Assignment 4
Topic Modeling

Ujjwal Yadav
2018CSB1127

Abstract

This Assignment deals with analysing a large set of text documents , processing it and extracting topics by means of an unsupervised Algorithm of Topic Modeling .

Here we learn about Topic Modeling Algorithm , which is an unsupervised algorithm . We use methods of Natural language Processing , so as to covert the textual data into a mathematical form to apply machine Learning algorithms on it . We learn about Latent Semantic Indexing (LSI) Topic Modeling and LDA Topic Modeling

1 Introduction

Here we use two Datasets -
1) The first dataset is the State of Union Dataset , which is a collection of Speeches along with the year in which they were delievered .
2) The second Dataset we use is AP Wire Stories dataset.

1.1 TASK 1

Task 1 deals with the processing of the State of Union dataset. The dataset is a csv file containing two columns , one with the year of speech and other containing the text of speech . Working with text data requires us to process it into a suitable mathematical form. First the csv file is read and the data is loaded into a DataFrame. Then we extract the text part of the dataset , i.e , we extract the speeches of the Presidents and tokenize the dataset by converting the strings into lower case and splitting words by space . We also pass a list of Stop words , which makes sure that we do not include those words under our set of words , which are there in the list . The stopwords are mostly those words , which are less meaningful and in general do not convey the meaning of the text . We then form a Bag of Words by counting the frequency of every token or word that we get in the above step. Those words which occur just once in the whole text of the Dataset are neglected as they do not contribute much in determining the Topic of speeches . In this way we generated the Dictionary of the words which occur in our Text and generate an unnormalized frequency of each word or token in the Token . Then , the bag of words are fed into the Genism Package , which is Used for Natural Language Processing and Text Interpretation . We generate the tfidf weighted vectors for each of the speeches in the dataset .

1.2 Task 2

Here we apply The Latent Semantic Indexing on our processed dataset of tf-idf weighted vectors, generated for each document(speech) , to find the Topics that can be interpreted from the entire text. It gives us a list of words along with their coefficients for each topic . The number of topics that we want need to be supplied to the LSI algorithm . Now , we can interpret the Topic that the text wants to convey from the list of words along with their coefficients given for every topic .
We randomply take words corresponding to any ten topics and then analyse them to guess a concept or theme behind thse words .
1) The words for the topics are :
-> "terror" , "terrorists" , "saddam" , "iraq" etc . which may want to convey some message related to "Terrorism In Iraq". We can relate it to as a real concept as the US President may have given a speech related to the problems in Iraq and their former President Saddam Hussein. There have been problems between US and Iraq in the past.
2) The words for the topics are :
->"economy" , "treaty" ,"silver" , "gold", "treasury" ,"program" etc . which may want to convey something about "Economic Policies" . Economic Policies are an important part of Government's job and there have been economic crisis like the Great Depression , World Economy Slowdown . It may be related to something like those conditions.
3) The words for the topics are :
->"soviet" , "japanese" ,"atomic","enemy","fighting" ,"communist" etc . are related to "World War II" , where US was on one side and Soviet Union and Japan were on other side . US then ,had also Atomic Bombed Japan.
4) The words for the topics are :
->"mexico" , "texas","kansas" , "tonight" ,"slavery" ,"gold","Mexicans" etc. These words hint slightly at the movement of "Immigrants from Mexico " , which happened around the beginning of 20th century . The Mexicans came to US in search of good livelihood and to overcome from poverty(indicated by Gold).
5) The words for the topics are :
->"vietnam" ,"dempcracy" ,"energy" ,"oil" ,"gentleman","production" ,"democratic" - These set of words do not specify clearly which topic is being talked about . It may be about the "Vietnam - US tussle" or it may be regarding the "Energy Resources" as suggested by Oil , Energy , Production .
6) The words for the topics are :
-> "soviet" ,"captures" ,"gentleman" ,"majesty" ,"Commisioners" etc suggest us about the "Soviet Afghan War" , where there were large differences between US and Soviet Union.
7) The words for the topics are :
"British" , "Minister" ,"silver" ,"reserved" ,"French" - They do not convey any meaningfull information about any particular topic . Something related to "European News" can be there .
8) The words for the topics are :
"Economy" ,"Industrial" ,"Income" ,"Growth" ,"California" etc describes abot the "Growth in US Economy" . These words are able to capture an important national topic like Economy .
9) The words for the topics are :
"Hitler" , "Germans" ,"Autocracy" , "Debatable" ,"Explaination " etc relate to the speech during the time when "Germany was ruled or dictated by Adolf Hitler" .

