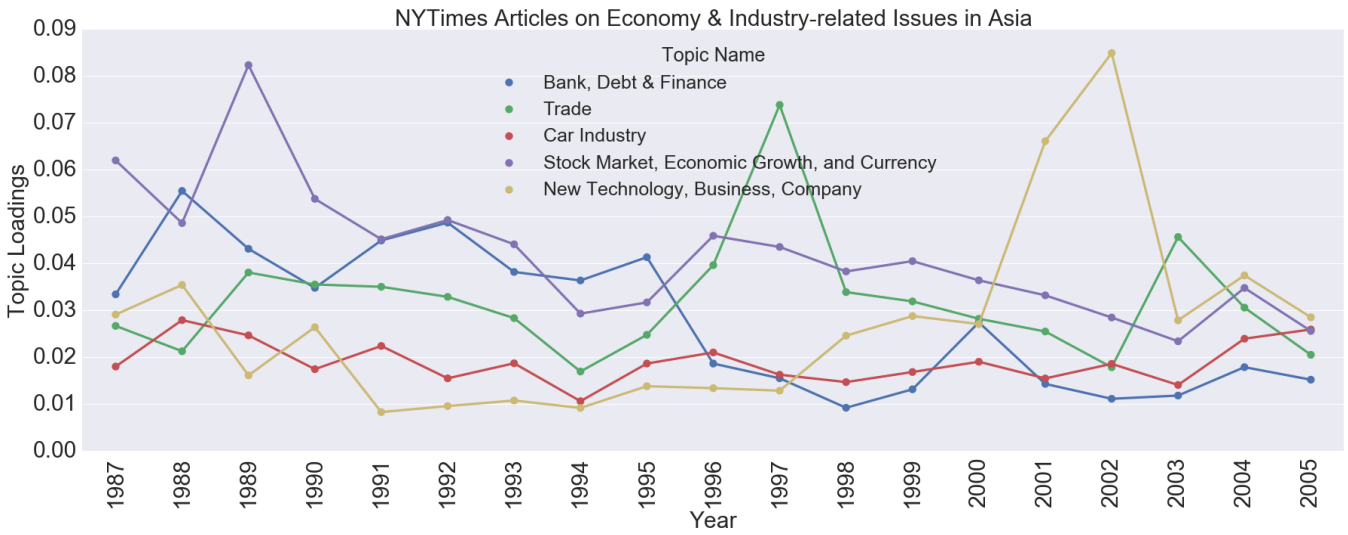
to represent the topic loadings over years. However, it is still not performing as we expect because some trends can hardly be explained. Finally we switch back to the first approach, which is counting the number of articles (instead of topic loadings) that mention a topic (i.e. 553 articles in our corpus mentions Topic A, so the value for Topic A is 553)

The last step is to visualize our topic scores by year and interpret the results.

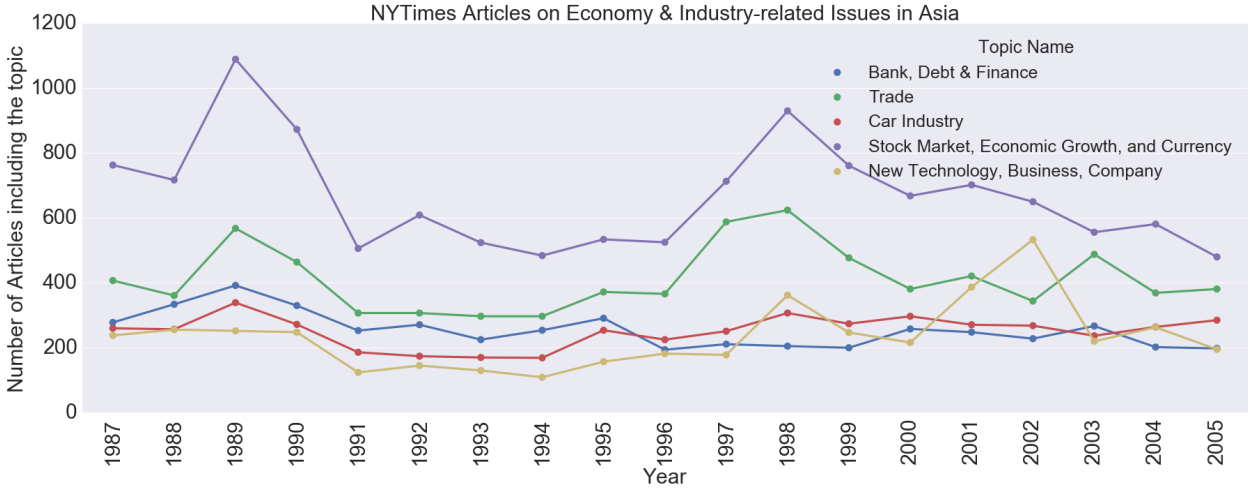
Discussion of results:

Here we include some interesting results for each region.

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.



