Big Data Analytics: Project 1

Adhiraj Sood
Rochester Institute of Technology
as9125@rit.edu

Shweta Nateshan Iyer
Rochester Institute of Technology
si9808@rit.edu

Sidhartha Amperyani
Rochester Institute of Technology
ga5310@rit.edu

1 Introduction

In this paper, we have collected new articles from The Hindu and performed topic modeling to analyze each article. We have used an LDA (Latent Dirichlet Allocation) model to classify text in the article to a particular topic [4, 5]. We have collected the required information such as the author's name, title, the date the article was published, and body from each article. The body text was then preprocessed. The process for this included Tokenization, words with length less than 3 were removed, Stopwords were removed, words were Lemmatized, and then Stemmed. To avoid the possibility of Stopwords being introduced after stemming, we have again checked for these here. All the above information regarding each article was then stored in the articles.json file. To train the LDA model, we have used the preprocessed body text and converted it to a bag of words. Gensim LDA model was used to build the model by giving the number of topics as a parameter. We have created a model against our data with a number of topics between 10 to 30. This trained data will help us find the topic for any unseen article we may see in the future. In this paper Section 2 discusses the methods we used, Section 3 shows the results and in Section 4 we have discussed our learning.

2 Methods

- tokenize(): This method uses the nltk library and divides the string into a list of words or tokens.
- lemmatize(): In our implementation, we have used the WordNetLemmatizer to lemmatize each word in the article. It combines the various inflected forms of a word together so that they can be analyzed as a single item [7].
- stem(): We have used the PorterStemmer().stem() function to remove the morphological variants from words, leaving only the stem of the word [3].
- gensim.corpora.Dictionary(): maps the words with integer ids and takes iterable strings as input. [1].
- filter_n_most_frequent(remove_n): Removes 'remove_n' most frequent words from the corpora.dictionary.
- doc2bow(document): Converts the list of strings given as input into a bag of words. It outputs a list where each item in the list is a tuple of (token_id, token_count) [1].
- gensim.models.ldamodel.LdaModel(): This function takes a variety of input arguments out of which we have used corpus, bow_corpus, num_topics, id2word, per_word_topics. Corpus is the bag of words obtained, id2word is the dictionary obtained using the corpora.Dictionary() method, num_topics is the number of topics to be extracted from the training corpus, and per_word_topics is a boolean value, which if True, computes a list of topics, sorted in descending order of most likely topics for each word, along with their phi values multiplied by the feature-length [2].
- gensim.models.CoherenceModel(): This function calculates the coherence for topic models.

3 Results

The performance of each topic model has been measured using perplexity and coherence. Figure 1 refers to the Coherence Graph and Figure 2 refers to the Perplexity Graph.

Result obtained from the projection of articles into the topic model are discussed in this section. For each topic, we have found the article where that topic is the most likely.

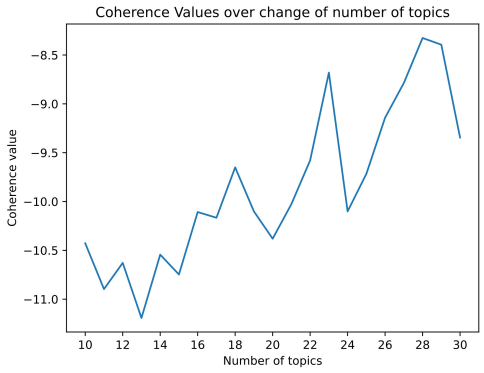

Figure 1. Coherence Graph.

