Quantitative Analysis of the Global Financial Crisis in The Economist

Jingying (Jane) Bi   jingyingb@uchicago.edu

I studied the full dataset of publications in The Economist from 2006 to 2010 to inspect how the articles in different sectors as well as different regions/ countries take response to the crisis. I explored how the responses compare and contrast with each other by looking at the inner sector semantic environment, the cross-sector comparisons, and the evolution of both of them. My paper contributes by doing quantitative analysis on texts about the Global Financial crisis. The main findings are intuitive and greatly align to the factual history. In particular, there exist a "response time lag" for any sector to react to the crisis event. This lag depends on the sectors as well as the regions/ countries where the articles come from. Moreover, the diversity in the contents cross-sectors have been observed before the crisis. Whereas as the crisis affects the world more deeply and widely, a convergence in the discussions relevant to the crisis appear among the sectors

The 2007-2008 Global Financial Crisis is the most serious and wide spread depression in history. It started from 2007 after the housing bubble burst. The severity was greatly leveraged by the mortgage backed securities which spread the panic and breakdown in the housing market to all the other industries and sectors, and finally the whole world. By conducting the comparative studies on The Economist articles during 2006 and 2010, I hope to explore how different sectors responded to the crisis. How did their response differ from each other? Similarly, how did the response differ across different regions or countries? Whether the articles' content tended to converge over time during the crisis? How did the responses evolve over time?

LITERATURE REVIEW

Text mining techniques have been widely used in studying the social and cultural issues. Chen (2013) conducted content analysis to study the relationship between the language and people's future-oriented behavior. He concludes that people speaking language grammatically associated with strong future-time reference tend to consider more about their future activities and thus end up with more savings, less obese, and huger wealth by the time of retirement. Bolukbasi *et al* (2016) studied the gender stereotypes by training the Google news articles with word-embedding. Michel *et al* (2010) quantitatively explored the culture trends (e.g. grammar evolution, fame, censorship) by analyzing a large-scale corpus of digital texts. Klingensteina *et al* (2014) analyzed the trial records from late 18th to the early 19th century in UK to inspect the civilizing process.

Inspired by such a variety of text mining studies, I hope to explore and understand the Global Financial Crisis with text mining. Conventionally, there are a number of papers working on the event. Many of them are about people's psychology, behavior, wellbeing during and after the crisis (Barberis, 2011; Deaton, 2012; Shapiro, 2010). Most of the time, they either do the qualitative analysis, or qualitatively study the effect of the financial crisis by looking at all kinds of statistics or

running regressions to explore the causality. I hope to study the crisis from a different methodology by exploring the corpus published before, during and after the crisis.

DATA COLLECTION

The .xml files of The Economist Archives from 2006 and 2010 are parsed with Python codes. Information of each article's corresponding published year, month, day, title, content, tag have been collected in the data frame and saved in the .csv files.

The tags are critical for the analysis. Each article published on The Economist has a tag. The tag could be assigned by *sectors*, or *region/country*. For instance, the sectors classification has four tags, including business, finance and economics, science & technology, books & arts. And the region/ country classification has six tags, including Britain, Europe, United States, The Americas, Middle East & Africa, and Asia. These tags, in addition to the time frame, allow me to conduct comparative studies cross sectors, cross regions and over time variance.

My corpus includes all the 18,754 articles published in The Economist during 2006 to 2010.

METHODOLOGY

I will apply a number of text mining techniques. For instance, I use the knowledge from the *corpus linguistic*s to coarse graining the corpus and reduce the word bad dimensionality. I also study the word occurrence and frequency changes over time by doing the word counting.

Since The Economist articles have been classified and assigned a tag. These articles could been visualized by *clustering* plots. Moreover, The evolution of the clustering pattern yields more informative findings.

The *word embedding* allows me to retrieve a certain word and its neighboring words which are with closeness and similarities to the word.

Last but not least, little discussion is about the Part of Speech (POS) which is analyzed via *information extraction*.

CORPUS PREPROCESSING

With the 18.754 articles stored in the data frame in python, I firstly coarse graining the corpus by tokenize the sentences and the words. Then the tokens have been normalized by removing the trivial but highly frequent words in the stop-word list. Meanwhile, the tokens have been stemmed by the standard stemming as well as weighted using the TF-IDF method. Finally, the total number of the words has been halved, which release the burden of intensive computations. More importantly, such way of reducing the word bag dimensionality will not largely incur the information loss.

DATA ANALYSIS

Observation 1: "crisis" word occurrence and word frequency

According to Michel et al (2011), the usage frequency of a word is ratio of its occurrence over the total number of words in the corpus in that year. However, I will release the restrictions on the timing frame, but define the word frequency as the incidence rate in a selected corpus. This is because I will focus on the word usage in each month's articles. For instance, in January 2006, there are 19 occurrences of the word "crisis" out of totally 208,005 words in this selected corpus. Hence the word frequency is 9.134e-05. I will scale it by 1,000 times. So, in the figures, it appears to be 0.09134 ‰. Comparatively, I also consider the word occurrence which is simply the number of the word.

Figure 1. displays the "crisis" word occurrence and the "crisis" word frequency. The pink highlighted area marks the financial crisis period from the mid of 2007 to the end of 2008. It contains three panels.

