Project Report

Anonymous Author(s)

1 INTRODUCTION

The goal of this project is to perform topic modeling on news articles collected from a website using the web scraping technique. The process of extracting data from websites is known as web scraping. For this project, we gathered 104 news articles from the official website of Smithsonian Magazine [4]. The text in the body of these articles has been preprocessed by stemming, lemmatization, and the removal of stop words, as well as frequent and infrequent words. We then created a bag of unique words from all of the articles and computed the frequency count of each word that appeared in an article. To extract the main topics in a corpus, we trained a number of LDA (Latent Dirichlet Allocation) Multicore models with varying numbers of topics ranging from 10 to 30 using the gensim library. To determine the optimal number of topics associated with the collection of articles, these models were evaluated using the performance measures coherence and perplexity. By analyzing the coherence and perplexity values we decided that 22 topics would be optimal for the chosen set of article. We also identified the keywords associated with each topic and the articles most appropriate for each topic using the trained model. Topic modeling is a powerful tool that uses statistical and probabilistic distributions of words in a given text for better comprehension.

2 METHODS

2.1 Topic Modelling

Topic modelling is an unsupervised machine learning technique which is mainly used to perform text mining and information retrieval. Given a collection of documents, the main aim of topic modelling is to draw out the patterns in words and sentences based on their frequency and probabilistic distributions in order to identify a group of words as belonging to a topic or list of topics from these documents. Overtime, many methods were proposed to implement topic modelling. The first topic model that was invented was Latent Semantic Indexing (LSI) that was proposed by Papadimitriou, Raghavan, Tamaki and Vempala [5]. Some of the other famous models or methods of Topic Modelling are :

- Latent Dirichlet Allocation (LDA)
- Non Negative Matrix Factorization (NMF)
- Latent Semantic Analysis (LSA)
- Parallel Latent Dirichlet Allocation (PLDA)
- Pachinko Allocation Model (PAM)

For our project, we used Latent Dirichlet Allocation (LDA) to develop our model for topic modelling.

2.2 Latent Dirichlet Allocation

LDA is an algorithm used to implement topic modelling. The main goal of LDA is to find topics in a collection of documents and associate each document to the topics that are most relevant to it. LDA was first proposed in 2000 by J. K. Pritchard, M. Stephens and P. Donnelly [2][3]. When provided with the number of topics as input, LDA can help determine the cluster of words that belong to each topic, with probability associated with each word and the topics that could be relevant to each document along with probability for each topic. The parameters [1] of LDA are listed as below :

- corpus: Collection of documents with frequency count of each word in the document.
- num_topics: Number of topics to be extracted from the corpus.
- id2word: A dictionary consisting of unique words in the collection of documents or in corpus where each word is associated with an integer id.
- workers: Number of processes used for parallelization.
- chunksize: It determines number of documents to be considered in each chunk while training the model.
- passes: Number of iterations through the collection of documents during training.
- decay: A number between the range 0.5 and 1 to measure the forgotten percentage of the previous lambda value for each new document.
- eval_every: This parameter represents the number of updates after which the log perplexity is estimated. When set to 1, this tends to slow down the training process by twice the amount of time.
- iterations: This represents maximum number of iterations through the collection of documents while determining the distribution of topics.
- gama_threshold: This parameter sets the minimum difference to be observed in the gamma parameter values to continue iteration.
- minimum_probability: This parameter provides the minimum probability for a topic to be considered. Topics with a probability less than this threshold are filtered out.
- random_state: This parameter can either be a randomState object or a seed value to generate a randomState object.
- minimum_phi_value: This parameter represents the minimum value of term probability to be considered when per_word_topics is true.
- per_word_topics: When true, this parameter indicates that the model must generate a list of most likely topics for each word in descending order multiplied by their word count.
- dtype: This parameter determines the data type to be used for all inputs during any calculations performed by the model.

The hyperparamters of lda method are:

- alpha: It represents the density of topics in a document.
- beta: It represents the density of words in a topic.
- offset: This parameter determines the speed of the first few steps in the first few iterations.