10) The words for the topics are :
"isthmus" ,"rebellion" ,"ocean" ,"Colambia" etc relate to the rebellion by the Panama nationlists , when US signed a "treaty with Colambia" .
Some more topics captured by the algorithm , which we can be identified with the historical events , policies of US government , world news . Thus we are able to get some idea of the historical events and topics about which something was said .

1.3 Task 3

Here we repeat the task done in previous section but now we use LDA or Latent Dirichlet Algorithm in place of Latent Semantic Indexing . We observe the various topics explained by our new new model employing the LDA algorithm .
-> First we understand the differences between the LDA and LSI algorithm .
-> LSI algorithm is a sort of Dimensionality Reduction Algorithm , which employs Dimensionality reduction of the the TF-IDF matrix of a large dimension , with the assumption that if a set of significant words are present in more than one documents , then the documents convey a sort of similar meaning .
->It employs SVD to reduce the dimensions of a large sparse matrix formed with the set of words as the rows and the different documents in the dataset as columns .
->It is relatively fast as it operates on a lower dimension space .
->LDA agorithm works with much larger dimension data as compared to LSI because it assumes that there are a number of topics in a document and further there are a number of words of significance for each topic .
->Thus LDA algorithm analyses a vast matrix of TF-IDF scores in our case to extract a specific distribution of topics and further extract a set of distribution of words corresponding to each topic .
LDA algorithms are slow as comapred to LSI algorithms as they deal with higher dimension space .
We now analyse the results or the topics obtained from LDA model : ->Many of the topics conveyed by LDA are quite similar to those conveyed by LSI .
->These 10 topics are randomly annotated from the topics obtained after applying LDA.
1) There is a set of words like "compassionate","hopeful" , "intiative" ,"competitive","america" etc which conveys the desire of the American leader for a "Modern America ".
2) The set of words like "Saddam" ,"Hussein" , "Terror" ,"Iraq" etc is related to the "Terror in Iraq".
3) Set of words like "financial" ,"exports" ,"sum" ,"debt" ,etc relates to condition of "Great Depression" which happened in 1920's and 1930's , and other economic slowdowns.
4) The words like "Soviet" ,"Afganistan" ,"weapons" , "war" takes us back to the time of "Soviet Afghan War" .
5) The words like "race" ,"democrats","trust","hopeful","1964","americans" etc highlight the "Civil Rights Movement in US" in 1960s against the Racial Discrimination.
6) The set of words like "Vietnam" ,"Soldiers" ,"Commisioners" ,"Treaty" etc gives us an hint about the "US-Vietnam War" and subsequent peace treaty.
7) Words like "wedding" ,"guests" ,"clouded" ,"served" ,"dishes" etc relate something to "Functions and Parties" .
8) Words like "chambers" ,"provision" ,"suspension" ,"french" etc do not provide a

clear topic , but something relating to "Suspension Measures of Officials" can be considered.
9) Words like "hovering", "commisions" ,"plundered" ,"complement" ,"peaceble","belligerer are indicator to "Crime Activity" , which may indicate to loot , assassination , etc.
10) Words like "labour-mangement" ,"wage" ,"program" ,"output" ,"subscriptions" ,"Economy"etc indicate about "The Economic Reforms".
-> The difference between results of LDA and LSI method is that in LDA method , the topics are more clearer as compared to LSI method . This is because the LSI tries to represent or convey the same meaning in a transformed matrix with much smaller dimensions as compared to the original word vs document matrix.
->So the significant words corresponding to different topics are more mixed within topics and are less separated as compared to LDA.
->While the LDA deals with much larger matrix and thus distributes each topic and further words in each topic by taking into account more amount of information , therefore the topic is slightly more clear in LDA as compared with LSI.