Panel (a) is based on the whole corpus, i.e. total 18,754 articles. The top plot corresponds to the word occurrence of the "crisis" over time. It shows that the occurrences only started to raise at the middle of the crisis period in around July, 2008. I will call this "response time lag" where the articles did take time to adjust the content according to the event. However, this is dubious because information is rather liquid and faster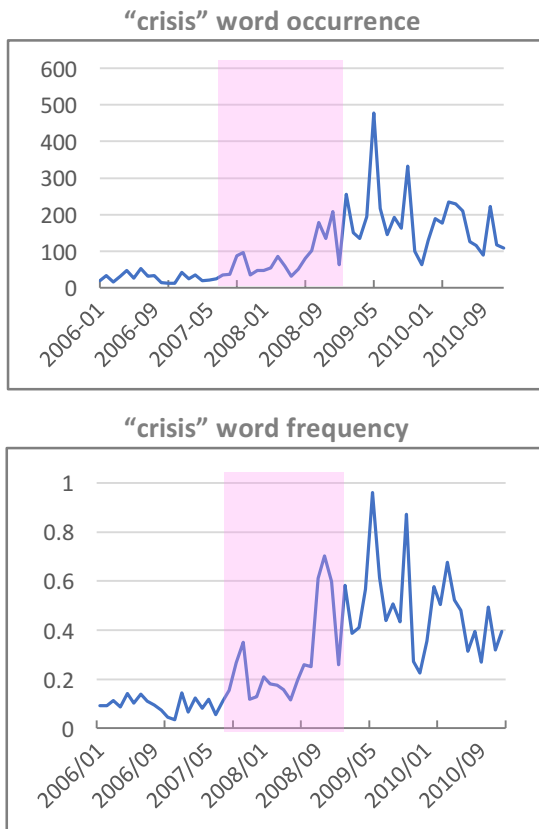 transferred, and the crisis is such a great economic event that affects everyone's daily life. Hence, the lag of almost half a year seems to be counterintuitive.

Hence, I generate the word frequency on the bottom plot. It leads to extra informative findings after being compared with the left one. Although the word occurrence at the beginning of the crisis is low, it could be caused by the little number of the total words. In fact, the bottom plot indicates that the proportion of "crisis" in the articles sharply increased when the crisis happened. Although it decreased for a while, this could be caused by the increase in the total number of words.

Moreover, both plots display a high-rising trend starting from the end of the crisis. This tells that people tended to talk and read more
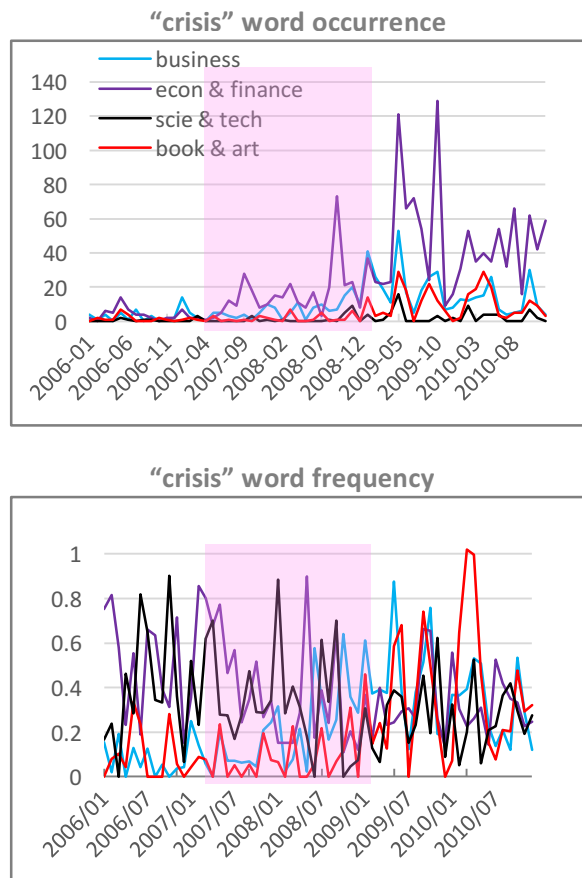

Fig 1. Panel (a) The word occurrence (top) and word frequency (bottom) of the word "crisis" based on the whole corpus. Note that the occurrence is the raw number of the times that the word appears in the selected corpus. The word frequency is ratio of the word's occurrence to the total number of words. Moreover, due to the sparsity of the words in the large corpus, I scale the frequency by 1,000 times. Hence, the unit of the word frequency is per thousand percent.

Fig 1. Panel (b) The word occurrence (top) and word frequency (bottom) of the word "crisis" by the four sectors.

about the crisis. The comments or analysis regarding the crisis is retrospective. In both plots, the "crisis" occurrence or frequency peak at the mid of 2009. Then it gradually decreased although with some volatilities. The relatively low levels during 2010 indicates the recovery of the economics and hence people talked less about the "crisis".

Panel (b) is based on the four sectors, namely business, economics & finance, science & tech, and books & art. In the right plot, neither the lines for economics &