For the model used in this project we used the following parameters:

- corpus: Frequency of all words for all articles provided.
- num_topics: Building various models for different number of topics varying from 10 to 30.

- id2word: Dictionary of words made from all the articles.
- passes: The value is set as 2.
- workers: The value is set as 2.

3 RESULT

The figure (1) represents the Coherence measure for different number of topics.

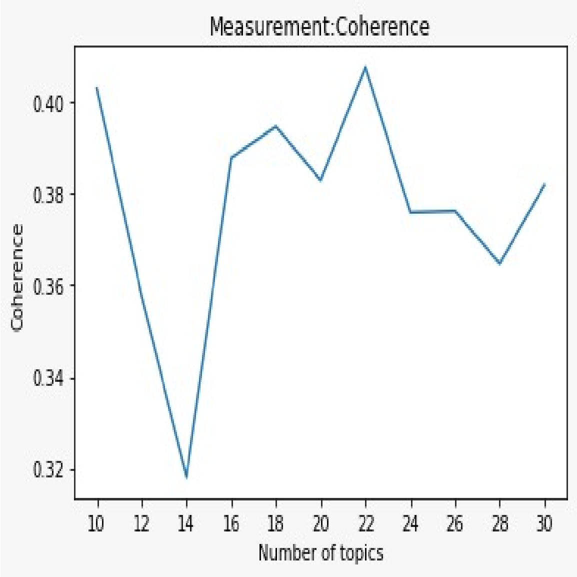


Figure 1: Coherence

The figure (2) represents the Perplexity measure for different number of topics.

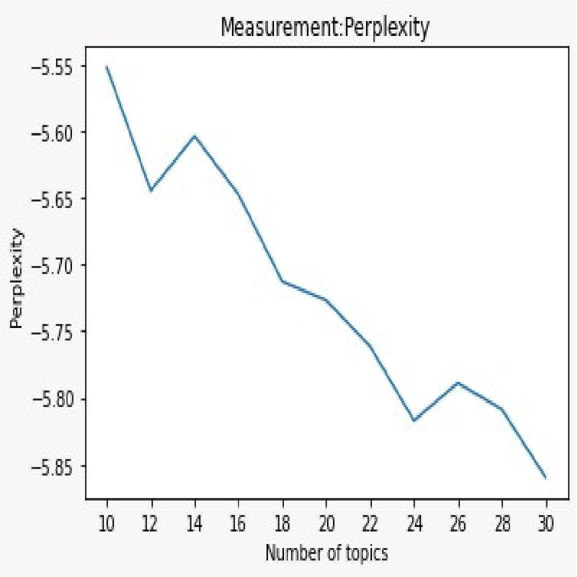
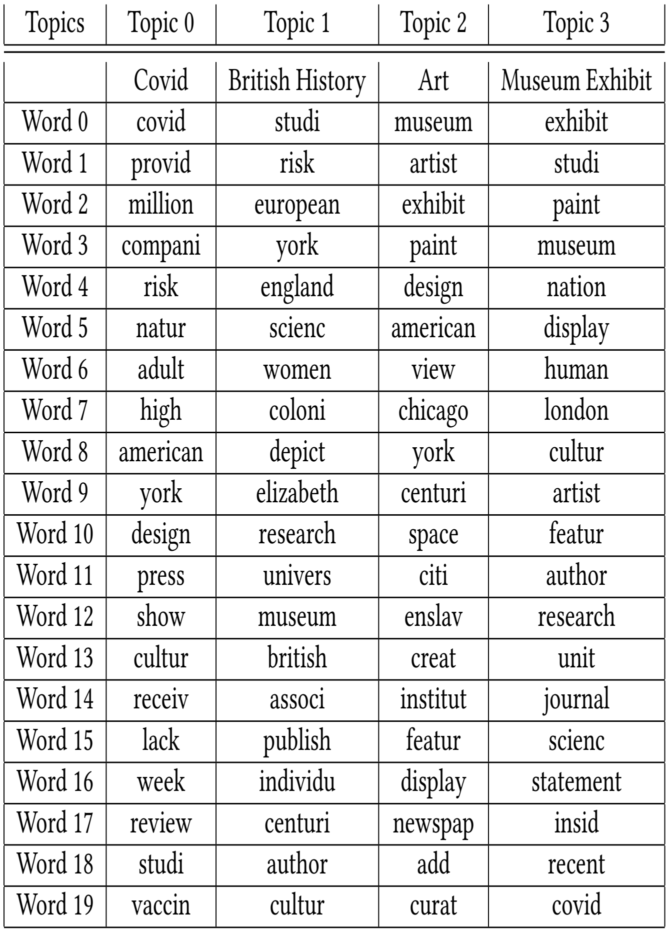


