การประยุกต์ใช้ทฤษฎีกราฟในการทำเหมืองข้อความ
The application of graph theory on text mining

สุรัสวดี นางแล[1*], ธนายุทธ ช่างเรือนงาม[1]
Suruswadee Nanglae[1*], Thanayut Changruenngam[1]

Received: 28 October 2019 ; Revised: 26 November 2019 ; Accepted: 9 January 2020

บทคัดย่อ

ในการวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาการประยุกต์ใช้ทฤษฎีกราฟในการหาความสัมพันธ์ระหว่างคำและจัดกลุ่มคำเพื่อค้นหาความหมายของกลุ่มคำนั้น โดยการยกตัวอย่างข้อมูลจากเครือข่ายออนไลน์ Twitter ในหัวข้อที่เกี่ยวข้องกับคำว่า “Amazon” ระหว่างวันที่ 11-18 กุมภาพันธ์ 2562 จากผลการศึกษาพบว่า หน่วยคำคู่ (bigrams) ที่มีความถี่มากที่สุดคือ ของเล่นหัวโต (funko pop), กำไรพันล้าน (billion profit) และ การจ่ายภาษี (pay tax) ตามลำดับ เมื่อทำการวิเคราะห์จัดกลุ่ม (Cluster analysis) โดยใช้เทคนิค Random Walk Algorithm พบว่า funko pop ที่นิยมมากที่สุดคือ funko pop ของนักร้องสมาชิกวง BTS ชื่อ Jungkook ส่วน billion profit และ pay tax มีความสัมพันธ์เชื่อมโยงกันอยู่ เพราะในช่วงระยะเวลาดังกล่าว Amazon ถูกตั้งข้อสังเกตเรื่องการไม่จ่ายภาษีให้รัฐบาลกลางทั้งๆ ที่สร้างผลกำไรในปีที่ผ่านมาในหลักพันล้านเหรียญสหรัฐ ซึ่งสอดคล้องกับการวิเคราะห์กราฟโดยการวัดค่าความเป็นศูนย์กลาง พบว่า ในเครือข่ายคำ คำว่า tax เป็นคำที่เป็นจุดศูนย์กลางของเครือข่ายคำ ทำหน้าที่เป็นทั้งโหนดศูนย์กลางของเครือข่ายคำ, โหนดศูนย์กลางในการเชื่อมโยงกับโหนดอื่น และเป็นโหนดศูนย์กลางในการเชื่อมโยงกับโหนดอื่นๆ ของเครือข่ายที่ไกลกันให้เข้าถึงกันผ่านการวัดค่า Degree centrality และ Betweenness centrality

คำสำคัญ: ทฤษฎีกราฟ เหมืองข้อความ เครือข่ายออนไลน์ การวิเคราะห์จัดกลุ่ม

Abstract

The objective of this research was to study the application of graph theory to seek the relationship between words and categorize groups of words for finding the meaning of the groups of words by analyzing messages sampled from social media like Twitter on the topic related to “Amazon” during 11-18 February 2019. The study revealed that the most frequent bigrams were “funko pop”, “billion profit”, and “pay tax” respectively. Moreover, cluster analysis using a random walk algorithm showed that the most popular “funko pop” was BTS-Funko Pop-Jungkook while billion profit and pay tax have a particular correlation because during that time Amazon was indicated to be a company did not pay any tax to the US government although they generated profits of a billion dollars. This is consistent with graph analysis which by measuring centrality found that in a word network, the word “tax” is the centrality of the word network and plays its role as a central node in the word network, a central node connecting to other nodes, and a central node connecting to other nodes of distant networks to be reachable through degree centrality and betweenness centrality measures.

Keywords: Graph Theory, Text Mining, Social Media, Cluster Analysis

[1] โปรแกรมคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงราย
โทร +668 1594 3105 E-mail: SNanglae@gmail.com

[1] Mathematics Program, Faculty of Science and Technology, Chiang Rai Rajabhat University, Thailand.
Tel: +668 1594 3105 E-mail: SNanglae@gmail.com

Introduction

Nowadays information technology plays an increasingly important role in people's daily lives, especially internet communication services that have gained much popularity. Multimedia is a platform spreading information that contains messages, photographs or videos and news in the form of text files that are widely publicized via online media such as Facebook or Twitter. Messages are considered unstructured data but the pattern of the message carries hidden contents, relationships and effects on what people are interested in studying. In the meantime, the varieties and the bulk of messages are input into the media systems at all times. Therefore, to investigate the content, the knowledge, the relationships or the influence on what people are interested in using various techniques called text mining are required.