finance, science & tech has any obvious trend. In comparative, "crisis" word frequencies in sectors of business and books & art changed in a more obvious way. The business sector greatly increased its discussion on the crisis at the beginning of the crisis, which takes a quick response to the event. This could be due to the need of the sector where people have to obtain the most updated and fresh information. Instead of commenting and analyzing the event retrospectively as many scholars and researchers do, they need to keep update and
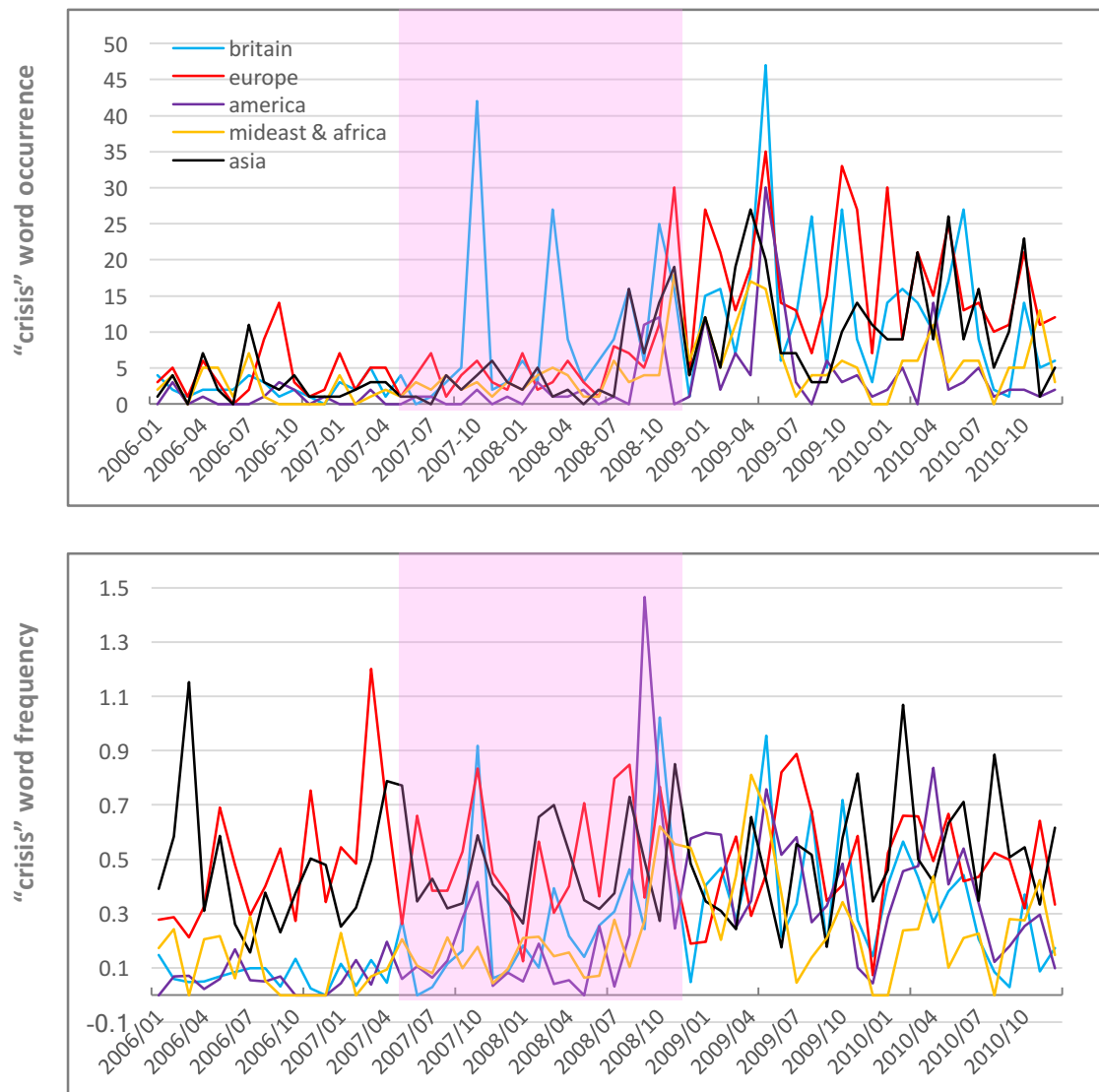

Fig 1. Panel (c) The word occurrence (top) and word frequency (bottom) of the word "crisis" by the six regions/ countries.

sometimes even predict the economic status. Compared with business sector's swift response, the books & art sector took a longer "response time lag". Moreover, the levels of the word occurrences infer a closeness relationship between the event and the sectors.

For instance, the directly related economics & finance and business sectors have more discussions as well as shorter "response time lag". However, the crisis discussion in the other two sectors is less frequent and requires longer "response time lag".

Panel (c) is based on the five regions or countries. In the top plot, all the regions or countries display a low level of discussion in "crisis" before the mid of 2008, except Britain. Britain took a rather immediate response right after the beginning of the crisis, but also a volatile response over the time, and this trend was maintained throughout the post-crisis period. Europe is the next country which started to increase its discussion in "crisis" since July 2008, followed by America and Asia but with lower levels. However, the word occurrences cannot fully reflect the regions or countries' care on the "crisis". The information from the frequency plot should also be integrated to capture a better understanding.

In the bottom graph, Europe and Asia are the two regions which have a continuously high level discussion in the "crisis", even before the beginning of the event. This could be due to that the two regions are in favor of talking about previous crisis, such as the Asia Financial Crisis in 1997.

Britain took a rather swift response in discussion on crisis, reflected by the sharp high-rise at the beginning of the event. Although America took a little bit longer "response time", its discussion overweighed the Britain's at the second half of the event. Generally, all the areas and regions involved

in a similar "crisis" word frequency level at the post crisis age, which is a convergence in the content relevant to "crisis". This is intuitive because the financial crisis is a globally impact economic recession.

**Observation 2: Interactions of the Sectors**

Figure 2. displays the clustering plots of the articles classified by the sector. There are five plots corresponding to 2006-2010, which shows the evolution of the clustering pattern.

The four clusters represent the four sectors, where the colors match the clusters with each sector. Regard the 2006 clustering plot as the status quo. The four clusters in this plot all have their own concentrated areas..

The locations of the dots show the closeness of the articles about each sector. So, the article content about the economics & finance (red dots) and the business (orange) sectors are close to each other. Different from the economics & finance sector articles which are almost isolated with those from the book & art and the science & tech sectors, the business sector touches the two marginal sectors. This is intuitive. One could easily imagine that the content about economics and finance are rather professional and requires more advanced understanding to those areas. In contrast, business sector is highly close to normal daily life. In addition, it need for updated information drives it to contact the other two sectors.

Interestingly, in addition to the four concentrated clusters, there exists a mixing area at the right-upper-corner where all the colorful dots mix with each other. The mixing area represents some overlapping discussion topics appearing in all four sectors' articles. The is little change in the 2007 and 2008 plots. From the 2009 plot, it is obviously seen that the local mixing area is diminishing and almost disappears. Instead, an almost global mingle among the colorful dots appears in the 2010 plot.