Figure 2: Perplexity

The tables below shows 22 topics and their names with 20 words that occur frequently in each topic.



Topics	Topic 0	Topic 1	Topic 2	Topic 3
	Covid	British History	Art	Museum Exhibit
Word 0	covid	studi	museum	exhibit
Word 1	provid	risk	artist	studi
Word 2	million	european	exhibit	paint
Word 3	compani	york	paint	museum
Word 4	risk	england	design	nation
Word 5	natur	scienc	american	display
Word 6	adult	women	view	human
Word 7	high	coloni	chicago	london
Word 8	american	depict	york	cultur
Word 9	york	elizabeth	centuri	artist
Word 10	design	research	space	featur
Word 11	press	univers	citi	author
Word 12	show	museum	enslav	research
Word 13	cultur	british	creat	unit
Word 14	receiv	associ	institut	journal
Word 15	lack	publish	featur	scienc
Word 16	week	individu	display	statement
Word 17	review	centuri	newspap	insid
Word 18	studi	author	add	recent
Word 19	vaccin	cultur	curat	covid

Topics	Topic 4	Topic 5	Topic 6	Topic 7
	Climate Change	Science Research	Slavery	Unclear
Word 0	percent	scienc	paint	bear
Word 1	climat	human	african	nation
Word 2	chang	plant	enslav	explor
Word 3	power	research	research	week
Word 4	gener	scientist	artist	octob
Word 5	scienc	gamillo	famili	septemb
Word 6	insid	elizabeth	british	success
Word 7	plant	african	center	fall
Word 8	statement	speci	depict	biolog
Word 9	book	know	california	care
Word 10	england	innov	slaveri	month
Word 11	histor	david	websit	know
Word 12	project	natur	group	brown
Word 13	california	american	univers	titl
Word 14	food	suggest	coloni	take
Word 15	water	reach	black	statu
Word 16	unit	help	identifi	summer
Word 17	creat	complet	david	washington
Word 18	countri	statement	life	best
Word 19	million	north	european	access

Topics	Topic 8	Topic 9	Topic 10	Topic 11
	Tourism	Artist	Ancient England	Fall Research
Word 0	citi	paint	histor	british
Word 1	york	artist	england	research
Word 2	museum	museum	research	fall
Word 3	artist	exhibit	sit	site
Word 4	visitor	life	featur	plan
Word 5	covid	hand	explor	play
Word 6	institut	studi	cultur	stori
Word 7	chicago	vaccin	grow	home
Word 8	photo	guardian	statement	take
Word 9	cultur	draw	period	histor
Word 10	anim	famili	centuri	land
Word 11	histor	receiv	david	come
Word 12	percent	featur	nation	octob
Word 13	gener	creat	statu	near
Word 14	note	figur	show	york
Word 15	scienc	plan	countri	england
Word 16	collect	technic	farm	writer
Word 17	nation	health	need	hous
Word 18	creat	show	websit	video
Word 19	mcgreevi	later	mark	david

Topics	Topic 12	Topic 13	Topic 14	Topic 15
	Medical Research	Black Culture	Revolution	Covid Study
Word 0	anim	photograph	statu	covid
Word 1	nation	anim	protest	vaccin
Word 2	washington	captur	citi	wetzel
Word 3	statement	museum	artist	american
Word 4	vaccin	artist	offici	corryn
Word 5	david	american	remov	studi
Word 6	speci	covid	octob	risk
Word 7	offici	enslav	globe	gamillo
Word 8	covid	women	major	elizabeth
Word 9	biolog	research	paint	scienc
Word 10	staff	number	univers	high
Word 11	institut	take	stand	receiv
Word 12	abl	natur	countri	nation
Word 13	farm	black	add	effect
Word 14	scientist	white	coloni	compani
Word 15	kindi	paint	south	caus
Word 16	receiv	statement	follow	expert
Word 17	start	studi	enslav	power
Word 18	research	want	press	nora
Word 19	review	compani	york	mcgreevi

