Knowledge Graph Based on the Impact of Trade War on Electronic Industry

2020

Team A9

Mei Yudong

Niu Yuke

Yuan Wengao

Zhang Yanling

Zhu Xin

Zou Xuyuan

Knowledge Graph Based on the Impact

# of Trade War on Electronic Industry

Group A9: MEI.Yudong, ZHU.Xin, ZHANG.Yanlin, ZOU.Xuyuan, NIU.Yuke, YUAN Wengao

1 Introduction

1.1 Project idea

The China–United States trade war is an ongoing economic conflict between the world's two largest national economies. Companies from China of electronic industry bore the brunt of trade war.

The catchall term “electrical machinery and equipment” is China’s biggest export category, accounting for 26.4 per cent of total exports in 2017. It is the sector that has been most heavily targeted by the Trump administration in the trade war.

The US government has also imposed great restriction on China’s high-tech electronic giants, such as Huawei and ZTE.

Electronics supply chains are generally long and complex, meaning that if one company closes, many others are likely to take a hit as a result. It’s important to dig out the underlying impact along each of the elements inside or outside the supply chains of electronic industry.

Based on these challenges, our team’s mission is to figure out the direct and indirect effects that trade war has imposed on electronic supply chain by applying the up-to-date social analytical techniques and visualize the logic links between events and targets by knowledge graph (KG).

1.2 Data and Methodology

In order to build the knowledge graph, we collected data from *South China Morning Post*[1]. Then we conducted text preprocessing with classic framework with output of innovative normalized Tf-Idf matrix. The Tf-Idf matrix and vectors generalized with Word2Vec were sent to LDA and Gaussian Mixture Model cluster to abstract the important ones and separate the words into different topics and clusters. Finally, the important nodes were connected to each other according to the specific logic, which generalized our final knowledge graph. The main procedure how our KG was built shown as Fig 1:

[1] http://www.scmp.com

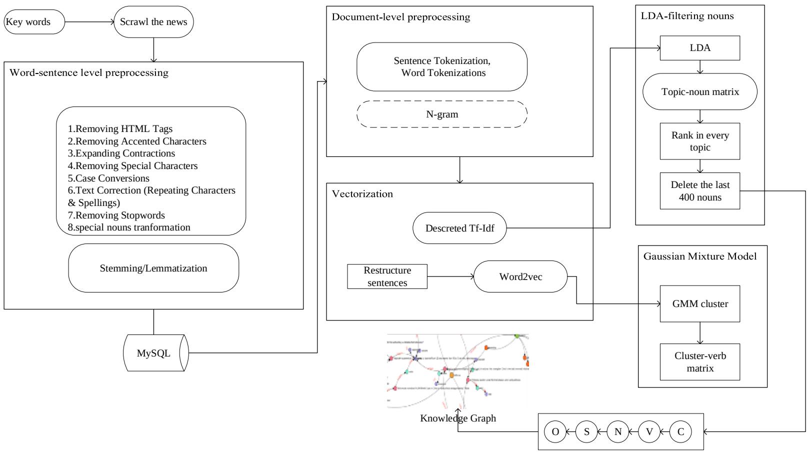

Fig 1. **The procedure of KG generalization**

As illustrated in the figure, considering that the n-gram is not so effective and efficient for our model, we finally decided to abandon the method in the step of Document-level preprocessing. In addition, The rule to connect the nodes is to following the sequence of cluster(C)→verb(V)→noun(N)→subject(S)→object(O), in which the cluster nodes is not only able to separate the nodes from root node to improve the efficiency in searching but also give every verbs a probabilistic trend on their types of cluster. The news node is nearly of the same use as clusters.

1.3 **Key Finding and Results**

The knowledge graph shows great ability to express the events into structured and human-like way. The nodes connecting to each other with the knowledge of is able to describe the relationship between each other in a simple and macro way.

The knowledge graph can not only show the relationship shown in news but also some logic connection between nodes such as the nodes of Huawei and HTC, both of which belong to phone manufacturer and they do connect to the node of “phone” and “company”.