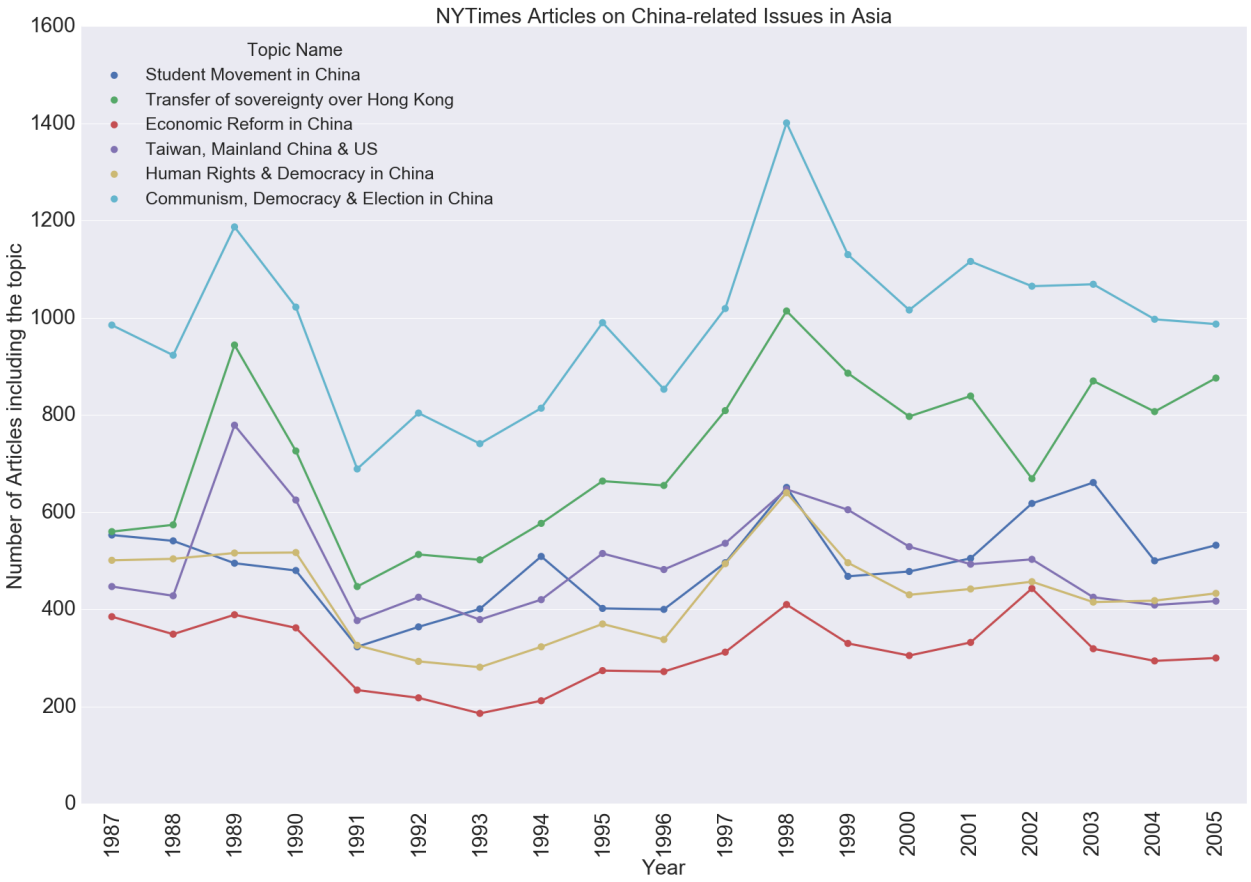
This figure uses method 1 to calculate topic value, which doesn't make much sense!



The figure above uses method 2 to calculate topic value, which makes more sense.

1. The purple line which is related to Stock Market and Economy *(top terms: percent yen dollar market stock markets economy currency investors rates growth stocks prices year rate economic exchange economists said quarter)* has a sudden peak in 1989. This might be related to The Black Friday: the 13th mini-crash of stock market that occurred on Friday, October 13, 1989. The peak of the purple line around 1997 and 1998 coincides with Asian financial crisis, which featured by currency devaluation.

2. The sudden peak in 2000-2003 represented by the yellow line explains Information Technology Bubble at the beginning of 21st century. *(top terms: company companies business said industry market technology million computer corporation percent american billion executives sales products new investment year research)* IT bubble witnessed stock markets having their equity value rise rapidly from growth in the Internet sector and related fields, and then the collapse of many IT companies.

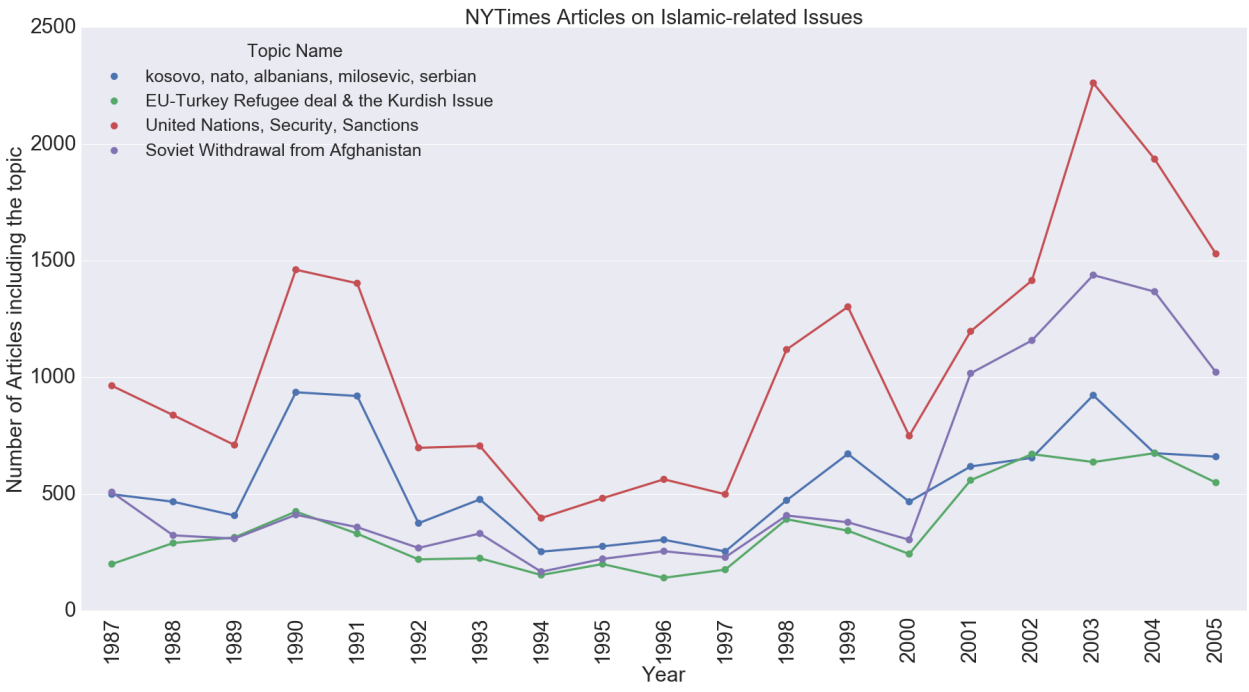


As we can see from the figure above, regarding China, NYT seems to be particularly focused on human rights, democracy, election, Hong Kong and Taiwan. For example:

1. The peak in blue line from coincides with two change of state leadership in China. In 1989, Jiang Zemin became the chairman of Central Military Commission as well as General Secretary of the Communist Party of China. In 1989. In 1998, Jiang Zemin was reelected as President and Zhu Rongji was appointed as the prime minister. Also, in 1989, China experienced a tide of student movements that ask for democracy. It seems that the New York Times cares more about the democratic appealing of the movement, rather than the movement itself.

