NLP- Topic Modelling

Instructions

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: Ayyappa Bharthwaj Nukala

Batch Id: DSWDMCOS 21012022

Topic: NLP- Topic Modelling

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Python codes perform:

3. Data Pre-processing

2.1 Data Cleaning, Feature Engineering, etc.

4. Exploratory Data Analysis (EDA)

5. Model Building

5.1 Perform Data Cleaning, Stemming, Lemmatization, Topic Modelling and Text Summarization

6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

Note:

The assignment should be submitted in the following format:

- Python code
- Documentation of the modules (elaborating on steps mentioned above)

Problem Statement-1

1) Perform NLP – Topic Modelling and Text summarization by following all the steps as mentioned below: -
2) Data Cleaning using regular expressions, Count Vectorizer, POS Tagging, NER, Topic Modelling (LDA, LSA) and Text summarization.
   Hint: - Use Data.csv file given in hands on material.

tweet_id	sentiment	text	tweet_created	tweet_location	user_timezone
1	neutral	What @dhepburn said.	24/02/2015 11:35		Eastern Time (US & Canada)
2	positive	plus you've added commercials to the experience... tacky.	24/02/2015 11:15		Pacific Time (US & Canada)
3	neutral	I didn't today... Must mean I need to take another trip!	24/02/2015 11:15	Lets Play	Central Time (US & Canada)
4	negative	it's really aggressive to blast obnoxious "entertainment" in your guests' faces	24/02/2015 11:15		Pacific Time (US & Canada)
5	negative	and it's a really big bad thing about it	24/02/2015 11:14		Pacific Time (US & Canada)
6	negative	seriously would pay $30 a flight for seats that didn't have this playing.	24/02/2015 11:14		Pacific Time (US & Canada)
7	positive	yes, nearly every time I fly VX this __ar worm_ won__ go away :)	24/02/2015 11:13	San Francisco CA	Pacific Time (US & Canada)
8	neutral	Really missed a prime opportunity for Men Without Hats parody, there. https:	24/02/2015 11:12	Los Angeles	Pacific Time (US & Canada)
9	positive	Well, I didn't_but NOW I DO! :-D	24/02/2015 11:11	San Diego	Pacific Time (US & Canada)
10	positive	it was amazing, and arrived an hour early. You're too good to me.	24/02/2015 10:53	Los Angeles	Eastern Time (US & Canada)
11	neutral	did you know that suicide is the second leading cause of death among teens 1	24/02/2015 10:48	1/1 loner squad	Eastern Time (US & Canada)
12	positive	I <3 pretty graphics. so much better than minimal iconography. :D	24/02/2015 10:30	NYC	America/New_York
13	positive	This is such a great deal! Already thinking about my 2nd trip to @Australia &an	24/02/2015 10:30	NYC	America/New_York
14	positive	@virginmedia I'm flying your #fabulous #Seductive skies again! U take all the	24/02/2015 10:21		Eastern Time (US & Canada)
15	positive	Thanks!	24/02/2015 10:15	San Francisco, CA	Pacific Time (US & Canada)
16	negative	SFO-PDX schedule is still MIA.	24/02/2015 10:01	palo alto, ca	Pacific Time (US & Canada)
17	positive	So excited for my first cross country flight LAX to MCO I've heard nothing but g	24/02/2015 9:42	west covina	Pacific Time (US & Canada)
18	negative	I flew from NYC to SFO last week and couldn't fully sit in my seat due to two la	24/02/2015 9:39	this place called M	Eastern Time (US & Canada)
19	positive	___ flying @VirginAmerica. _____	24/02/2015 9:15	Somewhere celeb	Atlantic Time (Canada)
20	positive	you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to	24/02/2015 9:04	Boston | Walthan	Quito
21	negative	why are your first fares in May over three times more than other carriers wher	24/02/2015 8:55		

Sol:

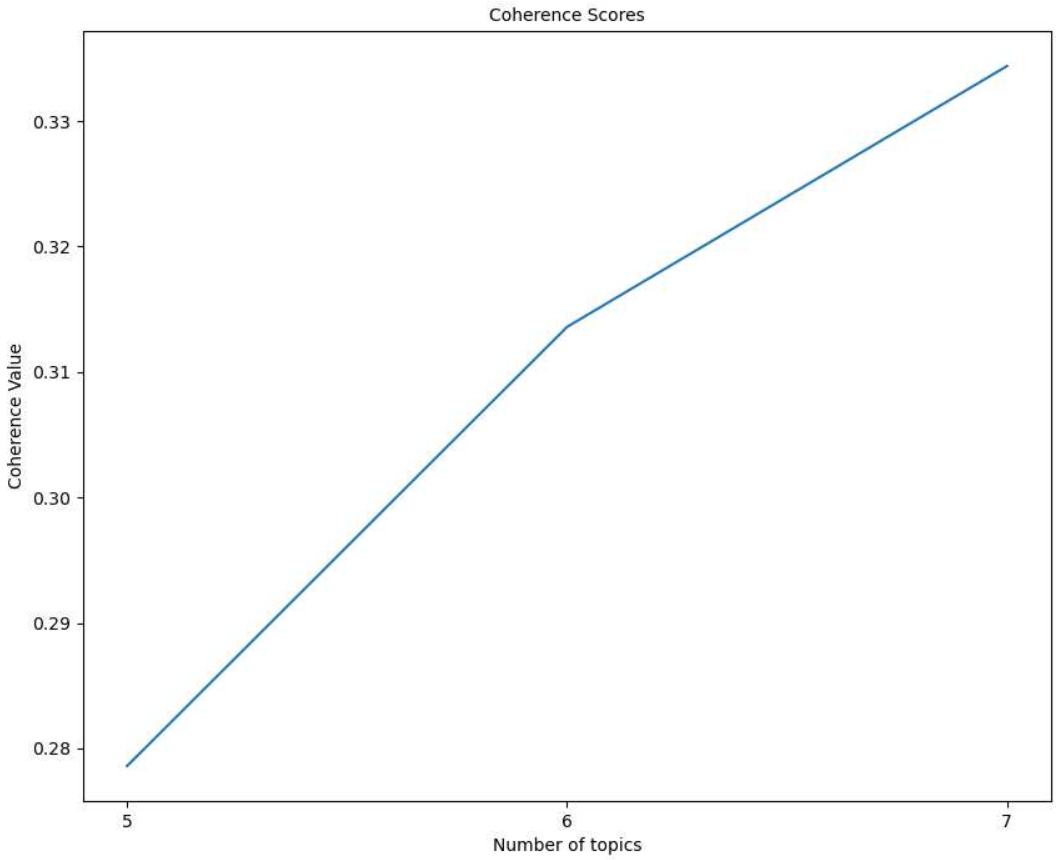
Business Objective: To create a topic modelling and text summarization for the tweets data

Data Type: The given are the tweets given by the different people having the particular tweet I'd

Data Pre Processing: I have done the data pre-processing on the data set in Python using custom functions. I have removed the punctuation marks, special character's, Tokenization etc. After cleaning the data I have done the analysis on the data.

Topic Modelling: I have done the LDA (Latent Direchlet Allocation) for the tweets data and I have made the model for the same using Python.

Text Summarization: Again I have cleaned the complete data set and then I have done the tokenization, removing the stop words, bag of words. After cleaning the data I have created a text summarization model and by using that I have summarized the data in Python.



Problem Statement-2

Perform topic modelling and text summarization on the given text data hint use NLP-TM text file.