For instance, the red dots, corresponding
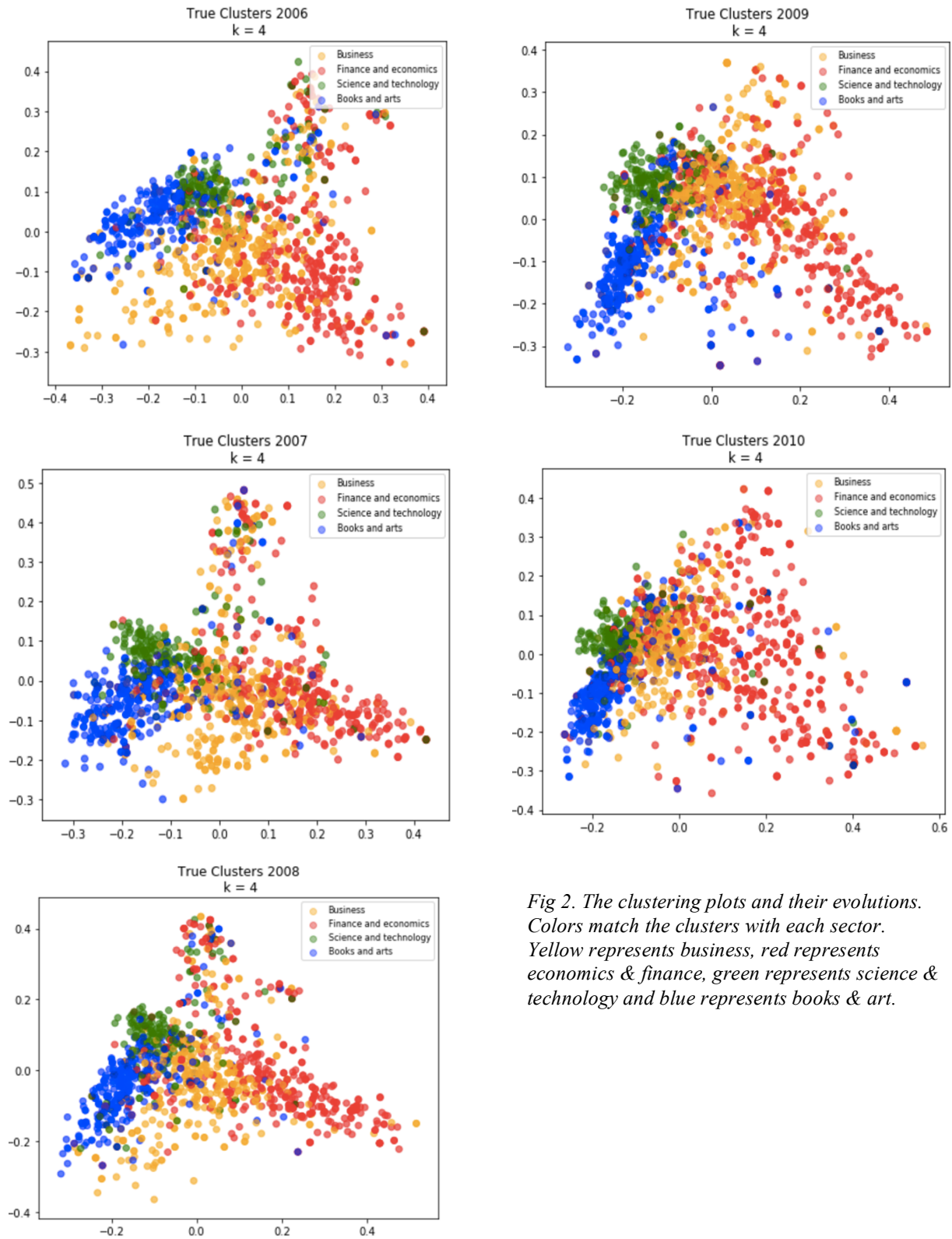
Fig 2. The clustering plots and their evolutions. Colors match the clusters with each sector. Yellow represents business, red represents economics & finance, green represents science & technology and blue represents books & art.

to the economics & finance sector, tend to expand themselves in the 2009 plot, and finally obtain a large space on the 2010 plot. This could be due to a topic divergence before the crisis came. Hence, each sector has their own commonly discussed topics as usual. And some their discussion could be cross-section. As the crisis came, all the sectors were affected significantly. The articles all had to sway their discussion toward the "crisis". This is the convergence of the topics cross-section. This is shown on the plot where the local mixing area diminishes and a global mixing patter emerges in 2009-2010.

Moreover, the clustering pattern started to change only since 2009, which infers to a "response time lag" align to the **Observation 1**. The blue dots (books & art) better follow the evolving pattern of the economics & finance than the green dots (science & tech). This also matches the **Observation 1** which shows a closeness between book & art and the economics & finance than the science & tech.

Observation 3 (a): The Evolution of "crisis" Word Embedding – by sector

Observation 1 tells that the discussion about the "crisis" changed over time, as well as affected by different sectors and countries or regions. But it tells little about how the specific semantic neighborhood of "crisis" looked like and how it evolved over time. By using the word embedding, I retrieved the word "crisis" in each sub-corpus for the four years, 2007~2010. I plotted the semantic neighborhood of the crisis, as shown in the Figure 3.

The size of the words is determined by its occurrences in the corresponding corpus. The locations of the words reflect the cosine similarities of them. For instance, two closed located words have higher similarity to each other, in terms of similar meanings, similar usage in the texts and more frequent

Fig 3. The semantic neighbor of the word "crisis" and its evolution over time – economics & finance sector

co-occurrence in the sub-corpus. These plots well display the relationship between the word "crisis" and other words close to it.

In 2007, the words closely located near the "crisis" are "regulators", "housing", "home", "liquidity", and "stock". This observation well substantiates the economic mechanism. The financial crisis has been greatly contributed by the mortgage backed securities. When the banks started to issue loans to more sub-prime borrowers whose repaying capability is relatively loan, the housing demand raised. This subsequently pushed up the housing price. The high-rising price drives those greedy investors to make more house purchasing orders not only for the residence but also for the investment. People will not worry about shirking their loans, as they could merely auction the house, gaining high revenues and clear up or the debts with luxury extra returns. However, when more and more people void their loans and sell the houses, the price of the houses dropped sharply, which triggers more underwater buyers, as their debt even exceeded the housing asset value. This tragedy has been leveraged through the mortgage backed security channels. When the bankrupt, loan defaults expanded to more areas, the crisis engulfed the whole society, and the whole world. So, the 2007 neighborhood provides "home" and "housing", which are the causal factors of the crisis.

In 2008, the words are "equity", "value", "private", "company", "finance", "public", "management", etc. This matches the second stage where the tragedy happening in the housing market expanded to other sectors such as financial sectors, private and public sectors.

When the crisis peaked in 2009, most topic are about "financial crisis". That why only the word "financial" exclusively takes the position near the word "crisis". This does not mean that no other words are close to

"crisis" at all, but tells such a strong bond-ship between "financial" and "crisis". This may be also due to that "financial crisis" has become one terminology for people to address this event.

