Tourism Destination Article Search Features using TF-IDF and Cosine similarity

Mizanul Ridho Aohana [1], Fitri Bimantoro [1]
[1]Teknik Informatika – Universitas Mataram, Jl. Majapahit 62, Mataram, 83115, Indonesia.

ARTICLE INFO

Article history:

Received July 20, 2023
Revised August 16, 2023
Accepted August 18, 2023

Keywords:

Term Frequency-Inverse Document Frequency (TF-IDF);
Cosine Similarity;
travel destinations;

ABSTRACT

In the current digital era, the increasing public interest in searching for information about travel destinations necessitates an effective and accurate search system. However, search results for travel destination articles often yield irrelevant or inadequate outcomes. To address this issue, this paper proposes applying the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm and Cosine Similarity in the search feature for travel destination articles. By employing these algorithms, the search system is anticipated to deliver more relevant and accurate results according to user needs. This research contributes to developing an effective search system for travel destination articles, assisting users in obtaining relevant and high-quality information about the destinations they are searching for. The methodology involves collecting data on travel destination articles, implementing the TF-IDF algorithm to evaluate term importance, and utilizing Cosine Similarity to measure the similarity between articles and user queries. The study results demonstrate that implementing the TF-IDF algorithm and Cosine Similarity in the search feature for travel destination articles enhances the accuracy and relevance of search results. Users can quickly discover articles that align with their queries, improving their search experience. In conclusion, this research highlights that applying the TF-IDF algorithm and Cosine Similarity in the search feature significantly improves the accuracy and relevance of search results for travel destination articles. This enhances the search experience for users seeking information about travel destinations.

Corresponding Author:

Mizanul Ridho Aohana, Teknik Informatika - Universitas Mataram, Jl. Majapahit 62, Mataram, 83115, Indonesia.
Email:mizanulridhoaohana@mhs.unram.ac.id

1. INTRODUCTION

Information Retrieval is the core of various real-world applications, such as digital libraries, expert discovery, web search, etc. Fundamentally, Information Retrieval obtains relevant information resources from an extensive collection based on information needs. Since multiple relevant resources exist, the returned results are typically ranked based on relevance [1]. In Information Retrieval, each word or term is assumed to be a different dimension, and documents are represented as vectors where the value of each dimension corresponds to the frequency of that term in the document [2].

Information retrieval aims to provide the most relevant results to the user's information needs from the available database. This process involves determining the best information that matches the user's request. Users can obtain relevant and valuable information with a good understanding and implementation of information retrieval [3].

The previous research on the development of an information system for the processing of research and community service activities (PKM) with the implementation of the Cosine Similarity algorithm has provided a strong foundation for the current research titled "Application of TF-IDF and Cosine Similarity Algorithms on Tourism Destination Article Search Features." The previous research demonstrated that using the Cosine Similarity algorithm in detecting content similarity successfully addressed the issue of proposal similarity in research [4].

In addition, this research also refers to the implementation of TF-IDF weighting and cosine similarity algorithms in the SiPaGa application [5], which has provided a strong foundation for the current research focusing on applying these algorithms in the tourism destination article search feature. By applying the TF-IDF and Cosine Similarity algorithms to the tourism destination article search feature, the current research aims to enhance the accuracy and relevance of the search results for tourism destination articles. Building upon the foundation laid by the previous research, it is anticipated that the current research will further contribute to developing improved tourism destination article search features.

The current research aims to develop and implement the TF-IDF and Cosine Similarity algorithms in the tourism destination article search feature. Building upon previous research, this study seeks to enhance the accuracy and relevance of search results for tourism destination articles. By applying TF-IDF weighting and Cosine Similarity calculations, this research will provide an effective method to identify and retrieve articles that align with user search queries. Additionally, this research is expected to contribute to developing improved tourism destination article search features, offering users a more efficient and accurate search experience. Through this study, it is anticipated that solutions can be found to enhance the effectiveness of tourism destination article searches and provide more significant benefits to users.

2. METHOD

The writers utilized various techniques in this study, such as text preprocessing, term weighting, Term Frequency-Inverse Document Frequency (TF-IDF), and cosine similarity.

2.1. Text Preprocessing

Unstructured textual data requires several initial stages in text mining to prepare the text into a more structured form. One of the implementations in text mining is the text preprocessing stage[6]. In Text Mining, data preprocessing is used to uncover exciting and significant knowledge from unstructured textual data [7]. The text preprocessing stage also involves selecting and eliminating meaningless words [8]. Four standard stages in text preprocessing are considered: stopword removal, stemming, case folding, and tokenization [9].

1. Case folding: Case folding is performed to convert all characters in the text to lowercase for uniformity [5].
2. Tokenizing: This process is carried out to separate the text into individual features (tokens) that will be processed by the system [5]. The removal step is performed on whitespace characters during this process as they do not influence the text preprocessing stage [10].
3. Stopword: This stage selects important words from tokenization results or removes words considered less important in text mining [5].
4. Stemming: Stemming aims to transform a word into its base form by removing all word affixes [5]. In other words, stemming is changing the word form into its base form by removing prefixes and suffixes, enabling a more optimal text mining process [10].

Text preprocessing plays a crucial role as the first stage in the workflow [11]. Each step in text preprocessing can be adjusted according to the type and condition of the data. The text preprocessing steps generate a collection of words that will be used as an index [10].

2.2. Term Weighting

Term weighting is a process of converting the indices or features resulting from preprocessing textual data into numerical data by assigning values or weights to each word. The output of the term weighting process can be used in the classification process. One of the methods used is raw-term frequency weighting, which measures the weight of a word in a document based on its frequency of occurrence in that document. This can be seen in Equation (1).

$$w_{t,d} = tf_{t,d} \tag{1}$$

Where $w_{t,d}$ represents the weight of word t in document d, and $tf_{t,d}$is the frequency of occurrence of word t in document d [10]. Furthermore, to obtain the weighted value of each word in the data being used, word weighting is performed based on Term Frequency-Inverse Document Frequency (TF-IDF) [12].

