Utilize Novel Algorithms to Acquire, Analyze, and Extract Data from TikTok Discover Page and Education-Related Topics

Jincheng Zhang* and Thada Jantakoon*

Received: March 26, 2024
Revised: October 1, 2024
Accepted: October 11, 2024

* Corresponding Author: Jincheng Zhang, E-mail: zjc1639834588@gmail.com

DOI: 10.14416/j.it/2025.v2.004

Abstract

Due to the swift advancement of research and technology, particularly in the fields of computer science and data science, individuals are progressively employing these technologies, along with others, in the realm of education. This study encompasses the development, creation, and utilization of a comprehensive range of techniques, spanning from data collecting to data analysis and mining. It introduces a novel algorithm and methodology for acquiring and refining data, as well as three innovative algorithms for data analysis and exploration. This project collects data on the topics featured on the TikTok Discover page for the purpose of doing data analysis and data mining. The research methodologies employed in this work encompass empirical research, experimental verification, algorithm design and optimization, system design, and implementation. Our study examined and extracted educational content from TikTok Discover pages. We studied the popularity of this data from various perspectives and levels. This allows users to easily and efficiently locate the specific information they are interested in for further investigation. Analysis, sentiment analysis, and potential anomalous data were discovered. The analysis and extraction of this data offer educational practitioners' significant insights that can be utilized to inform and direct educational practice.

Keywords: Novel Algorithms, Tiktok Discover Page, Data Analysis, Data Mining, Education-Related Topics, Education ICT.

1. Introduction

Individuals are progressively retrieving information from the Internet with the intention of doing data analysis and mining, as well as extracting important insights and knowledge for educational reasons [1], [2]. This can be attributed to the exponential expansion of scientific advancements and technological breakthroughs, namely in the realm of computer science. Data analysis and data mining technologies have the potential to greatly improve education by enhancing instruction quality, identifying weaknesses in the current educational system, improving instructional content design, and increasing student interest and efficiency [3], [4].

As a website with huge influence in the world, TikTok has a large number of users and data, including a large amount of valuable data waiting for people to mine and develop.

Computer science and data science have experienced significant advancements in recent years, leading to their extensive use across many domains. Within the education sector, there is a growing trend of utilizing computer science and data science technology to enhance and improve the industry. These encompass individuals' comprehensive and thorough utilization of data mining technology in the field of education. These strategies involve utilizing data mining to accomplish various tasks such as predicting students' academic performance [5], identifying tailored learning for students [6], evaluating existing courses [7], and predicting the likelihood of students dropping out [8].

This study devised and introduced a novel algorithm

* Faculty of Science and Technology, Rajabhat Maha Sarakham University

and methodology for efficiently and accurately acquiring data pertaining to the content featured on the TikTok Discover page. This technique and method are the first of their kind in the world, designed to efficiently and quickly gather data specifically related to the content found on the TikTok Discover page. Indeed, this novel technology and methodology can be applied to gather data from several other websites as well. This study is the first in the world to analyze and mine data from TikTok's discovery page topics. We developed this novel algorithm and methodology to get data from the Internet due to the presence of intricate anti-crawler obstacles encountered when individuals attempt to extract data from certain websites, such as TikTok, using crawlers. We attempted to employ a web crawler to extract data from the TikTok website, however we faced numerous intricate anti-crawler obstacles on TikTok. Typically, the software encounters errors after the crawler completes the data crawling process. Our algorithm offers a highly effective, straightforward, and efficient solution to problems.

In this study, we devised and suggested three novel algorithms for doing data analysis and data mining. The three methods are "Keyword Distance Weighted Frequency," "Keyword Distance Weighted Frequency-Inverse Document Frequency," and "Keyword Distance Weighted Frequency-Emotion Analysis Frequency." These three novel methods can be utilized for data analysis and data mining across several domains. We conducted a comprehensive analysis and data extraction of educational content on TikTok. Discover: doing comprehensive analysis of the prevalent data from many perspectives and levels, enabling users to effortlessly and effectively locate the pertinent facts of their interest for further investigation. Analysis, sentiment analysis, and potential anomalous data were detected. The analysis and mining of this data offer educational practitioners' significant insights that can be utilized to inform educational practice. This study develops new algorithms for general data mining purposes and conducts experiments in the education domain.

Recently, research in educational data mining has increasingly focused on integrating social media data, including platforms such as TikTok, to better understand educational trends and behaviors. Research has highlighted how data extraction and analysis algorithms can effectively process large amounts of content from such platforms to gain insights into user preferences, learning patterns, and popular educational topics. Our work follows this trend by integrating data from TikTok's Discover page, using novel algorithms to extract and analyze education-related content. This approach provides educators and researchers with actionable insights that enhance decision-making in educational technology.

1.1 Research framework and structure

The scope of this research is comprehensive and detailed, mainly including the following aspects:

1.1.1 Use the Python programming language to obtain and filter data on TikTok Discover page themes.

1.1.2 Apply data analysis and data mining technology, and use 11 codes to analyze the collected data on TikTok Discover page topics in different aspects and levels. This includes the use of 3 new algorithms we invented and proposed for data analysis and data mining.

1.1.3 Explain the results of data analysis and data mining so that they can be applied in educational practice.

1.1.4 Research Focus and Scope: Data mining algorithms can be used in education for personalized learning, grade prediction, and sentiment analysis. Clustering algorithms help identify students' learning styles, regression analysis predicts grades, and sentiment analysis processes feedback to improve courses. These applications improve learning outcomes, optimize teaching management, and closely integrate data mining with educational practice. In this study, we pay special attention to how to enhance the analytical capabilities of educational content through effective algorithm design. We invented and proposed keyword distance weighted frequency (KDWF), which enhances TF-IDF and sentiment analysis by considering the distance between keywords and surrounding words. Unlike traditional TF-IDF, which only focuses on word frequency, KDWF gives higher weights to words with close distances,

thereby more accurately reflecting the text context relationship. In sentiment analysis, KDWF can identify words related to the sentiment tendency of keywords and improve the accuracy of sentiment recognition. This method has broad application potential in social media, market analysis, and data mining in the field of education.

1.2 Benefits and beneficiaries related to this research

Through this research we expect to gain the following benefits:

1.2.1 Provide a complete set of methods from data acquisition to data analysis and data mining.

1.2.2 We invented and proposed an algorithm and method to solve the anti-crawler problem to obtain data simply and reliably.

1.2.3 We invented and proposed 3 algorithms for data analysis and mining. It has wide application value.

1.2.4 We have explained the results of data analysis and data mining on the topic data of education-related discovery pages on TikTok, which can be used by people in educational practice. Such as hot words, hot phrases, hot topics, related phrase data classification, sentiment analysis, etc. Provide reference for educational practice.