Topics	Topic 16	Topic 17	Topic 18	Topic 19
	American History	Women in Science	Media	Cinema
Word 0	research	women	risk	photograph
Word 1	black	scienc	york	anim
Word 2	take	receiv	scienc	captur
Word 3	american	past	studi	compani
Word 4	number	field	case	scene
Word 5	right	white	research	pictur
Word 6	univers	peer	current	home
Word 7	remov	percent	journal	know
Word 8	scienc	wetzel	press	artist
Word 9	studi	corryn	publish	take
Word 10	author	american	gamillo	natur
Word 11	white	decad	individu	final
Word 12	statu	scientist	high	museum
Word 13	associ	black	adult	paint
Word 14	paper	alongsid	final	statement
Word 15	find	begin	associ	make
Word 16	beauti	color	life	want
Word 17	claim	washington	measur	women
Word 18	center	research	member	director
Word 19	come	centuri	statement	larg

Topics	Topic 20	Topic 21
	Biology	Unclear
Word 0	innov	statu
Word 1	elizabeth	enslav
Word 2	gamillo	africa
Word 3	predict	british
Word 4	scientist	paint
Word 5	captur	remov
Word 6	futur	protest
Word 7	organ	coloni
Word 8	farm	figur
Word 9	health	member
Word 10	american	univers
Word 11	corryn	centuri
Word 12	yield	histor
Word 13	spot	south
Word 14	video	leav
Word 15	wetzel	famili
Word 16	scienc	african
Word 17	carbon	slaveri
Word 18	gershon	group
Word 19	environ	legaci

For each topic, the articles that have the highest likelihood have been listed below with their names and dates:

- Topic 0
  Article Names :
  - Roman-Era Statue of Venus, Goddess of Love, Discovered in England
  - 'Merck Asks FDA to Authorize Promising Covid-19 Pill',
  - 'Banksy Murals in England Defaced, Removed Just Days After Appearing',
  - 'Seafood Prices Soar Amid Supply Chain Issues and Worker Shortage',
  - 'Merck Asks FDA to Authorize Promising Covid-19 Pill',
  - 'Merck Asks FDA to Authorize Promising Covid-19 Pill'

  Published Dates:
  - '2021-10-05T14:54:36-04:00',
  - '2021-10-13T14:57:39-04:00',
  - '2021-08-16T13:49:43-04:00',
  - '2021-08-04T14:10:53-04:00',
  - '2021-10-13T14:57:39-04:00',
  - '2021-10-13T14:57:39-04:00'
- Topic 1
  Article Names :
  - "Obsidian 'Spirit Mirror' Used by Elizabeth I's Court Astrologer Has Aztec Origins"

Published Dates:
- ’2021-10-07T12:39:08-04:00’

- Topic 2
  Article Names :
  - "Don’t Just Look at These Paintings—Smell Them Too, Says New Dutch Exhibition",
  - ’Why Andy Warhol Peed on This Portrait of Jean-Michel Basquiat’,
  - ’Major Barbara Kruger Exhibition Spills Out Into the Streets of Chicago’,
  - "Artist Takes Museum’s $84,000, Returns With Blank Canvases Titled ’Take the Money and Run’",
  - "Tracing Christian Dior’s Evolution, From the Postwar ’New Look’ to Contemporary Feminism",
  - ’Major Barbara Kruger Exhibition Spills Out Into the Streets of Chicago’,
  - ’Inscribed VIP Seats Unearthed at Roman Amphitheater in Turkey’,
  - ’The Sights and Sounds of the Sea Have Inspired American Artists for Generations’,
  - ’First Museum Dedicated to American Arts and Crafts Movement Opens in Florida’,
  - ’New Maryland Museum Dives Into the Mythology of Mermaids’