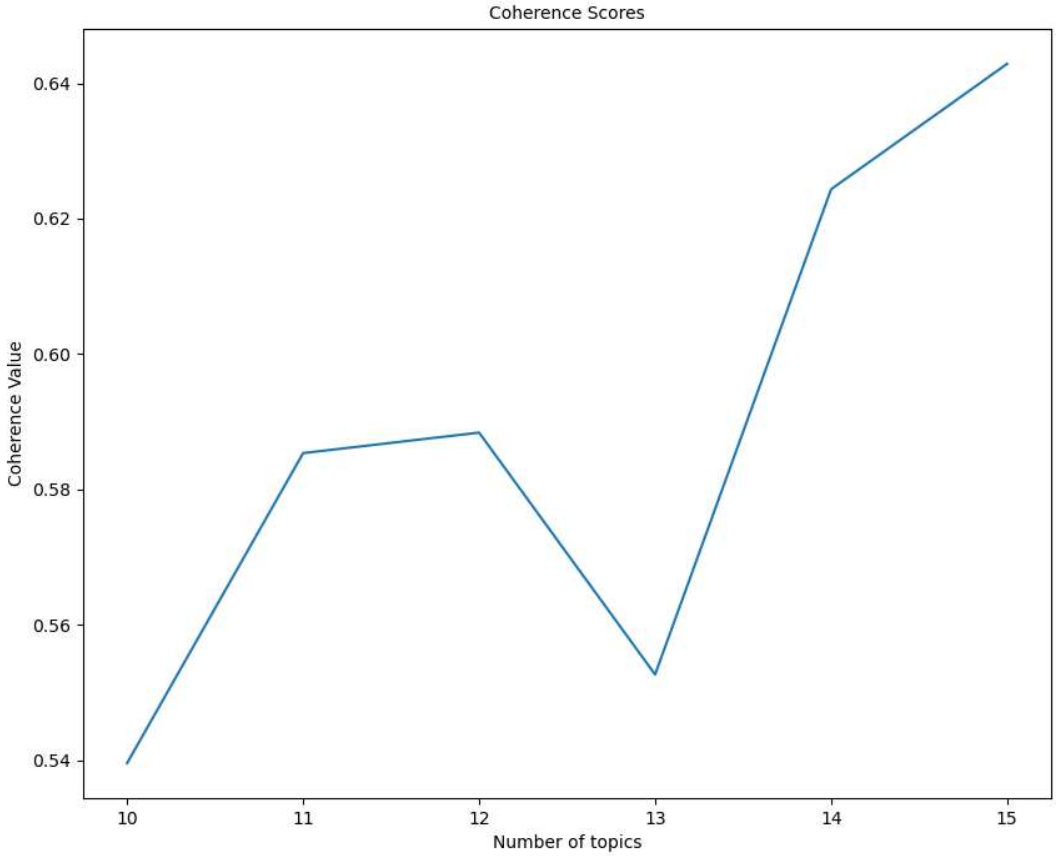
Business Objective: To create a topic modelling and text summarization for the text data.

Data Type: The text data of an article is given.

Data Pre Processing: I have done the data pre-processing on the data set in Python using custom functions. I have removed the punctuation marks, special character's, Tokenization etc. After cleaning the data I have done the analysis on the data.

Topic Modelling: I have done the LDA (Latent Direchlet Allocation) for the text data and I have made the model for the same using Python.

Text Summarization: Again I have cleaned the complete data set and then I have done the tokenization, removing the stop words, created bag of words. After cleaning the data I have created a text summarization model and by using that I have summarized the data in Python.



In an article in Cell, National Institutes of Health-funded researchers described how they used advanced genetic engineering techniques to transform a bacterial protein into a new research tool that may help monitor serotonin transmission with greater fidelity than current methods.Preclinical experiments, primarily in m showed that the sensor could detect subtle, real-time changes in brain serotonin levels during sleep, fear, and social interactions, as well as test the effecti of new psychoactive drugs.The study was funded, in part, by the NIH''s Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative which aims revolutionize our understanding of the brain under healthy and disease conditions.The study was led by researchers in the lab of Lin Tian, PhD, principal investiga at the University of California Davis School of Medicine. Current methods can only detect broad changes in serotonin signaling. In this study, the researchers transformed a nutrient-grabbing, Venus flytrap-shaped bacterial protein into a highly sensitive sensor that fluorescently lights up when it captures serotonin.Previously, scientists in the lab of Loren L. Looger, PhD, Howard Hughes Medical Institute Janelia Research Campus, Ashburn, Virginia, used traditional genetic engineering techniques to convert the bacterial protein into a sensor of the neurotransmitter acetylcholine.The protein, called OpuBC, normally snags the nutrient choline, which has a similar shape to acetylcholine. For this study, the Tian lab worked with Dr. Looger''s team and the lab of Viviana Gradinaru, Ph.D., Caltech, Pasadena, California, to show that they needed the added help of artificial intelligence to completely redesign OpuBC as a serotonin catcher.The researche used machine learning algorithms to help a computer ''think up'' 250,000 new designs. After three rounds of testing, the scientists settled on one. Initial experim suggested that the new sensor reliably detected serotonin at different levels in the brain while having little or no reaction to other neurotransmitters or similar shaped drugs.Experiments in mouse brain slices showed that the sensor responded to serotonin signals sent between neurons at synaptic communications points. Meanwh experiments on cells in petri dishes suggested that the sensor could effectively monitor changes in these signals caused by drugs, including cocaine, MDMA (also kn