In general, people can use the relevant content in this study for data acquisition, data analysis and data mining on any topic that people are interested in. People can use the relevant content in this study for research or commercial realization, etc. Beneficiaries include teachers, educational managers, managers in other fields, researchers in any field, managers in any field, and business people in any field, etc.

2. Related Theories and Research

This study uses some new algorithms to obtain, analyze and mine education-related data on TikTok. Among them, we invented and proposed a new algorithm and method to quickly and reliably obtain data on TikTok Discover page topics. Then we used 11 codes to conduct data analysis and data mining from different aspects and levels, including using 3 new algorithms that we invented and proposed for data analysis and data mining. They are "Keyword Distance Weighted Frequency", "Keyword Distance Weighted Frequency-Inverse Document Frequency" and "Keyword Distance Weighted Frequency-Emotion Analysis Frequency". Here we will introduce current theory and previous related research relevant to our study.

2.1 Theoretical Framework

Google is currently the world's largest search engine. A large amount of valuable data can be searched, providing users with massive data [9], [10]. Google uses crawlers to obtain data and then saves the obtained data to Google's servers. Then when the user searches on Google, Google provides the corresponding data to the user through the corresponding algorithm [11], [12]. In this study we use Google's advanced search function [9]. For example, we search intitle:education intext:the OR intitle:the site:https://www.tiktok.com/discover/ on Google to obtain data. It means that it is designed to retrieve two types of content from TikTok's Discover page: one is content that contains both "education" and "among" in the title, or content that contains "among" in the text and "education" in the title. First we prepared a word list of 500 common English words. Then use Python code to replace the keyword "the" in intitle:education intext:the OR intitle:the site:https://www.tiktok.com/discover/. Generates 500 search terms for advanced searches on Google. Then we search on Google using advanced search terms like intitle:education intext:the OR intitle:the site:https://www.tiktok.com/discover/. Then directly copy and paste these Google search results into a world file. Then we quickly extracted the topic name data of each TikTok discovery page topic through Python code, and used some filtering conditions to perform some corresponding screening. In this way we get the corresponding data. We then used Python code to quickly filter the collected data, such as deleting duplicate text and extracting relevant text. The restrictions imposed by the Google search engine, such as human-machine verification, can be easily resolved manually.

Searching and collecting a moderate amount of data on Google is reasonable and legal, but it still needs to follow legal

and ethical guidelines. According to the Digital Millennium Copyright Act (DMCA) and the General Data Protection Regulation (GDPR), when using public data, it is necessary to ensure that copyrights are not violated and personal privacy is respected.

Collecting data without the user's explicit consent may bring ethical issues. GDPR requires user consent when processing personal data and transparent notification of the purpose of data use. Therefore, data anonymization is crucial to avoid leaking personally identifiable information.

2.2 Key Concepts and Definitions

Google Advanced Search can filter Google's search results more precisely to find search results that better meet user requirements. And we use Python to quickly generate search terms for Google advanced search and use Python to quickly extract and filter out the data we need in the world file. Document data can be processed quickly and efficiently using Python code [13]. Data analysis and data mining techniques have a relatively long history, and are currently very popular and relevant technologies are developing rapidly [14]. Data analysis and data mining technology are currently widely used in various industries. Using these technologies can better understand the current situation and enable people to make better decisions about their businesses [15], [16].

2.3 Summarize

At present, digital technology is developing rapidly and is practical and powerful. All walks of life are using digital technology to upgrade. Digital technology brings many new opportunities and unlimited possibilities to all walks of life [17]. Several technologies invented and proposed in this study have strong practicality in the current environment. The method we proposed to obtain data can well avoid the anti-crawler problems that would exist if data were obtained through crawlers. And the targeted data people need can be obtained quickly and efficiently. Users can also replace the Google advanced search terms intitle:education intext:the OR intitle: the site:https://www.tiktok.com/discover/ with intitle:education intext:of OR intitle:of intext:"20.7M views" site: https://www.tiktok.com/discover/, etc., so as to be able to search for more detailed and more accurate data. Of course, this method and algorithm can also be used to collect data outside of Google. In this study we collect data on Google. And we proposed three new data analysis and data mining algorithms. These three new algorithms are very suitable for use in data analysis and data mining of data with keywords.

3. Research methods

3.1 Literature Review and Theoretical Analysis

In this study, we conducted a systematic literature review and theoretical analysis, including Google advanced search, using Python code to quickly and efficiently process document data, data analysis and data mining technology, etc. These related theoretical analyzes provide effective support for this research.

3.2 Interdisciplinary Research and Application

We applied it across disciplines in the field of education through some of the new algorithms we invented and proposed. We obtained, analyzed and mined education-related data on TikTok, conducted data analysis and data mining through 11 codes, and obtained information on education. Relevant and useful insights and knowledge. Such as hot words, hot phrases, hot topics, related phrase data classification, sentiment analysis, etc. Provide reference for educational practice. Make the education field more efficient through digital technology through an interdisciplinary approach, etc.

3.3 Empirical Research and Experimental Verification

Through experiments and empirical analysis, the effectiveness and reliability of the crawler and data mining methods in this scenario were verified. The experimental results were analyzed and discussed, and case analysis and conclusions were proposed, which provided important reference and guidance for subsequent research and practice.

3.4 System Design and Implementation

In this study, we used multiple new algorithms and methods we invented and proposed to obtain, analyze and mine education-related data on TikTok, and built a complete system to apply it in real application scenarios.

3.5 Algorithm Design and Optimization

In this study, we propose a new method of obtaining data on the Internet that is fast, simple, and reliable, which can well avoid the anti-crawler problems that would exist if data were obtained through crawlers. And the targeted data people need can be obtained quickly and efficiently. And we proposed 3 new data analysis and data mining algorithms. These 3 new algorithms are very suitable for use in data analysis and data mining of data with keywords.

3.6 Applications

In this study, we use multiple new algorithms and methods we invented and proposed to obtain, analyze and mine education-related data on TikTok, and apply them in real application scenarios. Provide reference and reference for people to use the multiple new algorithms and methods we invented and proposed in real application scenarios.

3.7 Dataset Construction and Annotation

We created a data set to acquire, analyze and mine education-related data on TikTok, including acquired data, filtered data, analyzed data, etc. Ensure the reliability of our research.

3.8 Comparison And Effectiveness Verification Of Keyword Distance Weighted Frequency (KDWF) With Existing Methods

In order to explore the method keyword distance weighted frequency (KDWF) and its advantages over existing methods, this section will analyze from multiple perspectives.