2.3. Term Frequency – Inverse Document Frequency (TF-IDF)

The TF-IDF (Term Frequency-Inverse Document Frequency) process involves assigning values to each term or feature by calculating the Term Frequency (TF), then calculating the Inverse Document Frequency

(IDF), and finally computing the TF-IDF value. The weight or value of each feature that has been calculated is used for the subsequent weight normalization step. The normalized weight values are then used to calculate cosine similarity [6].

The TF-IDF weight is calculated locally from the Post editing dataset using the word frequency as TF [13]. This method combines two concepts in calculating the weight: the frequency of a word appearing in a particular document and the inverse frequency of documents containing that word. The frequency of a word appearing in a specific document indicates its importance in that document [5]. The formulas for calculating TF-IDF are as follows [14]:

$$tf = 0.5 + 0.5\ x\ \frac{tf}{\max(tf)} \tag{2}$$

$$idf_t = \log(\frac{D}{df_t}) \tag{3}$$

$$W_{d.t} = tf_{d.t}\ x\ idf_{d.t} \tag{4}$$

Where D represents the total number of documents, d represents the specific document, t represents the specific term or feature in the document, W represents the weight of document d concerning term t, tf represents the frequency of term t in a document, idf represents the Inverse Document Frequency, df represents the number of documents containing term t.

2.4. Cosine Similarity

Cosine similarity is an algorithm that can be used to calculate the similarity level between documents. Its basic principle involves measuring similarity in a vector space (vector space similarity measure). The cosine similarity algorithm uses keywords in a document to calculate the similarity between those documents, expressed in vector form [4]. One of the drawbacks of the Cosine Similarity method is its failure to consider word frequency in the document. This can lead to inaccuracies in measuring the level of similarity between documents. However, these limitations can be overcome by applying TF-IDF (Term Frequency-Inverse Document Frequency) weighting [15].

Cosine similarity is used to measure the degree of similarity between two vectors. In this method, the dot product of the two vectors is normalized by dividing it by the product of the Euclidean lengths of the vectors. Cosine similarity can be implemented to calculate the similarity between sentences and is a popular technique for measuring text similarity [8]. The formula used in cosine similarity [14]:

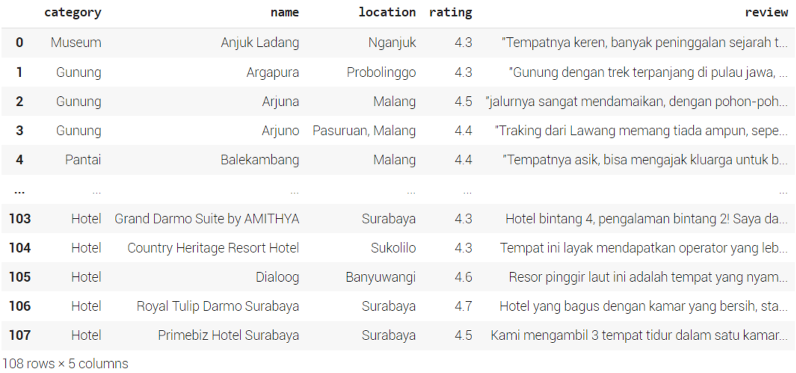
$$Cos\ a = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i\ x\ B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\ x\ \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{5}$$

Where, A is vector A, to be compared for similarity, B is vector B, to be compared for similarity, $A \cdot B$ is dot product between vector A and vector B, $|A|$ is length of vector A, $|B|$ length of vector B, and $|A||B|$ is cross product between |A| and |B|.

3. RESULTS AND DISCUSSION

3.1. Dataset

Prior to integrating the algorithm into the application, a dataset is essential as a critical component to assess the quality of the created recommendation system. In this project, the writers has curated two tourism area datasets in .csv file format from Kaggle as the data source. These datasets will be employed to evaluate the performance of the upcoming recommendation system within the application.

