The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Keyword-centered Collocating Topic Analysis

Yu-Lin Chang
National Taiwan University
b06701146@g.ntu.edu.tw

Yongfu Liao
National Taiwan University
liao961120@gmail.com

Po-Ya Angela Wang
National Taiwan University
diff@cmgsh.tp.edu.tw

Mao-Chang Ku
National Taiwan University
d08142002@ntu.edu.tw

Shu-Kai Hsieh
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

The rapid flow of information and the abundance of text data on the Internet have brought about the urgent demand for the construction of monitoring resources and techniques used for various purposes. To extract facets of information useful for particular domains from such large and dynamically growing corpora requires an unsupervised yet transparent ways of analyzing the textual data. This paper proposed a hybrid collocation analysis as a potential method to retrieve and summarize Taiwan-related topics posted on Weibo and PTT. By grouping collocates of 臺灣 'Taiwan' into clusters of topics via either word embeddings clustering or Latent Dirichlet allocation, lists of collocates can be converted to probability distributions such that distances and similarities can be defined and computed. With this method, we conduct a diachronic analysis of the similarity between Weibo and PTT, providing a way to pinpoint when and how the topic similarity between the two rises or falls. A fine-grained view on the grammatical behavior and political implications is attempted, too. This study thus sheds light on alternative explainable routes for future social media listening method on the understanding of cross-strait relationship.

Keywords: Social Media Listening, Collocation Analysis, Grammatical Collocation, Topic Modeling, Unsupervised Methods

1 Introduction

Social Media Listening (Social Media Monitoring) is a modern data science technique to the monitoring of social media. With the advances of NLP and text analytics, social mentions, i.e., keyword of key phrases referring to brand and product name, trending topic, etc., can be constantly tracked and analyzed in real-time.

The potentials of social media listening do not appear explicitly only in commercial and marketing domain, but also in other political agenda, and national security sphere. For instance, Taiwan and PRC are long known for their political rivalry since 1949. The tension of this rivalry changes from time to time and could be observed from news and social media. With the advent of the Internet, information flows instantly on social media. Monitoring corpora could thus be built to record and detect the latest issues hotly discussed on the web. The rivalry between Taiwan and mainland China demands applications that monitor tensions between the two political regimes. Since text data on the web accumulate rapidly, and only a subset of text data is relevant to particular applications, various methods need to be deployed to retrieve relevant information from the texts in an unsupervised yet transparent manner. One of such methods is collocation analysis.

Collocation analysis has been adopted in studies about how "Muslim" is represented in news media (Li and Zhang, 2021; Baker et al., 2013). These studies have shown that implicit political images of "Muslim" in news media can be revealed by analyzing the word choices, or collocates, around the target word. Collocates are not random companies but indicators of the context of the target word since collocations are the result of "mutual expectations" (Firth, 1957) between the two words. This expectation includes compatibility between the two units in grammatical aspects, semantic aspects, and knowledge about reality. Collocates can be further categorized according to their semantic information (Li and Zhang,

2021; Baker et al., 2013), and can be used to understand how a concept is described in the media. The original concept of collocation, which only captures associated word pairs occurring close in positions, could be extended. Grammatical collocations capture associated word pairs of a particular grammatical relationship. For example, Pearce (2008) has identified that "woman" often plays as the subject of verbs about annoyance.

In this study, we explore the potential of leveraging (grammatical) collocation analysis to monitor Taiwan-related topics that are posted on social media in mainland China and Taiwan. As a preliminary step, we compare text data collected from two representative social media, Weibo and PTT, aiming to provide a sketch of the differences and similarities between the two sources.

2 Data

To monitor changes over time, we construct two comparable corpora from Weibo and PTT respectively, using web crawlers to collect posts published between 2020-05-01 and 2020-10-01 (ranged 153 days). Posts with the word 臺灣 'Taiwan' and its form variants on Weibo and PTT are extracted respectively with weibo-search[1] and PTT corpus (Liu, 2014). Since the data collected from PTT is larger than that of Weibo, we balance the size of the two corpora by reducing the size of PTT corpus with random sampling of posts. Then, each of the corpora is split into nine time-sliced subcorpora, with each subcorpus containing post data spanning about 17 days. The resulting time-sliced subcorpora each contains about 0.4 to 1.25 million tokens. The corpora are word segmented with jieba[2]. After word segmentation, simplified Chinese are converted to traditional Chinese via OpenCC[3]. In addition, usage variations between Taiwan and mainland China, such as 台灣/臺灣, are normalized. These corpora are then used for collocation extraction and dependency relation parsing, which are described in Section 3 and 4 respectively.