Text mining is the process of exploring and analyzing patterns, concepts, and correlation hidden in text data by using fundamental knowledge such as statistics, mathematics, machine learning, etc., while obtained results are most likely on the basis of frequency of words, correlation within words, or patterns of words, etc.[1] With regard to the study on correlation among words, either the words occurring within the same document or the words accompanying other words, many methods, such as creating N-grams, are employed to generate a set of words. The words occurring at the same time are classified by means of probability to examine which words have a tendency to appear after other words and have correlations with surrounding words, or by using a diagram of word networks to explore correlation within words. One of the popular instruments creating a diagram of word networks is a graph theory showing correlation within words.

Graph theory is a model to show the relationship between objects. The objects are called nodes or vertex (vertices). An edge is a line that connects two nodes together, that the strength of their connection is represented by the weight value. A diagram or graph can be divided into 2 types. The first one is an undirected graph ; a graph in which the edges are not ordered. It has no arrows on edges indicating directions. The second one is a directed graph ; a graph that is made up of a set of nodes connected by edges, where the edges have a direction associated with them. Consequently, graph analysis can be used to identify both the strength of the relationship and the direction of relationship. There is a variety of analysis methods such as centrality measures. (i.e. degree centrality, betweenness centrality, and closeness centrality) [2]

This research studies the application of graph theory to investigate the relationship between words and categorize groups of words for interpreting the meaning of the groups of words using messages posted on social media like Twitter on the topic related to the subject "Amazon".

Methods

R software was used to collect, manipulate, and analyze the data according to the following procedures:

Data collection method

R software was used to gather data from Twitter social networking platform during 11-18 February 2019 for 100,000 messages on the topics related to the word "Amazon".

Data manipulation

The message data were manipulated by deleting website addresses (URLs), and gimmicks ; for example, words starting with @, #, Tag or RT including http, etc., The messages were also adjusted by removing other languages than English or characters that do not exist in English, replacing abbreviations with full words, converting the size of the letters to be small letters in English to obtain entire words in the same dimension, and deleting a full stop symbol, punctuation marks in English, spaces between words, and endings such as -ed, -es to ensure that all words have a unity and are in the same pattern.

Data analysis

After the message data were completely manipulated, a bigram was created to generate a pair of words that most commonly occur simultaneously and the frequency of the obtained bigrams was shown in the form of word cloud. Next, an undirected word graph was created. Nodes of the graph were represented by words. The weight value of edges was calculated by counting the number of occurrences of those two words which could be noticed from the frequency of the bigrams. Graph analysis was performed with the following methods:

1) The groups of words were divided by utilizing cluster analysis with a random walk algorithm. This is to find all subgraphs in a large graph using a short random walk technique. It is likely to occur within the same group or the community.[3]

2) In research, centrality measures were performed on degree centrality, betweenness centrality, and closeness centrality, detailed as follows:

- Degree centrality (D_C) refers to the number of ties linking a node to other nodes. A node with a high degree centrality will have a great influence within its network accordingly.[4] The equation can be written as

$$D_c = \sum_{j=1}^{n} n_{ij} \tag{1}$$

where i represents node i and j represents other nodes. When $n_{ij}=1$ means a connection between the two nodes and $n_{ij}=0$ means no connection between them.[5]

- Betweenness centrality (B_C) measures how important a node is to the shortest paths through the network. Nodes with high betweenness centrality are nodes that lie on communication paths. The equation can be written as

$$B_c = \sum_{jk} \frac{g_{ijk}}{g_{jk}} \tag{2}$$

where n is the total number of nodes, i, j, k represent different nodes, g_{jk} is the number of the shortest path from j to k and g_{ijk} is the number of the shortest paths from j to k through i.

- Closeness centrality (C_C) is a measure of the degree to which an individual is near all other individuals in a network. It is the inverse of the sum of the shortest distance between each node and other nodes in the network. The equation can be written as

$$C_c(n_i) = \sum_{j=1}^{n} d(n_i, n_j) \tag{3}$$

where $d(n_i,n_j)$ is the number of the shortest path form node i to node j through other nodes in a network.