1.4 Task 4

->This section describes how the topics of speech have changed with respect to time . We analyse the change in topics of speech with every decade i.e 10 years of time .
->First, the original dataset (before Topic Modeling) having two columns - one with the year and the other with the Speech in that year is taken .
->Starting with the first year of 1790 , we take the data for the next 10 years at one time and process only the data of that decade at once . We generate the dictionary and bag of words for the data of our decade .
->Then we generate the tfidf weighted vectors for our processed dataset , and then apply LDA algorithm on the data of current decade.
->We extract the set of words along with the weighted coefficients for each word of the topic extracted from our data of one decade.
->We then proceed further by taking the data for the next decade and repeat the same above steps for every decade data.
->We try to predict the different events with the help of topics generated and verify with the actual historical facts.
Decade-wise Change in Topics Observations
->We can connect with the major historical events by analysing the words of the topics extracted during that period.
->In the period between 1922 and 1932 , we can see the words which relate to the Great Depression like "emergencies" , "banking" ,"unemployment" ,"world-wide" ,"recovery" ,"expansion" ,"visas" etc . which are the economical terms and indiacte the crisis at that time.
-> In the period between 1910-1922 , we can see the words like "army" ,"germany" ,"foreign" ,"troops" etc. which are indicative factors of World War I which happened around that time.
-> Also , in the period between 1932-1952 , we can see the words like "autocracy" which mmay well indicate rule of Hitler ,"war" representing World War II , "reconstruction" which refers to recovery after war ,"Japan" and "Atomic" signalling the Atomic Bombing by U.S on Japan.
-> In the period between 1860-1890 , there are topics including words like "Emancipation" ,"depreciation" ,"1864" ,"kansas" ,"African" ,"liberty" etc signalling to the time of American

Civil War during that period.

1.5 Task 5

Here we perform Topic Modeling on the new dataset.
->The interpretation of 10 random samples from the topics obtained is shown below:
1) The words like "electricity" ,"adapt" ,"vietnames" ,"hydroelectric" ,"shark","collective" etc indicate "Collective Hydroelectricity Adaptation" topic which may imply that all whether Vietnames or Shark have adopted to new electricity called Hydroelectricity.
2) The words like "sentenced" ,"convicted" , "prison","navy" ,"troops" ,"european","nation etc indicate about something like "Convicted Sentenced Prison" and something about "US troops".
3) The words like "human" , "electoral" ,"abuses" ,"government" ,"collapsed" etc may indicate about the topic "Human Rights Disrespected".
4) The words like "low-water" , "recieve","hubbard" ,"California" etc may indicate about the "Lower Water Levels" around the region of California and Hubbard glacier.
5) The words like "assaults" ,"wrongdoing","torture","orphans","counseling" , "require" may indicate the need for "Counseling for Orphans".
6) The words like "natural" , "species" ,"week","per" ,"issue" may indicate the issue of "Conservation of Nature".
7) The words like "turnout" ,"democrats" ,"votes","municipal", "kennedy" ,"opposition","voter", etc give information of the "Election in US" topic.
8) The words like "senate" ,"legislation" ,"nomination" ,"ethics " etc may indicate about the "Nomination for Senate".
9) Words like "index" ,"financial" ,"economy" ,"wages" etc indicate about the "Economic Status".
10) Words like "toshiba" ,"resort" ,"korea" ,"lenders" etc indicate some information about the "Toshiba Traders".
-> The topics obtained by applying LDA on this dataset are more clearer than those obtained with States of Union dataset , as this dataset is not as vast subjected and large as compared with the States iof Union Dataset. The topics on which data is there in this dataset are not as vast as compared to previous dataset.
-> Due to this reason ,the segregation is better in this case as compared to previous case.
-> This dataset works better in topic extraction because each document of the text is not concentrated on a large number of topics , while in the State of Union Dataset , each document was speech given by the President of U.S , which included multiple topics.
-> So here in this dataset , better working means that after applying a Topic Modeling algorithm , when we obtain a set of words along with their coefficients for each topic , we can guess the topic or concept more efficiently about which text is written.