3.8.1 Comparison of KDWF with existing methods

When comparing KDWF with existing methods, we use the following non-traditional indicators for a more comprehensive evaluation:

Feature extraction ability:

KDWF: Combining the frequency of keywords with their distance in the text, KDWF can effectively capture contextual information and highlight the importance of keywords in different contexts. This method can better identify relevance and semantic relationships.

Existing methods: Many existing methods rely on simple frequency statistics and fail to fully utilize the distance relationship between words, which may lead to the neglect of some important information.

Information gain:

KDWF: By introducing distance weights, KDWF can better evaluate the information gain of keywords in the text. The relative position between words can reveal the hierarchy and relevance of the topic.

Existing methods: Traditional methods often only consider the number of occurrences of words, resulting in insufficient evaluation of information gain.

Robustness:

KDWF: This method is highly resistant to noise and redundant information in the text because it can distinguish the distance between important keywords and irrelevant words and reduce the impact of irrelevant information.

Existing methods: Many traditional methods perform poorly when faced with noisy data and cannot effectively extract useful features.

3.8.2 Verification of the effectiveness of KDWF

To verify the effectiveness of KDWF, we adopt a case analysis strategy:

By analyzing specific text instances, we show the keyword extraction results of KDWF and compare them with the results of existing methods. The advantages of KDWF in capturing contextual information and keyword importance can be illustrated by specific examples.

Through the above methods, we hope to be able to comprehensively evaluate the performance of KDWF and prove its advantages over existing methods in theory and practice. This will provide a solid foundation for future research and applications.

4. Collect, Analyze, and Mine Data

4.1 A New Method and Algorithm for Collecting Data

This research invents and proposes a new algorithm and method to obtain TikTok's discovery page topic data simply, quickly and efficiently. This is the world's first simple, fast and efficient algorithm and method to obtain data on the topic

of TikTok's discovery page. Of course, this new algorithm and method can also be used to collect data from other websites. And it is the world's first study to conduct data analysis and data mining on TikTok's discovery page topic data.

In this study we used Google's advanced search capabilities. For example, we search intitle:education intext:the OR intitle: the site:https://www.tiktok.com/discover/ on Google to obtain data. It means that it is designed to retrieve two types of content from TikTok's Discover page: one is content that contains both "education" and "among" in the title, or content that contains "among" in the text and "education" in the title.

For example, search for "intitle:education intext: the OR intitle:the site:https://www.tiktok.com/discover/" on Google Then you will get the following results:

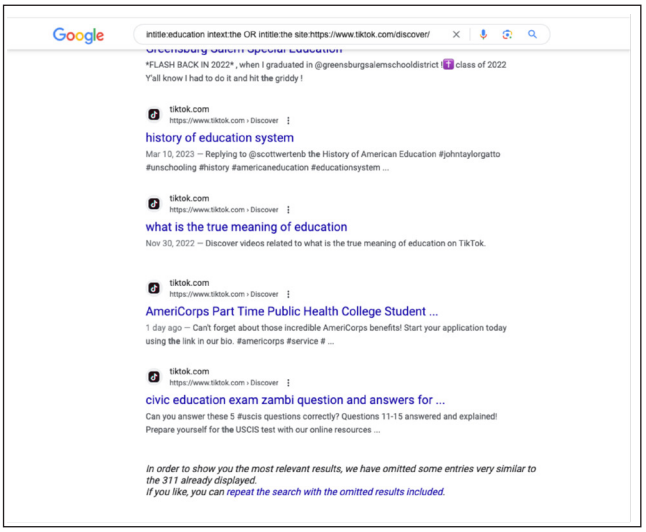

Figure 1. *We search on Google for "intitle:education intext: the OR intitle:the site:https://www.tiktok.com/discover/".*

First we prepared a word list of 500 common English words (500 common English words come from this website: https://www.smart-words.org/500-most-commonly-used-english-words.html) . (You can also use other word or phrase lists such as larger, smaller, other languages, keywords and strongly related word or phrase lists, etc.) Then replace intitle:education intext:the OR intitle with the Python code The keyword "the" in :the site:https://www.tiktok.com/discover/ generates 500 search terms for advanced search on Google. The code used is as follows (In this article, all Python code will use pseudo code.):

```
Import pandas as pd
Read CSV file into DataFrame (df)
Define text template with placeholder for word
Initialize empty list (replaced_texts)
For each word in df:
    Format template with word
    Append formatted text to replaced_texts
Create DataFrame (result_df) from replaced_texts
Save result_df to new CSV file
Print completion message
```

Then we search on Google using advanced search terms like intitle:education intext:the OR intitle:the site:https://www.tiktok.com/discover/. Then directly copy and paste these Google search results into a world file. (Of course, you can also use a crawler to automatically search on Google and then directly copy and paste these Google search results into a world file.) In this study, we did not log in to the Google account in guest mode when searching for data on Google. Use Google browser, and make the following settings for Google browser, set Display language to English, set Results language filter to English, set Results region to United States, and set Search customization to of. When we collect data and encounter "repeat the search with the omitted results included" shown at the bottom of the page in Figure 3, we click "repeat the search with the omitted results included". You can set the Google browser according to your needs to obtain popular search results in different countries, regions, languages, etc.

Then we used Python code to quickly extract the topic name data of each TikTok discovery page topic in the world file, and used some filtering conditions to perform some corresponding screening. In this way we get the corresponding data. We used this method to collect more than 10,000 rows of unique topic name data on TikTok's discovery page topics. Use Python code to quickly extract the topic name data of each TikTok discovery page topic. The code used is as follows:

```
Import libraries
```

Define function to check Chinese characters

Set folder and CSV paths

Initialize written_lines set

Open CSV for writing

For each .docx file:

Read document

For each paragraph:

If font size is 20pt and contains "education":

Truncate text at '|'

If valid line:

Write line to CSV

Add line to written_lines

Print message

Some of the data we obtained are as follows:

Table 1. *Some examples of data we obtained.*

text_column
Gerry Brooks Video in A Free Education
Education Alive School
john spencer education

Users can also replace the Google advanced search terms intitle:education intext:the OR intitle:the site:https://www.tiktok.com/discover/ with intitle:education intext:of OR intitle:of intext" 20.7M views " site: https://www.tiktok.com/discover/, etc., so as to be able to search for more detailed, more and more accurate data. Of course, this method and algorithm can also be used to collect data in search engines other than Google.

data analysis:

After we used our methods and algorithms to obtain data on the topic of the TikTok Discover page, we used the following 11 codes to analyze the collected data from different levels or aspects.