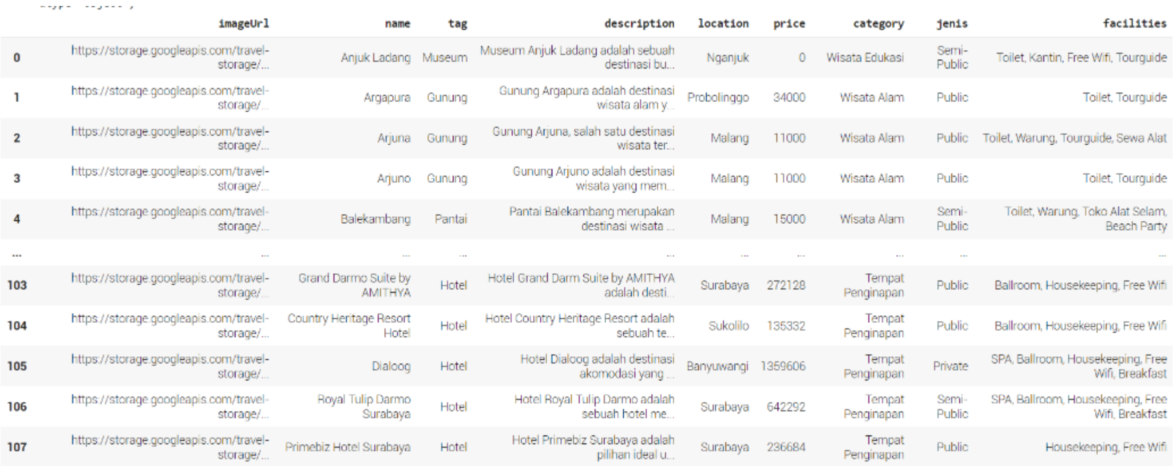


	category	name	location	rating	review
0	Museum	Anjuk Ladang	Nganjuk	4.3	"Tempatnya keren, banyak peninggalan sejarah t...
1	Gunung	Argapura	Probolinggo	4.3	"Gunung dengan trek terpanjang di pulau jawa, ...
2	Gunung	Arjuna	Malang	4.5	"jalurnya sangat mendamaikan, dengan pohon-poh...
3	Gunung	Arjuno	Pasuruan, Malang	4.4	"Traking dari Lawang memang tiada ampun, sepe...
4	Pantai	Balekambang	Malang	4.4	"Tempatnya asik, bisa mengajak kluarga untuk b...
...	...	...	...	...	...
103	Hotel	Grand Darmo Suite by AMITHYA	Surabaya	4.3	Hotel bintang 4, pengalaman bintang 2! Saya da...
104	Hotel	Country Heritage Resort Hotel	Sukolilo	4.3	Tempat ini layak mendapatkan operator yang leb...
105	Hotel	Dialoog	Banyuwangi	4.6	Resor pinggir laut ini adalah tempat yang nyam...
106	Hotel	Royal Tulip Darmo Surabaya	Surabaya	4.7	Hotel yang bagus dengan kamar yang bersih, sta...
107	Hotel	Primebiz Hotel Surabaya	Surabaya	4.5	Kami mengambil 3 tempat tidur dalam satu kamar...

108 rows × 5 columns

Fig. 1. Tourist Review and Rating Dataset

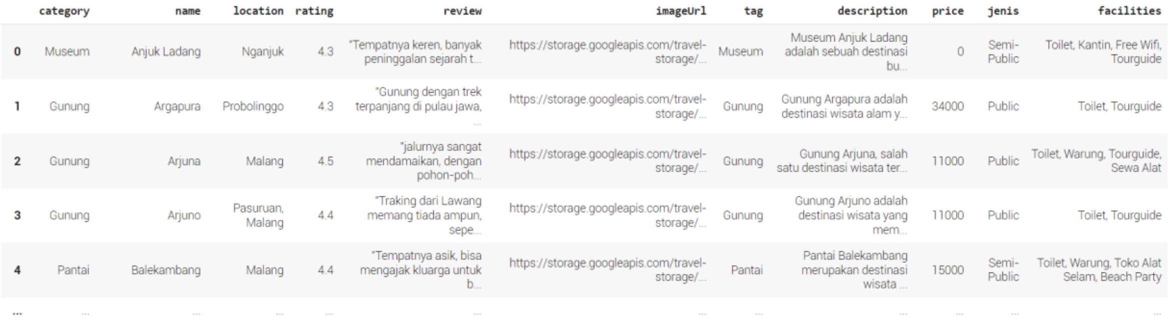
Meanwhile, the second dataset focuses more on reviews and ratings given by users to tourist destination locations. So that this dataset consists of 5 columns, namely, category, name, location, rating, and review.



	imageUrl	name	tag	description	location	price	category	jenis	facilities
0	https://storage.googleapis.com/travel-storage/...	Anjuk Ladang	Museum	Museum Anjuk Ladang adalah sebuah destinasi bu...	Nganjuk	0	Wisata Edukasi	Semi-Public	Toilet, Kantin, Free Wifi, Tourguide
1	https://storage.googleapis.com/travel-storage/...	Argapura	Gunung	Gunung Argapura adalah destinasi wisata alam y...	Probolinggo	34000	Wisata Alam	Public	Toilet, Tourguide
2	https://storage.googleapis.com/travel-storage/...	Arjuna	Gunung	Gunung Arjuna, salah satu destinasi wisata ter...	Malang	11000	Wisata Alam	Public	Toilet, Warung, Tourguide, Sewa Alat
3	https://storage.googleapis.com/travel-storage/...	Arjuno	Gunung	Gunung Arjuno adalah destinasi wisata yang mem...	Malang	11000	Wisata Alam	Public	Toilet, Tourguide
4	https://storage.googleapis.com/travel-storage/...	Balekambang	Pantai	Pantai Balekambang merupakan destinasi wisata ...	Malang	15000	Wisata Alam	Semi-Public	Toilet, Warung, Toko Alat Selam, Beach Party
...	...	...	...	...	...	...	...	...	...
103	https://storage.googleapis.com/travel-storage/...	Grand Darmo Suite by AMITHYA	Hotel	Hotel Grand Darm Suite by AMITHYA adalah desti...	Surabaya	272128	Tempat Penginapan	Public	Ballroom, Housekeeping, Free Wifi
104	https://storage.googleapis.com/travel-storage/...	Country Heritage Resort Hotel	Hotel	Hotel Country Heritage Resort adalah sebuah te...	Sukolilo	135332	Tempat Penginapan	Public	Ballroom, Housekeeping, Free Wifi
105	https://storage.googleapis.com/travel-storage/...	Dialoog	Hotel	Hotel Dialoog adalah destinasi akomodasi yang ...	Banyuwangi	1359606	Tempat Penginapan	Private	SPA, Ballroom, Housekeeping, Free Wifi, Breakfast
106	https://storage.googleapis.com/travel-storage/...	Royal Tulip Darmo Surabaya	Hotel	Hotel Royal Tulip Darmo adalah sebuah hotel me...	Surabaya	642292	Tempat Penginapan	Semi-Public	SPA, Ballroom, Housekeeping, Free Wifi, Breakfast
107	https://storage.googleapis.com/travel-storage/...	Primebiz Hotel Surabaya	Hotel	Hotel Primebiz Surabaya adalah pilihan ideal u...	Surabaya	236684	Tempat Penginapan	Public	Housekeeping, Free Wifi

Fig. 2. Tourism Destinations and Facilities Dataset

The above dataset focuses on describing each destination and its facilities. This dataset consists of several columns, namely, imgurl, name, tag, description, location, price category type and facilities.



