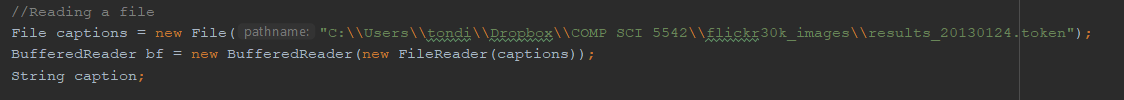
**Lab Report 1**

**Objective**

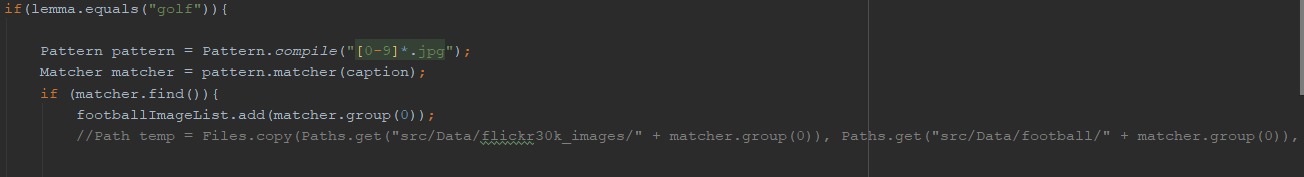
For this lab report we used the Flickr30 dataset provided by Mayanka in the documentation. The category we choose was Sports, which meant we had to scrub the Flickr30 dataset for pictures which made sense to our category. Then perform NLP tokenization and SIFT image feature extraction on the remaining dataset.

**NLP Features**

We went through the captions and searched for the keywords using the Stanford Core NLP tokenization, the words I choose where: baseball, golf and basketball. We modified the code so that it would read the captions file into the main class:



We also had to modify the code so that it would push our images according to their caption ID into tour desired categories for image extraction:



With this modified code we came up with the following image statistics:

**https://drive.google.com/file/d/1owxQKRWmcFSHZvWC9-UVy5zLtE9DoAw3/view?usp=sharing**

**SIFT Image feature extraction**

Using the SIFT extraction features provided by Mayanka, I did SIFT extraction on the baseball dataset with one of the training images. The results of which can be found below:





As you can see from the SIFT feature extraction for baseball, the feature extraction focuses on the persons clothing and posture, a little bit of attention is focused on the hard hat and ball, but unfortunately it looks like the bat is almost completely ignored.

The link to the github repo can be found here, in it you will find all the relevant code as well as the output of image statistics:

https://github.com/TondiToday/CS5542/tree/master/Lab%201