4.2 1st Code for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Read CSV and extract text column

Clean text (remove non-ASCII)

Compute TF-IDF values

Get total heat for each word

Convert to DataFrame, transpose, and rename columns

Save DataFrame to CSV

Print completion message

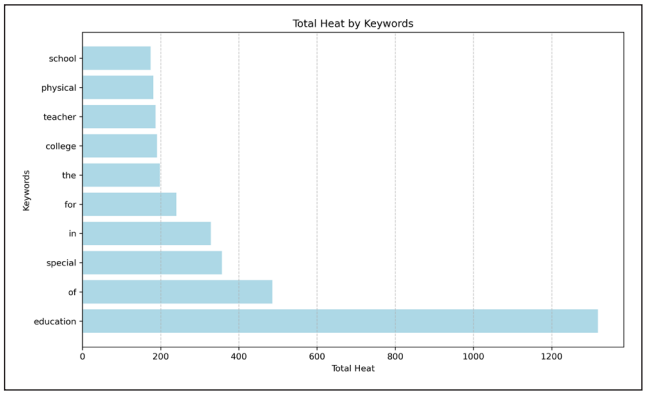

Figure 2. *The top 10 data output by the 1nd code.*

Analysis results:

We run this code [18]. In the output "File output by code 1.csv" file, there are a total of 9581 words and the "Total Heat" values corresponding to these 9581 words.

The top 10 are: education (1318.568818), of (485.6832716), special (356.6052461), in (328.508117), for (240.6092889), the (197.9038619), college (190.5411685), teacher (186.909698), physical (180.9941169), school (174.5185412).

Taking the ninth word "physical" as an example, we can predict that people may have relatively high interest or demand in knowledge related to "physical" or in the subject of physical. When we search and compare "biology" and "math" in the "File output by code 1.csv" file. Among them, "biology" is ranked 1106th among all words, and its "Total Heat" value is "4.824677856". Among them, "math" is ranked 208th among all words, and its "Total Heat" value is "19.94714983". Since the "Total Heat" value of "math" is greater than the "Total Heat" value of "biology". Therefore, we can predict that people may have higher interest or demand in knowledge related to "math" or corresponding subjects than to knowledge related to "biology" or related subjects.

4.3 Second Code for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Read CSV and select text column
Load BERT tokenizer and model

Compute BERT embeddings for each text
Calculate similarity matrix and heat scores
Add scores to DataFrame

Sort and reset index
Save DataFrame to CSV

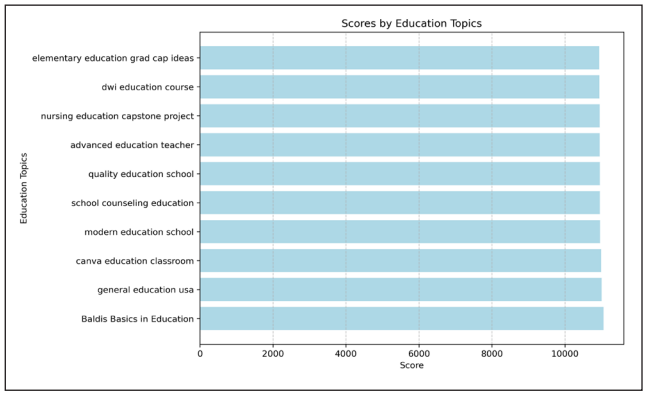

Figure 3. *The top 10 data output by the 2nd code.*

Analysis results:

We run this code [19]. In the "File output by code 2.csv" file it outputs, there are more than 16,000 rows of data and corresponding "score" values.

The top 10 are: Baldis Basics in Education (11058.722), general education usa (11003.078), canvas education classroom (10990.033), modern education school (10963.553), school counseling education (10956.008), quality education school (10955.442), advanced education teacher (10954.297), nursing education capstone project (10953.441), dwi education course (10944.557), elementary education grad cap ideas (10940.563).

Taking the data "quality education school" with the sixth-ranked "score" value as an example, we can predict that people value quality education, and people prefer schools that provide quality education.

4.4 3rd Code for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Read CSV and select text column
Split data into training and testing sets

Vectorize text data using TF-IDF
Train Naive Bayes classifier

Get predicted probabilities for test set
Rank texts by probabilities

Save ranking results to CSV

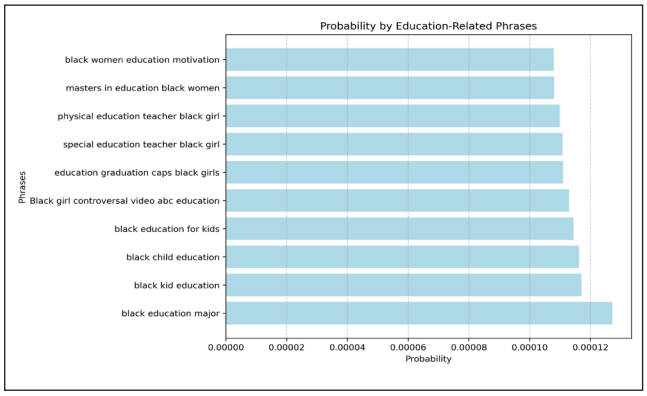

Figure 4. *The top 10 data output by the 3nd code.*

Analysis results:

We run this code [20]. In the "File output by code 3.csv" file it outputs, there are more than 3,000 rows of data and corresponding "Probability" values. The following are the first 3 rows of data sorted from high to low according to the "Probability" value:

Table 2. *The first 3 rows of data in the "File output by code 3.csv" file sorted from high to low according to the "Probability" value.*

Text	Probability
black education major	0.000127205
black kid education	0.000117005
black child education	0.000116268

The top 10 are: black education major (0.000127205), black kid education (0.000117005), black child education (0.000116268), black education for kids (0.000114405), Black girl controversal video abc education (0.000112992), education graduation caps black girls (0.000111029), special education teacher black girl (0.000110804), physical education teacher black girl (0.00010988), masters in education black women (0.000108051), black women education motivation (0.000107964).

From the first 3 rows of data, we can predict that educational content related to black will be very popular or concerned.

4.5 4rd Code for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Download stopwords
Read CSV file

Preprocess text data
Create dictionary and corpus

Run LDA model
Save model

Generate and save ranked topics

Analysis results:

We run this code [21], [22]. In the "File output by code 4.csv" file it outputs, there are a total of 10 subject data. The following is the data of the first topic:

Based on this topic data, we can predict that people may have a relatively high interest in information and content related to "major" in education.

4.6 Code 5 for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Table 3. *Data of the first topic in the "File output by code 4.csv" file.*

Topic	Words
0	department: 0.0837642252445221, major: 0.08170474320650101, system: 0.06884553283452988, higher: 0.0479467436671257, grade: 0.04165898263454437, program: 0.026927508413791656, institute: 0.024929115548729897, first: 0.02261144109070301, general: 0.01799621991813183, american: 0.016757341101765633

Read CSV file
Select text column

Preprocess text and extract features using TF-IDF
Apply KMeans clustering

Add cluster labels to dataframe
Save labeled dataframe to CSV

Reduce dimensions with t-SNE
Visualize clusters

Exit program

Analysis results:

We run this code [23], [24]. We use codes for classification. In the "File output by code 5.csv" file it outputs, the data is divided into 20 categories. You can find data with relatively high similarity in the same category for data analysis and data mining.