2. The green line coincides with discussion of the sovereignty transfer of Hong Kong in 1980s and the transfer in 1997.

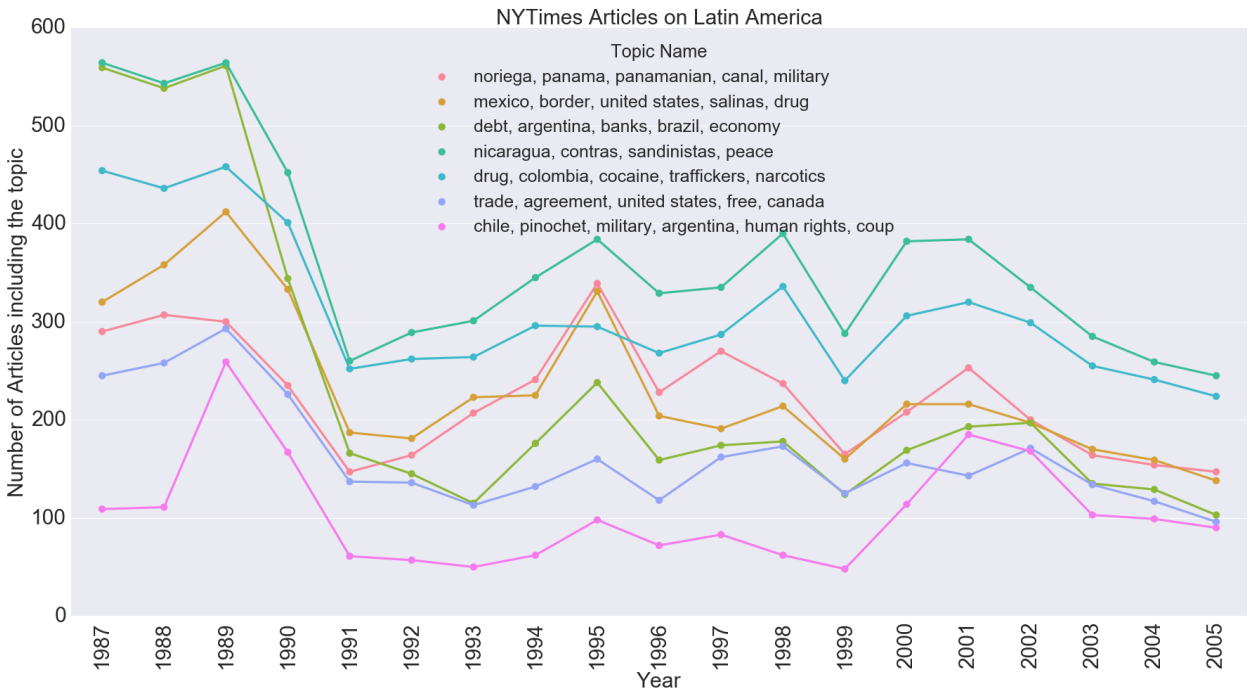
3. Top terms in the topic that related to Taiwan issues are: *taiwan mainland chen lee independence china beijing island relations military president united states diplomatic election washington missiles missile policy sides* (the purple line). We can find that US is concerned with Taiwan's independency from mainland China, and NYT also mentions military force (missile), which is quite interesting.



Looking at the purple line which represents the Soviet–US-Afghan *War (top terms: soviet moscow afghan union afghanistan kabul withdrawal guerrillas russian guerrilla government kuwait pakistan russia troops rebels gulf republic war state)* There is a peak in 2001, which could be explained by 2001 United States invasion of Afghanistan.

A sudden small peak in the blue line in 1998-1999 can be explained by the fact that NATO has been leading a peace-support operation in Kosovo since June 1999 in support of wider international efforts to build peace and stability in the area.[3]

[3] http://www.nato.int/cps/en/natolive/topics_48818.htm



The figure above depicts some of the important events in Latin American during the analyzed period.

1. The green line apparently shows some of the narrative of the debt crisis during the eighties in Latin America. It also can explain its small relevance after 1990.

2. It is interesting that the topic related to North America are divided in two, one related with Mexico, the NAFTA, and drugs, represented by the orange line. And the other one in the purple line (*terms: trade, agreement, free, Canada*) which clearly reflects than besides the physical proximity with both countries, the international relations focus in different topics.

3. The pink line represents South America, in special Chile and Argentina. It is possible to see that the relevance of this topic augment in parallel with the competitive election in Chile during 1999 – 2000.

4. The two green topics reflect the issues of Central America, as the channel of Panama, the sandinistas conflict, and some other regional wars.

Part 2: LDA topic modeling

Procedures:

For LDA topic modeling, we get the interactive map through the following steps:

1. Filter corpus for articles related to the specific topic (Asian, Islamic, or Hispanic) with different occurrence of keywords
2. Convert the raw documents into document-term matrix with TfidfVectorizer

3. Set the number of topics and fit Latent Dirichlet Allocation models
4. Test with different number of topics and features, until all topics can be clearly decoded
5. Visualzing the models with pyLDAvis
6. Interpret some interesting results

pyLDAvis is a useful package that helps users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization.[4] **The reason why I use this package is because it can visualize how prevalent each topic is and how topics are related to each other.**

The left panel presents a global view of the topic model. The topics as circles are plotted in the two-dimensional plane whose centers are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions.[5]