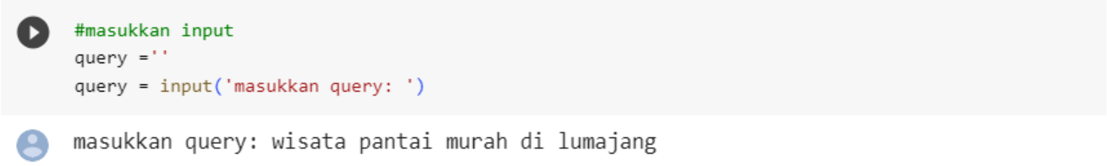
	category	name	location	rating	review	imageUrl	tag	description	price	jenis	facilities
0	Museum	Anjuk Ladang	Nganjuk	4.3	"Tempatnya keren, banyak peninggalan sejarah t...	https://storage.googleapis.com/travel-storage/...	Museum	Museum Anjuk Ladang adalah sebuah destinasi bu...	0	Semi-Public	Toilet, Kantin, Free Wifi, Tourguide
1	Gunung	Argapura	Probolinggo	4.3	"Gunung dengan trek terpanjang di pulau jawa, ...	https://storage.googleapis.com/travel-storage/...	Gunung	Gunung Argapura adalah destinasi wisata alam y...	34000	Public	Toilet, Tourguide
2	Gunung	Arjuna	Malang	4.5	"jalurnya sangat mendamaikan, dengan pohon-poh...	https://storage.googleapis.com/travel-storage/...	Gunung	Gunung Arjuna, salah satu destinasi wisata ter...	11000	Public	Toilet, Warung, Tourguide, Sewa Alat
3	Gunung	Arjuno	Pasuruan, Malang	4.4	"Traking dari Lawang memang tiada ampun, sepe...	https://storage.googleapis.com/travel-storage/...	Gunung	Gunung Arjuno adalah destinasi wisata yang mem...	11000	Public	Toilet, Tourguide
4	Pantai	Balekambang	Malang	4.4	"Tempatnya asik, bisa mengajak kluarga untuk b...	https://storage.googleapis.com/travel-storage/...	Pantai	Pantai Balekambang merupakan destinasi wisata ...	15000	Semi-Public	Toilet, Warung, Toko Alat Selam, Beach Party
...	...	...	...	...	...	...	...	...	...	...	...

Fig. 3. Merge Dataset

From the two datasets that have been collected above, the writers combine the datasets to facilitate the preprocessing and data processing process. Datasets that have the same labels or columns will be dropped, leaving unique columns. so that unique columns remain. These columns are category, name, location, rating, review, imgurl, tag, price, type, and facilities.

3.2. Experiment

Writers implementing the TF-IDF algorithm and cosine similarity for a destination travel article search feature, resulting in a picture as shown below. We have three conditions and three queries in the writer's experience. Condition 1 means that the program displays three search results articles, Condition 2 means five results articles, and Condition 3 means that the program displays ten search results articles. Here is an example of the utilization of TF-IDF and Cosine Similarity algorithms that have been implemented.

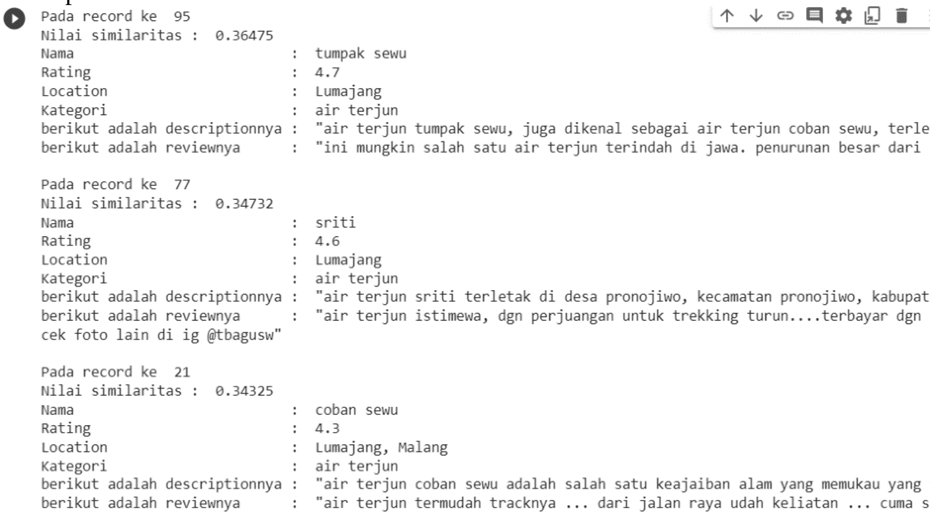


```
#masukkan input
query =''
query = input('masukkan query: ')
```

```
masukkan query: wisata pantai murah di lumajang
```