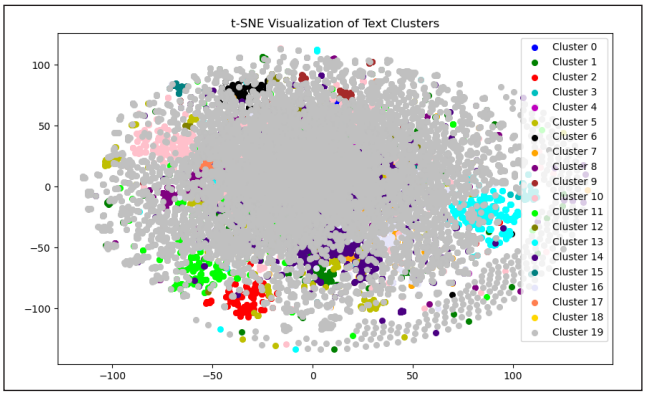

Figure 5. *Pictures of classification results.*

4.7 Code 6 for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Download models

Define functions for:
- Text processing
- Aspect extraction
- Sentiment determination

Read CSV, create pipeline, initialize counters

For each text:
Analyze, count, and print results

Save results to CSV

Calculate and print ratios
Analysis results:

We run this code [25]. The output is: Sentiment Ratios: {'positive': 0.07309252599691757, 'negative': 0.026165211075484064, 'neutral': 0.9007422629275984}. This result shows that people's attitude towards education is neutral to positive. And people can analyze changes in people's attitudes towards education in different time periods by collecting data in different time periods.

4.8 Code 7 for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Read CSV

Define functions:
- Build index
- Cosine similarity
- Term frequency

Main:
Create documents
Build index

For each document:
Calculate similarity

Sort and save results
Print similar documents
Analysis results:

We run this code [26]. We use the code to find the data we collected that are more similar to the text 'information and communication technology for education'. Of course, you can also search using other topics that interest you. In the "File output by code 7.csv" output file, we rank all the collected data according to the similarity with the text 'information and communication technology for education' from high to low. Here are the first 3 rows of data sorted by "Similarity" value from high to low:

Table 4. *The first 3 rows of data in the "File output by code 7.csv" file sorted from high to low according to the "Similarity" value.*

Document	Similarity
technology and education	0.707106781
technology education	0.577350269
communication education	0.577350269

You can use this method to quickly find the topic name data of the TikTok Discover page related to the topic you are interested in. Or used to analyze popular search words or popular search phrases, etc.

4.9 8th Code for Data Analysis and Data Mining:

The code used is shown below:

Import libraries

Read CSV and select text data

```
Convert text to bag-of-words

Define custom dataset class

Define Generator model

Define Discriminator model

Define GANomaly model:
    - Initialize generator and discriminator
    - Define training method

Main:
    Create dataset and data loader
    Initialize GANomaly model
    Train model

  Detect anomalies on test data

    Create results DataFrame
    Save results to CSV
    Analysis results:
```

We run this code [27], [28]. We used the code to find out and found 207 data that were marked as anomalies. This "anomaly" may mean that these sentences are different from other sentences, and we may have some unexpected gains when we analyze and mine these different data. The following is some of the data marked as anomalies:

Table 5. *Some of the data marked as anomalies.*

Text	Anomaly
What are you gonna even do with an education degree	TRUE
What are you going to do with a masters in education	TRUE
When to Apply for Uif for Education Assistants	TRUE

4.10 Code 9 for Data Analysis and Data Mining:

The code used is shown below:

```
Import libraries

Define weights for words

Read CSV and select text column

Set keyword

Initialize word scores

For each text in text data:
    Split text into words
    For each word in words:
        If word is keyword:
            For j from 1 to 10:
                If left word exists:
                    Update score for left word
                If right word exists:
                    Update score for right word

Create DataFrame from word scores
Round scores to one decimal

Save DataFrame to CSV
Print output file message
    Analysis results:
```

Here we invented and used the "Keyword Distance Weighted Frequency" algorithm for data analysis and mining. The "Keyword Distance Weighted Frequency" (KDWF) algorithm calculates the importance of words surrounding a keyword by assigning weights based on their proximity to the keyword in a sentence. Here's the general formula for this algorithm:

Weight calculation for words around a keyword:

Weight (wi, k) = max (0,1−0.1×Distance (wi, k))

wi is a word at position i,

k is the keyword,

Distance (wi, k) is the number of words between wi and k (to the left or right),

The weight decreases by 0.1 for each additional word distance from the keyword. (You can use different values according to different requirements.)

Score calculation for each word across multiple sentences: For a word www across multiple sentences S1, S2, ..., Sn containing the keyword k:

$$\textit{Total Score (w)} = \sum_{j=1}^{n} Weight\,(w, k\ in\ S_j)$$

where:

n is the number of sentences containing the keyword k,

Weight (w, k in Sj) is the weight of word w in the j^{th} sentence.

Handling multiple keywords: If a sentence contains multiple occurrences of the keyword k, the score for a word is the sum of its weights relative to each occurrence:

$$\textit{Total Score (w)} = \sum_{l=1}^{m} Weight\,(w, k_l)$$

where m is the number of keyword occurrences, and kl is the l^{th} occurrence of the keyword in the sentence.

In summary, the score of a word is determined by its proximity to the keyword, with closer words receiving higher weights. The total score of a word across sentences is the sum of its weighted scores.

The pseudo code of "Keyword Distance Weighted Frequency" (KDWF) is as follows:

```
Input: text_data, keyword, weights

Initialize: word_scores = empty dictionary

For each text in text_data:
words = text.lower().split()

For i, word in enumerate(words):
If word == keyword:
For j from 1 to 10:
If i - j >= 0:
left_word = words[i - j]
If left_word != keyword:
word_scores[left_word] += weights[j]

If i + j < len(words):
right_word = words[i + j]
If right_word != keyword:
word_scores[right_word] += weights[j]

Convert word_scores to DataFrame
Round scores
Save results to CSV
```

Because in a sentence, the words closest to a key word usually play an important role in the understanding and context of the key word. This is because adjacent words may provide more information, modification, or context about the keyword [29], [30].

