**Problem Statement**

The e-commerce business is quite popular today. Here, you do not need to take orders by going to each customer. A company launches its website to sell the items to the end consumer, and customers can order the products that they require from the same website. Famous examples of such e-commerce companies are Amazon, Flipkart, Myntra, Paytm and Snapdeal.

Suppose you are working as a Machine Learning Engineer in an e-commerce company named 'Ebuss'. Ebuss has captured a huge market share in many fields, and it sells the products in various categories such as household essentials, books, personal care products, medicines, cosmetic items, beauty products, electrical appliances, kitchen and dining products and health care products.

With the advancement in technology, it is imperative for Ebuss to grow quickly in the e-commerce market to become a major leader in the market because it has to compete with the likes of Amazon, Flipkart, etc., which are already market leaders.

As a senior ML Engineer, you are asked to build a model that will improve the recommendations given to the users given their past reviews and ratings.

In order to do this, you planned to build a **sentiment-based product recommendation system,**which includes the following tasks.

1. Data sourcing and sentiment analysis
2. Building a recommendation system
3. Improving the recommendations using the sentiment analysis model
4. Deploying the end-to-end project with a user interface

* **Data sourcing and sentiment analysis**

In this task, you have to analyse product reviews after some text preprocessing steps and build an ML model to get the sentiments corresponding to the users' reviews and ratings for multiple products.

The dataset that you are going to use is inspired by this [Kaggle competition](https://www.kaggle.com/datafiniti/grammar-and-online-product-reviews). We have made a subset of the original dataset, which has been provided below.

**Product Reviews Dataset**

**Download**

This dataset consists of 30,000 reviews for more than 200 different products. The reviews and ratings are given by more than 20,000 users. Please refer to the following attribute description file to get the details about the columns of the Review Dataset.

**Product Reviews Dataset- Attribute Description**

**Download**

The steps to be performed for the first task are given below.

1. **Exploratory data analysis**
2. **Data cleaning**
3. **Text preprocessing**
4. **Feature extraction:** In order to extract features from the text data, you may choose from any of the methods, including bag-of-words, TF-IDF vectorization or word embedding.
5. **Training a text classification model:** You need to build at least three ML models. You then need to analyse the performance of each of these models and choose the best model. At least three out of the following four models need to be built (Do not forget, if required, handle the class imbalance and perform hyperparameter tuning.).   
   1. Logistic regression  
   2. Random forest  
   3. XGBoost  
   4. Naive Bayes

Out of these four models, you need to select one classification model based on its performance.

* **Building a recommendation system**

As you learnt earlier, you can use the following types of recommendation systems.

1. User-based recommendation system

2. Item-based recommendation system

Your task is to analyse the recommendation systems and select the one that is best suited in this case.

Once you get the best-suited recommendation system, the next task is to recommend 20 products that a user is most likely to purchase based on the ratings. You can use the 'reviews\_username' (one of the columns in the dataset) to identify your user.

* **Improving the recommendations using the sentiment analysis model**

Now, the next task is to link this recommendation system with the sentiment analysis model that was built earlier (recall that we asked you to select one ML model out of the four options). Once you recommend 20 products to a particular user using the recommendation engine, you need to filter out the 5 best products based on the sentiments of the 20 recommended product reviews.

In this way, you will get an**ML model** (for sentiments) and the **best-suited recommendation system**. Next, you need to deploy the entire project publically.

* **Deployment of this end to end project with a user interface**

Once you get the ML model and the best-suited recommendation system, you will deploy the end-to-end project. You need to use the **Flask**framework, which is majorly used to create web applications to deploy machine learning models.

To make the web application public, you need to use **Heroku**, which works as the platform as a service (PaaS) that helps developers build, run and operate applications entirely on the cloud.

Next, you need to include the following features in the user interface.

1. Take any of the existing usernames as input.
2. Create a submit button to submit the username.
3. Once you press the submit button, it should recommend 5 products based on the entered username.