[1]https://github.com/dataabc/weibo-search
[2]https://github.com/fxsjy/jieba
[3]https://github.com/BYVoid/OpenCC

3 Exploring Collocating Topic of "Taiwan"

In this study, we are interested in exploring potential methods that could be applied to monitor topics discussed on different social media sources. As an exploratory step, we extract collocates of the term 臺灣 from different times on Weibo and PTT. This would allow us to compare these collocates across different dimensions (time and sources). These collocates could be seen as hints that provide information about what is being discussed about Taiwan. In addition to extracting collocates, we also need a (semi-)automatic method to cluster these collocates into meaningful groups to better interpret the results. To achieve this, we explored two different methods—the first uses word embeddings of the collocates to cluster them into discrete groups, and the second uses Latent Dirichlet allocation to derive topics from the corpora such that each collocate could be given a vector of weights across topics. Section 3.2 and 3.3 further describe the two methods respectively. Section 3.1 describes the procedure of collocation extraction.

3.1 Collocation Extraction

We focus here only on word pairs occurring in a running window of two (i.e., bigrams) as candidates of collocations. Since we are interested only in collocates of 臺灣, collocation extraction here is equivalent to finding word pairs containing 臺灣 and showing strong associations. To measure the strength of association between word pairs, we adopt a measure known as Log Likelihood (Dunning, 1993). Log Likelihood has several advantages over another widely used measure, mutual information (MI), in that (1) it is less subject to low-frequency bias, (2) it takes sampling variations into account, and (3) it best approximates Fisher's exact test, which is considered the most appropriate significance test for collocation contingency table (Evert, 2009). Equation (1) below shows how Log Likelihood and MI are calculated for a word pair. O_{ij} corresponds to the observed frequency and E_{ij} to the expected frequency of a cell in the contingency table. MI simply measures the log ratio of the observed frequency of a word pair (O_{11}) to its expected frequency (E_{11}). Log Likeli-

hood takes into account all four cells in the contingency table.

$$Log\ Likelihood = 2\sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}}$$
$$MI = log_2 \frac{O_{11}}{E_{11}} \quad (1)$$

By using the Log Likelihood measure (a.k.a, G^2), we extract 20 collocates of 臺灣 that have the largest G^2 values for each of the time-sliced subcorpus, resulting in 18 lists of collocates. In the next two sections, we describe how we convert these lists of collocates into probability distributions over topics.

3.2 Deriving a distribution of topics with word embeddings

We use the Weibo corpus and the PTT corpus to obtain three sets of word embeddings. The first and second sets were derived from the Weibo and PTT corpus respectively, and the third set is derived from data combined from both sources. The word embeddings were trained using the Word2Vec algorithm (Mikolov et al., 2013) implemented in gensim (Řehůřek and Sojka, 2010). The hyperparameters for the three sets are set to identical values—the window size is set to 5, the minimum frequency of occurrence is set to 5, and the resulting dimension of the word vectors is set to 100. Since during training, initializations of the Word2Vec models are random, the resulting vector spaces of the models need to be rotated and aligned before one can compute semantic distances across different models. We used orthogonal Procrustes, introduced in Hamilton et al. (2016), to align the first (Weibo) and second (PTT) model against the third model (Weibo + PTT).

We then assign word vectors from the Weibo and PTT models to the collocates[4] extracted from the time-sliced subcorpora. The vectors of the collocates could then be treated as a matrix. Principal component analysis is performed on the matrix to reduce the dimension of the word vectors from 100 to 4. K-Means clustering is then performed on the matrix to cluster the collocates into discrete groups. We tested different cluster numbers (k) by calculating the resulting *inertia* of each clustering. *Inertia* is defined as the sum of squared distance of the data points to their closest cluster center. Since there is an inverse relationship between *inertia* and k, and since we want both of them to be low, we adopt the elbow method by plotting *inertia* against k and look for the place where the decrease in *inertia* starts to slow down. Using this method, we arrive at an optimal k of 10.

Given these clusters of collocates, we could then derive a frequency distribution from a list of collocates. The idea is to assign each collocate in the list to its belonging cluster in order to obtain a frequency distribution of the clusters. Figure 1 contrasts the distribution of collocate clusters between Weibo and PTT. The cluster labels are generated by using the six closest words to the cluster center in each cluster. The distributions in Figure 1 leave out the time dimension and use collocates across all time steps to provide an overview of the topics discussed on Weibo and PTT. Contrasts considering the time dimension are discussed in Section 3.4, in which we will also consider distributions derived from another method—Latent Dirichlet allocation (Blei et al., 2003). We describe how distributions are derived with Latent Dirichlet Allocation next.