Fig. 4. Searching Query

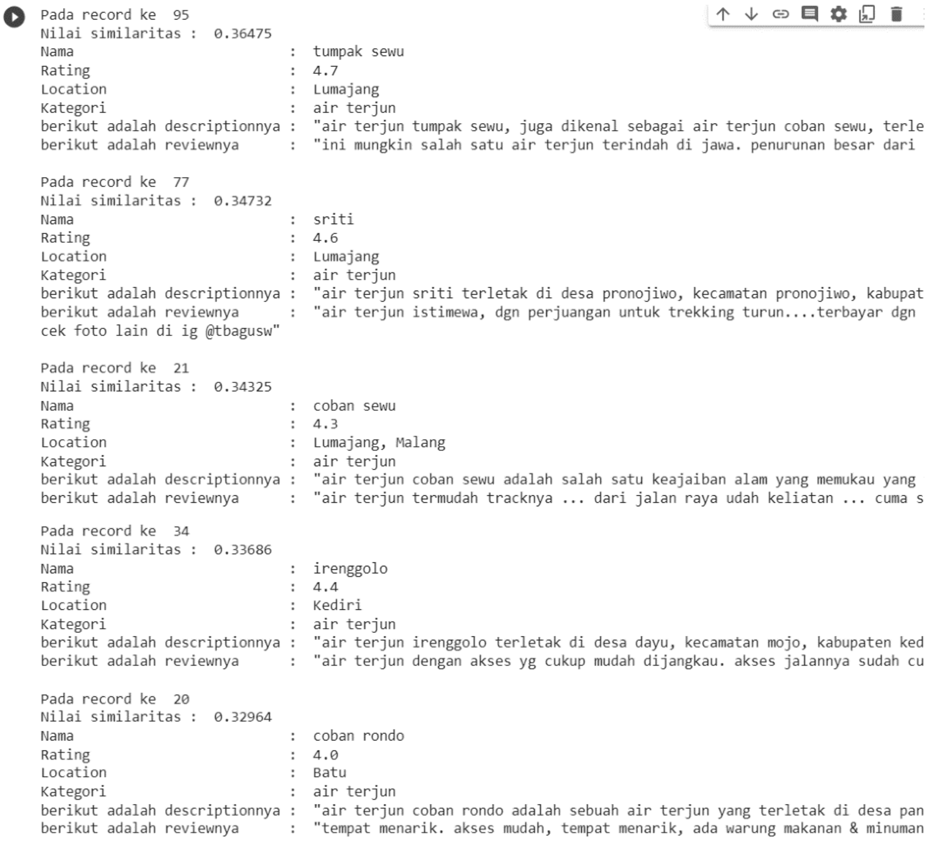
In Figure 1, the writers utilized the query "wisata air terjun keren di lumajang" to evaluate the quality of the algorithms implemented in the current search feature.



```
Pada record ke  95
Nilai similaritas :  0.36475
Nama                          :  tumpak sewu
Rating                        :  4.7
Location                      :  Lumajang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun tumpak sewu, juga dikenal sebagai air terjun coban sewu, terle
berikut adalah reviewnya      :  "ini mungkin salah satu air terjun terindah di jawa. penurunan besar dari

Pada record ke  77
Nilai similaritas :  0.34732
Nama                          :  sriti
Rating                        :  4.6
Location                      :  Lumajang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun sriti terletak di desa pronojiwo, kecamatan pronojiwo, kabupat
berikut adalah reviewnya      :  "air terjun istimewa, dgn perjuangan untuk trekking turun....terbayar dgn
cek foto lain di ig @tbagusw"

Pada record ke  21
Nilai similaritas :  0.34325
Nama                          :  coban sewu
Rating                        :  4.3
Location                      :  Lumajang, Malang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun coban sewu adalah salah satu keajaiban alam yang memukau yang
berikut adalah reviewnya      :  "air terjun termudah tracknya ... dari jalan raya udah keliatan ... cuma s
```

Fig. 5. Displaying 3 results

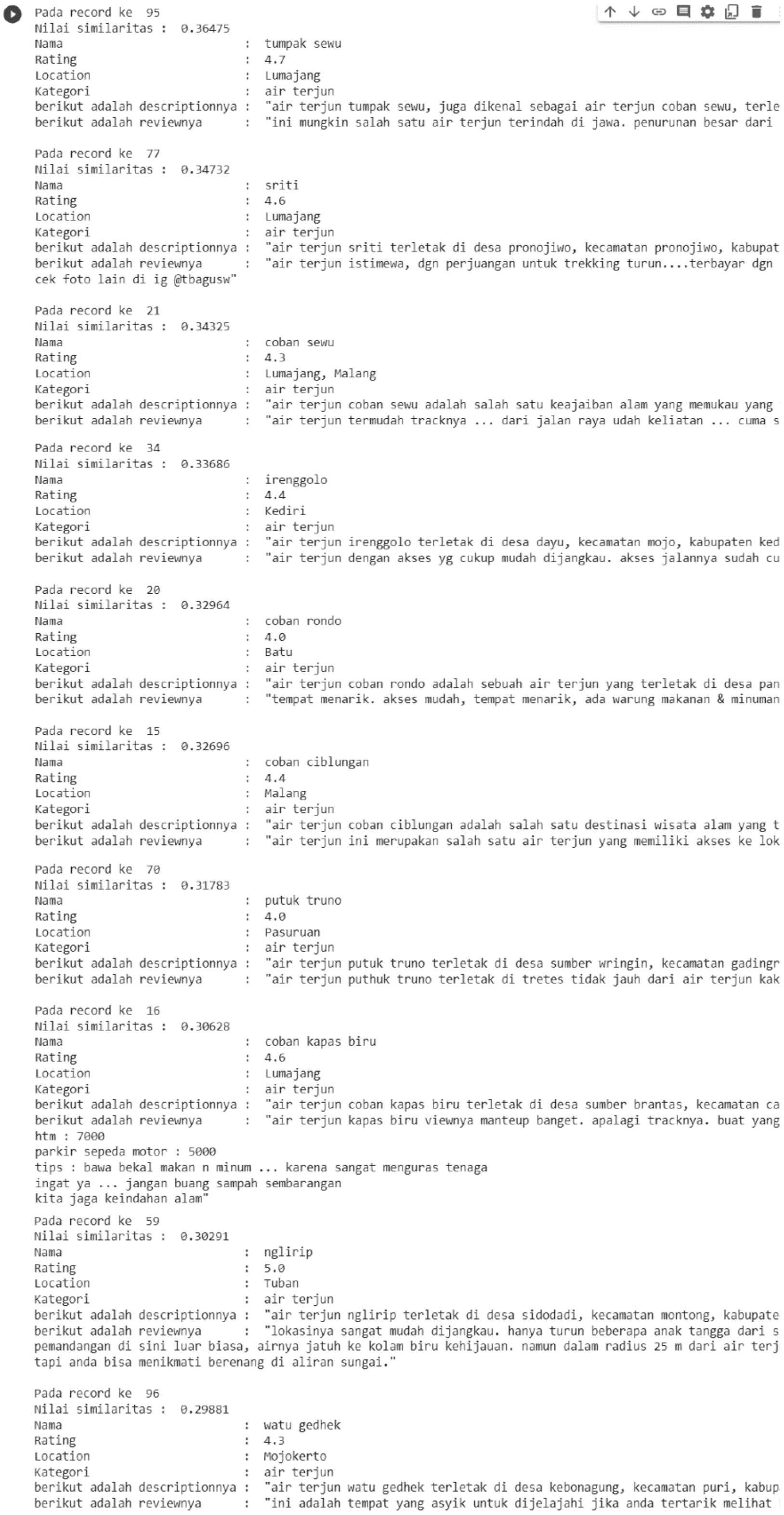
In the first experiment, the results obtained were as shown in the above figure. Based on the figure, the recommended articles align with what the user or query desires. The highest similarity value is 0.36475 in record 95.



```
Pada record ke  95
Nilai similaritas :  0.36475
Nama                         :  tumpak sewu
Rating                       :  4.7
Location                     :  Lumajang
Kategori                     :  air terjun
berikut adalah descriptionnya :  "air terjun tumpak sewu, juga dikenal sebagai air terjun coban sewu, terle
berikut adalah reviewnya     :  "ini mungkin salah satu air terjun terindah di jawa. penurunan besar dari

Pada record ke  77
Nilai similaritas :  0.34732
Nama                         :  sriti
Rating                       :  4.6
Location                     :  Lumajang
Kategori                     :  air terjun
berikut adalah descriptionnya :  "air terjun sriti terletak di desa pronojiwo, kecamatan pronojiwo, kabupat
berikut adalah reviewnya     :  "air terjun istimewa, dgn perjuangan untuk trekking turun....terbayar dgn
cek foto lain di ig @tbagusw"

Pada record ke  21
Nilai similaritas :  0.34325
Nama                         :  coban sewu
Rating                       :  4.3
Location                     :  Lumajang, Malang
Kategori                     :  air terjun
berikut adalah descriptionnya :  "air terjun coban sewu adalah salah satu keajaiban alam yang memukau yang
berikut adalah reviewnya     :  "air terjun termudah tracknya ... dari jalan raya udah keliatan ... cuma s

Pada record ke  34
Nilai similaritas :  0.33686
Nama                         :  irenggolo
Rating                       :  4.4
Location                     :  Kediri
Kategori                     :  air terjun
berikut adalah descriptionnya :  "air terjun irenggolo terletak di desa dayu, kecamatan mojo, kabupaten ked
berikut adalah reviewnya     :  "air terjun dengan akses yg cukup mudah dijangkau. akses jalannya sudah cu

Pada record ke  20
Nilai similaritas :  0.32964
Nama                         :  coban rondo
Rating                       :  4.0
Location                     :  Batu
Kategori                     :  air terjun
berikut adalah descriptionnya :  "air terjun coban rondo adalah sebuah air terjun yang terletak di desa pan
berikut adalah reviewnya     :  "tempat menarik. akses mudah, tempat menarik, ada warung makanan & minuman
```