  Published Dates:
  - ’2021-03-18T07:00:00-04:00’,
  - ’2021-10-06T14:48:11-04:00’,
  - ’2021-10-05T12:05:29-04:00’,
  - ’2021-10-01T13:40:24-04:00’,
  - ’2021-09-27T11:13:14-04:00’,
  - ’2021-10-05T12:05:29-04:00’,
  - ’2021-10-04T08:11:23-04:00’,
  - ’2021-09-17T06:30:00-04:00’,
  - ’2021-09-07T12:20:27-04:00’,
  - ’2021-08-03T10:07:49-04:00’
- Topic 3
  Article Names :
  - ’Dogs Sniff Out Answers to Bat and Bird Fatalities Near Wind Turbines ’,
  - ’Trove of Artifacts, Many Recovered From Abroad, Traces 4,000 Years of Mexican History’,
  - "Pop-Up Exhibition Brings Masterpieces From London’s National Gallery Outdoors"

  Published Dates:
  - ’2021-07-29T15:56:12-04:00’,
  - ’2021-10-04T13:20:32-04:00’,
  - ’2021-08-13T06:30:00-04:00’
- Topic 4
  Article Names :
  - "McDonald’s Will Offer More Sustainable Happy Meal Toys by 2025",
  - ’Western Drought Drives Decline in Hydroelectric Power Generation’,
  - ’Explore Stunning 360-Degree Panoramic Views of Mars in New NASA Video’,
  - ’United Kingdom Begins Large-Scale Carbon Removal Trials’

  Published Dates:
  - ’2021-10-15T12:57:21-04:00’,
  - ’2021-10-13T13:52:11-04:00’,
  - ’2021-08-24T14:57:21-04:00’,
  - ’2021-05-27T06:30:00-04:00’
- Topic 5
  Article Names :
  - "Humans’ Earliest Evidence of Tobacco Use Uncovered in Utah"

  Published Dates:
  - ’2021-10-12T16:45:17-04:00’
- Topic 6
  Article Names :
  - "Who Is the Enslaved Child in This Portrait of Yale University’s Namesake?",
  - "Who Is the Enslaved Child in This Portrait of Yale University’s Namesake?",
  - "Display of 100 Renaissance Portraits Underscores Humans’ Enduring Desire to Be Remembered",
  - ’The Best Place to Watch Monarch Butterflies Migrate Might Be This Little California Beach Town’,
  - "Who Is the Enslaved Child in This Portrait of Yale University’s Namesake?",
  - "Who Is the Enslaved Child in This Portrait of Yale University’s Namesake?"

  Published Dates:
  - ’2021-10-15T16:45:07-04:00’,
  - ’2021-10-15T16:45:07-04:00’,
  - ’2021-09-30T06:20:19-04:00’,
  - ’2021-09-22T17:09:13-04:00’,
  - ’2021-10-15T16:45:07-04:00’,
  - ’2021-10-15T16:45:07-04:00’
- Topic 7
  Article Names :
  - ’Meet the Bodaciously Bulky Bears of Fat Bear Week 2021’,
  - ’Five Cheetah Cubs Born at Smithsonian Conservation Biology Institute’,
  - ’Facebook Addresses Illegal Sales of Amazon Rainforest Lands on Its Platform’,
  - "World’s Widest Airplane Completes Successful Second Test Flight",
  - ’See When Fall Foliage Will Peak With This Interactive Map’,
  - ’Meet the Bodaciously Bulky Bears of Fat Bear Week 2021’,
  - ’Five Cheetah Cubs Born at Smithsonian Conservation Biology Institute’,
  - ’Meet the Bodaciously Bulky Bears of Fat Bear Week 2021’,
  - ’Five Cheetah Cubs Born at Smithsonian Conservation Biology Institute’