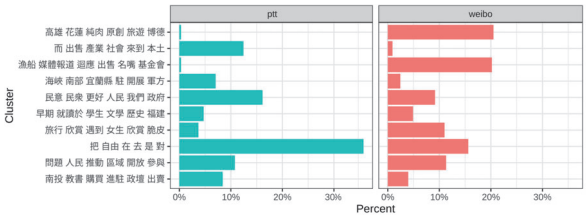

Figure 1: Distribution of collocate topics derived from word embeddings.

3.3 Deriving a distribution of topics with Latent Dirichlet Allocation

Topic modeling is a good unsupervised choice to explore unstructured data. Various topic models have been developed (Zhao et al., 2015), including Latent Semantic Indexing (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation (LDA) is one of the models popularly employed in mod-

[4] A word occurring in two different sources is treated as two different collocates.

eling topics of unorganized textual data. The purpose of adopting Latent Dirichlet Allocation (LDA) as another method to derive topic distributions is that, compared to "hard" clustering, LDA is more flexible in how it can be used to assign topics to a given word. Since a topic is a mixture of words weighted by probabilities, we can go in another direction and define a word as a mixture of topics, using relative probabilities of the word across all topics to derive a distribution.

Thus, we trained a single LDA model with all of the corpus data (Weibo + PTT). Words with frequencies below 5 are discarded and the number of topics is set to 20[5]. After obtaining the topics, two annotators independently assigned labels to the topics. Inconsistency between assigned labels were further discussed and resolved. The resulting labels for the twenty topics are shown in Table 1.

Given the topics generated from LDA, we are able to assign each occurrence of a word in the corpus a probability distribution. The remainder of this section describes the rationale behind this procedure. Suppose that we would like to assign a vector of topic probabilities to a word w_i, in vector form, this would be:

$$\begin{aligned}&\left[p(T_1|w_i), p(T_2|w_i), \cdots, p(T_{20}|w_i)\right] \\ &= \left[\frac{p(T_1 \cap w_i)}{p(w_i)}, \frac{p(T_2 \cap w_i)}{p(w_i)}, \cdots, \frac{p(T_{20} \cap w_i)}{p(w_i)}\right]\end{aligned} \quad (2)$$

Since we are interested in arriving at a probability distribution (i.e., a vector summing to one), we can discard the denominator $p(w_i)$, which gives us:

$$\left[p(T_1 \cap w_i), p(T_2 \cap w_i), \cdots, p(T_{20} \cap w_i)\right] \quad (3)$$

From the LDA model, we can obtain $p(w_i|T_j) = p(T_j \cap w_i)\ p(T_j)$. Plugging this into the equation gives us:

$$\left[p(w_i|T_1)p(T_1), p(w_i|T_2)p(T_2), \cdots, p(w_i|T_{20})p(T_{20})\right] \quad (4)$$

Finally, making the assumption that the marginal probability of all topics are equal $p(T_1) = p(T_2) = ... = p(T_{20})$ allows us to arrive at

$$\left[p(w_i|T_1), p(w_i|T_2), \cdots, p(w_i|T_{20})\right] \quad (5)$$

Normalizing the vector above such that it sums to one would then give us the probability distribution we want for each occurrence of a word. Thus, given a list of collocates with their frequencies of occurrence, we can sum the distribution derived from each occurrence together to arrive at a distribution of topics. For instance, Figure 2, which can be compared to Figure 1, contrasts the distribution of collocate topics between Weibo and PTT by adopting LDA to derive the distributions.

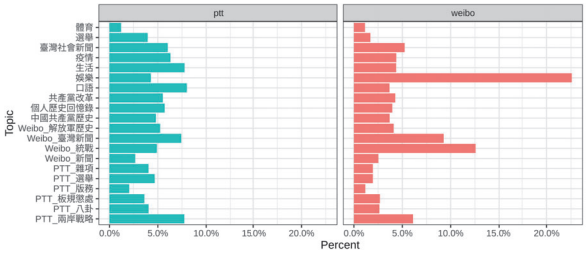

Figure 2: Distribution of collocate topics derived from Latent Dirichlet allocation.