Fig. 6. Displaying 5 results

In the second experiment, the writers limited the recommended articles to 5 and displayed the results in the above figure. Of the five recommendations, three align with what the user wants, which is "air terjun keren" in "Lumajang". However, the other two results slightly deviate from the user's intention. These results display articles with the term "air terjun keren" but in different locations that are not desired by the user, such as the cities of "Kediri" and "Batu". The highest similarity is found in record 95, with a value of 0.36475.



```
Pada record ke  95
Nilai similaritas :  0.36475
Nama                          :  tumpak sewu
Rating                        :  4.7
Location                      :  Lumajang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun tumpak sewu, juga dikenal sebagai air terjun coban sewu, terle
berikut adalah reviewnya      :  "ini mungkin salah satu air terjun terindah di jawa. penurunan besar dari

Pada record ke  77
Nilai similaritas :  0.34732
Nama                          :  sriti
Rating                        :  4.6
Location                      :  Lumajang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun sriti terletak di desa pronojiwo, kecamatan pronojiwo, kabupat
berikut adalah reviewnya      :  "air terjun istimewa, dgn perjuangan untuk trekking turun....terbayar dgn
cek foto lain di ig @tbagusw"

Pada record ke  21
Nilai similaritas :  0.34325
Nama                          :  coban sewu
Rating                        :  4.3
Location                      :  Lumajang, Malang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun coban sewu adalah salah satu keajaiban alam yang memukau yang
berikut adalah reviewnya      :  "air terjun termudah tracknya ... dari jalan raya udah keliatan ... cuma s

Pada record ke  34
Nilai similaritas :  0.33686
Nama                          :  irenggolo
Rating                        :  4.4
Location                      :  Kediri
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun irenggolo terletak di desa dayu, kecamatan mojo, kabupaten ked
berikut adalah reviewnya      :  "air terjun dengan akses yg cukup mudah dijangkau. akses jalannya sudah cu

Pada record ke  20
Nilai similaritas :  0.32964
Nama                          :  coban rondo
Rating                        :  4.0
Location                      :  Batu
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun coban rondo adalah sebuah air terjun yang terletak di desa pan
berikut adalah reviewnya      :  "tempat menarik. akses mudah, tempat menarik, ada warung makanan & minuman

Pada record ke  15
Nilai similaritas :  0.32696
Nama                          :  coban ciblungan
Rating                        :  4.4
Location                      :  Malang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun coban ciblungan adalah salah satu destinasi wisata alam yang t
berikut adalah reviewnya      :  "air terjun ini merupakan salah satu air terjun yang memiliki akses ke lok

Pada record ke  70
Nilai similaritas :  0.31783
Nama                          :  putuk truno
Rating                        :  4.0
Location                      :  Pasuruan
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun putuk truno terletak di desa sumber wringin, kecamatan gadingr
berikut adalah reviewnya      :  "air terjun puthuk truno terletak di tretes tidak jauh dari air terjun kak

Pada record ke  16
Nilai similaritas :  0.30628
Nama                          :  coban kapas biru
Rating                        :  4.6
Location                      :  Lumajang
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun coban kapas biru terletak di desa sumber brantas, kecamatan ca
berikut adalah reviewnya      :  "air terjun kapas biru viewnya manteup banget. apalagi tracknya. buat yang
htm : 7000
parkir sepeda motor : 5000
tips : bawa bekal makan n minum ... karena sangat menguras tenaga
ingat ya ... jangan buang sampah sembarangan
kita jaga keindahan alam"
Pada record ke  59
Nilai similaritas :  0.30291
Nama                          :  nglirip
Rating                        :  5.0
Location                      :  Tuban
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun nglirip terletak di desa sidodadi, kecamatan montong, kabupate
berikut adalah reviewnya      :  "lokasinya sangat mudah dijangkau. hanya turun beberapa anak tangga dari s
pemandangan di sini luar biasa, airnya jatuh ke kolam biru kehijauan. namun dalam radius 25 m dari air terj
tapi anda bisa menikmati berenang di aliran sungai."

Pada record ke  96
Nilai similaritas :  0.29881
Nama                          :  watu gedhek
Rating                        :  4.3
Location                      :  Mojokerto
Kategori                      :  air terjun
berikut adalah descriptionnya :  "air terjun watu gedhek terletak di desa kebonagung, kecamatan puri, kabup
berikut adalah reviewnya      :  "ini adalah tempat yang asyik untuk dijelajahi jika anda tertarik melihat
```

