VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Investigating Word Association Mining Techniques

Duygu BAĞCI DAŞ[1]*, Sevdanur GENÇ[2]

[1]*Ege University, Ege Vocational School, Department of Computer Technologies, İzmir*
[2]*Kastamonu University, Taşköprü Vocational School, Department of Computer Technologies, Kastamonu*

Abstract

This study presents the investigation of the effect of conditional entropy, mutual information (MI) values, log-likelihood ratio (LLR), and simple co-occurrences on extracting strong syntagmatic relationships. Experiments are conducted by using the Yelp Academic Dataset, which includes extracted 10.000 restaurant reviews. The mutual information values of word pairs are considered to extract the top syntagmatically related words from the corpus. For this purpose, Spyder 3.3.6 and Python Natural Language Toolkit (NLTK) Library are used. The mutual information values are then compared with simple co-occurrences count. The analysis results indicated that the three Word collocation techniques give similar results and therefore, all of those can be employed for Word collocations effectively.

Keywords: *Word Collocation, Collocation Mining, Collocation extraction, Mutual Information, Text Mining*

Makale Bilgisi

Başvuru:
07/10/2022
Kabul:
21/11/2022

Kelime Birliktelik Madenciliği Tekniklerinin İncelenmesi

Özet

Bu çalışma, koşullu entropi, ortak bilgi (MI) değerleri, log-birliktelik oranı (LLR) ve basit ortak oluşumların güçlü sözdizimsel ilişkilerin çıkarılması üzerindeki etkisinin araştırılmasını sunmaktadır. Deneyler, 10.000 restoran yorumunu içeren Yelp Akademik Veri Kümesi kullanılarak gerçekleştirilmiştir. Ortak bilgi değeri en yüksek sözcük çiftlerinin, söz dizimsel olarak ilişkili en üstteki sözcükleri derlemden çıkardığı kabul edilir. Bu amaçla Spyder 3.3.6 ve Python Natural Language Toolkit (NLTK) Library kullanılmıştır. Ortak bilgi değerleri daha sonra basit ortak oluşum sayısı ile karşılaştırılır. Analiz sonuçları, üç farklı kelime eşdizimleme tekniğinin benzer sonuçlar verdiğini ve bu nedenle, bunların hepsinin kelime eşdizimleri için etkili bir şekilde kullanılabileceğini göstermiştir.

Anahtar Kelimeler: *Kelime birlikteliği, Birliktelik madenciliği, Eşdizim çıkarma, Ortak bilgi, Kelime birliktelik oluşumu*

* İletişim e-posta: duygu.bagci.das@ege.edu.tr

1 Introduction

Word association is denoted as the relationship between words. The relation is examined in two categories namely, Paradigmatic and Syntagmatic. If the words are closely related to each other (i.e., they are in the same class), the relationship is Paradigmatic. If the words are able to be combined with each other, the relation becomes Syntagmatic. Word association techniques can be used to predict and analyze customer behavior to constitute a recommendation system for the customer [1]. In the literature, some of the different studies published in the last 30 years on Word Association Mining Techniques have been examined.