For example, education in different lines of text such as "What Are You Gonna Do with Education", "What are you gonna even do with an education degree", "what to do with a masters degree in education cry" is counted as a keyword, The weight of the nearest word to the left and right of a keyword word appearing once is calculated as 1, and the weight of the second nearest word to the left and right of a keyword word to appear once is calculated as 0.9. The weight of the third word appearing once to the left and right of the keyword word is calculated as 0.8. The weight of the 10th word appearing once to the right and left of the keyword word is calculated as 0.1. But the word education is not counted. For example, if you use the three lines "What Are You Gonna Do with Education", "What are you gonna even do with an education degree" and "what to do with a masters degree in education cry", then in "What Are You Gonna Do with an education degree"? The score for "What" in "Do with Education" is 0.5. The score for "What" in "What are you gonna even do with an education degree" is 0.3. In "what to do

with a masters degree in education cry" the score of "What" is 0.3. The total score of "What" among these three pieces of data is 0.5+0.3+0.3 = 1.1.

In "What Are You Gonna Do with Education" the score for "are" is 0.6. The score for "are" in "What are you gonna even do with an education degree" is 0.4. There is no word "are" in "what to do with a masters degree in education cry", so the score is 0. The total score of "are" in these three pieces of data is 0.6+0.4+0 = 1.

There is no word "degree" in "What Are You Gonna Do with Education", so the score is 0. The score for "degree" in "What are you gonna even do with an education degree" is 1. There is no word "degree" in "what to do with a masters degree in education cry", so the score is 0. The total score of "degree" in these three pieces of data is 0+1+0 = 1.

In this way, the scores of all the words in the three lines of data "What Are You Gonna Do with Education", "What are you gonna even do with an education degree" and "what to do with a masters degree in education cry" are calculated in sequence.

If the input piece of data contains multiple keywords "education", in this case, we will consider the words surrounding each "education" keyword and accumulate the scores of each keyword. For example, in "what to do with a masters degree in education cry education", the score of "in" in the first "education" is 1, and the score of "in" in the second "education" is 0.8, then in "what The total score of "in" in "to do with a masters degree in education cry education" is: 1+0.8 = 1.8.

This algorithm is called "Keyword Distance Weighted Frequency".

Let's run this code. In the output "File output by code 9.csv" file, there are a total of more than 9,000 words and their corresponding "Score" values.

Taking the 10th word "minecraft" sorted from high to low according to the "Score" value as an example, we can predict that people may have relatively high interest or demand for "minecraft1"-related content, and can use "minecraft1"-related content in education. When we search and compare "art" and "science" in the "File output by code 1.csv" file. Among them, "art" is ranked 156th among all words, and its "Total Heat" value is "44.9". Among them, "science" is ranked 125th among all words, and its "Total Heat" value is "53.5". Since the "Score" value of "science" is greater than the "Score" value of "art". Therefore, we can predict that people may have higher interest or demand in knowledge related to "science" or corresponding subjects than to knowledge related to "art" or corresponding subjects.

4.11 Code 10 for Data Analysis and Data Mining:

The code used is shown below:

Import library

Read CSV file

Extract Score column

Calculate sum of Score column

Calculate result as Score divided by sum

Insert result into new column in DataFrame

Save DataFrame to new CSV file

Print success message

Import libraries

Read CSV file

Select text column for analysis

Initialize CountVectorizer

Transform text to word count vector

Initialize TfidfTransformer

Fit transformer to word count vector

Get list of words

Get IDF values

Create DataFrame with words and IDF values

Save DataFrame to CSV file

Print success message

Import pandas

Read first CSV file

Read second CSV file

Merge both DataFrames on 'Word'

Calculate product of specific columns

Save merged DataFrame to CSV file

Print success message

Analysis results:

The "Keyword Distance Weighted Frequency" algorithm we invented and proposed can be used alone or in combination with other algorithms. For example, here we use the "Keyword Distance Weighted Frequency" algorithm in combination with the "Inverse Document Frequency". A new algorithm "Keyword Distance Weighted Frequency-Inverse Document Frequency" is formed. Inverse Document Frequency measures the importance of words in the entire document collection. This algorithm combines the advantages of Keyword Distance Weighted Frequency and Inverse Document Frequency.

Let's run this code. In the output "File output by code 10.csv" file, there are a total of more than 9,000 words and their corresponding "Keyword Distance Weighted Frequency-Inverse Document Frequency" values, etc.

Taking the data "2023" with the 8th ranked "Keyword Distance Weighted Frequency-Inverse Document Frequency" value as an example, we can predict that the data we collect contains a large amount of data related to 2023. Taking the 7th ranked data "bad" as an example, we can predict that some people may be dissatisfied with educational or education-related content.

Comparison with Existing Methods

Although the KDWF-IDF (Keyword Distance Weighted Frequency-Inverse Document Frequency) algorithm and its variants are regarded as novel feature representation methods, a rigorous quantitative comparison with traditional text processing algorithms is crucial. To evaluate the performance of KDWF-IDF compared to classical algorithms, this study compares it with the traditional TF-IDF algorithm. In the process, cosine similarity is used to quantify the similarity between the text feature vectors generated by the two algorithms.

The cosine similarity calculated by the feature vectors of KDWF-IDF and TF-IDF is 0.9978, indicating that the two algorithms have extremely high consistency in text feature representation. A cosine similarity close to 1 means that the feature vectors generated by KDWF-IDF and TF-IDF are almost exactly the same in direction. This result shows that the KDWF-IDF method is basically consistent with TF-IDF in maintaining the core information of text features, while introducing a mechanism based on keyword distance weighting, which may provide additional fine-tuning capabilities in specific application scenarios.

Although the cosine similarity results show a high similarity between KDWF-IDF and TF-IDF, KDWF-IDF can still improve the flexibility and accuracy of feature expression in specific situations through its unique weighting mechanism. Therefore, KDWF-IDF not only retains the advantages of the traditional TF-IDF algorithm, but also has the potential to further improve performance in text analysis tasks. This comparison provides quantitative evidence for the effectiveness of KDWF-IDF, proves the robustness of the algorithm on traditional benchmarks such as TF-IDF, and lays the foundation for further research on its advantages in practical applications.

4.12 11th Code for Data Analysis and Data Mining:

The code used is shown below:

Import pandas and Counter

Read CSV file

Convert DataFrame to list of tuples

Count occurrences of rows

Add count to DataFrame

Drop duplicate rows

Save updated DataFrame to CSV file

Import pandas

Read CSV file

Filter out rows with 'education' in 'aspect' column

Extract 'Count' column

Calculate sum of 'Count'

Normalize 'Count' values

Insert normalized results into DataFrame

Save updated DataFrame to new CSV file

Import pandas

Read first CSV file

Read second CSV file

Merge DataFrames on the specified columns

Multiply specific columns and store result in a new column

Save the modified DataFrame to a new CSV file

Import pandas

Read CSV file

Select sentiment and frequency columns

Initialize counters and total frequency

For each row in sentiment column:

Add frequency to total

Update sentiment counters

Calculate total count of sentiments

Calculate percentages for each sentiment

Print total frequency and sentiment sums

Print sentiment percentages

Analysis results:

Here we use the "Keyword Distance Weighted Frequency" algorithm we invented and proposed in conjunction with sentiment analysis. A new algorithm "Keyword Distance Weighted Frequency-Emotion Analysis Frequency" is formed. This algorithm can perform more accurate sentiment analysis when there are keywords.