Words in the 2010 neighborhood become "economic", "policy", "system", and "price". At the post-crisis era, people tend to think more about the root cause of the crisis, whether there is any system problem. So, they spent more time in reviewing the crisis and learned from the lesson. On the other hand, governments and all the other institutes had to take actions. For instance, government will try to refine or pass new policies to regulate the financial market. The central bank also need to curve the price inflation and stabilize the market.

In conclusion, the word embedding provides a good way to study the eco-semantic environment of the word "crisis". Its surrounding words kept changing over time and well followed the contemporary events. At the begging, the surroundings are words about the cause of the "crisis". The middle stage has more words in the subsequently affected areas. And during the post crisis era, more actions and learnings have been taken.

Observation 3 (b): The Evolution of "crisis" Word Embedding – by regions

Similarly, I also generate the semantic neighbor of the word "crisis" by different sectors in 2009. The result is shown in Figure 4.

Recall that in **Observation 1**, compared with the articles about the science & tech sector, the articles about the book & art tends to be closer to those about the economics & finance. This could be verified here again. In the semantic environment of the book & art plot, the words surrounding "crisis" are "system", "national", "banks", "central", "show", etc. It seems to study the system problems contributing to this global tragedy.

Finance and Economics 2009

Science and Technology 2009

Books and Art 2009

Fig 4. The semantic neighbor of the word "crisis" and its evolution over time – by sectors in 2009

In contrast, those around the "crisis" in the science & tech are "however", "become", "place", "go", "things", etc. These are rather trivial words with no critical meanings.

Therefore, by comparing the semantic surroundings of the word "crisis" in the 2009 articles, I find that the articles from different sectors paid different amount of attention to the crisis.

Observation 4: Most Similar Words

Besides using the wed embedding and plotting the closely related words, a list of words could also provide useful information. The words on the list are the top ones which have a large cosine similarities with the targeting word "crisis" Table 1 displays such a list of words. They seem to be the different results as shown in the semantic neighbor plots. Hence, they provide some new information. For instance, the first panel "crisis" lists the highly close words to the "crisis" based on the whole corpus dataset. The words are "bust", "turmoil", "recession"

Crisis				
2006	2007	2008	2009	2010
nationalism	('concern'	('crunch'	('meltdown'	('loaned'
instability	('audited'	('derivatives'	('crunch'	('meltdown'
crises	('stability'	('turmoil'	('stability'	('collapse'
budgeting	('contagion'	('bubble'	('turmoil'	('panic'
	('economies'	('bust'	('downturn'	('recession'
	('lender'		('collapse'	
Britain -- Crisis				
2006	2007	2008	2009	2010
('policies'	('june'	('global'	('markets'	('western'
('problems'	('war'	('energy'	('system'	('global'
('trying'	('problems'	('markets'	('global'	('inflation'
('head'	('run'	('american'	('policy'	('demand'
('land'		('price'	('market'	('chile'
		('source'	('companies'	
Europe -- Crisis				
2006	2007	2008	2009	2010
('western'	('voters'	('western'	('land'	('financial'
('national'	('western'	('global'	('model'	('markets'
('source'	('national'	('research'	('turn'	('costs'
('price'	('annual'	('firms'	('history'	('fund'
('global'	('energy'	('political'		('fiscal'
('russia'		('national'		('imf'
America -- Crisis				
2006	2007	2008	2009	2010
('immigrants'	('price'	('india'	('problems'	('western'
('energy'	('island'	('rise'	('recession'	('problems'
('firms'	('place'	('market'	('enough'	('chinese'
('problems'	('rather'	('group'	('chinese'	('price'
('racial'	('energy'	('source'	('size'	('international'
('research'	('americans'	('company'		
('land'	('land'	('prices'		

Table 1. The words that are most similar to the word "crisis" – by regions/ countries

"contagion" and so on. Some of them are the words to describe the crisis or a synonym of it. Some have deeper meanings. "Bubble", "lender", "derivatives" vividly depict the cause of the crisis. "Contagion" reflect the property of the crisis. Moreover, across the four panels, the word "global" is a highly frequent word. This reflects that the 2008 crisis has a global effect.

Observation 5: More Discussions on Evolutions by Sectors and Areas

I further inspect how the correlation of different sectors evolve over time. In this section, I will talk about the findings on their evolutions.

Figure 5 shows the evolution of the content similarities among articles in different sectors. The similarities are the cosine similarities between two compared corpus. For instance, in the 2006 plot, the four corpus are the articles from the four sectors. All four documents have the similarity to be 1 with themselves. Hence, the four cells on the diagonal are white.