3.4 Quantitative analysis of collocate topic distributions

Based on the distribution extraction methods described in previous sections, we are able to quantify the distance (or similarity) between Weibo and PTT across different dimensions. One of such dimensions is *time*. By splitting the corpus into time-sliced subcorpora, we could retrieve a list of collocates for each time step. These lists of collocates are then converted to distributions and thus can be compared directly using quantitative metrics.

For the first analysis, we look at topical variations *within* a source. This analysis allows us to notice, for instance, whether there are particular times when the topic distribution deviated from the grand mean, either on Weibo or PTT. To achieve this, for each of the nine time steps and each of the sources (Weibo and PTT), we extract 20 collocates of 臺灣. This results in 18 distributions (9 for PTT and 9 for Weibo). The mean distribution for each of the sources is calculated by summing up the

[5]The number of topics (k) is determined through the elbow method described in Section 3.2. We test LDA with different values of k and compute corresponding values of perplexity and arrive at an optimal k of 20 with the elbow method.

Table 1: Labels assigned to the 20 topics generated by Latent Dirichlet allocation.

Topic	Translation	Topic	Translation
PTT 選舉	PTT election	PTT 版務	PTT affairs
口語	Spoken terms	PTT 板規懲處	PTT Regulations
中國共產黨歷史	History of Chinese Communist Party	生活	Daily life
Weibo 統戰	Weibo Cross-Strait unification	選舉	Election in Taiwan
Weibo 臺灣新聞	News about Taiwan on Weibo	Weibo 新聞	News on Weibo
PTT 兩岸戰略	Cross-Strait Military strategy	娛樂	Entertainment
臺灣社會新聞	Social News in Taiwan	PTT 八卦	PTT Gossiping
共產黨改革	Chinese Communist Party reform	個人歷史回憶錄	Personal memoir
Weibo 解放軍歷史	History of People's Liberation Army	體育	Sports
PTT 雜項	Miscellaneous topics on PTT	疫情	Coronal virus

nine time-step distributions and normalizing it to a probability distribution. The distance between a particular time step and the mean then is defined as the cosine distance between the two distributions:

$$CosDist = 1 - CosSim(Distr_t, Distr_{mean})$$
$$CosSim(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \quad (6)$$

This would give us nine distance measures for each source, indicating how deviated the topic discussed is from the mean in a particular time step. Figure 3 shows the results derived from both word embedding clustering (Section 3.2) and LDA (Section 3.3). A quick glance at the plots shows that topic variations on Weibo are larger compared to PTT. This seems to result from certain topics bursting in particular time steps on Weibo, as evident in Figure 4.

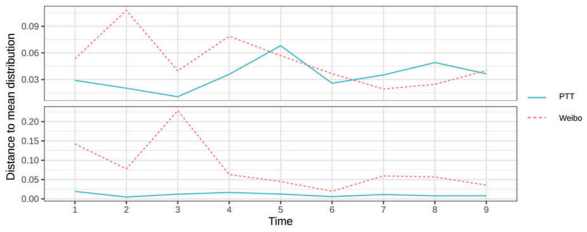

Figure 3: Topical variation with a source. The plot on the top results from adopting word embedding clusters to derive distributions for distance calculations; the plot on the bottom results from adopting LDA to derive distributions for distance calculations.

The second analysis directly compares Weibo and PTT across time. With the same set of distributions, we could calculate the similarity between Weibo and PTT for each time step. This allows us to notice, for instance, when the topics discussed are similar and when the topics seem to deviate between the two sources. The results are shown in Figure 5.

4 Dependency Relations

To provide a finer-grained view of discussions of Taiwan-related topics, we resort to a more linguistically-involved dependency relation analysis. As mentioned in Section 3.1, it has been widely attested that Log Likelihood (G^2) is a better association measure of collocates, as it always provides higher precision values at the same recall percentage. However, it is also found that for some specific types of grammatical collocation, other measures might work better (Uhrig et al., 2018). For instance, while G^2 provides the best performance for almost all relations (such as verb-object, adverb-adjective, etc), `t-score` surpasses it for adjective-noun collocation. Based on BNC, Uhrig et al. also did a comprehensive evaluation of different dependency parser and annotation schemes as a filter on the collocation candidate extraction task, and show that SpaCy (Honnibal et al., 2020) is a robust parser with good results on all grammatical relations. So in this study, we utilize SpaCy as the syntactic dependency parser to further extract grammatical collocations of the verb-directObject pairs, with all verbs that take 臺灣 'Taiwan' as the direct object in both PTT and Weibo subcorpora across different time-sliced periods.