Church and Hanks estimated the word association norms from computer-readable corpora by using an objective measure based on the information-theoretic notion of mutual information [2]. Damani studied the source of the improvement of the performance of pointwise mutual information by investigating the co-occurrence levels (corpus and document). According to the outcomes, the corpus level significance was responsible for the improvement, whereas the document level had no impact on performance [3]. Jain and Pandey proposed a sentiwordnet-based algorithm to find the polarity of a given sentence more efficiently. Their work, called Sentiwordnet, is the main tool for calculating the score of a particular word in a sentence, as well as taking into account words that somehow influence it. Thanks to the developed algorithm, successful results were obtained on randomly selected normal input sentences [5]. In the study of Xu et al. they proposed a new method to determine the semantic orientation of subjective terms to perform sentiment analysis. The method adopts a classification approach based on a new semantic orientation representation model called S-HAL (Emotion Hyperspace Analog to Language). It basically generates a set of weighted features based on surrounding words and characterizes the semantic orientation information of words through a specific feature space. This method, which performed well, was able to quickly and accurately identify the semantic orientation of terms without using an Internet search engine [6]. Khan et al. presented SentiWordNet, which is a labeled corpus for training. They define SentiMI as a sentiment dictionary based on mutual information values. They developed a complete framework by utilizing feature selection and extracting mutual information via SentiMI for chosen features. Their investigation comprises a large dataset of 50.000 movie reviews [4]. Garrett et al. presented a study to assess population sensitivity to disaster relief efforts immediately after Hurricane Maria in 2017. They leveraged geo-located Tweets from Twitter in Puerto Rico and used a general purpose Multi Perspective Question Answering (MPQA) dictionary and a common word polarity scoring method to extract sense analysis of each Tweet. They also used measurement techniques such as Pointwise Mutual Information (PMI) and Mutual Information (MI) in their studies. They compared the sentiment results using MPQA' with MPQA, the results showing that MPQA detected a significantly higher number of negative Tweets. They observed that the number of negative Tweets identified by the MPQA' was much closer to human-verified results [8]. Kang examined the relationship between grammar and language use by comparing word association and collocation. Among the measures of collocation, the (simple) log probability and t-score were more consistent with association with leading log probability by a small margin than MI or MI3. Among the collocation measures, log likelihood (simple-ll) found word association closest to duplication, with the t-score trailing by a small margin, while MI had the worst outcome, especially for higher-frequency stimulus words. In general, he predicted that word association and collocation were quite close, but not exactly close due to differences in related sources and characteristics of lexical/semantic relationships [9]. Lai examined the use of an ethnic term in news discourse from linguistic, discursive, and social-cultural perspectives and used the point mutual information (PMI) method. The results showed diversified distributions of collocations according to frequency, distance, and semantic connections. The findings show that some collocations occur with high frequency and show strong semantic associations; some occur over a long distance and have shown strong semantic connections; others occurred at a high frequency but over a long distance and showed weak semantic connections; still others showed stronger semantic connections but occurred at a low frequency and over a long distance. In short, some variations were observed showing interesting correlations [10]. Liu et al. developed the distributional semantics-based collocation extraction method by introducing collocation models in both the candidate discovery stage and the candidate filtering stage. At the same time, they have made improvements in bigram noise filtering. They specified four different methods to take full advantage of the complementarity between them. They improved the multi-gram collocation extraction performance of the system. They incorporated the collocation framework into the system and recursively extended the bigram collocation subtraction results according to certain collocation rules. Experimental results have shown that the proposed method does a very good job of extracting multigram collocations and shows significant improvement in all values such as metrics, precision, recall, and f-value compared to the baseline [11]. Krenn developed computational linguistic methods and tools for determining collocations from arbitrary text, and methods and tools for representing collocations in a relational database integrating competence (collocation type-specific linguistic analysis) and performance information (clause sentences). They reported that PP-entropy is a good alternative to association criteria for identifying FVG and figurative expressions from high and medium frequency full-form data, and also defining FVG from high frequency fundamental form data and medium-frequency fundamental form data [13]. Williams has published a study on Doubtful Coincidences and Point Mutual Information. He showed that when marginal effects are removed, MI and PMI behave similarly to Y as functions of λ. Point reciprocal information has been widely used in some research communities to flag suspicious coincidences, but highlighted the importance of keeping in mind the sensitivity of PMI to marginals, with increasing scores for less frequent events. He considered crossover information and point crossover information, along with their normalized versions, as association criteria [12]. Zhang et al. proposed a method for constructing a corpus in the field of electric power based on multi-method collaboration. With the Jieba word segmentation method, they aimed to eliminate the disadvantage of the word segmentation results being excessively small, and they used the TF-IDF method to extract keywords from the Jieba word segmentation results. At the same time, the entropy word segmentation method for Information and the rule of forming strict phrases that can reduce the number of words created is used. Compared with Jieba word segmentation method, the information entropy word combination algorithm (IEWCA), information entropy word segmentation algorithm (IEWSA), experimental results, and richer vocabulary has proven to be more successful [14].

In general, according to literature, using the lexical resource and pointwise mutual information is the most efficient way for word association [6, 7].

This study presents the investigation of the effect of conditional entropy and mutual information values on extracting strong syntagmatic relationships. The study also takes the effect of the LLR, and simple co-occurrences values on the extraction of syntagmatic relations into account.

2 Material and Method

In this study, a dataset, which includes 10000 restaurant reviews (1.043.05 words) and belongs to Yelp Corpus [15], is used. A sample of the dataset is given in Fig.1. In the pre-processing stage, words, which exist in the English stop words list of Python NLTK and have three or fewer strings, are removed.