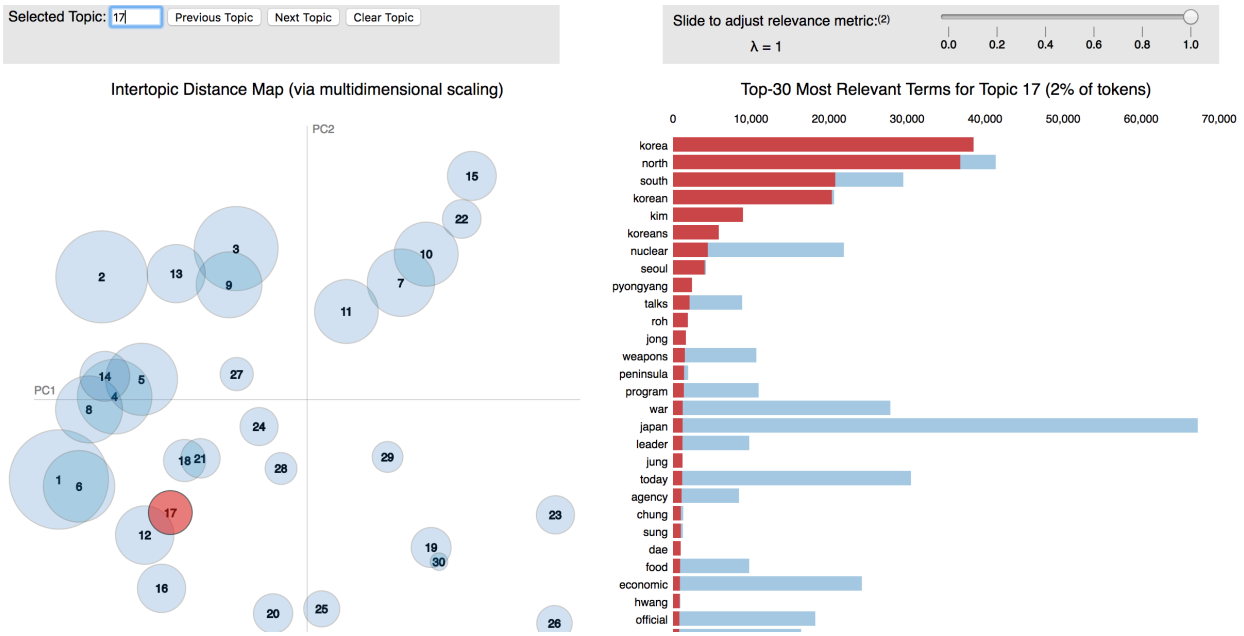
The right panel depicts a horizontal bar chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left. A pair of overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term.

The left and right panels of our visualization are linked such that selecting a topic (on the left) reveals the most useful terms (on the right) for interpreting the selected topic. In addition, selecting a term (on the right) reveals the conditional distribution over topics (on the left) for the selected term.

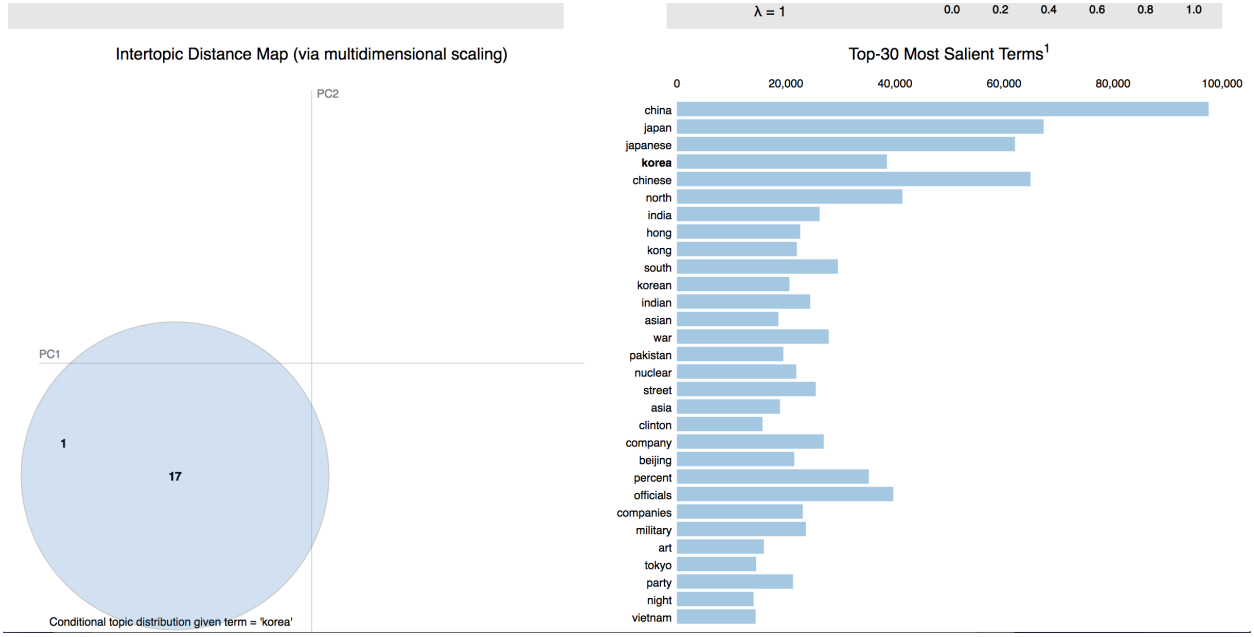
[4] https://github.com/bmabey/pyLDAvis
[5] http://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

(It is also important to know that the biggest circles (Topic 1 and Topic 2) are usually garbage topics.)

