Linyi Chen
11791 HW4 Report

Design Choices

1. Code Architecture Design

I followed the general structure provided in the starter code. However, I made some changes regarding the usage of ksTest(self, listA, listB), tTest(self, listA, listB) and wilcoxonTest(self, listA, listB). Originally, they were called in the main function and supposedly will return p-values for all pairs of systems. I modified the usage so that they only return a single p-value. In addition, I added three methods doAllwilcoxonTest(self), doAllksTest(self) and doAlltTest(self) to calculate p-values for these significance tests and they were called in the main function

2. Library Selections

Instead of implementing the significance test myself, I chose to use functions from the scipy library - scipy.stats.ks_2samp, scipy.stats.wilcoxon and scipy.stats.ttest_ind. For calculating basic statistical features, I used numpy for mean and median, scipy.stats for mode and python built-in min and max.

System Performance Analysis

I noticed that in ROUGE_SCORES.csv, ROUGE-2 metrics and ROUGE-SU4 metrics are with exactly the same value for every system. Therefore, the following analysis will not compare same systems on different metrics. Instead, I will compare system performance under the same metrics.

1. Basic Statistics

metric	model	mean	median	mode	min	max
ROUGE-2	Baseline	0.4165538	0.33333	[1.]	0	1
ROUGE-2	Baseline+Fusion	0.3029308	0.218255	[0.]	0	1
ROUGE-2	Baseline+Orderir	0.4124338	0.318015	[0.]	0	1
ROUGE-2	Baseline+Orderir	0.2929442	0.207615	[0.]	0	1
ROUGE-SU4	Baseline	0.4165538	0.33333	[1.]	0	1
ROUGE-SU4	Baseline+Fusion	0.3029308	0.218255	[0.]	0	1
ROUGE-SU4	Baseline+Orderir	0.4124338	0.318015	[0.]	0	1
ROUGE-SU4	Baseline+Orderir	0.2929442	0.207615	[0.]	0	1

I first collected mean, median, mode, min and max for each system. I found that the baseline system have the highest mean and median, and it is the only one with positive mode. At the same time, all systems share the same min = 0 and max = 1. From these basic statistics

above, we can tell that baseline might actually be the best system. And each system have good score on some metrics (all have max = 1) and perform bad on some other metrics (all have min = 0).

2. Significance Test Results

metric	model	mean diff	P(T test)	P(wilcoxon test)	P(ks test)
ROUGE-2	Baseline & Fusion	-0.113623	0.010449189	0.006592269	0.13997518
ROUGE-2	Baseline & Ordering	-0.00412	0.934514037	0.621489949	0.193041652
ROUGE-2	Fusion & Ordering	0.109503	0.01870175	0.005647045	0.069092435
ROUGE-2	Baseline & Ordering+Fusion	-0.1236096	0.005105056	0.004456609	0.069092435
ROUGE-2	Ordering & Ordering+Fusion	-0.1194896	0.009930215	0.000983699	0.031376652
ROUGE-2	Fusion & Ordering+Fusion	-0.0099866	0.798775473	0.753345684	0.999996899
ROUGE-SU4	Baseline & Fusion	-0.113623	0.010449189	0.006592269	0.13997518
ROUGE-SU4	Baseline & Ordering	-0.00412	0.934514037	0.621489949	0.193041652
ROUGE-SU4	Fusion & Ordering	0.109503	0.01870175	0.005647045	0.069092435
ROUGE-SU4	Baseline & Ordering+Fusion	-0.1236096	0.005105056	0.004456609	0.069092435
ROUGE-SU4	Ordering & Ordering+Fusion	-0.1194896	0.009930215	0.000983699	0.031376652
ROUGE-SU4	Fusion & Ordering+Fusion	-0.0099866	0.798775473	0.753345684	0.999996899

In the table above, I highlighted positive mean difference with dark green and negative mean difference with dark red. At the same time, I highlighted p-values <= 0.05 with light green, and those >0.05 with light red.

We can see that among all pairs, only ordering has improved when compared to Fusion. However, the difference is insignificant according to ks test.

Also, there were only two rows with all “green” p-vaules, that is Ordering & Ordering+Fusion. It means that Ordering+Fusion is significantly worse than only Ordering itself. Thus, adding Fusion might might be the reason the system performance is worse. Meanwhile, there are four rows with all “red” p-values - Baseline & Ordering and Fusion & Ordering + Fusion. It means that the negative mean difference is not statistically significant. In other words, adding ordering does not necessarily degrade the system performance.

To sum up, from the significance test results above, we can tell that adding Fusion may make the system performance worse, given that the p-values are all < 0.05. In addition, although the mean differences with ordering added are negative, the p-values are all > 0.05, thus the differences are not significant.