Results

Based on generating a pair of words and measuring the frequency of the obtained double morpheme, the result is shown in Figure 1

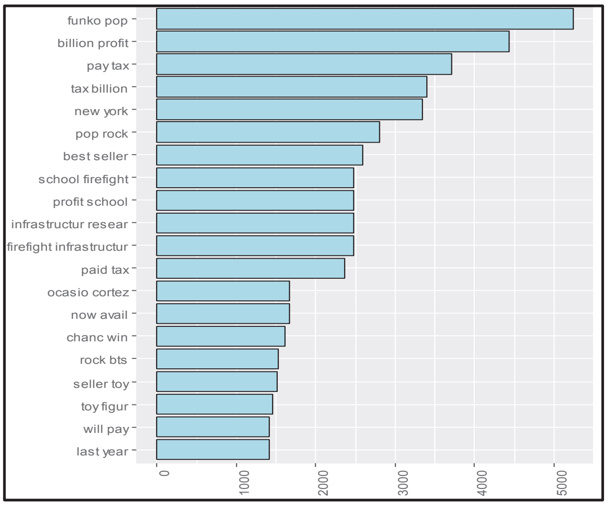


Figure 1 The top 20 frequency of bigrams as arranged in descending order.

Figure 1 shows the top 20 frequencies of the bigram in the form of a bar graph arranged in descending order. It can be seen that the most frequently occurring bigrams are “funko pop”, “billion profit”, and “pay tax”, respectively. All bigrams will be generated groups of words (Word Cloud) later as seen in Figure 2.

Figure 2 shows how the word cloud is generated and it can be seen that bigrams consist of large and small scales depending on the frequency. The bigrams with large sizes are “funko pop”, “billion profit”, “pay tax”, “tax billion”, and so on. The bigrams and the frequency generate a set of nodes and a set of edges in the word graph. The graph nodes are represented by words while edges and the weight value are represented by the frequency of occurrences of those two words obtained from the frequency of bigrams as shown in Figure 3.

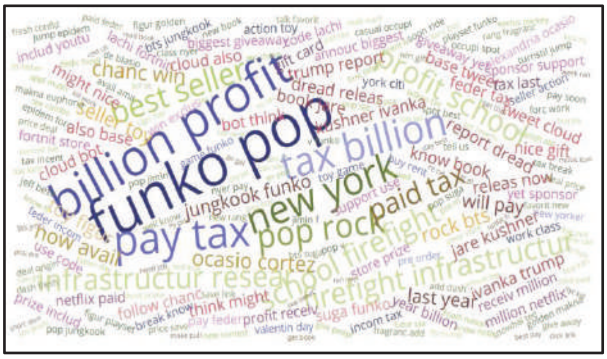


Figure 2 Word clouds of bigrams

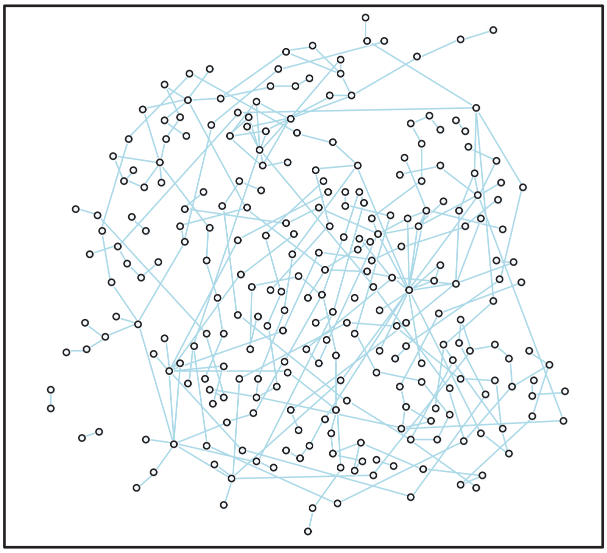


Figure 3 Word Graph

Figure 3 shows the big picture of the network in which nodes represent words and edges that connect those nodes representing the connection between those two words. The weights of edges are the frequency of occurrence of those two words. It can be noticeable that the word graph has a connection as a complex network. Therefore, network analysis is necessary to seek the correlation within occurring words.