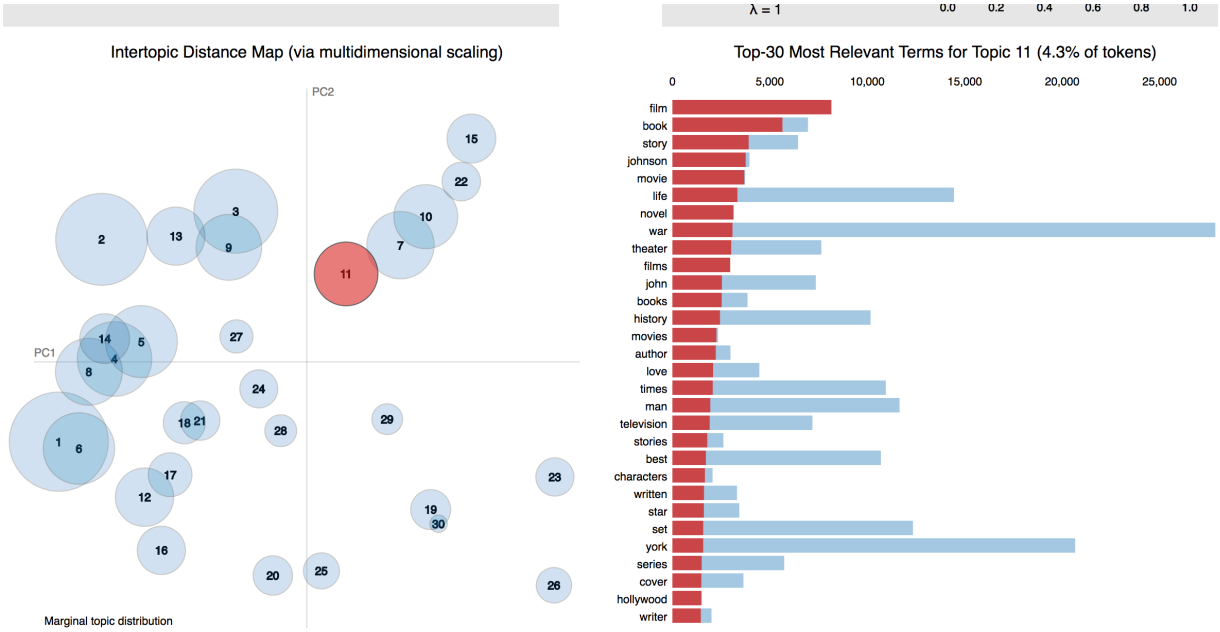
One example using Asian corpus:



For Asian corpus, when you select topic 17 on the left panel, top-30 most relevant terms in this topic will appear on the right panel. **We might detect that this topic is related to North-South Korea Relation and nuclear issues.**



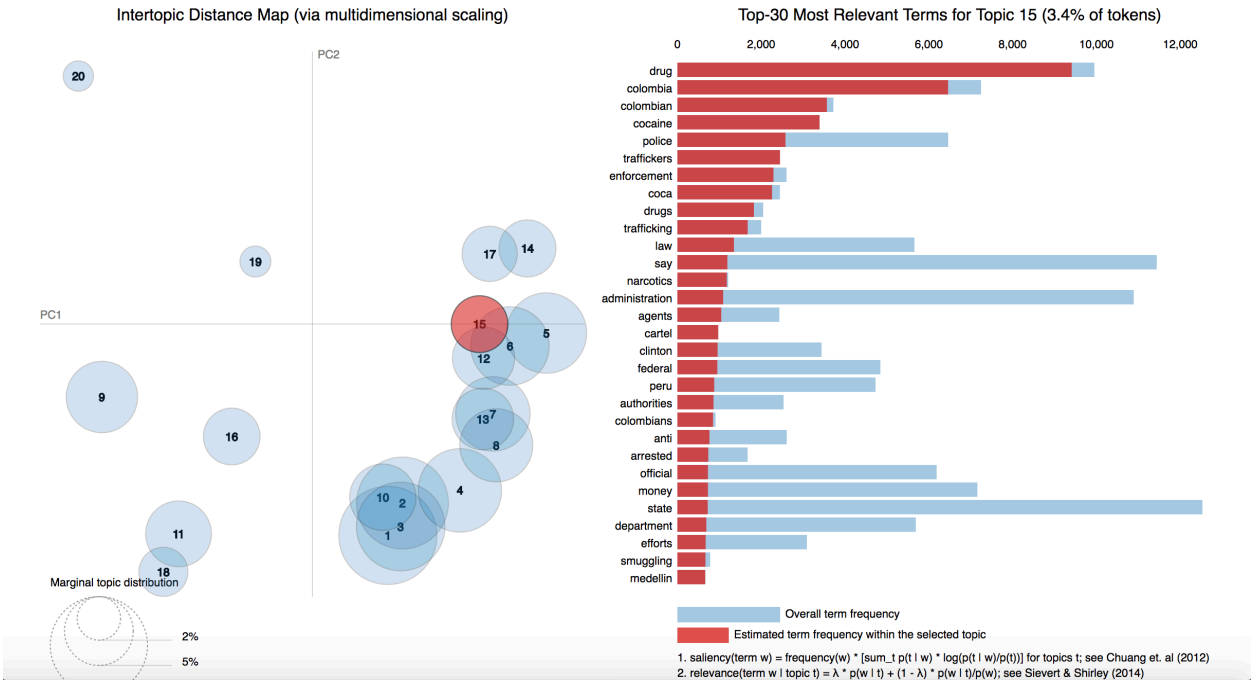
Then when you hover on the term ‘korea’ in the right panel (top-30 most salient terms in the whole Asian corpus), it will tell you which topics are the most relevant in terms of this token: **Topic 17, which is correct.**



Another interesting fact is that the closer these circles are, the more related their topics are. For example, Topic 11, 7, 10, 22, 15 are all related to food, arts, life, film, etc. You can play around this plot by yourself to confirm that, and you can also explore more about the corpus. [6]

Another interesting example from Hispanic corpus:

[6] Since it is an interactive plot, we cannot insert it in the final report. Please check html files in the submission (or topic modeling notebook) to review the plot.



'drug' appears to be one of the most salient terms within Hispanic corpus. Topic 15 might represent illegal drug trade in Colombia, since its top terms include police, enforcement, narcotics, and arrest.

Discussion and Limitations

From what we have explored so far, New York Times depicts Asian world focusing in business issues, since a lot of its topics are related to economic growth, economic reform, stock market, currency, industry, and trade, while Islamic and Hispanic corpus do not have this feature. And NYT is also concerned about democracy, human rights, election and social movement in China, Hong Kong, and Taiwan.