  Published Dates:
  - ’2021-09-29T14:45:29-04:00’,
  - ’2021-10-13T15:03:59-04:00’,
  - ’2021-10-13T12:05:52-04:00’,
  - ’2021-05-03T09:30:00-04:00’,
  - ’2021-08-31T12:58:09-04:00’,
  - ’2021-09-29T14:45:29-04:00’,
  - ’2021-10-13T15:03:59-04:00’,

  - ’2021-09-29T14:45:29-04:00’,
  - ’2021-10-13T15:03:59-04:00’
- Topic 8
  Article Names :
  - "Colonial-Era Papers Stolen From Mexico’s National Archive Return Home",
  - "Why a String Quartet Set Sail on a Giant Violin in Venice’s Grand Canal",
  - ’Starting Next Summer, Day-Trippers Will Have to Pay to Enter Venice’

  Published Dates:
  - ’2021-09-28T07:44:00-04:00’,
  - ’2021-09-24T08:28:13-04:00’,
  - ’2021-08-27T12:15:34-04:00’
- Topic 9
  Article Names :
  - "The Untold Story of van Gogh’s Once-Maligned Masterpiece,
  - ’The Potato Eaters’",
  - "The Untold Story of van Gogh’s Once-Maligned Masterpiece,
  - ’The Potato Eaters’"

  Published Dates:
  - ’2021-10-07T14:48:05-04:00’,
  - ’2021-10-07T14:48:05-04:00’
- Topic 10 Article Names :
  - "This Interactive Map Lets Users Explore England’s Hidden Archaeological Landscape",
  - ’Lab-Grown Coffee Passes Taste Test’,
  - "This Interactive Map Lets Users Explore England’s Hidden Archaeological Landscape"

  : Published Dates
  - ’2021-10-12T14:26:26-04:00’,
  - ’2021-09-21T14:00:30-04:00’,
  - ’2021-10-12T14:26:26-04:00’
- Topic 11
  Article Names :
  - ’Underwater Museum Allows Divers to Explore Shipwrecks From the Battle of Gallipoli’,
  - "Meteorite Crash-Landed in a Canada Woman’s Bed While She Slept",
  - "You Could Own the Landmark That Inspired Winnie-the-Pooh’s ’Poohsticks Bridge’",
  - ’Underwater Museum Allows Divers to Explore Shipwrecks From the Battle of Gallipoli’,
  - "You Could Own the Landmark That Inspired Winnie-the-Pooh’s ’Poohsticks Bridge’",
  - "Meteorite Crash-Landed in a Canada Woman’s Bed While She Slept",
  - "Meteorite Crash-Landed in a Canada Woman’s Bed While She Slept"

  Published Dates:
  - ’2021-10-08T16:43:16-04:00’,
  - ’2021-10-15T16:23:28-04:00’,
  - ’2021-10-06T15:33:34-04:00’,
  - ’2021-10-08T16:43:16-04:00’,
  - ’2021-10-06T15:33:34-04:00’,
  - ’2021-10-15T16:23:28-04:00’,
  - ’2021-10-15T16:23:28-04:00’
- Topic 12
  Article Names :
  - "Lions and Tigers at the Smithsonian’s National Zoo Test Positive for Covid-19",
  - ’Two Escaped Zebras Are Still Roaming the Suburbs of Maryland’,
  - ’With a Nearly Foot-Long Proboscis, This New Moth Species Holds Record for Longest Insect Tongue’,
  - ’Researchers Potty Trained Young Cows, a Promising Measure to Reduce Greenhouse Gases’,
  - ’Bruce the Parrot Uses Tools to Survive Despite a Broken Beak’,
  - "’Star Wars’–Like Running Robot Finishes 5K on Two Legs",
  - "Permafrost Thaw in Siberia Creates a Ticking ’Methane Bomb’ of Greenhouse Gases, Scientists Warn",
  - "Lions and Tigers at the Smithsonian’s National Zoo Test Positive for Covid-19",
  - ’Two Escaped Zebras Are Still Roaming the Suburbs of Maryland’,
  - "Lions and Tigers at the Smithsonian’s National Zoo Test Positive for Covid-19",
  - ’Two Escaped Zebras Are Still Roaming the Suburbs of Maryland’