Classifying groups with cluster analysis

To classify the groups of words, we use cluster analysis with a random walk algorithm. This is to find all subgraphs in a large graph using a random walk algorithm as shown in Figure 4. From Figure 4, the interesting groups of words are obtained as shown in Table 1.

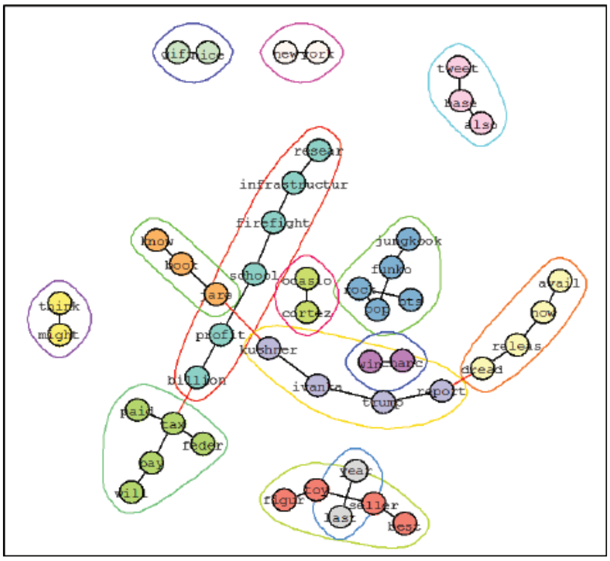
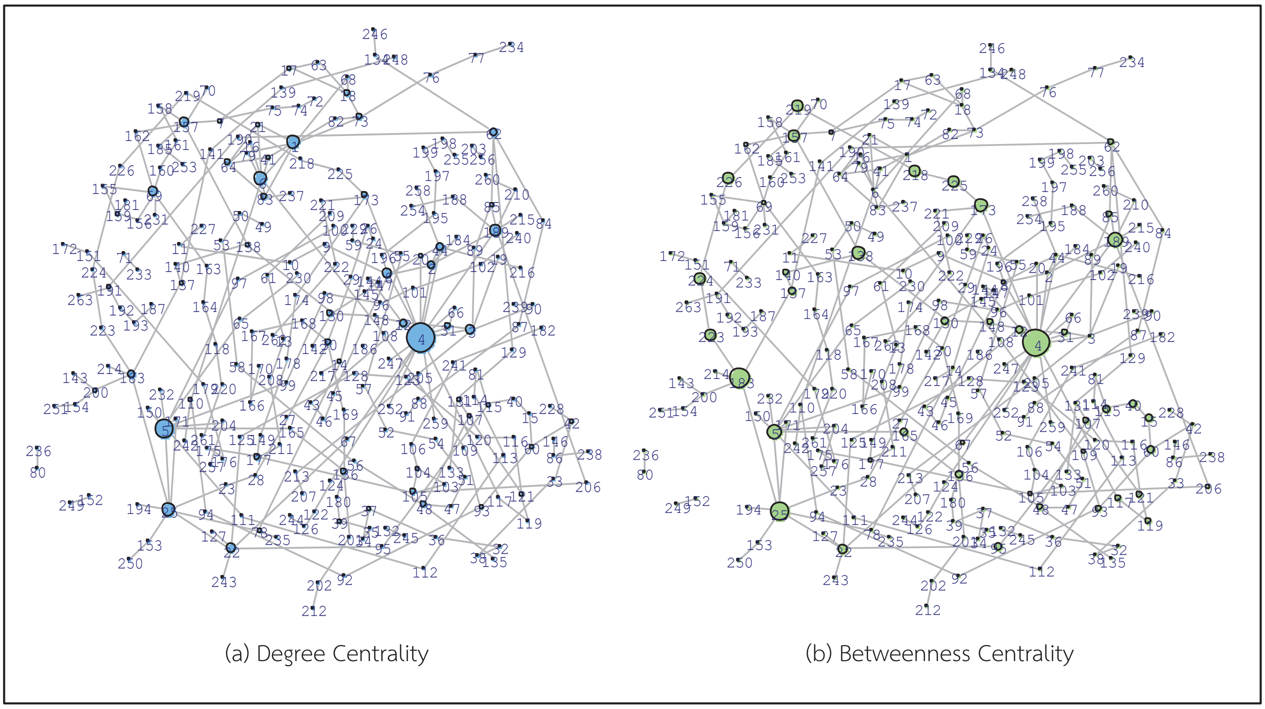


Figure 4 Dividing groups of words using cluster analysis with a random walk algorithm

Table 1 Groups of words from cluster analysis using a random walk algorithm.