NYT tends to describe Islamic world as a place of terrorism, wars, insecurity and weapon. And it tends to focus on topics of drug and border conflicts in Latin America.

With regard to methodologies, LDA prefers to form topics from words that co-occur frequently. NMF learns more incoherent topics than LDA, but it can often create low quality topics from completely unrelated words[7], and it might be the cause of some weird trends in our NMF modeling visualization.

For example, the model puts these terms in one topic: *kong hong british britain 1997 territory mainland beijing people democracy singapore china asia rule chinese government residents executive hotel local*, which makes us assign this topic as 'Transfer of sovereignty over Hong Kong' because it mentions Hong Kong, Britain, Beijing, and 1997. And the green line in the first figure has a small peak in 1996-1997 (when Hong Kong is transferred back to China). However, we still cannot explain another two peaks in 1989 and 2002-2005. It is possible that the model

[7] http://aclweb.org/anthology/D/D12/D12-1087.pdf

puts unrelated words together and this 'Hong Kong topic' consists of several other topics that we do not detect, which probably means that we should extract more topics (i.e. set number of topics = 40).

We also wish to try SVD topic modeling (Singular Value Decomposition) in the future, to compare the performance of all three models. We also consider changing the number of topics, and the number of features, and changing the method we use to calculate 'value' of topic loadings to improve our classification.

4.Word Embedding

To explore the perceptions of New York Times toward Islamic, Asian and Latin American societies, we apply the word embedding method. Specifically, we project words vectors to an arbitrary semantic dimension and to see how semantic categories can help us understand those societies in the context of New York Times.

Method:
For word embedding, we use the genism implementation of Word2Vec to produce word vectors. The method specifies an underlying number of dimensions, and trains a model with a neural network auto-coder that best describes corpus words in their local linguistic context. Exploring their locations in the resulting space helps us to learn about the socio-economic positions of certain words in the corpus. We have not fully analyzed the entire corpus. Only results from 1987, 1997 and 2006 are shown.

Procedures:

1. Tokenize and Normalize our filtered sample. Tokenizing requires two steps. Word2Vec needs to retain the sentence structure so as to capture a "continuous bag of words (CBOW)" and all of the skip-grams within a word window. The algorithm tries to preserve the distances induced by one of these two local structures.
2. Load our data and give all the sentences to the trainer.
3. Calculate three dimensions: education (educated-uneducated), safety(peaceful-war) and class(wealthy-poor). Specify a list of words that we want to project. One thing to notice here is that we do not project the term "Latino" in 1987, since the word was officially adopted in 1997 by the US government.
4. Get the projections and plotting.

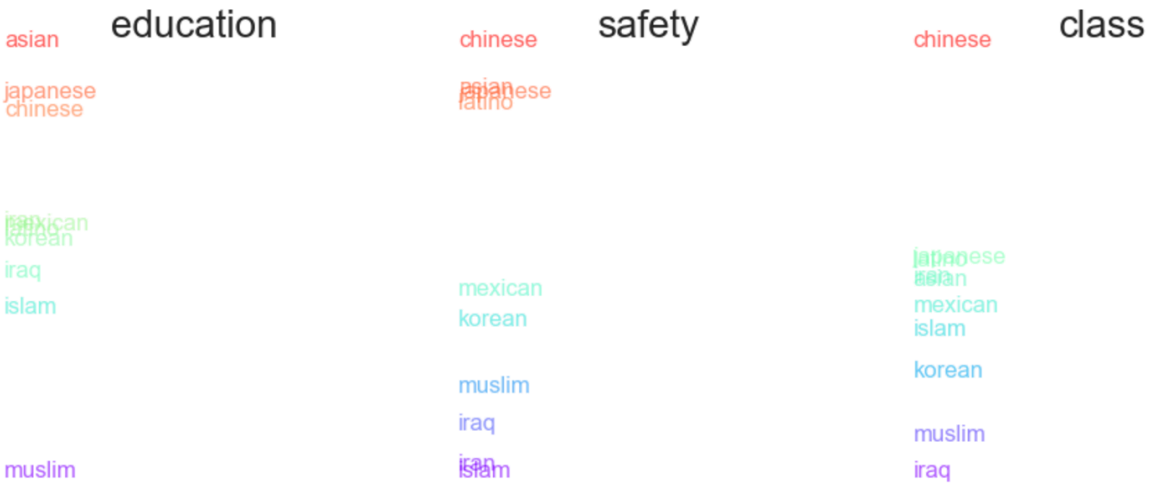
Discussion of Results:
Here we have some interesting results of each year.

Projection in 1987

In 1987, we have several results:

1. Asians, especially Japanese and Korean, are classified as more educated than other groups. While Islamic people in NYT corpus are closer to lower education.
2. Asia countries as well as Latin American countries are viewed as more safe and peaceful than Islamic countries.
3. Japanese is the wealthiest, while Hispanic people are viewed as relatively poor. Muslims are relatively wealthy. This reflect the facts that many Islamic countries are oil-rich.

Projection in 1997



In 1997, we find that

1. Asians are still perceived as relatively more educated than other groups.
2. China become more peaceful, while Korea, Iraq and Iran become less peaceful.
3. Chinese becomes more wealthy while Japanese and Korean become poorer. This change may due to the financial crisis happened in Asia in 1997, which hit Korean and Japan's economy heavily but has limited impact on China.

Projection in 2006

In 2006, the projection shows that

1. On the dimension of education, Japanese, Chinese and Asians still occupy the high end.
2. On the safety dimension, Islam is still related to more warfare and risk.
3. In the class dimension, Japanese becomes the wealthiest again. One astonishing result is about Korean, which becomes the poorest. Our guess is that this reflects a mixed result of North Korea and South Korea.

In sum, during the 20-year period (1987-2006). New York Times reports Asians as more educated, wealthier and more peaceful. While Latinos are at a relative middle position of educational level and economic status. And Islamic people are rich (but with a decreasing trend), poor-educated, and suffered from warfare.