Let's run this code. The result is Total Sum: 0.0014664594825907539, Sum of Negative Sentiment: 9.278729754058212e-05, Sum of Neutral Sentiment: 0.0007849887007069968, Sum of Positive Sentiment: 0.0005886834843431 The proportion of various emotional results calculated by 047 is: Percentage of Negative Sentiment: 0.0632730045678874, Percentage of Neutral Sentiment: 0.5352951854627558, Percentage of Positive Sentiment: 0.40143180996935685.

From this we can know that people's emotions towards education are positive. And the Percentage of Positive Sentiment is as high as: 0.40143180996935685.

And we can use two or more of the above 11 codes to comprehensively perform data analysis and data mining. For example, we can first use the 5th code or the 7th code to find the topic data strip that interests you, and then use the 2nd code or the 3rd code to check the popularity of the topic data strip that interests you, etc.

The insights gleaned from our data analysis of TikTok's discovery page topics reveal significant trends that can be instrumental in shaping educational practices and policies. For instance, the prominence of terms related to "quality education" suggests an increasing public interest in high-standard educational environments. This insight could guide educators and administrators in developing programs that emphasize quality and effectiveness, ultimately improving student outcomes.

Moreover, our findings related to the popularity of certain subjects, such as "math" over "biology," indicate where educational resources might be allocated for greater impact. Educational institutions could leverage this data to enhance their curriculum offerings, prioritize professional development for educators in high-demand subjects, and create targeted marketing strategies to attract students.

Furthermore, the neutral-to-positive sentiment observed in attitudes toward education highlights an opportunity for educational stakeholders to cultivate this positive perception. Engaging with communities through social media platforms like TikTok can foster an interactive learning environment, encouraging student participation and community support.

Additionally, the versatility of our proposed data collection method allows for its application beyond TikTok, making it

a valuable tool for researchers and educators in various fields. By adapting this methodology to different platforms and content types, we can continuously refine our understanding of public interest in education and related topics.

Ultimately, the insights derived from this study underscore the importance of data-driven decision-making in education. By systematically analyzing trending topics and public sentiment, educators and policymakers can make informed choices that enhance learning experiences and address the evolving needs of students and communities.

5. Results and analysis

In this section, we introduce the relevant results and analysis of our use of some new algorithms to obtain, analyze and mine the topic data of education-related pages on TikTok Discover. We used a new algorithm and method we invented to quickly and reliably obtain data on TikTok Discover page topics, and used 11 codes to conduct data analysis and data mining. This includes using 3 new algorithms we invented and proposed for data analysis and data mining. They are "Keyword Distance Weighted Frequency", "Keyword Distance Weighted Frequency-Inverse Document Frequency" and "Keyword Distance Weighted Frequency-Emotion Analysis Frequency".

We employed 11 codes to perform comprehensive data analysis and mining on various aspects and levels of the acquired data. These encompass the following and additional items:

5.1 Analyze popular data from different aspects and levels:

This includes finding popular search words, popular search phrases, popular data bars, popular topics, and more in the data. These include using codes 1, 2, 3, 4, 9, 10, etc. Including using the "Keyword Distance Weighted Frequency" and "Keyword Distance Weighted Frequency-Inverse Document Frequency" algorithms we invented and proposed for data analysis and mining.

5.2 Allow users to find relevant data they are interested in simply and efficiently:

This involves utilizing classification and search data to enable consumers to easily and effectively locate pertinent information that aligns with their interests, including the utilization of codes such as 5 and 7.

5.3 Emotion analysis:

We employ sentiment analysis to assess the sentiment expressed in the gathered data. After careful analysis, we have determined that individuals generally hold favorable sentiments towards education. These include utilizing codes 6, 11, and so on. It encompasses the utilization of our self-developed "Keyword Distance Weighted Frequency-Emotion Analysis Frequency" technique for data analysis and mining.

5.4 Find possible anomalous data:

We examine the gathered data to detect potential irregularities. Individuals have the ability to observe and examine this potential irregularity data in order to identify any distinct patterns or deviations. These include utilizing code 8, among other methods.

6. Conclusions and future directions

In this section, we will provide a concise overview of the research and propose potential future research areas that involve the utilization of novel algorithms for acquiring, analyzing, and extracting education-related page topic data from TikTok Discover.

6.1 Summary of Findings

We conducted a comprehensive analysis and extraction of educational content on TikTok Discover. By examining popular data from many perspectives and levels, our research aimed to facilitate users in efficiently locating important information for future investigation. Analysis, sentiment analysis, and potential anomalous data were detected. The analysis and extraction of this data offer educational practitioners' significant insights that can be utilized to direct educational practice. We present novel algorithms and methodologies for the collection, analysis, and extraction of data.

6.2 A Complete Set of Methods from Data Collection to Data Analysis and Mining

In this study, we developed and used a complete set of

methods from data collection to data analysis and mining to study the acquisition, analysis and mining of education-related page topic data on TikTok Discover. This complete set of data collected Data analysis and mining methods can not only be used for education-related research, this complete set of methods and processes can be used in almost every industry. People can use this complete set of methods and processes to use data for research or commercial monetization.

6.3 A New Algorithm and Method for Obtaining Data

This research invented and proposed a new algorithm and method that can quickly and reliably obtain data on the theme of the TikTok Discover page. This is the world's first simple, fast, and efficient algorithm and method to obtain data on the theme of the TikTok Discover page. Of course, this new algorithm and method can also be used to collect data from other websites. The reason why we invented this new algorithm and method to obtain data on the Internet is because if people want to obtain data from some websites such as TikTok through crawlers, they may encounter many complex anti-crawler problems. The algorithm and method we invented can solve this problem very well, simple and efficient.

6.4 3 New Algorithms for Data Analysis and Data Mining

In this study, we invented and proposed three new algorithms for data analysis and data mining. They are "Keyword Distance Weighted Frequency", "Keyword Distance Weighted Frequency-Inverse Document Frequency" and "Keyword Distance Weighted Frequency-Emotion Analysis Frequency". People can use these three new algorithms for data analysis and data mining in various fields. These three new algorithms are particularly suitable for data analysis and data mining in data with keywords.