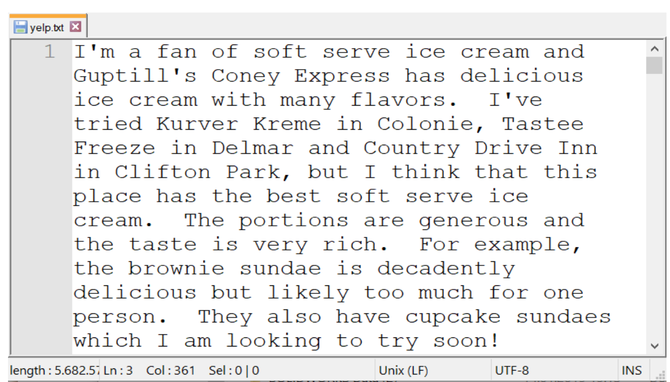

Figure 1. Yelp Restaurant Review Dataset[15]

Text mining applications often require processing on unstructured data. To make sense of unstructured data, the data needs to be made workable. The flowchart in Fig.2. shows the operations to make the data workable.

It was processed with the Corpus model with the data obtained from the Yelp Restaurant Review Dataset. More than a million words of data have been reviewed 10,000 times. Then, for the data preparation stage, the process of separating each word that makes up a whole text was carried out. For this, Line Split was used in the tokenization process. With this process, the text is fragmented as desired and saved in arrays. Because texts are often broken-down word by word.

In the first step of the data preprocessing step, to extract punctuation marks and numbers from the text; "Removing punctuation and digits" operations were applied. In order to discard the words in the text that do not make any changes in the meaning; "the Removing stop words" operation has been applied.

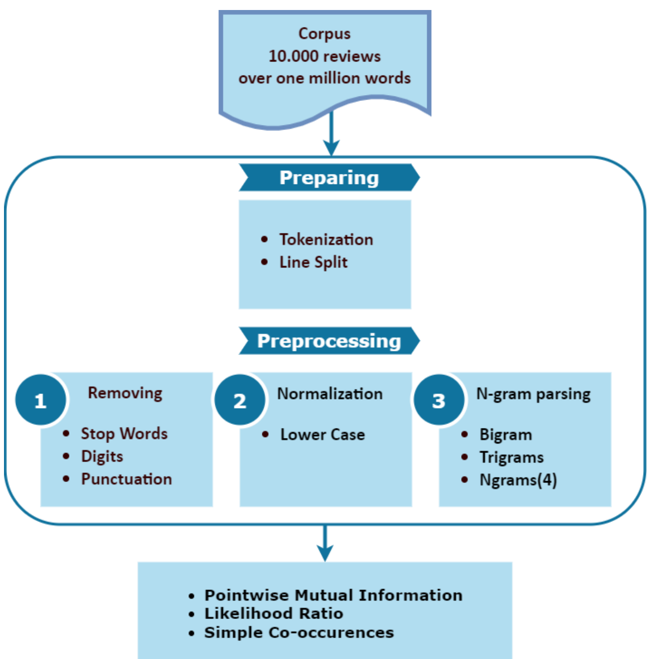

Figure 2. Flow Chart of NLTK Word-Collocation Analysis

In the second step of data preprocessing, the Normalization process is applied. The reason is that the different formats and erroneous discourses of the texts need to be converted into a canonic (standard) format. For this, "lower-case" is used.

In the last step of data preprocessing, the N-gram decomposition algorithm is used. N-gram is a general name for sequential arrays of n elements. In the context of natural language processing and computational linguistics, the elements that makeup n-grams can be selected as words, syllables, phonemes, or letters in a spoken text or written text, depending on the need and application area. It is usually selected from among the models also known as corpus. A Bigram is a special variant of n-gram. Some n-grams are given special names according to the size of the n-number, bigram is one of these names. In this study, Bigram, Trigrams, and Ngrams(4) parsing algorithms are used.

In the study, Pointwise Mutual Information was applied to measure the probability of two words occurring together, taking into account the fact that it may be caused by the frequency of single words.

There are several approaches to calculating word association values. Approaches such as Chi-Square (x’), Dice, Jaccard, Log Likelihood Ratio (LLR), Pointwise Mutual Information (PMI), and T-Test are given in Fig.3.

