PRODUCT REVIEW SUMMARIZATION

1. INTRODUCTION:

Talking about the past decade, online shopping market has reached new heights. The e-commerce industry has grown at a rapid pace in this period and now days it's an essential part of human lives all over the world. Online shopping now has a vast variety of products which are being purchased on a simple web application hassle free. The continuous improvement in the whole process by use of growing technology has brought a wide range of services to give the customer real time shopping experience. One of these services include customer reviews which are found to be very helpful in providing a full-fledged information of the product. Not only the images, prices, specifications but the reviews also are a part of a product catalogue now days. In this project we will focus on improving review readability for the customer using Natural Language Processing Techniques.

2. PROBLEM STATEMENT:

In this project we have to build a system which summarizes the customer reviews of a particular product into a bunch of keywords, so that when a customer goes to a product page, he/she doesn't have to read long reviews. Instead he/she can make up his/her mind based on the product average rating and summarized keywords of the review.

For this problem you may use any tools and techniques you like. The data consists of reviews and ratings information of the products which are being sold by the client via online website. The data description is as follows:

DATA DESCRIPTION:

You are given a file named **"Cell_Phones_and_Accessories.json".** This file contains review information under following columns:

- **IC** – Item Code of the product, e.g. B016MF3P3K
- **Reviewer_Name** - Name of the reviewer
- **Useful**- Number of useful votes (upvotes) of the review
- **Prod_meta**- a dictionary of the product metadata. It contains only additional information about the product, if any available.
- **Review**- text of the review
- **Rating**- rating given to the product by the reviewer.
- **Rev_summ**- summary of the review
- **Review_timestamp**- time when the review has been posted (unix time format)
- **Review_Date**- Date when the review has been posted
- **Prod_img**- images that users post after they have received the product
- **Rev_verify**- Flag to represent whether the review has been verified or not. (True/False)

Now, since you have understood the features present in the dataset, you have to do a proper data cleaning for the same. You may remove all the rows where no review is present. You may choose any column(s) to perform this task. You may perform EDA, feature engineering if you are able to find any important new feature.

Once you have done data pre-processing for all the products, you have to predict the important words which summarize the reviews for each product and thus return those words. Number of words extracted for each topic depends on your understanding, you need to give a suitable reason for the number you choose. The summary keywords should not contain more than 30 words.

OUTPUT:

The output should be a dataframe containing following information:

- The Item Code: Must be unique. There should not be any duplicates.
- The maximum rating given by the users to the product.
- The average rating given by the users to the product.
- The minimum rating given by the user to the product.

- Review summary keywords extracted for each product.
- You may include any other relevant information you may feel like.

Your code should be well commented and also mention the brief reasoning behind all the steps and assumptions you have made in order to arrive to your solution. Save the resultant dataframe to json file. Also save the trained model used for this task.

SUBMISSION:

Submit your jupyter notebook or python file, resultant dataframe and saved model in a zipped file.