The 2006 plot is still regarded as the status quo. In this plot, the documents among different sectors are very similar to the others. Although, the heatplot is colorful, the scale ruler shows very small deviation, scaling from 0.995 to 1.000.
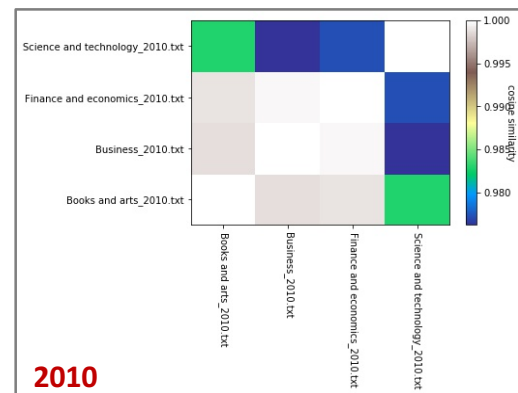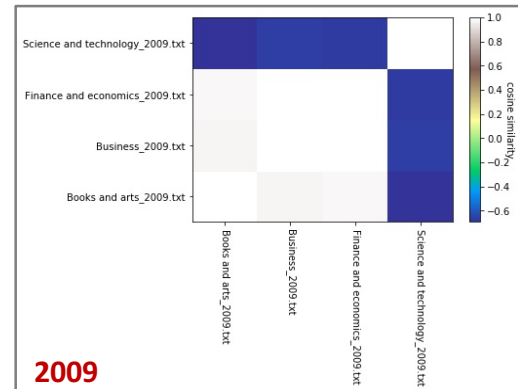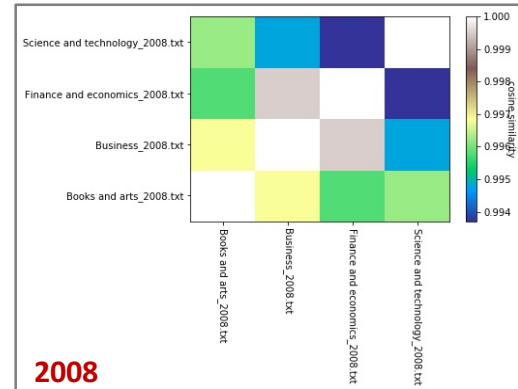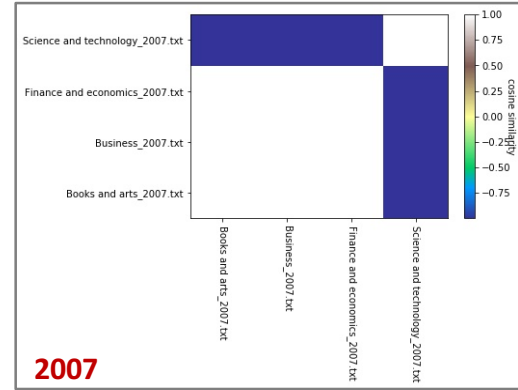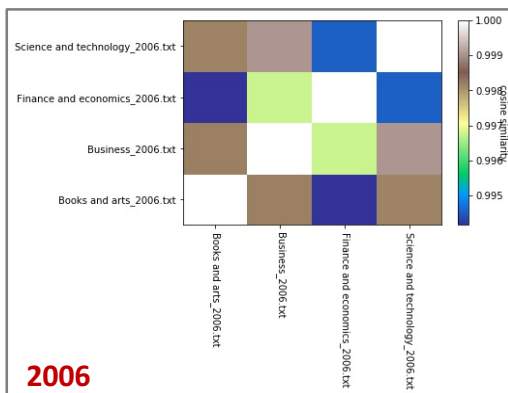
Fig 5. Similarities between the documents and its evolution – by sector.
The color represents the scale of the cosine similarities of the four documents classified by the sectors and years.

In 2007, the content tends to be uniformed except that from the science and technology sector. This could be told from the large scale of white and the only dark blue strip of the science & tech. The 2008 plot becomes diverse again but as the plot in 2006, the differences are in a small extent. The 2009 plot jumps back to uniform similarity except for the science and technology sector. Then 2010 is about a mixture of general similarity and local dissimilarities.

The observations are interesting in terms of the pattern evolution. From the 2006 to 2007 plot, it gets more uniform. But one is unable to declare that this is due to the crisis, because the "time response lag" has been observed in the Observation 1. Hence the sectors need to take some time before taking a uniform response to the crisis. When the plot goes back to the diversified pattern, I could not stop wondering whether any confounding factors in 2007 cover the "truth" that its plot should also be as diverse as the 2006 and 2008. As discussed in the Observation 1, there exists a convergence in content from different sectors when all of them started to care and talk more about the crisis. Hence, this could be reason of the uniformity display in the 2009 plot.

However, the science & technology sector articles' content is very inelastic. This may be due to people in that sector cares less about the economic status, and especially their profits and interests would not be greatly affected by the crisis. During the post crisis era, I observe an inertia in maintaining the discussions on crisis. Meanwhile, as the hotness of the topic attenuates, it also allows more divergence in the contents as shown in green, blue and purple in the plot.
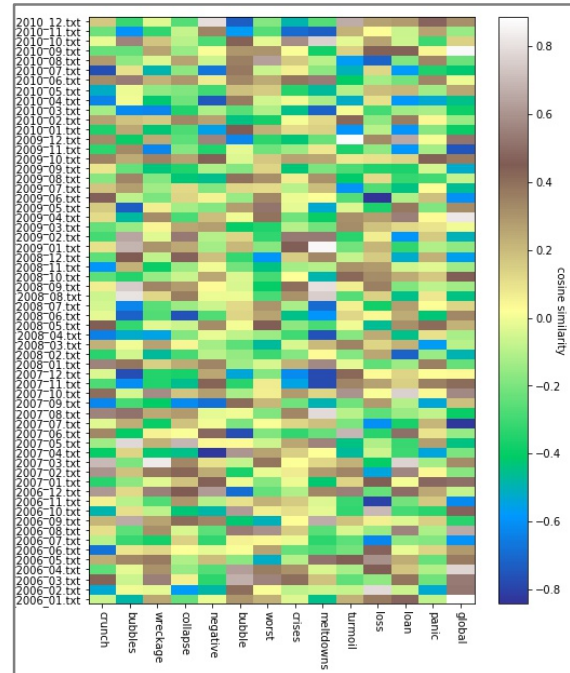
Observation 6: The similarities between documents and keywords
On the other hand, I discover how the evolution unfolds within each sector, as

shown in Figure 6.

First of all, the mechanism of the plotting is as follows. I manually selected a list of words which are highly related to the topic "crisis". These words are 'crunch', 'bubbles', 'wreckage', 'collapse', 'negative', 'bubble', 'worst', 'crises', 'turmoil', 'loss', 'loan', 'panic', and 'global'. As for the corpus, the whole corpus is divided based on the sectors as well as the date of publication. Hence, there are four sectors. Each sector has 5*12 = 60 months of sub-corpus. The heatmap displays the cosine similarities of each monthly corpus and each keyword. One could read the plot and draw conclusions on the overtime changes in the "crisis" discussion happening in each sector. The color changes in each column tell us the evolution of the similarities between the corpus and the corresponding keyword.