Measure	Definition
Chi-Square(χ^2)	$\sum_{\substack{x' \in \{x, \neg x\} \\ y' \in \{y, \neg y\}}} \frac{\left(f(x',y')-Ef(x',y')\right)^2}{Ef(x',y')}$
Dice	$\frac{2f(x,y)}{f(x)+f(y)}$
Jaccard	$\frac{f(x,y)}{f(x)+f(y)-f(x,y)}$
Log Likelihood Ratio(LLR)	$\sum_{\substack{x' \in \{x, \neg x\} \\ y' \in \{y, \neg y\}}} p(x',y') log \frac{p(x',y')}{p(x')p(y')}$
Pointwise Mutual Information(PMI)	$log \frac{f(x,y)}{f(x)*f(y)/W}$
T-test	$\frac{f(x,y)-Ef(x,y)}{\sqrt{f(x,y)\left(1-\frac{f(x,y)}{W}\right)}}$

W	Total number of tokens in the corpus
$f(x), f(y)$	unigram frequencies of x, y in the corpus
$p(x), p(y)$	$f(x)/W, f(y)/W$
$f(x,y)$	Span-constrained (x, y) word pair frequency in corpus
$p(x,y)$	$f(x,y)/W$

Figure 3. Definition of some co-occurrence based word association measures [3]

2.1 Entropy

Entropy is a mathematical concept that enables us to calculate the randomness of a variable. The lesser the entropy of a word, the lesser randomness it has. Hence, it has more significance in word relations. Eq.(1) represents the entropy of a word in a text segment. $H(X_{w1})$ denotes the entropy of the word w1, and p(x) is the probability of the existence of the word w1.

$$H(X_{w1}) = -\sum_{u\in[0,1]} p(x)log_2p(x) \tag{1}$$

2.2 Conditional Entropy

Conditional entropy, which is given in Eq.(2), gives information about whether a word pair tend to be together. Besides, it also makes us understand whether these words represent a stronger meaning than when they are individuals. $H(X_{w1}|Y_{w2})$ represents the conditional entropy, and p(x,y) is the conditional probability of the existence of the words w1 and w2.

$$H(Y_{w2}|X_{w1}) = -\sum_{u\in[0,1]} p(u)H(Y_{w2}|X_{w1}=u) = -\sum_{u\in[0,1]}\sum_{v\in[0,1]} p(u,v)log_2p(u,v) \tag{2}$$

2.3 Mutual Information

Mutual Information (MI) is a standard method to extract the Word Association. Even though there are several mathematical formulations to calculate MI, Eq.(3) is used within the scope of this study. I (Y|X) denotes the mutual information value between w1 and w2.

$$I(Y_{w2}|X_{w1}) = H(X_{w1}) - H(X_{w1}|Y_{w2}) = H(Y_{w2}) - H(Y_{w2}|X_{w1}) \tag{3}$$

To compute mutual information, we often use a different form of mutual information that we can mathematically rewrite as Eq.(4) [7].

$$I(Y_{w2}|X_{w1}) = -\sum_{x\in u}\sum_{y\in v} p(X_{w1}=u, Y_{w2}=v)log_2\frac{p(X_{w1}=u, Y_{w2}=v)}{p(X_{w1}=u)p(Y_{w2}=v)} \tag{4}$$

Fig.4 represents the simple co-occurrences of w1 and w2.

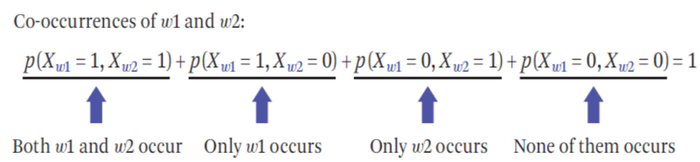

Figure 4. Simple co-occurrences of w1 and w2 [1]

The probability of the existence of the related words are calculated by using Eq.(5).

$$p(X_{w1}=1) = \frac{count(w1)}{N} \tag{5}$$

3 Results

The mutual information values are calculated by using Eq.(4) by defining the window size as 4. Smoothing is applied by using Eq.(6) to accommodate zero counts.