Group	Groups of words	Information retrieval on the internet
1	paid tax feder will pay	Amazon will pay federal taxes on billion profits.[6,7]
2	billion profit school firefight infrastructur resear	
3	avail now releas dread	The Extraordinary Story of Jared Kushner and Ivanka Trump that will be for sale on Amazon website.[8]
4	know book jare	
5	kushner ivanka trump report	
7	jungkook funko pop rock bts	BTS Funko Pop figure-Jungkook for sale on Amazon website.[9]
8	figure toy seller best	Top-selling figure toy on Amazon website.[9]

It shows that, the messages on Twitter in February 2019 can be concluded into 3 major topics. The first topic is about Amazon itself that generated its profit of greater than 11.2 billion US dollars but did not pay federal taxes,[7] causing Amazon to be attacked by criticism from multiple sources on social media like Twitter. The second topic is the story of Ivanka Trump, the daughter of the US President, Donald Trump, who is married to Jared Kushner, a businessman. Both of them are White House Senior Advisors. There will be an explosive book about their greed, ambition, and corruption,[10] causing this book to become interesting for people in general and it later became a bestseller on the Amazon website.[11] The last topic is the story about Funko Pop figures or figures with a large head. BTS Funko Pop figure-Jungkook is the most popular on social media. The website www.allkpop.com publicized its article that Jungkook Funko Pop marked the best -selling figure.[12]



(a) Degree Centrality

(b) Betweenness Centrality

Figure 5 Degree Centrality and Betweenness Centrality measures

Graph analysis

Graph analysis is performed through measuring centrality in 3 characteristics as degree centrality, betweenness centrality, and closeness centrality shown in Figure 5 and Figure 6. Figure 5 (a) is a degree centrality measure. It is found that the word with the largest node is node 4 which is the word “tax”. That means tax is the central node performing its duty as a hub and has the highest power in the network, followed by node 5 that is the word “new”, node 25 that is the word “book”, node 6 that is the word “pop”, and node 1 that is the word “funko” respectively. As observed in Figure 5 (b), it is a betweenness centrality measure. It is found that the largest node is node 4 which is the word “tax” again. This means that tax is the central node performing its duty as a connection to other nodes of distant networks enabling them to be reachable, followed by node 183 which is the word “get”, and node 25 which is the word “book”.

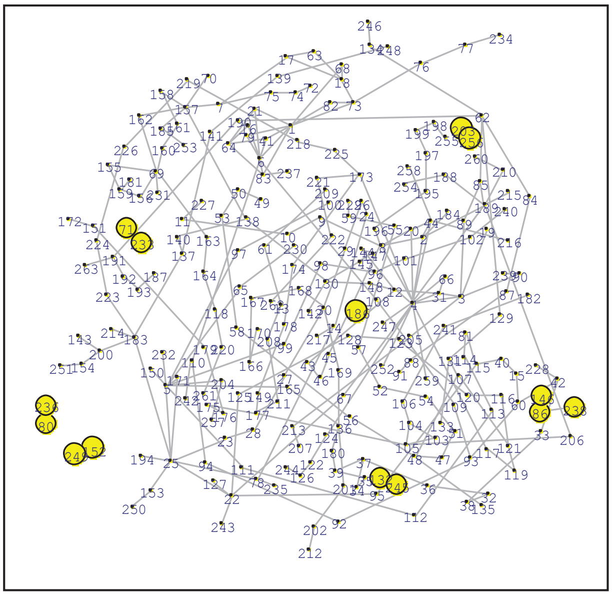


Figure 6 Closeness Centrality measure

According to Figure 6, the closeness centrality is measured. It was found that large nodes have similar scales ; for instance, node 203 which is the word "short", node 256 which is the word "story". Nodes that are close to each other like this are considered the central nodes in the networks that are close to other nodes using the shortest paths as a measuring standard.