The panel (a) in Figure 6 displays the evolution of similarities between corpus
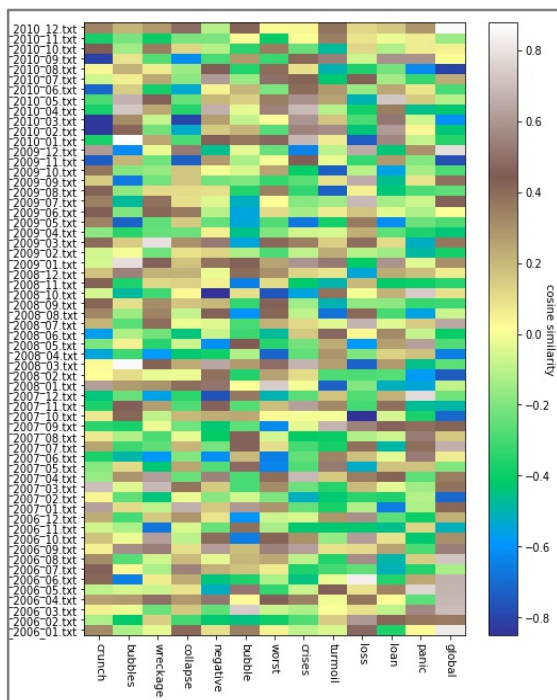

Economics & Finance 2006 ~ 2010

Fig 6. Similarities between documents and keywords and the evolution – by sectors
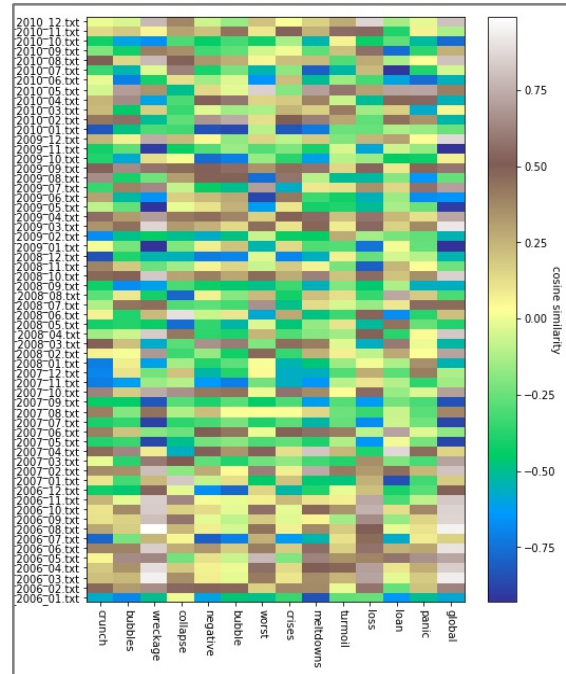*Panel (a): the economics & finance sector*

from the sector economics & finance and the list of keywords. For instance, let's see the column of the "negative" first. Before April 2007, the colors are mostly green, blue or at most a little bit light yellow. This indicates that the sentiment of the corpus before the crisis relatively opposite to the "negative". The colors changes to yellow and brown during the crisis, and returns back to green and blue again since the mid of 2009. This tells that the crisis renders the content to be more "negative", and this sadness faked gradually during the recovery. The column of "turmoil" starts with blues and greens until October 2007, changes to the light yellow and brown until April 2009, and return to the green and blue again after the crisis.

The panel (b) in Figure 6 displays the evolution of similarities between corpus from the sector business and the list of the
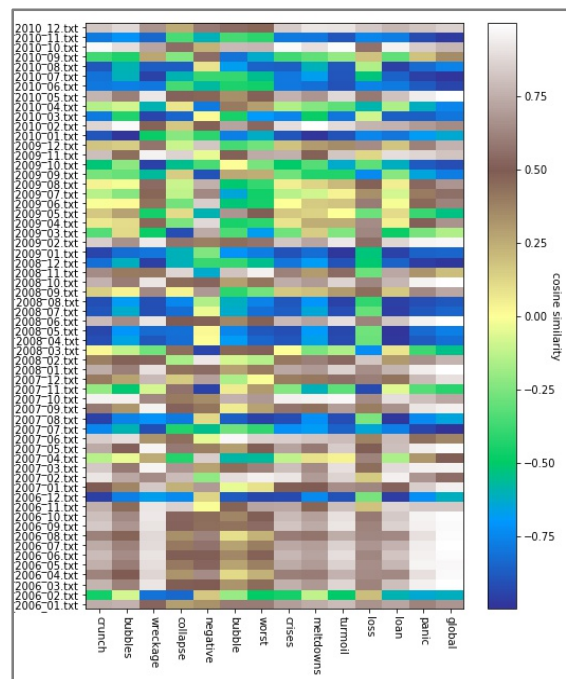

Books & Art 2006 ~ 2010

Science & Technology 2006 ~2010

Fig 6. Similarities between documents and keywords and the evolution – by sectors
Panel (c): the books & art sector (upper)
Panel (d): the science & tech sector (lower)

Business 2006 ~ 2010

Fig 6. Similarities between documents and keywords and the evolution – by sectors
Panel (b): the business sector

keywords. Compared with the heatplot of the econimcs & finance sector, the business

sector seems to be slightly less affected by the crisis. For instance, the "negative" and "turmoil" columns starts with blues and greens and maintains the colors until October 2008, after which it turns to some light yellow and brown. The colors return to more green and blue by September 2010. Therefore, the "response time lag" of this sector is indeed longer than the economics & finance sector. The effect fades slower and the returns to the pre-crisis status later as well.

The panel (c) and (d) in Figure 6 displays the evolution of similarities between corpus from the sector books & art and science & tech and the list of the keywords. The latter one obviously has more green and ever darker blue. This indicates that the articles about the sector science & tech is less affected by the crisis. This perfectly matches the previous observations.

