Justin Oakley and Micah Roberts

CSC 526-01

Prof. Manda

April 8th, 2019

CSC 526 Assignment 5 – Hypothesis Testing Report

The focus of this assignment was to perform hypothesis testing utilizing a dataset that contains both semantic similarity scores of machine performances and a Gold Standard dataset that represents the “correct” Gene Ontology annotations, and scores of human performances and the same Gold Standard data. It should be noted that in this semantic similarity dataset, only two types of metrics were used on machine and human performances, Jaccard and Resnik metrics, which means that there were only four columns utilized in the testing phase. The purpose of going through the testing process was to formulate hypotheses around a comparison of the machine performance and human performance, then conducting statistical tests to determine which of the hypotheses was correct and which should be rejected.

The first step taken in the hypothesis testing was to formulate two hypotheses: the null hypothesis and the alternative hypothesis. For the null hypothesis, it was theorized that the mean of the similarity between the machine annotation output and the Gold Standard output would be equal to the mean of human annotation output and the Gold Standard output. The alternative hypothesis was that there would be a noticeable difference between both semantic similarity score means.

Now that the hypotheses had been formulated, the second step was to select an appropriate statistical testing method and use its results to prove which hypothesis was correct. Since there were only two types of semantic similarity metrics and two types of performances conducted on each letter to form the dataset, one statistical test needed to be conducted on a sample of the Jaccard scores of machine and human performances while the other test was to be focused on a sample of the Resnik scores of both performances. With this information in mind, the paired t-testing method was selected for the statistical tests and the samples used consisted, individually, of fifty random observations.

Using the t-testing function built for paired samples from SciPy’s statistical module, fifty observations of Jaccard scores of machines and humans were computed in one test while the same observations’ Resnik scores of machines and humans were conducted in the other. The results of both Jaccard similarity and Resnik similarity t-tests showed that the t-statistical scores were somewhat high, and the probability values were lower than 0.05. This means is that the semantic similarity scores of machine and human performance are different and the results were not obtained by chance. This also means that the null hypothesis should be rejected because of the low *p­*-values and the alternative hypothesis should be accepted. In other words, the testing of hypotheses proves that there is a noticeable difference between the mean of similarity scores of machine performances and the mean of similarity scores of human performances.