For the sake of brevity, we only consider verbs that consist of two or more characters, and the minimum frequency is set to 2. After collecting the verbs, we compute G^2 values for every verb, extract the top 10 verbs that have the highest G^2 values in each time step, and

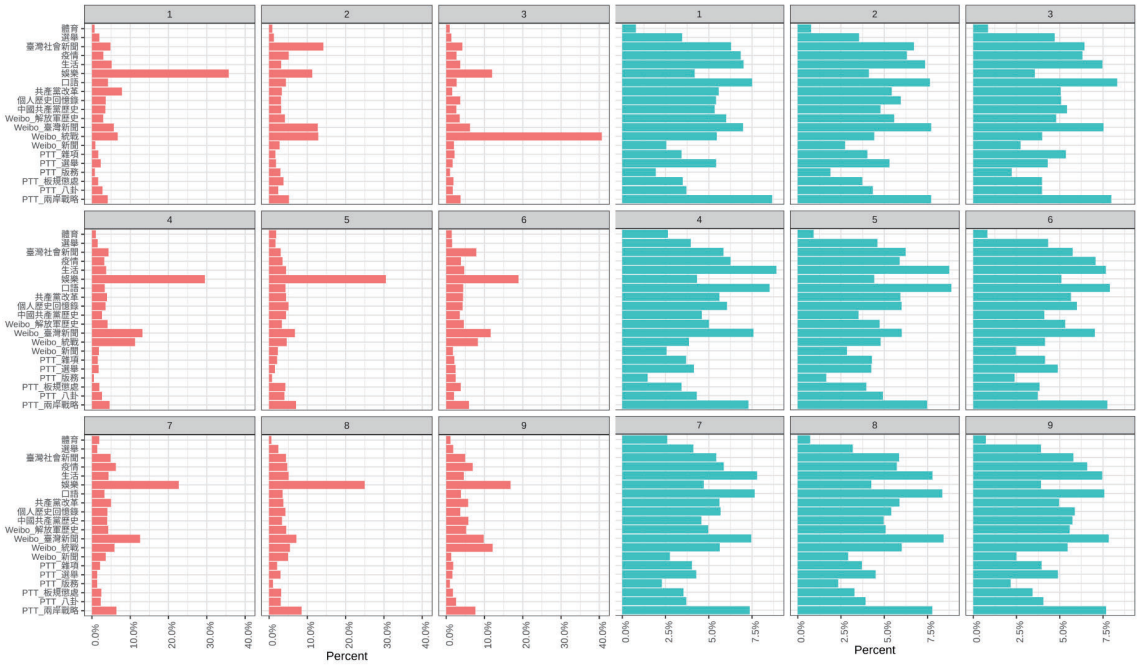

Figure 4: Distributions of topics across time using topics derived from LDA. The left part plots the results of Weibo and the right part plots the results of PTT.

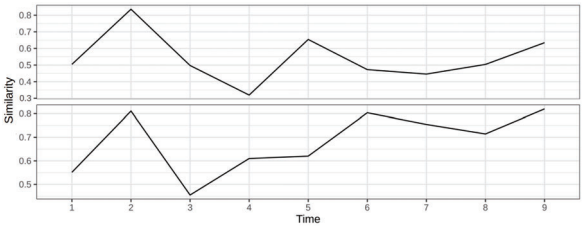

Figure 5: Similarity of topic distributions between Weibo and PTT across time. The top row results from word embedding clustering; the bottom row results from LDA.

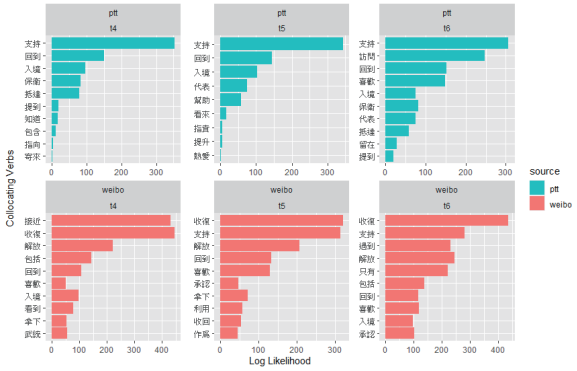

Figure 6: Collocating Verbs of the Direct Object 臺灣 'Taiwan'

finally rank them in descending order. The charts in Figure 6 demonstrate the results of dependency relations from time step 4 to 6 in the two corpora respectively. It is shown that compared with PTT, Weibo users frequently choose military terms such as 收復 'recapture' and 解放 'liberate'. PTT users, in contrast, consistently use Taiwan-oriented verbs, e.g. 入境 'enter', 回到 'return', and 抵達 'arrive'.