To dig deeper, the relationship found from the knowledge graph may give us some inspirit on the potential risks and business opportunities. For example, we are able to see the event in a different vision and found some point that was ignored before is pretty risky, so we would have relatively enough time to deal with the risks before it come.

Thus, the main findings while constructed our knowledge graph including:

- KG is capable enough to tell stories in a data-driven and smarter way.
- The stories told by KG may give the key factors about the business valuable point and potential risks to help companies to run smoothly.

2 **Data Description**

Source of data

Essentially, we were planning to scrape news articles related to US-China trade war from *scmp.com*, a news media platform distributed by the leading global news company, *South China Morning Post*, that has a special focus on reporting China and Asia affairs.

The flow chart of web scraping processes are shown below:

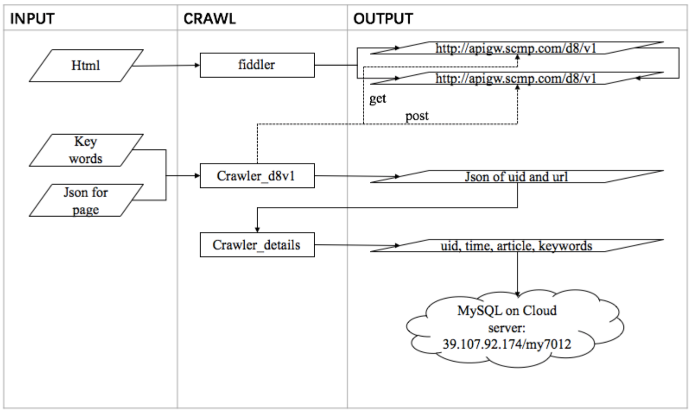
INPUT
CRAWL
OUTPUT
Html
fiddler
http://apigw.scmp.com/d8/v1
http://apigw.scmp.com/d8/v1
get
post
Key words
Json for page
Crawler_d8v1
Json of uid and url
Crawler_details
uid, time, article, keywords
MySQL on Cloud server: 39.107.92.174/my7012

Fig 2.**procedure of crawler**

2.1 **Raw data**

Since *scmp.com* is designed to only load new contents on a scrolling basis, and selenium commands failed to control the webpage. We switched to first download json files that contain various information about those articles, and then extract useful information of each article from json files. We collected 260 json files under the search word ‘electronic’, 230 json files under ‘chip’ and 517 json files under ‘huawei’ from the first stage web scraping.

Most of json file contains information about 20 articles each. We extracted the ‘url’ of each article in prepare for further web scraping, along with the ‘entitiUuid’ for identifying and locating the unique article and the ‘title’ for recording and readability. After iterating through all json files in the folder, we got nearly 20,000 records that are ready to serve as the input of the second stage web scraping.

2.2 **Structured data**

On the second stage web scraping, we opened detailed webpages of each article with the url and uid scrapped previously, and further scraped the time the article written, the content and keywords of each article. After filtering articles with massive missing information and after eliminating duplicated news articles, more than 10,000 clean records, with ‘title’, ’url’, ‘uid’, ‘time’, ‘article’, ‘keywords’ of each article listed respectively, were stored into MySQL Workbench that are ready for further analysis.

3 **Methodology**

3.1 **Vectorization**

3.1.1.Tf-Idf

There might be some terms that occur frequently across all documents and these may tend to overshadow other terms in the feature set, especially those don’t occur less frequently, but might be more interesting and effective as features to identify specific categories. That’s why TF-IDF is more practical than simple Bag Of Words.

TF-IDF stands for term frequency-inverse document frequency, which, as its name, is a combination of two metrics. Term frequency denotes frequency for word w in D, while inverse document frequency denoted by idf is the inverse of the document frequency for each term and is computed by dividing the total number of documents in our corpus by the document frequency for each term and then applying logarithmic scaling to the result, i.e. "$\mathrm{idf}\,(\mathrm{w}, \mathrm{D}) = 1 + log\frac{N}{1+df(w)}$. Usually, we normalize the tfidf matrix by dividing it with its L2 norm, which is the square root of the sum of the square of each term's tfidf weight, i.e.$\frac{tf-idf}{||tf-idf||}$.