  Published Dates:
  - ’2021-09-17T11:51:53-04:00’,
  - ’2021-10-06T14:39:59-04:00’,
  - ’2021-10-07T16:44:24-04:00’,
  - ’2021-09-15T15:28:28-04:00’,
  - ’2021-09-14T16:18:45-04:00’,
  - ’2021-07-30T14:19:31-04:00’,
  - ’2021-08-05T14:50:53-04:00’,
  - ’2021-09-17T11:51:53-04:00’,
  - ’2021-10-06T14:39:59-04:00’,
  - ’2021-09-17T11:51:53-04:00’,
  - ’2021-10-06T14:39:59-04:00’
- Topic 13: None
- Topic 14
  Article Names:
  - ’Statue of Pre-Hispanic Woman Will Replace Columbus Sculpture in Mexico City’,
  - ’Statue of Pre-Hispanic Woman Will Replace Columbus Sculpture in Mexico City’,
  - ’Statue of Pre-Hispanic Woman Will Replace Columbus Sculpture in Mexico City’

  Published Dates:
  - ’2021-09-09T10:13:55-04:00’,
  - ’2021-09-09T10:13:55-04:00’,
  - ’2021-09-09T10:13:55-04:00’
- Topic 15
  Article Names :
  - "Powerful Immune System Response May Be Behind ’Covid Toes’",
  - ’Receiving a Flu Shot and Covid Vaccine at the Same Time Is Safe, Study Finds’,

  - ’Covid-19’,
  - ’Covid-19’

  Published Dates:
  - ’2021-10-12T14:59:57-04:00’,
  - ’2021-10-08T15:59:23-04:00’,
  - ’2020-03-25T10:30:00-04:00’,
  - ’2020-03-25T10:30:00-04:00’
- Topic 16
  Article Names:
  - ’Survey Identifies Correlation Between Confederate Monuments and Lynchings’,
  - ’Survey Identifies Correlation Between Confederate Monuments and Lynchings’,
  - ’Survey Identifies Correlation Between Confederate Monuments and Lynchings’

  Published Dates:
  - ’2021-10-14T13:35:31-04:00’,
  - ’2021-10-14T13:35:31-04:00’,
  - ’2021-10-14T13:35:31-04:00’
- Topic 17
  Article Names:
  - "No Nobel Prizes in Science Went to Women This Year, Widening the Awards’ Gender Gap"

  Published Dates:
  - ’2021-10-08T15:28:36-04:00’
- Topic 18
  Article Names:
  - "Aspirin No Longer Recommended as a Preventative Measure Against Heart Attacks and Strokes in Older Individuals’,
  - ’Aspirin No Longer Recommended as a Preventative Measure Against Heart Attacks and Strokes in Older Individuals’,
  - ’Aspirin No Longer Recommended as a Preventative Measure Against Heart Attacks and Strokes in Older Individuals’

  Published Dates:
  - ’2021-10-14T09:55:41-04:00’,
  - ’2021-10-14T09:55:41-04:00’,
  - ’2021-10-14T09:55:41-04:00’
- Topic 19
  Article Names:
  - "Ruthie Tompson, Who Shaped Disney’s Most Beloved Films, Dies at 111",
  - ’Ten Breathtaking Images From the 2021 Nature Wildlife Photographer of the Year Awards’,
  - "Ruthie Tompson, Who Shaped Disney’s Most Beloved Films, Dies at 111",
  - ’Ten Breathtaking Images From the 2021 Nature Wildlife Photographer of the Year Awards’,
  - "Ruthie Tompson, Who Shaped Disney’s Most Beloved Films, Dies at 111",
  - ’Ten Breathtaking Images From the 2021 Nature Wildlife Photographer of the Year Awards’,
  - "Ruthie Tompson, Who Shaped Disney’s Most Beloved Films, Dies at 111"

  Published Dates:
  - ’2021-10-13T15:43:20-04:00’,
  - ’2021-10-15T13:19:39-04:00’,
  - ’2021-10-13T15:43:20-04:00’,
  - ’2021-10-15T13:19:39-04:00’,
  - ’2021-10-13T15:43:20-04:00’,
  - ’2021-10-15T13:19:39-04:00’,
  - ’2021-10-13T15:43:20-04:00’
- Topic 20
  Article Names:
  - ’Innovation for Good’