Conclusion and Discussion

The study on the application of graph theory to seek the relationship between words found that making an undirected graph by generating bigrams and counting the frequency can only explore the words that are most mentioned at that time but cannot indicate how those words correlate with other words. For example, the word "funko pop" has the highest frequency but it does not indicate that which funko pop is the most mentioned or the best-selling. When a word graph is made for cluster analysis, it can indicate that the most popular funko pop is BTS Funko Pop-Jungkook, the best seller compared to other Funko Pop figures of other members.[9,12] The number of the frequency is followed by the word "billion profit" and "pay tax" respectively. Considering the cluster analysis, both words relate to each other because by that time Amazon had been noticed about not paying federal taxes even though they could generate billions of dollars in profits.[6,7] This is consistent with graph analysis by centrality measures. The word tax is the central node of the network, performing its duty as the central node of the word network and the central node to connect to other nodes, and the central node connecting to other nodes of distant networks to be reachable through degree centrality and betweenness centrality measures. With regard to closeness centrality analysis, it is an analysis of the central node of the network close to other nodes using the shortest paths as a measuring standard. This means that the node of words has the least number of active connections or there are a small number of people who posted this word a few times. The objective of this analysis is to seek the relationship of the words that are mentioned most frequently in the network. Therefore, the closeness centrality analysis is not necessary. The use of graph theory to analyze word networks is useful to find opinions hidden on a large scale of messages but it cannot indicate whether or not they are positive or negative opinions. Thus, to acquire both the quantity and quality of the message posted on social media, the obtained results need to be performed by sentiment analysis. Moreover, understanding the trend or the impact of the messages can lead to a new branch of knowledge which can be applied or extended into other aspects such as marketing or communication, etc.

References

1. EDUCBA. (2018). Difference between Text Mining and Natural Language Processing. Available from: https://www.educba.com/important-text-mining-vs-natural-language-processing/Accessed 27 December, 2018
2. Chonnikarn Rodmorn and Maturos Panmueang. Social Network Analysis using Graph Theory: Case Study of Instructors of Faculty of Business Administration, Bangkokthonburi University. The 10th National Conference on Computing and Information Technology: 2014.
3. Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks (long version), Available from: https://arxiv.org/abs/physics/0512106 Accessed 4 January, 2018.
4. Wasserman, S., and Faust, K. Social Network Analysis Methods and Applications, (Vol. 8), Cambridge Cambridge University Press. 1994.

5. Durland, M. M., and K. Fredericks (eds). New Directions For Evaluation: Social Network Analysis in Program Evaluation, (Vol 107). San Francisco: Jossey-Bass. 2005.

6. Laura Stampler. Amazon Will Pay a Whopping $0 in Federal Taxes on $11.2 Billion Profits. Available from: https://fortune.com/2019/02/14/amazon-doesnt-pay-federal-taxes-2019/, Accessed 7 May, 2019.

7. Christopher Ingraham. Amazon paid no federal taxes on $11.2 billion in profits last year. Available from: https://www.washingtonpost.com/us-policy/2019/02/16/amazon-paid-no-federal-taxes-billion-profits-last-year/?noredirect=on, Accessed 7 May, 2019.

8. Richard Johnson. Jared Kushner, Ivanka Trump bracing for release of Vicky Ward book. Available from: https://pagesix.com/2019/02/08/jared-kushner-ivanka-trump-bracing-for-release-of-vicky-ward-book/, Accessed 7 May, 2019.

9. Amazon. Customers who viewed Funko Pop! Rocks: BTS-Jungkook also viewed. Available from: https://www.amazon.com/Funko-Pop-Rocks-BTS-Jungkook/dp/ B07KPSYW7Y, Accessed 7 May, 2019.

10. Allison Pecorin. Ivanka Trump says she and Jared Kushner got no special treatment for security clearances. Available from: https://abcnews.go.com/Politics/ivanka-trump-jared-kushner-special-treatment-security-clearances/story?id=60940398 Feb 8, Accessed 7 May, 2019.

11. Amazon. Kushner, Inc.: Greed. Ambition. Corruption. The Extraordinary Story of Jared Kushner and Ivanka Trump. Available from: https://www.allkpop.com/article/2019/03/bts-jungkook-funko-pop-become-the-number-1-best-sellerworldwide, Accessed 7 May, 2019.

12. Allkpop. BTS Jungkook Funko Pop become the Number #1 Best Seller, worldwide. Available from: https://www.allkpop.com/article/2019/03/bts-jungkook-funko-pop-become-the-number-1-best-sellerworldwide, Accessed 7 May, 2019.