3.1.2.Word2Vec

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

It can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words (CBOW) or continuous skip-gram. The size of the context window determines how many words before and after a given word would be included as context words of the given word. The recommended value is 10 for skip-gram and 5 for CBOW. Word2vec can also be trained with hierarchical softmax and/or negative sampling.

3.2 **Filtering**

3.2.1.LDA for nouns

Topic mining is the process of identifying topics in a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important.

Latent Dirichlet Allocation (LDA) is a form of unsupervised learning that views documents as bags of words (ie order does not matter). LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. LDA is mainly developed with EM algorithm, and the details of LDA is shown below.

A. The EM algorithm

The EM algorithm is a repeated alternation between Step b (E-step) and Step c (M-step). After each M-Step, log $(P|\theta)$ guaranteed to increase (unless it is already at a maximum). The graphics are shown as below:

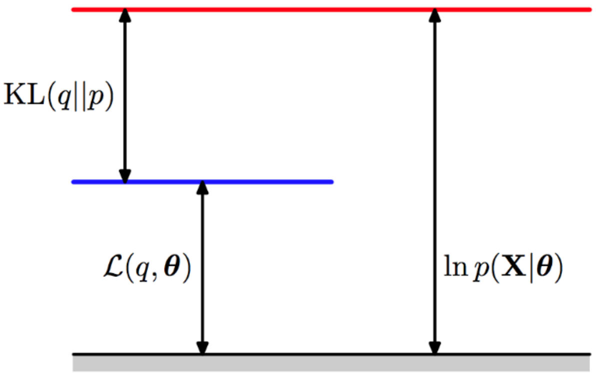
$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

Fig 3.**The graphic of the principle of EM algorithm**

a) Initial decomposition

Illustration of the decomposition given by Fig 2, which holds for any choice of distribution q(Z). Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(\mathrm{q}||p) \geq 0$, we see that the quantity $\mathcal{L}(q,\theta) \geq 0$ is lower bound on the log likelihood function $\ln p(X\,|\theta)$.

b) E-step

Holding the parameters θ constant, minimize KL(q(Z)||p(Z|X,θ)) with respect to q(Z). Remember, as is a distribution with a fixed functional form, this amounts to updating its parameters λ.

The caption implies that we can always compute q(Z)=p(Z|X,θ). We will below that this is not the case for LDA, nor for many interesting models.

c) M-step

In the M-step, maximize the ELBO with respect to the model parameters θ.

$$E_{q(z)}[\log(\frac{p(X,Z|\theta)}{q(z)})] = E_{q(z)}[\log p(X,Z|\theta)] - E_{q(z)}[\log\,(q(z))]$$
$$= E_{q(z)}\big[\log\big(p(X,Z|\theta)\big)\big] + H[q(z)]$$

Maximizing this expression with respect to θ, we treat the latter as a constant.

d) LDA for filtering nouns

LDA is utilized in our experiment to abstract the nouns and find the most apparent combination of the nouns and topics. Then rank all the nouns according to the descend order of the word’s contribution or probability to its corresponding topic. Then the last few words are deleted to filter out the more important ones.

3.2.2.GMM for filtering verbs

The Gaussian Mixture Model (GMM) is also modeled with EM algorithm. The main idea of GMM is to model the mixed Gaussian distribution of randomized variables into several clusters (E-step) and adjust the border of the clusters according to the maximum likelihood estimation of the distributions (M-step).

Considering the words come from different news, which has different distribution of words, GMM is one of the most ideal models for clustering the verbs. Moreover, The number of clusters is chosen according to the Silhouette Coefficients. When the Silhouette Coefficient reaches the minimum of the series of the number, we chose the number of cluster as our number of GMM cluster. The application of word2vec and GMM cluster is shown as below.

