**Topic Modelling:**

**LDA Implementation:**

1. Loading data
2. Data cleaning
3. Tokenize and Clean-up using gensim’s simple\_preprocess():
4. Phrase Modeling: Bi-grams and Tri-grams:
5. Data Transformation: Corpus and Dictionary:
6. Preparing data for LDA analysis
7. LDA model training
8. LSA
9. PLSA

# Loading data:

# we’ll use the dataset of Amazon phone reviews available in the following location <https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews#20190928-reviews.csv>

# The CSV data file contains information on the different users reviews of different phones and the rating regarding that review.

# Let’s start by looking at the content of the file:

# 

# Data Cleaning:

# Since the goal of this analysis is to perform topic modeling, we will solely focus on the text data from each paper, and drop other metadata columns.

# 

# Now Let’s Remove the NULL values from the data by dropping the rows having NULL values.

# 

# Tokenize and Clean-up using gensim’s simple\_preprocess():

# Tokenize each sentence into a list of words, removing punctuations and unnecessary characters altogether.

# Gensim’s simple\_preprocess() is great for this. Additionally, I have set deacc=True to remove the punctuations.

# 

# Loading Stop Words:

# 

# Phrase Modeling: Bi-grams and Tri-grams:

Bigrams are two words frequently occurring together in the document. Trigrams are 3 words frequently occurring. Some examples in our example are: ‘back\_bumper’, ‘oil\_leakage’, ‘maryland\_college\_park’ etc.

[Gensim’s Phrases model](https://radimrehurek.com/gensim/models/phrases.html) can build and implement the bigrams, trigrams, quadgrams and more. The two important arguments to Phrases are min\_count and threshold.



# Let’s call the functions in order.

# 

# 

# Data Transformation: Corpus and Dictionary:

# The two main inputs to the LDA topic model are the dictionary(id2word) and the corpus. Let’s create them.

# 

# 

# Base Model:

# We have everything required to train the base LDA model. In addition to the corpus and dictionary, you need to provide the number of topics as well. Here the no of topics are 5 as we have 1-5 ratings.

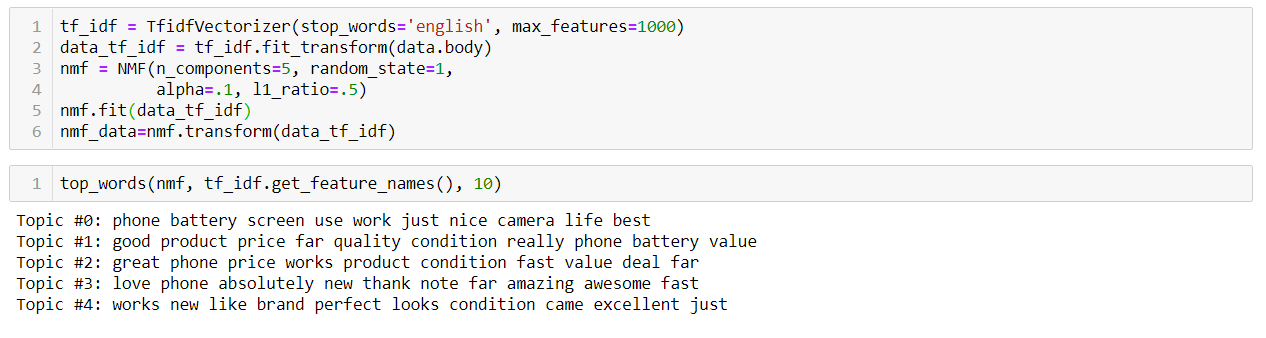
**LDA model training**

# 

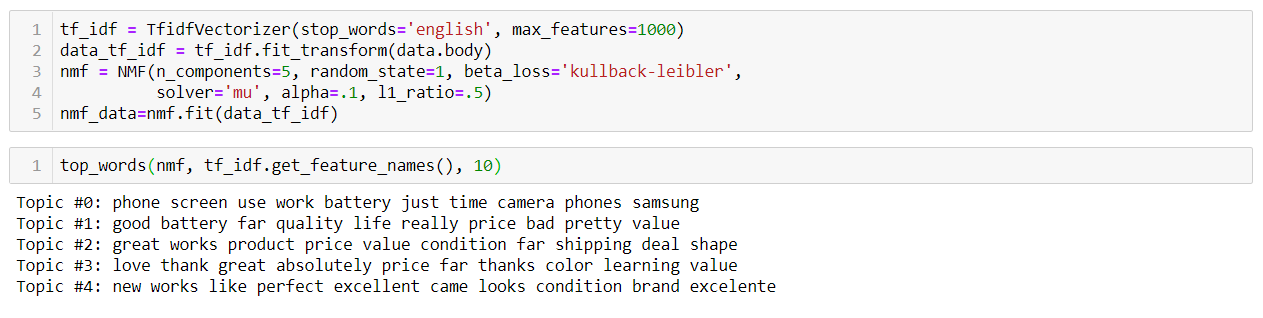
# Coherence of the Model:

# 

**LSA:**



**PLSA:**



**LDA:**

