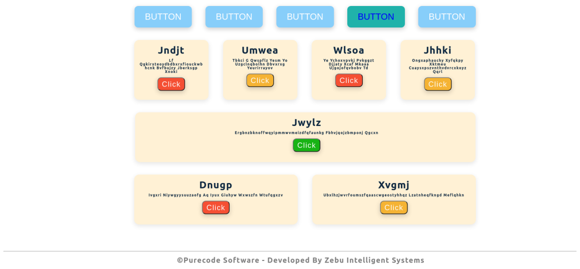
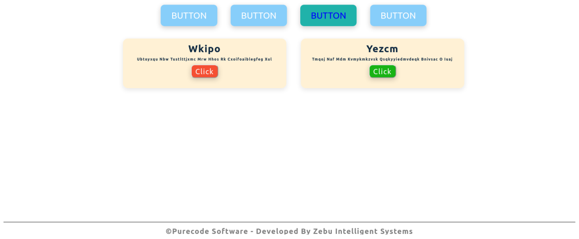
Dataset 1:

This dataset consists of considerably simpler UI, with primarily various types of buttons that are either active or inactive. The following are a couple of samples from the set:



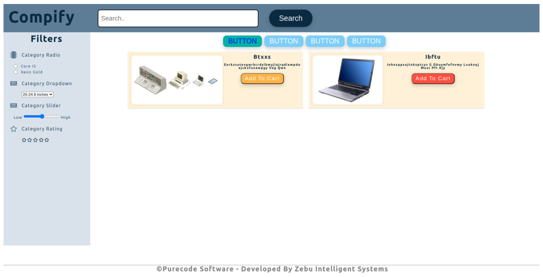


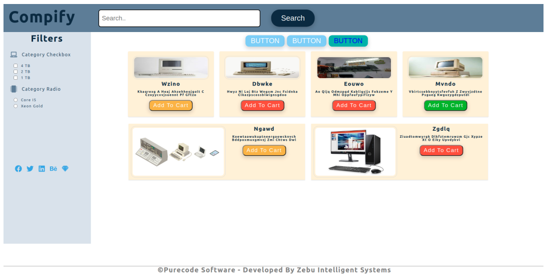
The label statistics in terms of the tokens present in it are mentioned below:

Maximum token length: 87
Minimum token length: 14
Average token length: 50.92
Number of unique tokens: 18

Dataset 2:

The second dataset is significantly more complex than the first one, as it has UI images consisting of not just buttons, but also several other types of components like checkboxes, search-bars, radio-buttons, drop-down menus, star-ratings, etc. Couple of sample images from this dataset are shown below depicting this complexity:



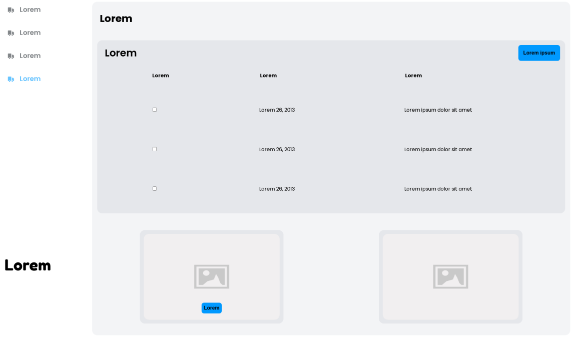


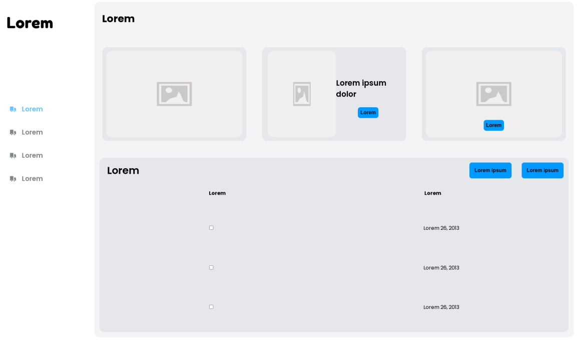
The label statistics in terms of the tokens present in it are mentioned below:

Maximum token length: 73
Minimum token length: 14
Average token length: 44.63
Number of unique tokens: 57

Dataset 3:

The third dataset has UI images with more complex structures to be identified like tables with rows and columns containing button cards. The types of images are somewhat different from the earlier two datasets. Couple of samples from this dataset are shown below:





The label statistics in terms of the tokens present in it are mentioned below:

Maximum token length: 58
Minimum token length: 34
Average token length: 51.72

Number of unique tokens: 55

Overall impression:

Out of the three datasets, the first one seems to be the easiest for a trained model to comprehend, since it consists of simpler structures and also the vocabulary size for an image captioning model is quite small. On the other hand, the other two datasets seem to be much more complicated for a trained model to comprehend due to their much more complex components. Moreover, the number of unique tokens is also 3x more than the first dataset which will introduce even more complexity in the decoding part of the model.