Topic Modeling for Olympic and Fashion News

Gabriella Wolf **Sai Subathra Anbarasu** **Vama Shah**

1 Introduction

In today's world, data being collected is growing by leaps every day and it is becoming difficult to make sense out of it. As a result, we require robust algorithms and techniques that can be used to obtain a better understanding and categorization of such tremendous amounts of data. Topic modeling is an unsupervised text mining technique that can be used to identify the distinct topics that occur in large sets of documents based on the frequency of words present in them. In this project, we are going categorize a sizable amount data into distinct topics based on the density of the words present in the entire corpus. In doing so, we make use of articles that were ethically scrapped from, https://www.cbssports.com/olympics/ and https://www.vogue.com/, thus contributing to Sports and Fashion related data.

The process itself consists of first cleaning the data by identifying the important parts of the article such as the title, author, date of publishing, article body etc, followed by removal of stop words. We then use the gensim LDA model to perform topic modeling. Perplexity and coherence are used to measure the performance. After obtaining the topics, we manually give names for each of the topics. Lastly, we identify the article that contributes the most towards each topic and compare the expected and actual results.

2 Methods

2.1 Topic Modeling

Topic modeling can be used to gain insight on a set of documents by finding the main topics that occur (Dwivedi, 2018). We start with a collection of words and documents containing these words. We may randomly populate the topics with the words from the document by some probability distribution and we may randomly choose a list of words according to the documents. We repeat this many times with new topic models and probability distributions for each topic. We choose the best topic model between iterations, which is the document that is more likely. In the end, each topic model should have a list of words following a dirichlet distribution.

To create these topic models, we may use the help of gensim, which is a python library that creates and analyzes topic models. First, we must convert the documents into an acceptable format for gensim. We preprocess the text by removing stop words, lemmatizing, and removing frequent words. We may take the preprocessed text and create a bag of words and a corpus. The bag of words is a dictionary that stores each word and the number of times the word appears in all documents. The corpus stores each word and the number of times the word appears within each document. The last parameter required is the number of topics we would like to find.

2.2 Choosing the Number of Topics

We use perplexity and coherence when we choose the optimal number of topics. Perplexity is the log-likelihood of some corpus of words (Kapadia, 2019). We would like to minimize this value. Perplexity is expressed by the following:

$$\exp \frac{-logP(\beta_1...\beta_k, \theta_1, ..., \theta_D | Z_i, ..., Z_d)}{tokens}$$

Topic coherence measures the semantic similarity between the likely words within a topic. We use the coherence "u_mass," which takes into account the ordering of the topic words (Michael Röder, 2015). We would like to maximize this value. Coherence umass is expressed by the following:

$$C_{umass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=i}^{i-1} log(\frac{P(w_i, w_j) + \epsilon}{P(w_j)})$$

We may iterate over different number of topics and graph the perplexity and coherence values. Based on these graphs we can choose the best topic model with the associated number of topics.

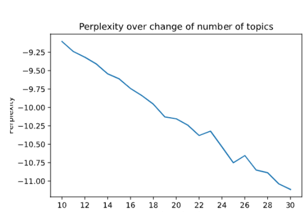

Figure 1: 25 words

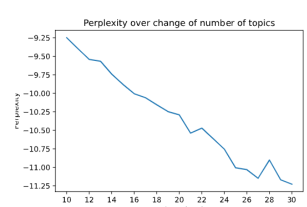

Figure 2: 50 words

Figure 3: Perplexity values after removing most frequent words.

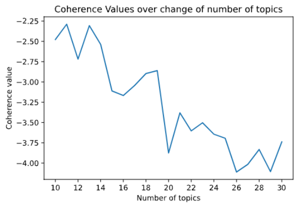

Figure 4: 25 words

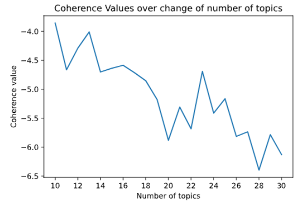

Figure 5: 50 words

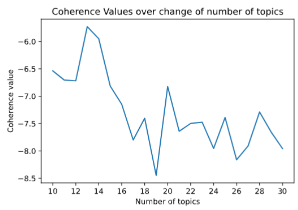

Figure 6: 75 words

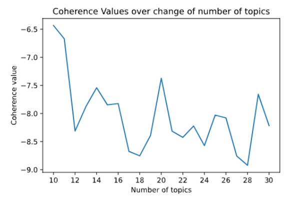

Figure 7: 100 words

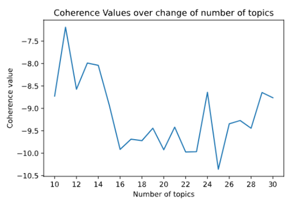

Figure 8: 125 words

Figure 9: Coherence values after removing most frequent words.