5.Conclusion

The utilized methods are useful to detect some features that reflect the political and economic situation of the regions. For instance, it was possible to identify the Asian financial crisis with topic modelling, as weel as the independence of Taiwan, the United States - Afgan war, the issues related to drugs in Colombia, the Latin American debt crisis in the eighties, and the conflict between North and South Korea.

The word frequency and topic modeling method helps us to understand what kind of international issues and topics that the New York Times care about.

The tagging methods helped us to identify the big topics related to each region, such as economic growth in Asia, armed conflict in the Islamic World, and economic, drugs, and immigration issues in Latin America. One of the significant trends identified was the change in leadership in Asia, where Japan lost relevance during the nineties and China emerged.

The word embedding method help us to understand better the perceptions and social-economic roles of different groups and countries in the eye of the New York Times.

Appendix

Part 1: The list of keywords for each of the groups is the following

For Asians World:

Asia|Asian|Asia's|Asian's|China|Chinese|China's|India|Indian|Indian's|Japan|Japanese|Japan's|Korea|Korean|Korea's|Indonesia|Indonesian|Indonesia's|Hong Kong|Singapore|Singaporean|Singapore's|Macao|Taiwan|Tokyo|Seoul|Shanghai|Beijing|Malaysia|Malaysia's|Malaysian|Pakistan|Pakistan's|Pakistani|Thailand|Thailand's|Thai|Vietnam|Vietnam's|Vietnamese|Philippines|Philippines's|Philippine.

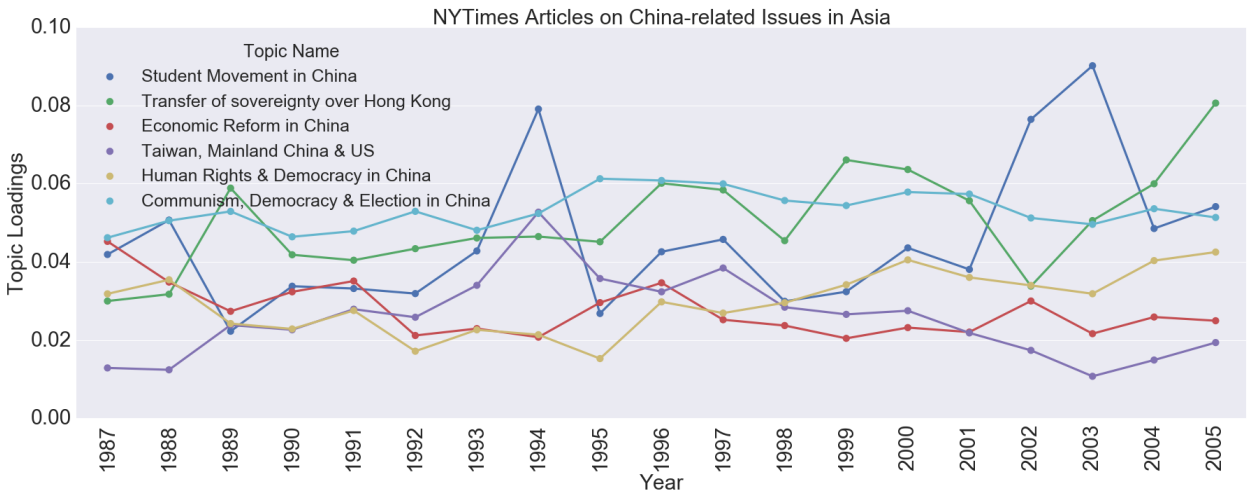
For Islamic World:

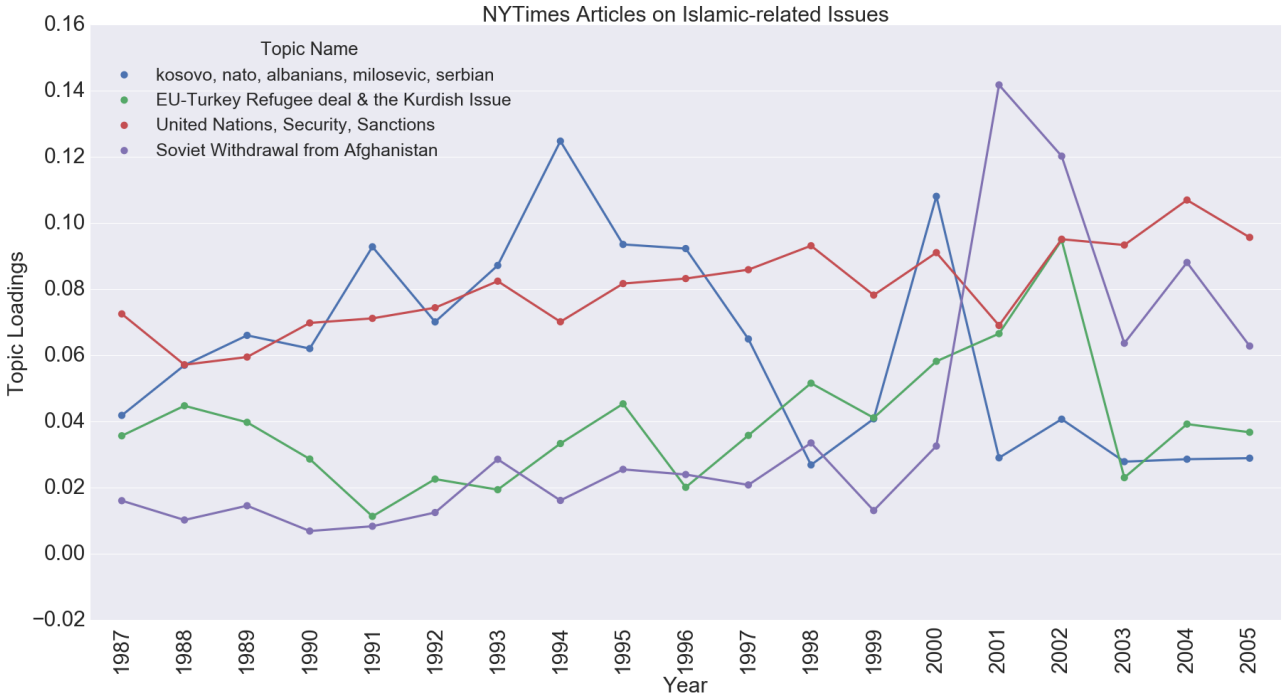
muslim|Muslim|Muslim\'s|Islam|islam|Islam\'s|Afghanistan|Islamic|Oman|Omanis|Iran|Iranian|Saudi Arab|Saudis|Mauritania|Yemen|Yemeni|Yemenis|Indonesia|Indonesian|Pakistan|Pakistani|Pakistanis|Bangladesh|Bangladeshis|Bengalis|Nigeria|Nigerian|Egypt|Egyptian|Turkey|Turkish|Sudan|Sudanese|Algeria|Algerian|Morocco|Moroccan|Iraq|Iraqis|Iraqi|Malaysia|Malaysian|Malay|Uzbekistan|Uzbek|Syria|Syrian|Kazakhstan|Kazakh|Burkina Faso|Burkinabe|Mali|Tunisia|Tunisian|Guinea|Guinean|Somalia|Somalis|Azerbaijan|Azerbaijanis|Tajikistan|Tajiks|Sierra Leone|Libya|Libyan|Jordanian|United Arab Emirates|Emiratis|Kyrgyzstan|Kyrgyz|Turkmenistan|Turkmen|Lebanon|Lebanese|Kuwait|Kuwaiti|Albania|AlbanianMauritania|Kosovo|Gambia|Bahrain|Bahrani|Comoros|Comorian|Qatar|Qatari|Djibouti|Brunei|Bruneian|Maldives|Maldivian.