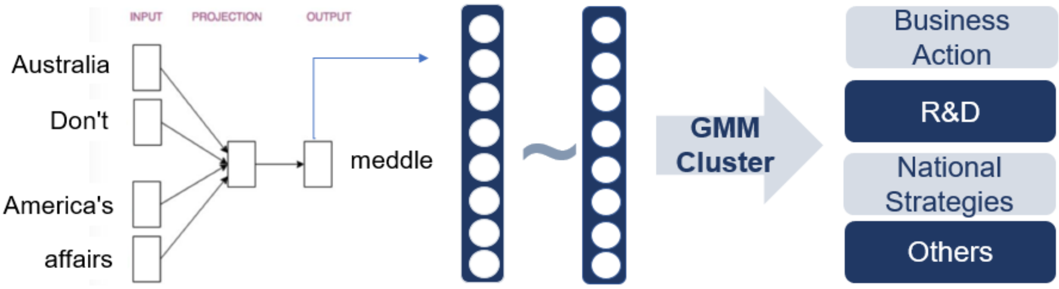

Fig 4. **The application of word2vec and GMM cluster**

Verbs are vectorized using Word2vec and processed through GMM cluster, which divides them into four categories named business action, R&D, national strategies and others.

4 Experiments and Results

4.1 Data Preprocessing

4.1.1.Normalization

Normalization aims to put all texts on a level playing field. In this part, we first lowercase all test and expand contraction to standardize text. Moreover, stemming and lemmatization are essential process that reduce the complexity of data by cutting words back to their base form.

4.1.2.Removing useless information

The aim of the part is to filter some useless information to increase the accuracy of the NLP tasks. First, It is likely that the data scraped online still contain HTML tags, which is useless for our further work, so we use `remove_html_tag(sentence)` function to remove the HTML tag of text data. Besides, since the original article contains punctuation and stop words that have no connection with possible topics and occur with a high frequency, we use the` `removePunctuation (sentence)` and the `remove_stopwords (sentence)` function to exclude these useless information in order to save computing time and efforts in processing large volumes of text.

4.1.3.Special treatment for unrecognized terminology

As we performed topic mining, we find that some newly-appeared terminologies related to electronic industry, such as 5G, have not been included in the thesaurus, so the computer cannot identify these newly appeared terminologies and processed them in a wrong way. Therefore, in this part, the main strategy is to find these unrecognized terminologies and use the ‘specialNN (sentence)` function to transform them in a proper way, like “fiveg”.

4.2 **Modeling**

4.2.1.LDA

For this part, we found that some stop words ignored by NLTK was left such as “is”. Additionally, if we just delete the words with three-time standard deviation of Tf-Idf, we may delete too many words considering that most of the words are of low Tf-Idf. In order to solve the problem, we transformed the Tf-Idf matrix into logarithmical form and the data of the matrix was transformed into normalized distribution. Then we deleted the words with Tf-Idf outside the range of mean plus three-time standard deviation and those appeared only once in our corpus. Then we used the left words to form a count matrix, which is a standard form of input for LDA function.

Using our processed data, the output from the LDA model is 5 topics each categorized by a series of words after 150 iteration. LDA model doesn’t give a topic name to those words and it is for us humans to interpret them. See below sample output from the model with words descend ranked according to their contribution to their own topics.

Table 1.The LDA Results (part)

Topic-1 Energy	Topic-2 International society	Topic-3 Lives of people	Topic-4 Internal society	Topic-5 Hi-tech
energyefficiency	merge	nan	people	las
watts	exports	music	way	handsets
currentaccount	per	exhibition	get	vegas
volts	china	theatre	long	server
citigroup	cent	dance	work	ces
		……		

Therefore, we have three observations below:

- The model did impressively well in extracting the unique topics in the data set which we can confirm given we know the target names.
- The model runs very quickly. I could extract topics from data set in minutes.
- It does assume that there are distinct topics in the data set. So if the data set is a bunch of random tweets than the model results may not be as interpretable.