Before we start looking at the change in values over the number of topics, we may look into removing frequent words. Over our trials, we removed words in increments of 25. Despite the number of words we remove the perplexity trend remains consistent as we see in figure 3. The values tend to range from around -9.5 and -11.5 and there is a negative correlation. So the perplexity plots will not be helpful in finding the optimal number of words to remove. We can see a difference in coherence values when we remove more words in figure 8. From 25 to 50 words and from 50 to 75 words the maximum coherence values decreases by roughly two. From 75 to 100 words the maximum coherence value decreases around 0.5. From 100 to 125 words the maximum coherence value decreases around 1. There is an interesting point at 100, where the change in the coherence value is a minimum. When we look at the topic models when we remove too few of the frequent words, we see that there are many repeating words. When we only remove 25 words, we see that words like "come," "make," and "like" appear in the majority of the topics. These words do not give insight into what the topic can be. When we remove too many of the frequent words, we start to loose good words. When we remove 125 words, we see words like "tournament," "prediction," and "race" start to disappear. The words look more random and it is harder to determine the topics. So we have determined to remove 100 words.

We may create the documents and remove 100 words

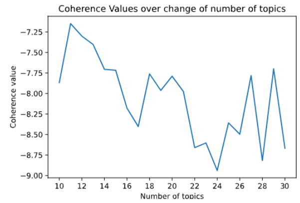

Figure 10: Coherence plot

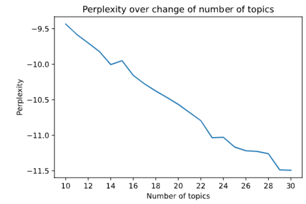

Figure 11: Perplexity plot

when we preprocess the text and use this to create our topic models. We may look at different numbers of topics from 10 to 30, and determine from the plots what may be a good topic model. In the coherence graph we look for the point before a big drop. In the perplexity graph we look for a points before a positive slope or before the negative slope becomes more gradual. Then we may look at the topic models corresponding to these points and determine if the topics are clear.

In figure 10 the interesting points are 15, 21, 27, and 29. These are values right before there is a big dip in the graph. In figure 11 from those points 15 and 21 are the most interesting. Looking at the corresponding topic models, 15 topic looks to have more defined topics. So, overall we choose 15 topics in our topic model.

3 Data/Results

The data consisted of equal number of articles from https://www.cbssports.com/olympics/ and https://www.vogue.com/. Thus contributing to data belonging to the Sports and Fashion field.

3.1 Topics and highest frequency words

	0	1	2
	trackevents	trends	unclear
0	richardson	season	instagram
1	race	goal	view
2	penalty	trend	prediction
3	herah	2022	sportsline
4	classic	spring	russian
5	medalist	bag	life
6	late	bet	tournament
7	post	mexico	include
8	steveson	love	side
9	minute	gainsbourg	week
10	kick	medalist	place
11	100	home	average
12	career	sportsline	work
13	thompson	slovenia	hat
14	end	pandemic	summer
15	korda	jean	total
16	history	four	big
17	canada	bile	2016
18	deal	tournament	experience
19	prefontaine	try	part

	3	4	5
	people	numbers	stats
0	parchment	6	slovenia
1	end	7	sportsline
2	thing	5	prediction
3	name	10	tournament
4	hat	12	percent
5	people	8	high
6	never	11	side
7	xueying	south	king
8	earn	country	big
9	give	20	trend
10	place	republic	spain
11	committee	9	bag
12	report	bile	share
13	global	charge	average
14	trijana	people	assist
15	great	keller	bet
16	dream	14	total
17	offer	21	count
18	little	korea	nine
19	need	seven	offensive

	6	7	8
	fencing	fashionweek	musicfashion
0	2022	chanel	n’t
1	coach	jawad	hat
2	spring	pair	airpod
3	leach	dress	headphone
4	anderson	paralympic	dress
5	bring	red	director
6	foil	house	love
7	designer	\’	harris
8	we	never	wire
9	boutique	week	vogue
10	life	put	aldridge
11	work	paris	model
12	country	add	thing
13	dame	around	’
14	rank	stewart	piece
15	2016	’	set
16	vogue	pentathlon	bring
17	notre	friend	talk
18	fencing	side	many
19	championship	much	michael

	9	10	11
	gymnastic	running	unclear
0	paralympic	seidel	springsteen
1	history	marathon	aug
2	andrejczyk	set	record
3	afghanistan	double	july
4	country	country	10
5	work	third	state
6	dress	ross	become
7	never	even	richardson
8	vogue	12	pergolini
9	2016	mazdzer	jump
10	even	dress	rousteing
11	female	medalist	b
12	part	history	race
13	still	rebound	spain
14	committee	tell	pagoni
15	tell	love	love
16	six	straight	6
17	much	10	100
18	stylist	place	canada
19	four	race	heel

	12	13	14
	winners	dateandtime	unclear
0	mclaughlin	much	chelimo
1	steveson	storey	king
2	percent	spain	goal
3	medalist	race	score
4	bile	july	brazil
5	chelimo	history	life
6	record	four	big
7	hat	bet	include
8	balance	set	season
9	race	paralympic	career
10	sportsline	coach	give
11	2016	late	hard
12	tournament	aug	medalist
13	8	mexico	balenciaga
14	home	south	steveson
15	big	week	player
16	view	include	record
17	part	friday	spain
18	face	place	earn
19	competition	republic	lot