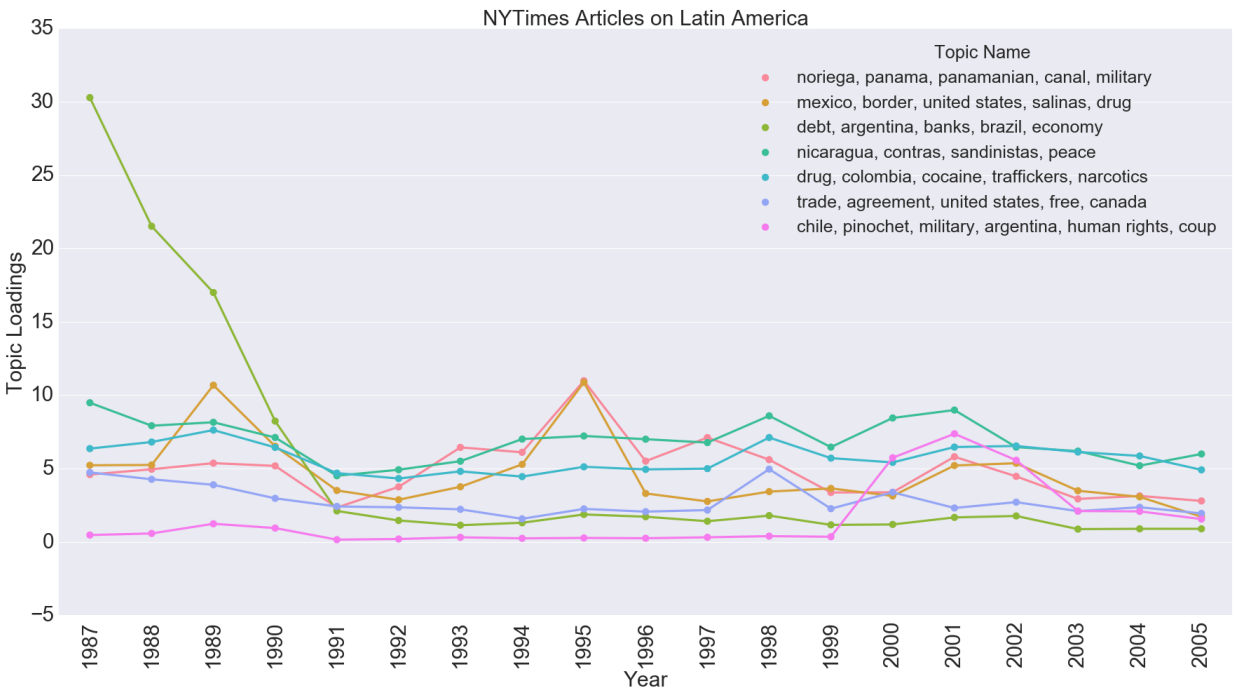
For the Latin American World:

hispanic|Hispanic|Hispanic\'s|latino|latino\'s|latinos|latinos\'|latina|latina\'s|latinas|latinas\'|Cuban|Mexican|Puerto Rican|Dominican|Cuba|Mexico|Puerto Rico|Dominican Republic|Central America|South America|Belize|Costa Rica|El Salvador|Guatemala|Honduras|Nicaragua|Panama|Argentina|Bolivia|Chile|Colombia|Ecuador|Paraguay|Peru|Uruguay|Venezuela.

Part 2: Figures of Topic Modelling that Use Topic Loading as the Y-axis.


NYTimes Articles on China-related Issues in Asia
Topic Name
Student Movement in China
Transfer of sovereignty over Hong Kong
Economic Reform in China
Taiwan, Mainland China & US
Human Rights & Democracy in China
Communism, Democracy & Election in China
Topic Loadings
0.10
0.08
0.06
0.04
0.02
0.00
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Year


NYTimes Articles on Islamic-related Issues
Topic Name
kosovo, nato, albanians, milosevic, serbian
EU-Turkey Refugee deal & the Kurdish Issue
United Nations, Security, Sanctions
Soviet Withdrawal from Afghanistan
Topic Loadings
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Year


NYTimes Articles on Latin America
Topic Name
noriega, panama, panamanian, canal, military
mexico, border, united states, salinas, drug
debt, argentina, banks, brazil, economy
nicaragua, contras, sandinistas, peace
drug, colombia, cocaine, traffickers, narcotics
trade, agreement, united states, free, canada
chile, pinochet, military, argentina, human rights, coup
Topic Loadings
35
30
25
20
15
10
5
0
−5
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Year