6.5 future Direction

In future research, we can collect this complete set of data into data analysis and mining methods, this new algorithm and method for obtaining data, and these three new algorithms for data analysis and data mining. Research in more detailed education-related directions. Such as "Computer Science Education", "Science Education", "Application of Artificial Intelligence in Education", "Science Education" and other more detailed research directions related to education. This complete set of data collection to data analysis and mining methods and these 3 new algorithms for data analysis and data mining can also be used in other fields such as "art", "entertainment", "artificial intelligence", "Science" and other research directions in other fields. And we can collect data in different time periods, conduct data analysis and data mining in certain fields in different time periods, and compare changes in different time periods.

The multiple algorithms invented and proposed in this study can be applied not only to the field of education but also to many other fields such as medicine, entertainment, art, and culture, agriculture, manufacturing, etc.

In future work, people can test on larger data sets to improve the accuracy of the algorithm. In addition, if a large amount of relevant data of a website is obtained through Google, there may be potential ethical and legal issues, which is also worthy of attention.

In future research directions, in addition to continuing to deepen research in the field of education, we can also apply this method and algorithm to the following specific directions:

Detailed research on educational topics: Future research can focus on more specific directions such as "user behavior analysis of online education platforms", "the application effect of virtual reality in education", and "differences in educational needs under different cultural backgrounds". This will help improve the pertinence and efficiency of educational content.

Cross-domain application: The application of algorithms should not be limited to the field of education, but can also be extended to fields such as healthcare (such as sentiment analysis of patient health data), art and entertainment (such as popular trend prediction), and agriculture (such as analysis of market demand for agricultural products). This will greatly improve the accuracy and operability of data-driven decision-making.

In terms of potential applications, this set of methods and algorithms has a wide range of application potential and can be used by various industries for data-driven tasks such

as market analysis, user behavior prediction, and trend discovery. These tools are not only suitable for academic research, but also can provide business insights for enterprises, thereby improving the accuracy of business decisions.

As for research limitations, our current work is mainly based on educational topic data from the TikTok Discover page, and may face limitations of different data sets and platforms in broader applications. In addition, algorithm performance may encounter bottlenecks when processing larger-scale data, especially in the analysis and mining of real-time data. Therefore, future research needs to focus on how to optimize algorithm performance on larger data sets while ensuring that data acquisition and processing comply with relevant ethical and legal regulations.

7. References

[1] R. S. J. D. Baker and K. Yacef. "The state of educational data mining in 2009: A review and future visions." *Journal of Educational Data Mining,* Vol. 1, No. 1, pp. 3-17, 2009.

[2] G. Siemens and R. S. J. d. Baker. "Learning analytics and educational data mining: towards communication and collaboration." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge,* Vancouver, Canada, pp. 252-254, 2012.

[3] R. S. J. d. Baker. "Data mining for education." *In: McGaw, B., Baker, E., Peterson, P. (eds.) International Encyclopedia of Education,* 3rd edn., Vol. 7, Elsevier, Oxford, pp. 112-118, 2010.

[4] P. Baepler and C. J. Murdoch. "Academic analytics and data mining in higher education." *International Journal for the Scholarship of Teaching & Learning,* Vol. 4, No. 2, 2010.

[5] R. S.Baker, A. T. Corbett, and K. R. Koedinger. "Detecting student misuse of intelligent tutoring systems." *Proceedings of 7th International Conference (ITS2004),* Maceió, Alagoas, Brazil, pp. 531-540, 2004.

[6] P. Long and G. Siemens. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE Review,* Vol. 46, No. 5, pp. 30-40, 2011.

[7] K. E.Arnold and M. D. Pistilli. "Course Signals at Purdue: Using Learning Analytics to Increase Student Success." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge,* pp. 267-270, 2012.

[8] D. Gašević, S. Dawson, and G. Siemens. "Let's Not Forget: Learning Analytics Are About Learning." *TechTrends,* Vol. 59, No. 1, pp. 64-71, 2015.

[9] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. "Analysis of a Very Large Web Search Engine Query Log." *ACM SIGIR Forum,* Vol. 33, No. 1, pp. 6-12, 1999.

[10] R. Jones and K. L. Klinkner. "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs." *Proceedings of the 17th ACM conference on Information and knowledge management,* pp. 699-708, 2008.

[11] T. Y. Liu. "Learning to Rank for Information Retrieval." *Foundations and Trends® in Information Retrieval,* Vol. 3, No. 3, pp. 225-331, 2009.

[12] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems,* Vol. 30, pp. 107-117, 1998.

[13] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc., 2009.

[14] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[15] H. Chen, R. H. Chiang, and V. C. Storey. "Business intelligence and analytics: From big data to big impact." *MIS Quarterly,* Vol. 36, No. 4, pp. 1165-1188, 2012.

[16] U. Fayyad, G. P. Shapiro, and P. Smyth. "From data mining to knowledge discovery in databases." *AI Magazine,* Vol. 17, No. 3, pp. 37-54, 1996.

[17] E. Brynjolfsson and A. McAfee. *The Second Machine*

Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton & Company, New York, London, 2014.

[18] K. S. Jones. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation,* Vol. 28, No. 1, pp. 11-21, 1972.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is All You Need." *Advances in Neural Information Processing Systems,* pp. 5998-6008, 2017.

[20] T. Bayes. "An essay towards solving a problem in the doctrine of chances." *Biometrika,* Vol. 45, No. 3-4, pp. 296-315, 1958.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research,* Vol. 3, pp. 993-1022, 2003.

[22] M. Hoffman, F. Bach, and D. Blei. "Online learning for latent dirichlet allocation." *Advances in Neural Information Processing Systems,* Vol. 23, pp. 1-9, 2010.

[23] J. MacQueen. "Some methods for classification and analysis of multivariate observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, California, USA: University of California Press, pp. 281-297, 1967.

[24] L. v. d. Maaten and G. Hinton. "Visualizing Data using t-SNE." *Journal of Machine Learning Research,* Vol. 9, pp. 2579-2605, November, 2008.

[25] W. Xue and T. Li. "Aspect Based Sentiment Analysis with Gated Convolutional Networks." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* Melbourne, Australia, pp. 2514-2523, 2018.

[26] H. E. Williams, J. Zobel, and D. Bahle. "Fast Phrase Querying With Combined Indexes." *ACM Transactions on Information Systems,* Vol. 22, No. 4, pp. 573-594, 2004. doi.org/10.1145/1028099.1028102.

[27] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR),* Vol. 41, No. 3, pp. 1-58, 2009.

[28] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets." *Advances in Neural Information Processing Systems,* Vol. 27, 2014.

[29] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science,* Vol. 41, No. 6, pp. 391-407, 1990.

[30] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval.* Cambridge University Press, Cambridge, 2009.