  Published Dates:
  - ’2021-06-01T09:28:51-04:00’
- Topic 21
  Article Names
  - "Why a New Plaque Next to Oxford’s Cecil Rhodes Statue Is So Controversial",
  - ’New A.I. Tool Makes Historic Photos Move, Blink and Smile’,
  - "Why a New Plaque Next to Oxford’s Cecil Rhodes Statue Is So Controversial",
  - ’Abdulrazak Gurnah, Chronicler of Migrant Experience, Wins 2021 Nobel Prize in Literature’,
  - "Why a New Plaque Next to Oxford’s Cecil Rhodes Statue Is So Controversial",
  - "Why a New Plaque Next to Oxford’s Cecil Rhodes Statue Is So Controversial"

  Published Dates:
  - ’2021-10-13T13:36:11-04:00’,
  - ’2021-03-09T07:30:00-05:00’,
  - ’2021-10-13T13:36:11-04:00’,
  - ’2021-10-07T17:05:50-04:00’,
  - ’2021-10-13T13:36:11-04:00’,
  - ’2021-10-13T13:36:11-04:00’

4 DISCUSSION

As discussed above, we generated a number of LDA models with varying number of topics between the range of 10 to 30. To find the optimal number of topics that can best represent the collection of articles, we evaluated the performance of each model using the the performance measures Perplexity and Coherence. Perplexity provides a measure of how accurately a LDA model predicts for a given sample text.The lower the value of the Perplexity score, the better the performance of the model. Coherence gives a measure of the quality of the cluster of words in a topic. It helps us understand if the words grouped in a topic are similar to each other.The higher the Coherence value, the better the performance of the model. The Coherence value is the main determinant of the optimal number of topics to be From Figure 1, it can be observed that the Coherence value is the highest when the number of topics is 22. It can be observed from Figure 2, that the Perplexity value is on the lower end when the number of topics is 22. Hence we decided to choose 22 as the suitable number of topics for our LDA model.

The most time-consuming part of the project was gathering articles from the website. This was due to the fact that this method necessitated us comprehending the website's format and identifying the tags that give links to content. Then we had to parse each article link to get information about each article, such as the title, author, publication date, and article text body.Using BeautifulSoup, we first pulled the content from the Smithsonian website's homepage. After that, we looked at the website's structure and decided on the Smart News category of articles to use for topic modeling. We found links to sub-categories of the Smart News section using the class name that provided category-related information. These included History, Science, Innovation, ArtsCulture, and Travel. We parsed the content of the sub-category links with another BeautifulSoup object to get the URLs to the articles in these sections. We then retrieved article information from each link obtained via a json script, which included the title, author, date, and body of the article.

From the titles of articles relevant to each topic it can be observed that the articles are mostly similar to each other. However, getting the proper distribution of articles is majorly dependent on choosing the right number of topics. Topic modelling is very powerful as it helps to distribute the words for the given document in a meaningful way without attaching any labels to the document manually.

REFERENCES

[1] 2021. Gensim: Topic modelling for humans. https://radimrehurek.com/gensim/models/ldamulticore.html#module-gensim.models.ldamulticore

[2] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164, 4 (08 2003), 1567–1587. https://doi.org/10.1093/genetics/164.4.1567 arXiv:https://academic.oup.com/genetics/article-pdf/164/4/1567/37401990/genetics1567.pdf

[3] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164, 4 (08 2003), 1567–1587. https://doi.org/10.1093/genetics/164.4.1567 arXiv:https://academic.oup.com/genetics/article-pdf/164/4/1567/37401990/genetics1567.pdf

[4] Smithsonian Magazine. 2021. https://www.smithsonianmag.com/

[5] Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. 1998. Latent Semantic Indexing: A Probabilistic Analysis. *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, May 1998* (May 1998), 159–168 pages. https://dl.acm.org/doi/10.1145/275487.275505.