**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: Avishek kumar verma**

**Batch Id: 05012021\_10A.M**

**Topic: NLP- Topic Modelling**

**1. Business Problem**

* 1. **Objective**
  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.

Text

Description automatically generated

**Ans:-**

**Business Problem**

**Objective :-** Maximize customer positive sentiment

**Constraints :-** minimize customers inconvenience

**Data types:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| tweet\_id | Tweet ID | Nominal | Not Relevent |
| sentiment | Sentiment | Nominal | Relevant |
| text | Tweet text | Nominal | Relevant |
| tweet\_created | Date | Count | Not relevant |
| Tweet\_location | Location | Nominal | Not relevant |
| user\_timezone | Time zone | Nominal | Not relevant |

**Data Pre-processing**

Performed data pre- processing. Removal of tweeter handle name, links in text, Special character.

**Exploratory Data Analysis (EDA)**

Performed EDA

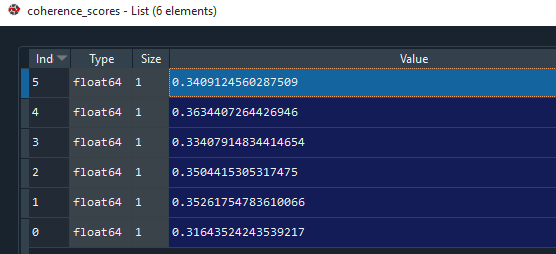
**Model Building**

Built model for LDA(Latent Dirichlet Allocation), and Text Summarization. Can refer Python code

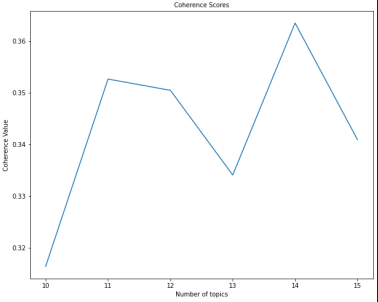
**Results:-**

**Results of LDA Model:-**

Coherence scores for 6 elements

****

Graph between coherence value and Number of topics



Problem Statement-2

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

**A picture containing letter

Description automatically generated**

**Ans:-**

**Business Problem**

Objective :- minimize time to understand

Constraints :- keep summary relevant

**Data Pre-processing :-**

Pre-process the text file with removal of stop words and punctuation marks.

**Exploratory Data Analysis (EDA)**

Performed EDA, can refer Python code.

**Model Building**

Topic modeling done in R code

Built model for Text Summarization in Python code.

**Result:-**

This is our final summary for out text data.

Out[72]: ['example technologies might include crispr-based methods, high-throughput use of compact (adeno-associated virus (aav)-sized) enhancers that can control hundreds or thousands of specific cell types; monoclonal antibodies and/or nanobodies against cell type-specific surface proteins for pseudotyping lentiviruses; aav serotypes with novel cell specificities; permanent, activity-dependent cell-marking methods; and methods that combine approaches and targets (e.g., split-gal4 with two enhancers, split-gal4 with pseudotyped lentivirus).reversible, cell type-based manipulation of brain activity would advance understanding of fundamental principles of brain function, but also guide novel therapies for brain disorders through the use of animal models.']