1. **Product opinion mining application:**

   We generally read product reviews before we decide to buy a product online, but it is practically difficult to go through all the reviews and conclude as the number of reviews is way high and also some reviews are either spam or duplicates which a person cannot easily identify. It will be very helpful for the buyers if they get a mined customers' opinion on the features of the product they are looking for.

   Application: Building an opinion mining application for the products based on the product reviews by identifying the polarity of the opinions, removing the biased reviews, spam reviews, duplicate reviews and provide a summarized review using the machine learning and natural language processing techniques.

   Data: Amazon Customer Reviews dataset from s3.amazonaws.com has millions of reviews. For this project, Amazon US reviews for wireless product categories will be used. Dataset size: ~6GB and File Format: tsv.

2. **Invasive Ductal Carcinoma – Breast Cancer – Auto Classifier:**

   Breast cancer is one of the most caused cancers among women in the world. Breast cancer has many subtypes and Invasive Ductal Carcinoma (IDC) is the most common subtype of all breast cancers subtypes. To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. Common pre-processing steps for automatic aggressiveness grading is to describe the exact regions of IDC inside of a whole mount slide.

   Application: As IDC is the most common subtypes of breast cancer caused in women in the world, it is very important to accurately identify and classify the breast cancer type. And automating this critical classification activity on breast cancer histopathology images can be used to prevent human error and save time.

   Data: Breast cancer histopathology images from Kaggle. The dataset contains 277K patch images of breast cancers in 50 X 50 size in .png format.

3. **Tag Prediction for StackOverflow Questions:**

   Stack Overflow is the largest, most trusted online community for developers to learn, share their programming knowledge, and build their careers. It is something which every programmer uses one way or another. Each month, over 50 million developers come to

Stack Overflow to learn, share their knowledge, and build their careers. It features questions and answers on a wide range of topics in computer programming.

Application: StackOverflow recommends the users to add tags to their questions for easy classification and notifying the subject area experts who already answered similar kind of questions. It also helps to suggest similar questions & answers to the users based on their previous post. In case no tags are given to the questions, StackOverflow needs to predict the correct tags which is very critical for the business as wrong tagging can build mistrust among the users and it cannot efficiently notify the appropriate audience or experts. To handle the accurate tag prediction for the StackOverflow questions based on the question title and body, machine learning algorithm and natural language processing concepts can be used.

Data: Found following two datasets of StackOverflow posts in Kaggle:

- Stack Overflow questions & tags [Size: 2 GB & Format: csv]
- Sample Stack Overflow Q & A [Size 3 GB & Format: csv]