$$p(X_{w1}, Y_{w2},) = \frac{count(w1,w2)+0.25}{N+1} \tag{6}$$

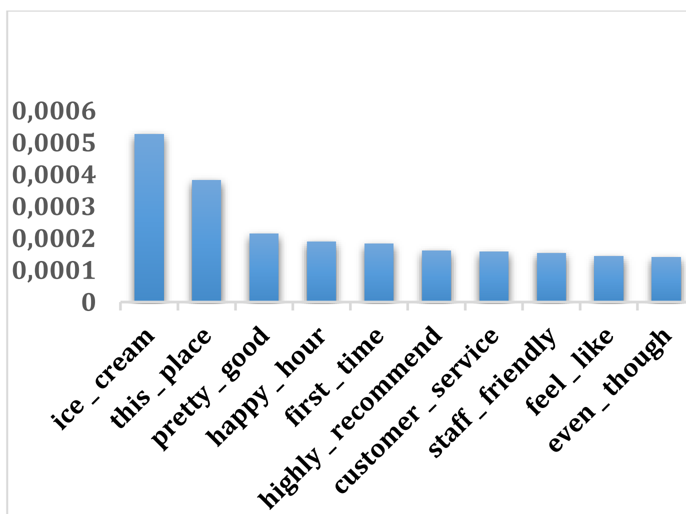

Figure 5. Top 10 word-pairs in terms of MI values

Fig.5 gives the top 10 word-pairs with their mutual information values. It is concluded from Fig.5 and Table 1, the word pair “ice cream” has the highest MI score with 0.00052676. After that, the word pairs with the highest mutual information values were “this place” and “pretty good”, respectively.

Table 1 gives the top 50-word pairs with their MI value. It is seen that the 50-word pairs obtained contain meaningful information about the restaurant, service, personnel or the products served.

Table 1. The top 50 word-pairs with their MI value

	Word Pair	MI Values		Word Pair	MI Values
1	ice cream	0.000526760	**26**	sweet potato	9.44E-05
2	this place	0.000382468	**27**	nothing special	9.32E-05
3	pretty good	0.000214195	**28**	onion rings	9.00E-05
4	happy hour	0.000189699	**29**	french toast	9.00E-05
5	first time	0.000182915	**30**	give stars	8.55E-05
6	highly recommend	0.000161285	**31**	definitely back	8.39E-05
7	customer service	0.000157912	**32**	they also	8.35E-05
8	staff friendly	0.000153544	**33**	much better	8.32E-05
9	feel like	0.000143897	**34**	chocolate chip	8.27E-05
10	even though	0.000140336	**35**	pleasantly surprised	8.19E-05
11	harvard square	0.000137714	**36**	food good	8.16E-05
12	come back	0.000129861	**37**	best ever	7.78E-05
13	every time	0.000129274	**38**	last night	7.78E-05
14	ann arbor	0.000128854	**39**	pretty much	7.75E-05
15	pad thai	0.000124948	**40**	bubble tea	7.71E-05
16	late night	0.000122238	**41**	love place	7.70E-05
17	next time	0.000121801	**42**	palo alto	7.61E-05
18	years ago	0.000119032	**43**	new york	7.56E-05
19	red velvet	0.000117535	**44**	mac cheese	7.47E-05
20	reasonably priced	0.000117416	**45**	san diego	7.41E-05
21	across street	0.000109788	**46**	coming back	7.35E-05
22	make sure	0.000109624	**47**	frozen yogurt	7.09E-05
23	really good	0.000107108	**48**	peanut butter	7.09E-05
24	behind counter	9.81E-05	**49**	great place	7.08E-05
25	would recommend	9.66E-05	**50**	top notch	7.05E-05

Table 2. Top words based on simple co-occurrences

w1	w2	Count (w1)	Count (w2)	count of co-occurrences	MI Values
pretty	good	1857	5715	434	0.0002142
great	place	3634	6487	275	0.0000708
love	place	2372	6487	192	0.0000770
pretty	much	1857	1594	161	0.0000775
staff	friendly	1176	1073	121	0.0001535
French	toast	236	238	76	0.0000900

Table 3. Comparison of the three different word collocation metrics

	Simple co-occurence	Mutual Informations	nltk.collocations likelihood
1	ice cream	ice cream	ice cream
2	pretty good	this place	happy hour
3	food good	pretty good	pretty good
4	really good	happy hour	first time
5	first time	first time	highly recommend
6	great place	highly recommend	customer service
7	good food	customer service	staff friendly
8	feel like	staff friendly	harvard square
9	love place	feel like	ann arbor
10	come back	even though	feel like
11	even though	harvard square	pad thai
12	every time	come back	even though
13	staff friendly	every time	reasonably priced
14	good place	ann arbor	red velvet
15	thy also	pad thai	late night
16	customer service	late night	years ago
17	next time	next time	every time
18	happy hour	years ago	come back
19	great food	red velvet	next time
20	really like	reasonably priced	across street
21	food service	across street	make sure
22	one best	make sure	behind counter
23	definitely back	really good	onion rings
24	food great	behind counter	sweet potato
25	service good	would recommend	french toast
26	pretty much	sweet potato	nothing special
27	would recommend	nothing special	would recommend
28	make sure	onion rings	pleasantly surprised
29	place good	french toast	chocolate chip
30	like place	give stars	palo alto
31	place great	definitely back	give stars
32	great service	thy also	really good
33	highly recommend	much better	bubble tea
34	much better	chocolate chip	san diego
35	good good	pleasantly surprised	new york
36	good service	food good	mac cheese
37	late night	best ever	peanut butter
38	place get	last night	much better
39	one places	pretty much	top notch
40	best ever	bubble tea	definitely back
41	great great	love place	credit card
42	really nice	palo alto	frozen yogurt
43	last time	new york	last night
44	service great	mac cheese	chapel hill
45	years ago	san diego	goat cheese
46	one favorite	coming back	best ever