3.2 Mapping between the topics and the most likely article

Topic: trackevents
No. of occurrences: 38
Article name: LOOK: Sha’Carri Richardson sounds off after her first race following her suspension at Prefontaine Classic
Article Date: 8/21/2021

Topic: trends
No. of occurrences: 40
Article name: Article name : Mexico vs. Japan odds, predictions: Soccer expert reveals picks for Tokyo Olympics 2020 bronze medal match

Article Date: 8/6/2021

Topic: unclear1
No. of occurrences: 29
Article name: Olympics 2020 basketball odds, picks: Australia vs. Slovenia bronze medal game predictions from proven expert
Article Date: 8/7/2021

Topic: people
No. of occurrences: 48
Article name: 2020 Tokyo Olympics: The cost of athletes achieving dreams has become extraordinarily high
Article Date: 8/9/2021

Topic: numbers
No. of occurrences: 81
Article name: 2020 Tokyo Olympics medal count: USA tops China in gold, silver, bronze and overall medal totals
Article Date: 8/8/2021

Topic: stats
No. of occurrences: 64
Article name: Olympics 2020 basketball odds, picks: Australia vs. Slovenia bronze medal game predictions from proven expert
Article Date: 8/7/2021

Topic: fencing
No. of occurrences: 38
Article name: Spring 2022's Most Viewed Shows on Vogue Runway
Article Date: October 8, 2021

Topic: fashionweek
No. of occurrences: 22
Article name: Tokyo Paralympics 2021: British powerlifter Ali Jawad self-isolated for three years in preparation for Games
Article Date: 8/26/2021

Topic: musicfashion
No. of occurrences: 31
Article name: Ruby Aldridge Takes Vogue Behind the Scenes of the Rodarte Show
Article Date: October 6, 2021

Topic: gymnastic
No. of occurrences: 25
Article name: Paralympic athletes from Afghanistan, not able to leave country, will miss 2020 Games
Article Date: 8/17/2021

Topic: running
No. of occurences: 36
Article name: 2020 Tokyo Olympics top 10 Team USA moments: Allyson Felix makes history, Caeleb Dressel dominates and more
Article Date: 8/8/2021

Topic: unclear2
No. of occurrences: 35
Article name: 2020 Tokyo Olympics: Men's basketball tournament TV schedule, live stream, start times, group standings
Article Date: 8/7/2021

Topic: winners
No. of occurrences: 41
Article name: 2020 Tokyo Olympics top 10 Team USA moments: Allyson Felix makes history, Caeleb Dressel dominates and more
Article Date: 8/8/2021

Topic: dateandtime
No. of occurrences: 36
Article name: 2020 Tokyo Olympics: Men's basketball tournament TV schedule, live stream, start times, group standings
Article Date: 8/7/2021

Topic: unclear3
No. of occurrences: 53
Article name: 'Running saved me': Long-distance runner Paul Chelimo's path from Kenya to the U.S. Army to the Olympic podium
Article Date: 9/9/2021

4 Discussion/Conclusions

The initial expectation out of this project was to obtain clear and distinct topics. The use of equal number of articles related to olympics and fashion led to the expectation of creating equal number of topics focused on olympics and fashion. However, the actual results obtained were definitely varying with the expected results.

The phase of identifying the names for the topics after the identification of the 20 most frequent words in each topic was not straight forward and distinct. It was rather a more trial and error method, which involved performing the analysis multiple times to obtain a consistent result.

The actual results consisted of topics that were not so clear to categorize. For example, the words in the topic named "fencing" consisted of words like "coach", "country", "fencing", "championship" which give a perspective of being related to the olympics. Whereas, there were also words like "spring", "designer", "boutique", "vogue" etc. which give an idea of it being related to fashion. This intermix of words from both fields were not expected.

Despite the complication involved in identifying the accurate names for these topics, the identification of the most likely article to contain this topic was straight forward. Topics such as trackevents, people, numbers, stats, musicfashion, running, winners and dateandtime matched with articles whose titles were relevant to the topic names. Whereas, topics like trends, fencing, gymnastics were matched to articles whose titles did not have much of a relation with the topic names. The unclear1 topic could actually be named as predictions, unclear2 could have been named schedule and unclear3 could have been named as biography.

An interesting observation was that the topics with a strong correlation between the title and topic name had large number of occurrence of the words belonging to the topic as compared to the topics where the title did not coincide with the topic name.

We can therefore conclude the project by saying that topic modeling is an relatively quick and efficient method for identifying a set of words from a large set of documents that represent the information with high accuracy. It would otherwise be extremely time consuming and inefficient to go through each document and identify the set of topics a corpus is centered towards.

References

Priya Dwivedi. 2018. Nlp: Extracting the main topics from your dataset using lda in minutes, August.

Shashank Kapadia. 2019. Evaluate topic models: Latent dirichlet allocation (lda), August.

Alexander Hinneburg Michael Röder, Andreas Both. 2015. Exploring the space of topic coherence measures. *WSDM 2015: Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408.