Topic: 0 with header 'Energy source'
Article name: Explained | What is the extent of India's coal crisis? - The Hindu
Article Date: October 17, 2021 03:45 IST

Topic: 1 with header 'Finance'
Article name: Credit score myths that can harm financial health - The Hindu
Article Date: October 17, 2021 22:48 IST

Topic: 2 with header 'Documentaries'

	0	1	2	3	4	5	6
0	Energy source	Finance	Documentries	Defence	Entertainment	State of Punjab	Biology
1	coal	credit	cast	china	film	singh	child
2	film	score	coal	armi	vaccin	murder	polic
3	unit	kill	film	aviat	eleph	polic	cancer
4	energi	card	russia	bhutan	death	arrest	screen
5	product	milit	china	helicopt	water	narain	understand
6	seawe	civilian	militari	border	congress	regist	dlx1
7	fund	polic	trip	chines	gandhi	film	cell
8	invest	loan	congress	oper	turkey	punjab	protein
9	price	oper	pakistan	cancer	compani	border	kill
10	plant	account	gandhi	gandhi	malaria	sikh	prostat
11	water	vaccin	travel	singh	servic	court	film
12	cover	singh	afghanistan	heron	visit	singhu	delay
13	bank	murder	round	arecanut	bollywood	crime	televis
14	save	tata	august	system	tata	worker	expos
15	child	travel	russian	capabl	drain	litr	kashmir
16	bitcoin	histori	visit	negoti	singh	lakhbir	leopard
17	allianc	product	price	missil	reserv	commit	express
18	reader	settl	leopard	infrastructur	citi	surrend	labour
19	tamil	gandhi	carri	deal	presid	pradesh	local
20	observ	lender	hous	presid	forest	daili	mother

Table 1. Table with 20 most likely words in that topic (Part 1).

	7	8	9	10	11	12	13
0	Terror	Vaccine	Climate	Poverty	Credit	Sports	Religion
1	singh	cancer	monson	child	credit	protest	film
2	cast	vaccin	china	indic	monson	singh	protest
3	polic	protein	reader	undernourish	hasan	wicket	singh
4	film	turkey	leopard	index	card	farm	polic
5	sikh	dlx1	polic	hunger	product	pakistan	border
6	terror	film	taiwan	wast	player	prime	local
7	nihang	cell	chines	stunt	account	district	oper
8	vaccin	brand	antiqu	rank	loan	nihang	administr
9	hunger	malaria	editor	food	scotland	photograph	aviat
10	administr	thing	award	film	greav	film	china
11	kill	child	prize	energi	fleme	cricket	2018
12	attack	prostat	nobel	border	coal	afghanistan	sector
13	valley	china	paper	russia	wicket	idukki	drug
14	arrest	chines	kill	russian	crore	drug	namaz
15	credit	higher	california	preval	singh	farmer	armi
16	forc	taiwan	film	mortal	score	china	lalit
17	commit	research	research	methodolog	punch	kerala	muslim
18	allegedli	tata	account	europ	viswanathan	border	deploy
19	situat	captain	taipei	2018	udham	incid	ahmad
20	malaria	singh	water	calori	bangladesh	captain	site

Table 2. Table with 20 most likely words in that topic (Part 2).

	14	15	16	17	18	19	20
0	Flights	International	Elections	Films	Bollywood	Games	Money
1	europ	coach	cast	film	film	praggnanandhaa	vaccin
2	crore	korea	viswanathan	price	seawe	credit	malaria
3	russia	board	elector	child	product	film	travel
4	tata	project	vote	china	china	keymer	film
5	credit	singh	entrepreneur	diesel	compani	game	coin
6	energi	head	success	petrol	insur	reader	singh
7	price	telangana	small	indic	bollywood	editori	credit
8	venu	polic	support	monson	taiwan	gandhi	debt
9	water	gandhi	strategi	polic	danc	compani	child
10	aviat	turkey	credit	bank	technolog	price	trillion
11	demand	kill	seat	undernourish	muslim	pawn	gandhi
12	monson	rocket	student	actor	energi	vaccin	health
13	muslim	technolog	elect	rate	food	drug	platinum
14	oper	film	mark	award	allianc	protest	ceil
15	airlin	unit	belong	litr	drug	gain	hous
16	suppli	schedul	match	cost	cinema	advanc	capit
17	atmospher	peddavagu	learn	hunger	observ	posit	cancer
18	product	hope	winner	game	coal	rate	congress
19	fashion	author	ball	debt	khan	final	insur
20	servic	congress	achiev	director	actor	valu	novemb

Table 3. Table with 20 most likely words in that topic (Part 3).