47	would back	frozen yogurt	front desk
48	get food	peanut butter	http www
49	give stars	great place	thy also
50	really place	top notch	coming back

Some top words based on simple co-occurrences, count (w1), and count (w2) values with their MI values are given in Table 2.

It is seen from Table 2 that the highest count of co-occurrence is obtained for w1=*pretty* and w2=*good*.

Table3 shows that the three Word collocation techniques give similar results and therefore, all of those can be employed for Word collocations effectively.

4 Conclusions and Discussions

This study presents the evaluation of MI based on conditional entropy by using Eq.(4). Additionally, simple co-occurrence values of word pairs are also calculated.

It is seen from Table 1 that the top 50 word pairs can summarize the dataset with 1000 paragraphs and over one million words well. By considering these word pairs, the place, the menu of a restaurant, and the manners of an employee can be inferred.

The window size enables us to analyze the document on different scales. While smaller window sizes will identify fixed expressions (i.e., idioms), larger window sizes will show us the semantic concepts and other relationship characteristics.

Table 3 gives the top 50-word pairs with MI, LLR MI, and co-occurrences, which are obtained via the Python NLTK library. The pairs of MI and LLR MI are close to each other. It can be concluded conditional entropy, which is based on probability, is one of the best important tools for extracting syntagmatic relations.

References

[1] Zhai, C. X., Massung, S., Text Data Management and Analysis- A Practical Introduction to Information Retrieval and Text Mining, ACM Books , 2016.

[2] Church, KW., Hanks, P., Word Association norms, mutual information and lexicography. Computational Linguistics, ACM Books , 1990.

[3] Damani, O.P., Improving Pointwise Mutual Information (PMI) by incorporating Significant Co-occurrence. 17th Conference on Computational Natural Language Learning , 2013.

[4] F. H. Khan, U.Qamar, S. Bashir, SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection, Applied Soft Computing 39, 140–153, 2016.

[5] A.K. Jain, Y. Pandey, Analysis and implementation of sentiment classification using lexical POS markers, Int. J. Comput. Commun. Netw. 2 (1) , 36-40, 2013.

[6] T. Xu, Q. Peng, Y. Cheng, Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. 35, 279–289, 2012.

[7] Manning, C.D., Raghavan, R. and Schütze, H., Introduction to Information Retrieval, Cambridge University Press (2008).

[8] Garrett, Michael, et al. "Leveraging mutual information to generate domain specific lexicons." Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation, Washington DC, USA. 2018.

[9] Kang, Beom-mo. "Collocation and word association: Comparing collocation measuring methods." International journal of corpus linguistics 23.1, 85-113, 2018.

[10] Lai, Huei-ling. "Collocation analysis of news discourse and its ideological implications." Pragmatics 29.4 ,545-570, 2019.

[11] Liu, Xiaoxia, et al. "Recognition of collocation frames from sentences." IEICE TRANSACTIONS on Information and Systems 102.3, 620-627, 2019.

[12] Williams, Christopher KI. "On Suspicious Coincidences and Pointwise Mutual Information." arXiv preprint arXiv:2203.08089, 2022.

[13] Krenn, Brigitte. "Collocation mining: Exploiting corpora for collocation identification and representation." Entropy 1, 2000.

[14] Zhang, Ke, et al. "A Construction Method of Electric Power Professional Domain Corpus Based on Multi-model Collaboration." 2022 4th Asia Energy and Electrical Engineering Symposium (AEEES). IEEE, 2022.

[15] https://www.yelp.com/dataset

[16] L. R. Dice. 1945. Measures of the amount of ecological association between species. Ecology, 26:297– 302.