Fig. 7. Displaying 10 results

In the third experiment, the writers attempted to display the top 10 most relevant data to the user's query. The above figure shows that the top 5 article data produce the same results as in the second experiment. This indicates the algorithm's consistency in generating the displayed recommendations.

For the following five results, it can be observed that 4 out of 5 recommendations still deviate from what the user wants. Upon further analysis, these articles are related to the user's search query in terms of description and review available in the dataset. However, the critical parameter of location is overlooked in the generated recommendations.

An extension of this finding shows that while the algorithm can recognize content relationships and provide matching recommendations regarding descriptions and reviews, it still needs to consider the location factor. This parameter should be critical in generating recommendations more appropriate to users' needs and preferences. Thus, further efforts in integrating location aspects into recommendation algorithms can be a valuable step in improving the quality of recommendations.

3.3. Result Comparison

Here is an explanation and summary of the comparison, as well as the conclusions from the cosine similarity experiment that has been conducted:

Table 1. Result Comparison

Query	***Condition 1***	***Condition 2***	***Condition 3***
	True	***True***	***True***
1	3/3	3/5	4/10
2	3/3	4/5	9/10
3	3/3	4/5	5/10

In Condition 1, the program presents three articles for each query. In this scenario, the program consistently displays all relevant articles for the given queries. This indicates that, for every query, all three displayed articles are relevant.

Moving to Condition 2, the program exhibits five articles for each query. In this condition, the program still produces relatively satisfactory results, but there is some variation in the level of relevance. For Query 1, the program showcases three relevant articles out of the five displayed. As for Query 2 and Query 3, the program displays four relevant articles out of the total five exhibited.

Transitioning to Condition 3, the program displays ten articles for each query. In this situation, there is an increase in the number of articles displayed, but with this increase comes a decrease in the level of relevance. For Query 1, the program only manages to show four relevant articles out of the ten displayed. Similarly, for Query 2, the program displays four relevant articles out of the total ten exhibited. Additionally, with Query 3, the program presents five relevant articles out of the ten displayed.

Taken together, these experiment outcomes illustrate how the program provides article recommendations based on the number of articles displayed within a query. While the quantity of articles can influence the level of relevance, there is also variability that needs to be considered.

3.4. Similarity Analysis

Based on previous results, the writers analyzed the minimum and maximum similarity values obtained previously. In this analysis, the writers observed the lowest and highest similarity values that emerged in the previous experiments.

After observing the lowest and highest similarity values, the writers noted patterns that emerged from these results. These findings provide further insights into how variations in similarity values can influence the generated recommendations. Furthermore, the writers will delve deeper into analyzing the contributing factors to these differences, such as content characteristics, similarity calculation methods, and the number of displayed articles.

Table 2. Similarity Analysis

Query	***Condition 1***		***Condition 2***		***Condition 3***	
	Max	***Min***	***Max***	***Min***	***Max***	***Min***
1	0.36475	0.34325	0.36474	0.32964	0.36475	0.29881
2	0.12536	0.10401	0.12536	0.07865	0.12537	0.06966
3	0.24671	0.23275	0.24670	0.21649	0.24671	0.18393

Based on the data table above, there is a decrease in the minimum similarity value as the number of displayed articles increases from Condition 1 to Condition 3. This indicates that the more articles are displayed, the lower the minimum level of relevance achieved.

Although there is a decrease in the minimum similarity value, the maximum similarity value tends to remain stable or only fluctuates slightly between Condition 1 and Condition 3. This indicates that articles with the highest level of relevance remain relevant in all conditions.

The difference between the maximum and minimum similarity values reflects the variation in the level of relevance among the displayed articles. The larger the difference, the higher the level of variation in relevance among those articles.

4. CONCLUSION

Based on the analysis, it can be concluded that the more articles displayed in the search results, the lower the minimum level of relevance achieved. However, articles with the highest level of relevance remain relevant in all conditions. It is important to note that other factors, such as the quantity and quality of the dataset used, can also influence the level of relevance. The larger and more complete the dataset, the more likely the model can find articles with high similarity.

In the context of using the TF-IDF and cosine similarity algorithms in the destination travel article search feature, the analysis results indicate that conditions with fewer articles tend to provide more accurate and relevant results. However, deciding the number of articles to be displayed should consider user preferences and specific system requirements. This allows for appropriate adjustments to meet the desired needs and results in quality. Therefore, adjustments need to be made to achieve optimal relevance in searching destination travel articles, considering these factors.