4.2.2.GMM

Then we need to filter important verbs in different clusters to simplify our graph and make it more interpretable. We firstly use a pattern of noun-verb-noun to abstract the stem of the sentence, and vectorize the verbs in their original corpus with word2vec considering its supervised CBOW structure that may be helpful in expressing the relationship between words.

The vectors were then put into Gaussian Mixture Model, which measure the distance between words not only with their geographic distance but also according to their distribution. We chose the number of clusters with the index of silhouette coefficients. So we are able to figure out that the differences in numbers of words are not so large between them. We put the data under the clusters as their child nodes.

The parameters of Word2Vec is set as below:

Table 2. **The parameters of Word2Vec**

name	value
Vector_size	100
Window	5
Min_count	1
Iteration	200

In the table, the vector_size means the size of the vector gotten from Word2Vec, window means the size of the moving window, Min_count means the least time word appears. After 200 iterations, our model got loss function of 0.092.

4.2.3.Knowledge Graph Generalization

Our knowledge graph finally generalized by connecting the nodes of clusters, verbs, news, subjects, objects together. The model shows that the verbs is able to represent an event while an event is not only a verb, which is the same as human logistic.

Meanwhile, the news nodes is able to block the verb and its subjects and objects, considering that it is quite easy to judge a verb according to the subjects and objects and that the deeper the searched nodes are, the specific the events we are able to find. So there are two kinds of methods to search things in our knowledge graph: "before news" and "after news".

For example, if the system receives the key words like "cluster-1", it will output the verbs related to this kind of event. That is the method of "before news".

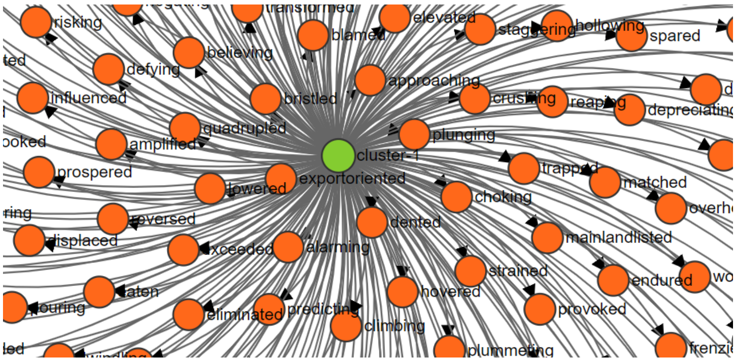

Fig 5.**Example for "after news" method**

But if you search key words like "huawei" and "chips", the results will be different.

4.3 **Examples of Application**

Here is an example to illustrate how the KG works. When we searched 'Huawei' and 'chip', these two words were automatically connected through a logic link of 'chips > authority > bankers > Huawei > apple'. The underlying sentence is 'The innovative chip, checked by corresponding authorities, provides a platform for Huawei to adapt itself to both IOS (Apple) and Android together, which is a threaten to the market of smartphones in the US, especially to Apple. The investment banks’ suggestions to invest to electronic stocks can also reflect the situation.’ The subject ‘chip’, the objects ‘authorities’, ’Huawei’, ‘Apple’, and ‘bank’ were clearly identified. Our logarithm thought like human and tried to connect these words within a logic link, which made sense to some extent that the chips for Huawei were somehow controlled through authorities, and Huawei were related to Apple. It might not produce a perfect explanation of the event, but it still provided clues that all these areas were connected.