	21	22	23	24	25
0	Crime	Healthcare	Unclear	Editorial	Cinema
1	crime	dlx1	hunger	film	film
2	venu	cancer	singh	china	coal
3	water	protein	worker	editori	monson
4	forc	prostat	regist	compani	actor
5	final	research	sector	bhutan	crore
6	titl	cell	udham	actor	cinema
7	trip	reader	score	seri	crisi
8	dhoni	seawe	polic	home	energi
9	chhetri	hunger	child	coal	borgohain
10	never	child	progress	valu	reader
11	order	product	vaccin	cancer	demand
12	reader	mous	protest	dlx1	templ
13	total	russia	drug	interest	antiqu
14	border	europ	rank	cinema	director
15	fashion	editor	index	cover	price
16	atmospher	posit	crore	insur	award
17	murder	food	debt	arya	claim
18	turkey	indic	punch	subscrib	domest
19	film	undernourish	brand	hasan	china
20	trophi	higher	death	weekli	bollywood

Table 4. Table and 20 most likely words in that topic (Part 4).

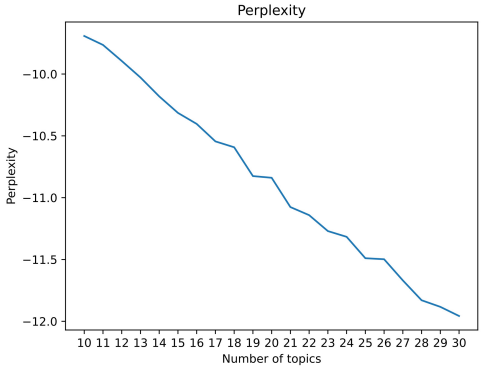

Figure 2. Perplexity Graph.

Article name: A precis on Dr. S. Krishnaswamy, the quintessential Indian documentary filmmaker - The Hindu
Article Date: None

Topic: 3 with header 'Defence'
Article name: UAVs, Aviation unit boost Army surveillance in eastern sector - The Hindu
Article Date: October 17, 2021 21:11 IST

Topic: 4 with header 'Entertainment'
Article name: A precis on Dr. S. Krishnaswamy, the quintessential Indian documentary filmmaker - The Hindu
Article Date: None

Topic: 5 with header 'State of Punjab'
Article name: Singhu murder: 1 more arrested, 2 'surrender' - The Hindu
Article Date: October 16, 2021 22:51 IST

Topic: 6 with header 'Biology'
Article name: IIT Kanpur team identifies a novel target to treat prostate cancer - The Hindu
Article Date: October 16, 2021 21:07 IST

Topic: 7 with header 'Terror'
Article name: Credit score myths that can harm financial health - The Hindu
Article Date: October 17, 2021 22:48 IST

Topic: 8 with header 'Vaccine'
Article name: IIT Kanpur team identifies a novel target to treat prostate cancer - The Hindu
Article Date: October 16, 2021 21:07 IST

Topic: 9 with header 'Climate'
Article name: How the law caught up with an artful dodger in Kerala - The Hindu
Article Date: October 16, 2021 00:04 IST

Topic: 10 with header 'Poverty'
Article name: Alarming hunger or statistical artefact? - The Hindu
Article Date: October 18, 2021 00:15 IST

Topic: 11 with header 'Credit'
Article name: Credit score myths that can harm financial health - The Hindu
Article Date: October 17, 2021 22:48 IST

Topic: 12 with header 'Sports'
Article name: ICC T20 World Cup 2021 — history, full squads and fixtures of all teams - The Hindu
Article Date: October 10, 2021 15:23 IST

Topic: 13 with header 'Religion'
Article name: Jumma Namaz: Efforts to find amicable solution in Gurugram prove to be futile - The Hindu
Article Date: October 17, 2021 22:24 IST

