

Paris 2024 Olympics: A Big Data Analysis

Namkil Kang

Professor, College of Liberal Arts, Far East University, South Korea

Abstract : This paper aims to analyze 26 articles of Google written in early 2024 concerning Paris 2024 Olympics. A major point of this paper is that *Olympics* was the most widely occurred one, followed by *Paris, games, summer, events, men, and athletes*, in that order. A point to note is that the biggest keyword is *Olympics* in the word cloud. This seems to suggest that the keyword *Olympics* is the most noteworthy and pivotal one in 26 articles of Google. A further point to note is that the keywords *gold, basketball, man, games, guide, channel, country, and ceremony* are the core ones that constitute 26 articles of Google. It is vital that the so-called cluster is different from the so-called topics in that the former is somewhat larger than the latter. Finally, particularly noteworthy is that the word *Paris* is the closest to the word *Olympics*, followed by the word *Olympic*, the word *games*, and the word *athletes*, in descending order.

Keywords: big data, network, token, term frequency, cluster, word cloud

1. Introduction

This paper aims to analyze 26 articles of Google written in early 2024 concerning Paris 2024 Olympics. Our research was carried out by python. First, we aim at contemplating 26 articles whose samples are provided for expository purposes. Second, we go over the so-called term frequency of 40 major keywords in descending order. This enables us to see what kinds of keywords are the representative ones in 26 articles of Google. Third, we inquire into a word cloud showing which keywords are the noteworthy ones in 26 articles. Fourth, we look into the core networks of 26 articles of Google in which the noteworthy keywords show up around the center of the map.

Fifth, we attempt to investigate the so-called cluster that is dubbed as a group with similarities. Python generated 6 groups, but we provide one of them. Sixth, we attempt to investigate the similarity between the word *Olympics* and many other words. This analysis enables us to see what keywords are close to the keyword *Olympics*. Note that the more there are similarities between two words, the closer they are.

2. Results

2.1. Samples of 26 Articles and Pretreatment

In what follows, we aim to provide the samples of 26 articles of Google. We only provide the first sentences for convenience:

Table 1 Samples of 26 Articles

Number	The First Sentence
1	The highly-anticipated 2024 Summer Olympic Gam...
2	"It's really difficult and we're not even at t...
3	With less than a year before the event kicks o...
4	The water quality of the Seine has improved, t...
5	Although designed to be eco-friendly and free ...
6	The far-right National Rally (RN) party of Mar...
7	Dutch golfer Joost Luiten will not be compet...
8	For more than fifty years, athletes competing ...
9	The 2024 Paris Olympics will introduce fresh f...
10	The men's tournament will get underway before ...
11	All Cooper Flagg had to do was show he belonge...
12	Bam Adebayo has two primary goals to accomplis...
13	Though he both qualified to be there and won i...

14	The only real surprise when the US track and f...
15	You can share an article by clicking on the sh...
16	The opening ceremony of the Paris Olympics wil...
17	The Paris 2024 Olympics will officially launch...
18	From Simone Biles to Steph Curry, some of the ...
19	That's the Paris 2024 Summer Olympics slogan, ...
20	Summer is almost here, and with it comes an ev...
21	The organisers of Paris 2024 say the parliamen...
22	Every four years, the Olympic Games spotlight ...
23	As the world turns its gaze towards Paris for ...
24	If you're like me, the summer Olympics is a mu...
25	Paris is ready for the Olympics with a rich hi...
26	Inside Gymnastics will be on the scene in Pari...

It is vital that we got rid of adverbs, adjectives, pronouns, verbs, relative pronouns, to-infinitives, gerunds, etc. from the original texts. Put differently, we preprocessed the original texts and obtained only nouns for six analyses. Only nouns suffice for six analyses.

2.2. 40 Keywords

The goal of this section is to contemplate 40 representative keywords that consist of 26 articles of Google. Note that these 40 representative keywords enable us to see what words are the noteworthy ones in 26 articles of Google:

Table 2 40 Keywords

Number	Word	Frequency
1	Olympics	220
2	Paris	219
3	games	117
4	summer	80
5	events	68
6	men	68
7	athletes	61
8	women	61
9	sports	58
10	team	49
11	ceremony	48
12	July	48
13	world	45
14	day	41
15	final	39
16	event	38
17	time	38
18	year	35
19	August	34
20	Place	34
21	Tokyo	34
22	France	33
23	opening	33
24	round	33

ones that constitute 26 articles of Google.

2.5. Cluster

This section focuses on providing a cluster, which is dubbed as a group with similarities. Python

generated 6 clusters, but we provide one cluster.

This cluster is supposed to have a lot of keywords in common since they are related to one another in 26 articles of Google:

Table 3 Cluster

Cluster
ability absence ac access access hospitality accommodation accor account achievements act action activities actors adapters add addition additions adebayo admission advantage adventure adventures advertisement affectionately affiliates afghanistan afp africa afternoon age agency agility aim air airbnbs aires airfare airline airlines airport airports airways albar albeit alberta alberto albertville alexandre algonquin aliens aline all allman ally alone altercation altercations america american ami amik amount ampezzo amsterdam analysis angeles angelesyuichi anne anthems anthony anticipation antonio anyone anything ap app apparatus appearance appointment approach aquatics arabesque archery architecture ardor are area areas aren arena arent argentina arms army arrangements arrival arrivals arrondissement arrondissements art article artist arts as asap asia aspects at athena athens athing athlete athletes athleticism athletics atlanta atlantic atmosphere attack attempt attempts attendance attendees attention attitudes attraction attractions audience audiences aug august augustin austerlitz australia australian australias austria authorities authority authorization authors avail averages award bach backdrop background bacteria bag baggage bakeries balance ball ballgame bam banevic banks banter bar barber barcelona barges barometers barriers bars baseball basis basketball battle battles batum bbc beam bear bearing bears beats beaune beautiful beaver bed bednarek begin beijing Beijings bel belarus belief belongings bercy berlin best bid big bike biking bilal biles bing bit bite black blend bleus block blue bmx bmxing boat boats body bolt boltholes bonus book booker bordeaux boston bottger boulder bouldering boundaries bourget boutique bowl boyle boys br bradley brainer brand brandin bravo brazil brazils break breakdancing breaker breakers breakfast breaking bridge briefing brisbane britain broadcast broadcasts bronx bronze brooklyn brouillet bryce bucket budapest budget buenos build building buildings built bulgari bundle bureau burns bus buses business butterfly buttes button buzz by cable cafes cafés calgary call calls camaraderie camp campaigns camus canada canal canoe canoes cap capable capacity capital capitals capture carbon card career carew caricatures carissa carri carriers carroll cars cartoon cartoons case castex categories category caters cbs cdg celebration celebrations celsius center centre centric centrist centrists century ceo ceremonies ceremony ceremonys cet cha chagall chairs challenge challengers challenges champion champions championship championships champs chance change changes channel

Notice that python generated a lot of keywords, but we included the part of them for the reason of space. Notice, furthermore, that the so-called cluster is different from the so-called topics in that the former is somewhat larger than the latter. To put it differently, the so-called cluster is meant to

be a group with similarities, but the so-called topics are not.

2.6. Similarities

This section is focused on probing into the similarity between *Olympics* and many other keywords. Table 4 shows the similarity between *Olympics* and many other words:

Table 4 The Similarity between Olympics and Many Other Words

Number	Keyword	Similarity
1	Paris	0.9998929500579834
2	Olympic	0.9998766779899597
3	Games	0.9998663067817688
4	athletes	0.9998421669006348
5	July	0.9998318552970886
6	event	0.9998022317886353

7	competition	0.999791145324707
8	ceremony	0.9997768998146057
9	Team	0.999776303768158
9	events	0.9997740387916565

What is interesting is that the word *Paris* is the closest to the word *Olympics*, as indicated in Table 4. More specifically, the proportion of the similarity between *Paris* and *Olympics* is 99.989% (the highest), which seems to suggest that these two words are meant to be the closest. Perhaps it is worthwhile pointing out that the word *Olympic* is the second closest one to the word *Olympics*. This in turn implies that these two words appear to be the second most widely used. It must be stressed that the word *games* is the third closest to *Olympics*. To be more specific, the proportion of the similarity between *Olympics* and *games* is 99.986%. Quite interestingly, the word *athletes* ranks fourth, which in turn suggests that this keyword is the fourth closest to the word *Olympics*. Put differently, this word is the fourth most widely used with the word *Olympics*. To sum up, the word *Paris* is the closest to the word *Olympics*, followed by the word *Olympic*, the word *games*, and the word *athletes*, in descending order. It should also be pointed out that the word *ceremony* ranks eighth. This in turn indicates that the word *ceremony* is deemed to be the eighth closest to the word *Olympics*. We thus conclude that the word *Paris* is the closest to the word *Olympics*. For the analysis of artificial intelligence and big data, see Kang (2024a, 2024b, 2024c, 2024d, 2024e, 2024f, 2024g).

3. Conclusion

To sum up, we have analyzed 26 articles of Google written in early 2024 concerning Paris 2024 Olympics. In section 2.2, we have contended that *Olympics* was the most widely occurred one, followed by *Paris*, *games*, *summer*, *events*, *men*, and *athletes*, in that order. In section 2.3, we have argued that the biggest keyword is *Olympics* in the word cloud. This seems to suggest that the keyword *Olympics* is the most noteworthy and pivotal one in 26 articles of Google. In section 2.4, we have further argued that the keywords *gold*, *basketball*, *man*, *games*, *guide*, *channel*, *country*,

and *ceremony* are the core ones that constitute 26 articles of Google. In section 2.5, we have maintained that the so-called cluster is different from the so-called topics in that the former is somewhat larger than the latter. In section 2.6, we have argued that the word *Paris* is the closest to the word *Olympics*, followed by the word *Olympic*, the word *games*, and the word *athletes*, in descending order.

References

- [1] Kang, N (2024a). A Big Data Analysis of a Hot Political Issue. *Studies in Linguistics* 70, 149-165
- [2] Kang, N (2024b). The Ukraine War in BBC News: A Big Data Analysis. *International Journal of Applied Engineering and Technology* 6(1), 1548-1555
- [3] Kang, N (2024c). A Study of the Prediction of the Model Logistic Regression in Machine Learning: Focusing on a Survey. *Semiconductor Optoelectronics* 43(1), 86-94
- [4] Kang, N (2024d). The Spring of Seoul: A Big Data Analysis. *Semiconductor Optoelectronics* 43(1), 264-272
- [5] Kang, N (2024e). A Study of the Prediction of the Three Models Logistic Regression, Gaussian Naive Bayes ,and Multi-Layer Perceptron Classifier in Machine Learning. *Semiconductor Optoelectronics* 43(1), 253-263
- [6] Kang, N (2024f). An Analysis of the Prediction Power of Four Models in Machine Learning and Deep Learning: Focusing on the Reviews of the Movie Noryang. *A Jungang Journal of English Language and Literature* 66(1), 101-122
- [7] Kang, N(2024g). An Association Analysis of Shopping Items in Machine Learning. *Stochastic Modelling and Computational Sciences* 4(1), 97-109