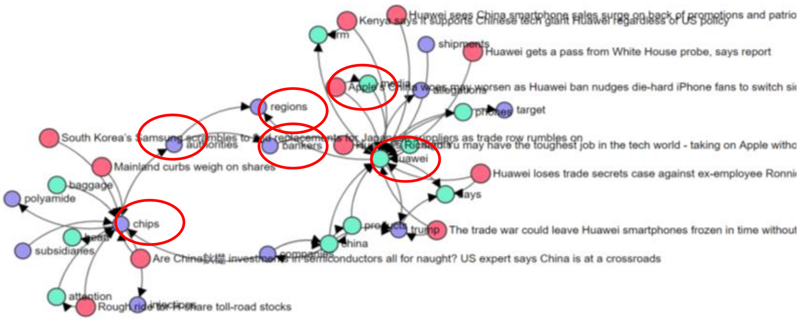
Huawei sees China smartphone sales surge on back of promotions and patrio
Kenya says it supports Chinese tech giant Huawei regardless of US policy
rm
shipments
Huawei gets a pass from White House probe, says report
Apple's China woes may worsen as Huawei ban nudges die-hard iPhone fans to switch si
media
allegations
phones
target
regions
South Korea's Samsung scrambles to find replacements for Japanese suppliers as trade row rumbles on
authorities
bankers
Richard Yu may have the toughest job in the tech world - taking on Apple witho
Huawei
Mainland curbs weigh on shares
Huawei loses trade secrets case against ex-employee Ronni
baggage
polyamide
says
chips
products
trump
The trade war could leave Huawei smartphones frozen in time without
china
subsidiaries
Are China's investments in semiconductors all for naught? US expert says China is at a crossroads
companies
attention
Rough ride for H-share toll-road stocks
injections

Fig 6.**Example of knowledge graph searching**

5 **Limitations and Future Work**

5.1 **Limitations**

1. **Unable to exhaust all unrecognized terminologies**

As we mentioned above, only when we discover the mishandled terminology during the data processing process, we can modify them and thus improve our performance. However, there are so many such unrecognized terminologies that are highly correlated to our topic that we cannot identify all of them.

2. **Invalid Sentence extraction (Entities extraction)**

In the event mining part, we desire to discover events by extracting the main components of the sentence (subject, verb and object). However, our current extraction work is not always valid and it may extract some useless components, which is difficult for us to mine events.

3. **Topic visualization**

The current knowledge graph we obtained has millions of nodes and involves some noisy and unobserved data, so it is hard to find useful information and relationships based on the current graph. Besides, the main idea of building knowledge graph is to go through a sentence and extract the subject and the object when they encounter. However, there are a few challenges— an entity can span across multiple words, for example, “trade war”, and the dependency parser only tags the individual words as subjects or objects.

5.2 **Future Work**

5.2.1.Use more scientific algorithms to improve the performance of sentence extraction

The current logic that we used to extract the primary component of the sentence is not based on the grammar and does not take clauses into account, so one of further improvements is using more scientific algorithms based on English grammar so that it can identify the primary subject, verb and object more precisely.

5.2.2.Improvement of the knowledge graph

The main reason for the messy of the current knowledge graph is that it contains many unrelated information. Our further improvement is to filter some noisy before building knowledge graphs. Besides, extracting the primary components more precisely is also a useful method that can be conducted in the future.

5.2.3.Application of knowlege graph

Due to the time limit, the current complication of our knowledge graph is to discover some event between a series of subjects. Some further analysis can be conducted to use the knowledge more efficiently. One application is causal inference. Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect

6 **Work allocation**

MEI Yudong	Desgning the model, allocating work, designing the codes of crawler, designing the structure of MySQL, integrating the csv into MySQL, constructing the cloud server, designing the framework of experiments, designing the codes of word2vec, designing the codes of GMM model fitting and predicting, designing the codes of building the KG, designing the framework of Tornado of the KG, designing the contents of PPT, presentation, writing the part of 1.2, 1.3, 3.1.2, 3.2.2, 4.2.2, 4.2.3 of the final report.
NIU Yuke	Designing the codes of integrating data in the form of JSON, Integrating the data into csv, writing the part of Data Description of the final report
YUAN Wengao	Designing the codes of Tf-Idf vectorizer, writing the part of 3.1 int the final report, Designing the style of PPT
ZHANG Yanling	Crawling data, Designing the codes of Data Preprocessing, writing the part of 4.1, 5 in the final report, Designing the style of PPT
ZHU Xin	Crawling data, Designing the codes of LDA, writing the part of 3.2.1, 4.2.1in the final report, Designing the style of PPT, Defining the topics
ZOU Xuyuan	Crawling data, Integrating the contents of PPT, Writing the part of 1.1, 4.3, 5