Topic: 14 with header 'Flights'
Article name: Explained | Will the Tatas be able to turn around Air India? - The Hindu
Article Date: October 17, 2021 03:30 IST

Topic: 15 with header 'International'
Article name: A precis on Dr. S. Krishnaswamy, the quintessential Indian documentary filmmaker - The Hindu
Article Date: None

Topic: 16 with header 'Elections'
Article name: State elections, limits of caste-based strategies - The Hindu
Article Date: October 18, 2021 00:02 IST

Topic: 17 with header 'Films'
Article name: How the law caught up with an artful dodger in Kerala - The Hindu
Article Date: October 16, 2021 00:04 IST

Topic: 18 with header 'Bollywood'
Article name: It is time to rise in defence of Bollywood - The Hindu
Article Date: October 16, 2021 00:02 IST

Topic: 19 with header 'Games'
Article name: Credit score myths that can harm financial health - The Hindu
Article Date: October 17, 2021 22:48 IST

Topic: 20 with header 'Money'
Article name: Explained | U.S. plan for $1 trillion platinum coin to address debt crisis - The Hindu

Article Date: October 13, 2021 12:43 IST

Topic: 21 with header 'Crime'
Article name: A precis on Dr. S. Krishnaswamy, the quintessential Indian documentary filmmaker - The Hindu
Article Date: None

Topic: 22 with header 'Healthcare'
Article name: IIT Kanpur team identifies a novel target to treat prostate cancer - The Hindu
Article Date: October 16, 2021 21:07 IST

Topic: 23 with header 'Unclear'
Article name: Alarming hunger or statistical artefact? - The Hindu
Article Date: October 18, 2021 00:15 IST

Topic: 24 with header 'Editorial'
Article name: A precis on Dr. S. Krishnaswamy, the quintessential Indian documentary filmmaker - The Hindu
Article Date: None

Topic: 25 with header 'Cinema'
Article name: How the law caught up with an artful dodger in Kerala - The Hindu
Article Date: October 16, 2021 00:04 IST

4 Discussion

While processing the body of various news articles, we found out that there are many words which even after our processing was stopped words, so we added an extra stop word check, and also before adding we checked that the word is greater than 3. This helped us in building good preprocessed data. Once our data was stored we used gensim libraries methods mention in section 2 to do topic modeling. In this, we created models based on a different number of topics. During modeling, we calculated the Perplexity and Coherence. Perplexity is the measure of how surprised the LDA model is by the new data. It measures how likely any previously unseen data is, given the previously learned model. However, recent research has shown that perplexity and human judgment are frequently not correlated [6]. The degree of similarity between high-scoring words in a topic is measured by coherence. Based on this argument we choose to use coherence and found out that the optimal number of topics was 26 in our case. Once we found that we created a model and stored the top 20 words in our data frame. We used the dataframe to_latex() method to directly covert this to our LateX. In the last part of our analysis, using our trained model we ran a process where we found out the article which is closest to the topics we have. The complications we had during the processing were during scrapping the data as the classes were varying for different input points. Although once the data was processed the topic modeling was clearly explained in the gensim documentation and based on some tests we were able to understand the modeling. Our trained model was 60 percent accurate in guessing the correct topic. We choose the topics for our data frame based on multiple iterations and passing through different articles. In the end, we were sure that more the data we train, we are more likely to find the correct topic for any unseen document.

References

[1] [n. d.]. Gensim Corpora Dictionary. https://radimrehurek.com/gensim/corpora/dictionary.html

[2] [n. d.]. LDA Model. https://radimrehurek.com/gensim/models/ldamodel.html

[3] [n. d.]. Stemming. https://www.nltk.org/howto/stem.html

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[5] Priya Dwivedi. 2018. NLP: Extracting the main topics from your dataset using LDA in minutes. https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

[6] Shashank Kapadia. 2019. Evaluate Topic Models: Latent Dirichlet Allocation (LDA). https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[7] Yash R. 2021. Python | Lemmatization with NLTK. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/