5 Discussion

In Section 3.2 and 3.3, two different methods are used to derive collocate topic distributions of 臺灣 'Taiwan'. The two methods each have advantages and drawbacks. The word embedding clustering method is fully automatic, in that labels of the clusters are generated from picking out the words that are closest to the cluster center. In addition, compared to LDA, word embedding clustering is more computationally efficient, which may be a crucial property for developing large-scale monitoring applications since the corpus data are subject to frequent updates in such applications. On the other hand, results generated from the LDA method match human intuition more since the labels of the topic are manually given and word topic distributions can be interpreted easily due to inherent properties of LDA. Below, we focus the discussion on the results of distributions derived from LDA across time.

The distributions of topics across nine time

steps on Weibo and PTT are summarized in Figure 4. The most prominent distinction between the two sources is that distributions from PTT are flat whereas distributions from Weibo often peak in particular topics. This shows that in general, the focuses of discussion of Taiwan on PTT seems to be more widespread, whereas, on Weibo, discussions of Taiwan are focused on particular topics such as Entertainment (time step 1 and 4~9), Cross-Strait unification (time step 2~4 and 9), and News about Taiwan (time step 2, 4, 7). The focus-shifting on Weibo seems to be the primary force that drives the fluctuation in the similarity between Weibo and PTT. As evident in the bottom part of Figure 5, the sharp drop in similarity at time step 3 corresponds to Weibo's extreme focused discussion about Cross-Strait unification, which is estimated to account for roughly 40% of the top-ranked collocates of Taiwan. On the other hand, the peaking of similarity at time step 2, 6, and 9 corresponds to a flatter distribution of topics on Weibo, where the focuses of the discussion seem much less prominent.

In addition to utilizing simple co-occurrence in texts to extract collocates (Section 3), incorporating grammatical relations also shows promising results. In Section 4, by narrowing down the collocates of Taiwan to verbs only, we are able to see some interesting contrasts between Weibo and PTT. Since a verb often expresses *intentions* toward its object, we could immediately spot that most of the top collocating verbs of Taiwan on Weibo exhibit intentions to assault, whereas such intentions are absent in the top collocating verbs on PTT. This points to a direction for future research, in which clustering methods (Section 3.2 and 3.3) may be applied to these collocating verbs, and the annotation of the resulting clusters could focus on the intentions of the verbs. This could potentially facilitate the development of monitoring applications that focus more on sensitive topics (e.g., whether there may be a rise in political or military tensions between mainland China and Taiwan).

6 Conclusion

In this study, we adopted collocation analysis as a method to explore Taiwan-related topics posted on social media. To summarize a large number of extracted collocates, we utilized clustering and topic modeling to derive a probability distribution over topics from a list of collocates. Similarities between Weibo and PTT across time were characterized in terms of these distributions. Finer-grained analysis of grammatical collocates hints on future work to apply clustering methods to verbal collocates, which may reveal details such as intentions toward the object.

References

Paul Baker, Costas Gabrielatos, and Tony McEnery. 2013. Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998—2009. *Applied Linguistics*, 34(3):255–278.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(null):993–1022. Number of pages: 30 Publisher: JMLR.org tex.issue_date: 3/1/2003.

Stefan Evert. 2009. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Volume 2*, pages 1212–1248. De Gruyter Mouton.

John Rupert Firth. 1957. *A Synopsis of Linguistic Theory, 1930-1955.*

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Ke Li and Qiang Zhang. 2021. A corpus-based study of representation of Islam and Muslims in American media: Critical Discourse Analysis Approach. *International Communication Gazette*, page 1748048520987440. Publisher: SAGE Publications Ltd.

Tsun-Ju Liu. 2014. PTT Corpus: Construction and Applications. Master's thesis, National Taiwan University, Taipei, Taiwan, January.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Michael Pearce. 2008. Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, 3(1):1–29.

Peter Uhrig, Stefan Evert, and Thomas Proisl. 2018. Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes. In Pascual Cantos-Gómez and Moisés Almela-Sánchez, editors, *Lexical collocation analysis: Advances and applications*, pages 111–140. Springer, Cham.

Weizhong Zhao, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13):S8.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop new challenges for NLP frameworks*, pages 46–50, Valletta, Malta. University of Malta.