Observation 7: More Discussions on Evolutions by Regions and Countries

I also inspect the correlation of the articles about different regions or countries and how it evolves over time. Similar to the heatplots in the **Observation 5**, the colors represent the cosine similarities between the documents.
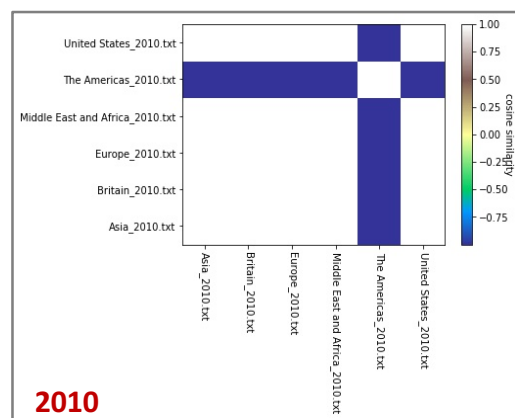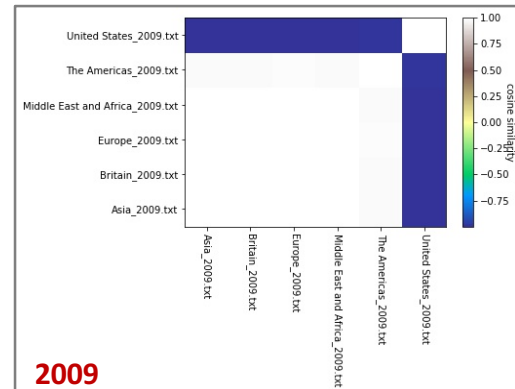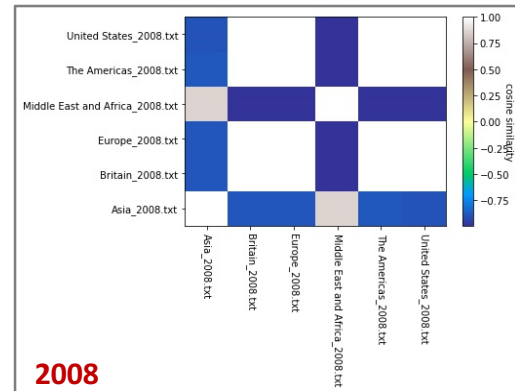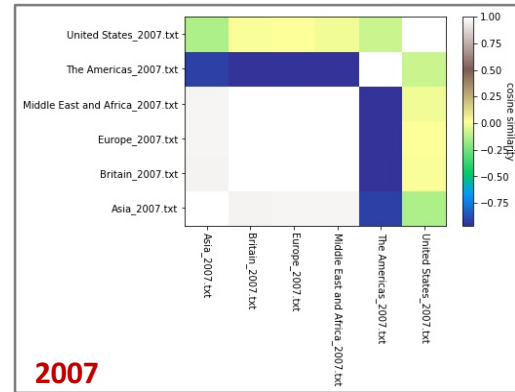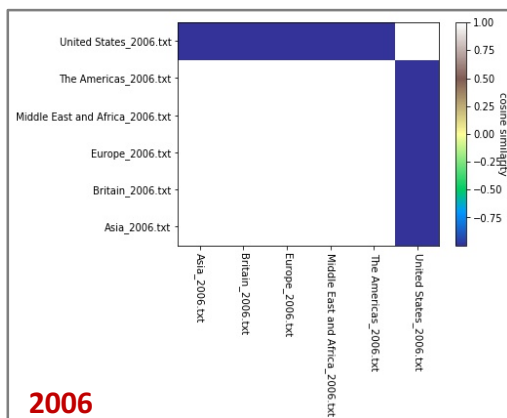

Fig 7. Documents similarities and its evolution – by regions/ countries

The evolution of the heatplots tells some interesting trends. Holistically speaking, Most documents are very similar in 2006 except the United States. This uniform similarity is reduced during the 2007 where some light green and gray appear. The America and the United States documents tend to be most dissimilar to others. In 2008, the dissimilarities are more obvious in the documents of Middle East & Africa and Asia. But these two documents are of high similarity. In 2009 and 2010, the United States and the Americas are still the two documents which are most dissimilar to the others.

To interpret my observations, I compare the pattern changes. It seems that the America document are usually different from others. Hence if we could ignore it, we will see the dissimilarity of the United States in 2006, 2008 and 2009, and a global similarity in 2007 and 2010. The 2006 dissimilarities could be explained by the distinct topics which are talked in the documents. When the crisis happens, the documents from all the regions tend to respond to the events and hence display a global similarity in 2007 (without accounting for the Americas). The dissimilarities appear in 2008 may be due to that in 2008 the Middle East & Africa and Asia have less discussions about the crisis compared with other regions, recalling the panel (c) in Figure 1. The 2009 distinction in the United States may be due to its above-average attention in the crisis, after all, the crisis started from U.S.. When it comes to 2010, during the recovery, the global similarities appear again (without accounting for the Americas). I hope to explain this as that all the regions and areas paid similar extent of attentions to the crisis. They may consider what lessons have been learned from the tragedy, what actions could be taken to better recover the economy, or what are the root problems that led to the crisis.

Conclusion

The qualitative analysis of The Economist provides deeper insights on how the articles respond to the financial crisis. The findings are informative and mostly well align to the factual history. For instance, the "response time lag" is displayed on the word frequency/ occurrence distributions. And the rate of the "response time" depends on the sectors as well as the regions/ countries which the article is affiliated to. Moreover, the cross-sector and cross-region documents relevant to the Global Financial Crisis tend to converge over time. It indicates that the crisis is a global event which affects every countries and sectors.

Some weakness of the analysis are (1) lack of the robustness test; (2) most analysis are based on the local texts, such as the word "crisis", which may be biased.

Future research could be the comparative studies between the 1929-1933 recession and the 2007-2008 crisis. After all, the crisis happening every time may be different and due to some deeply rooted mechanism. Inspections on how the available corpus reflects these mechanisms will be interesting.

Reference

Barberis, N. (2011, August). Psychology and the Financial Crisis of 2007-2008.

Chen, M. K. (2013). The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets. American Economic Review, 103 (2).

Deaton, A. (2012, January). The financial crisis and the well-being of Americans. Oxf Econ Pap.

Jean-Baptiste Michel, Y. K. (2010, December 16). Quantitative Analysis of Culture Using Millions of Digitized Books. Sciencexpress.

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) Quantitative analysis of culture using millions of digitized books. Science 331: 176–182.

Sara Klingensteina, T. H. (2014, May 22). The civilizing process in London's Old Bailey. PNAS.

Shapiro, M. D. (2010). The Effects of the Financial Crisis on the Well- Being of Older Americans: Evidence from the Cognitive Economics Study.

Tolga Bolukbasi, K.-W. C. (2016, July 21). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.