Acknowledgments

I express my heartfelt gratitude to the Travelyours team from the Bangkit 2023 program for their invaluable discussions, ideas, and support throughout this research. I also extend my thanks to the Computer Science Department, Universitas Mataram, for providing the facilities that supported this research.

REFERENCES

[1] J. Guo *et al.*, “A Deep Look into neural ranking models for information retrieval,” *Inf Process Manag*, vol. 57, no. 6, pp. 1–20, Nov. 2020, doi: 10.1016/j.ipm.2019.102067.

[2] M. D. R. Wahyudi, “Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur’an Berdasarkan Keterkaitan Topik,” *Semesta Teknika*, vol. 22, no. 1, May 2019, doi: 10.18196/st.221235.

[3] Ahmad N, Prasetyo A, and Masruri A, “PENERAPAN INFORMATION RETRIEVALPADA SEARCH ENGINE,” *Jurnal Inovasi Hasil Penelitian dan Pengembangan UIN Kalijaga Yogyakarta*, vol. 1, no. 1, pp. 16–23, Dec. 2021, Accessed: Jul. 06, 2023. [Online]. Available: https://digilib.uin-suka.ac.id/id/eprint/49155/

[4] F. A. Nugroho, F. Septian, D. A. Pungkastyo, and J. Riyanto, “Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat,” *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, pp. 529–536, Dec. 2021, doi: 10.32493/informatika.v5i4.7126.

[5] Sintia, S. Defit, and G. Widi Nurcahyo, “PRODUCT CODEFICATION ACCURACY WITH COSINE SIMILARITY AND WEIGHTED TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY (TF-IDF),” *Journal of Applied Engineering and Technological Science*, vol. 2, no. 2, pp. 14–21, May 2021, doi: https://doi.org/10.37385/jaets.v2i2.210.

[6] J. Adlinnas, K. Muslim Lhaksmana, and D. Richasdy, “Implementasi Metode TF-IDF dan K-Nearest Neighbor untuk Seleksi Pelamar Kerja,” *e-Proceeding of Engineering*, vol. 7, no. 3, pp. 1–11, Dec. 2020, Accessed: Jul. 06, 2023. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14224

[7] L. Hermawan and M. B. Ismiati, “Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval,” *TRANSFORMATIKA*, vol. 17, no. 2, pp. 188–199, Jan. 2020, doi: http://dx.doi.org/10.26623/transformatika.v17i2.1705.

[8] F. B. Sejati, P. Hendradi, and B. Pujiarto, “DETEKSI PLAGIARISME KARYA ILMIAH DENGAN PEMANFAATAN DAFTAR PUSTAKA DALAM PENCARIAN KEMIRIPAN TEMA MENGGUNAKAN METODE COSINE SIMILARITY (Studi Kasus: Di Universitas Muhammadiyah Magelang),” *Jurnal Komtika-Komputasi dan Informatika*, vol. 85, no. 2, pp. 85–94, Jan. 2019, doi: https://doi.org/10.31603/komtika.v2i2.2594.

[9] B. PARLAK, “The Effects of Preprocessing on Turkish and English News Data,” *Sakarya University Journal of Computer and Information Sciences*, vol. 6, no. 1, pp. 59–66, Apr. 2023, doi: 10.35377/saucis...1207742.

[10] A. Kurniawan, Indriati, and S. Adinugroho, “Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dan Lexicon Based Features,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 9, pp. 8335–8342, Sep. 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[11] C. Mugisha and I. Paik, “Comparison of Neural Language Modeling Pipelines for Outcome Prediction from Unstructured Medical Text Notes,” *IEEE Access*, vol. 10, pp. 16489–16498, Feb. 2022, doi: 10.1109/ACCESS.2022.3148279.

[12] R. Ulgasesa and A. Bijaksana Putra Negara, “Pengaruh Stemming Terhadap Performa Klasifikasi Sentimen Masyarakat Tentang Kebijakan New Normal The Impact of Stemming on the Classifications Performance of Public Sentiments About New Normal Policy,” *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, vol. 10, no. 3, pp. 286–293, Jul. 2022, doi: 10.26418/justin.v10i3.53880.

[13] I. Arroyo-Fernández, C. F. Méndez-Cruz, G. Sierra, J. M. Torres-Moreno, and G. Sidorov, “Unsupervised sentence representations as word information series: Revisiting TF–IDF,” *Comput Speech Lang*, vol. 56, pp. 107–129, Jul. 2019, doi: 10.1016/j.csl.2019.01.005.

[14] R. T. Wahyuni, D. Prastiyanto, and D. E. Supraptono, “Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi,” *Jurnal Teknik Elektro*, vol. 9, no. 1, pp. 18–23, Jun. 2017, doi: https://doi.org/10.15294/jte.v9i1.10955.

[15] A. Apriani, H. Zakiyudin, and K. Marzuki, “Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta,” *Jurnal Bumigora Information Technology (BITe)*, vol. 3, no. 1, pp. 19–27, Jul. 2021, doi: 10.30812/bite.v3i1.1110.

BIOGRAPHY OF AUTHORS

Mizanul Ridho Aohana is a student pursuing a Bachelor's degree in Informatics at the University of Mataram. Currently in the sixth semester or third year of study, His interest is in artificial intelligence, specifically in image processing and natural language processing.

Email: mizanulridhoaohana@mhs.unram.ac.id

Fitri Bimantoro is a lecturer in Informatics at University of Mataram since 2015. he gets his bachelor's at Electrical Engineering from UNRAM in 2010 and Master Degree of informatic at the Institut Teknologi Sepuluh Nopember in 2014. His interest is in the areas of computer vision, image processing, pattern recognition, and image retrieval.

Email: bimo@unram.ac.id