# Summary Report (11/20/2014)

## Overview

In this week, I have done the following things:

* K=4
* Significant Test

## K=4



The last two rows are the results with setting K=4.

However, it is comparable but slightly worse than K=sqrt(V). In fact, for LP, it is better.

Although the median number of entries in the TA’s summary is 4, it doesn’t mean K=4 is the best option. Because, the TA doesn’t cover all the topics. Thus, the number of topic in the students’ answers are usually more than 4.

## Significant Test

Here are the raw values.



I chose the bigram and PhraseMeadLexRank\_syntax as baselines to do the significant test. Although PhraseMeadMMR\_syntax has better ROUGE scores, they are a combination of best performances with different parameters. Thus, it is not chosen.

For the proposed model, I used Clustering\_lexrank\_LSA\_syntax as the model.

### Paired T-Test (p-values)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | cluster vs. bigram | cluster vs. PhraseMead+LexRank |
| LP | R1-R | 0.000 | 0.186 |
| R1-P | 0.110 | 0.026 |
| R1-F | 0.000 | 0.089 |
| R2-R | 0.014 | 0.088 |
| R2-P | 0.394 | 0.078 |
| R2-F | 0.015 | 0.076 |
| RSU4-R | 0.005 | 0.074 |
| RSU4-P | 0.042 | 0.022 |
| RSU4-F | 0.003 | 0.081 |
| MP | R1-R | 0.000 | 0.316 |
| R1-P | 0.082 | 0.046 |
| R1-F | 0.005 | 0.073 |
| R2-R | 0.003 | 0.131 |
| R2-P | 0.188 | 0.061 |
| R2-F | 0.003 | 0.094 |
| RSU4-R | 0.003 | 0.167 |
| RSU4-P | 0.018 | 0.375 |
| RSU4-F | 0.033 | 0.352 |
| LP | R1-R | 0.003 | 0.329 |
| R1-P | 0.480 | 0.306 |
| R1-F | 0.073 | 0.489 |
| R2-R | 0.058 | 0.115 |
| R2-P | 0.047 | 0.131 |
| R2-F | 0.047 | 0.116 |
| RSU4-R | 0.012 | 0.125 |
| RSU4-P | 0.198 | 0.478 |
| RSU4-F | 0.277 | 0.257 |

#### Observations:

* Compared to bigram model, Cluster based method is significant different except for R1 and R-SU4 for LP
* Unfortunately, compared to the PhraseMead+LexRank model, no significant different result is observed. However, the p-values of R1, R2, R-SU4 for LP and R1, R2 for MP are close to 0.05.
* Although the F-Measure scores are not significant different, the recall of R1 for LP and MP and the Recall of R-SU4 showed a significant differences.

### Combined the three questions together.

Since there are only 12 documents for each question, combing them together which has 36 documents which might have a better chance to see significant difference.

Here is the result:



#### Observations:

* This time, we do observe significant different at R1-F for cluster vs. PhraseMead+LexRank
* This table gives some interesting findings
  + The clustering based method have a better Recall, but a lower precision (It makes sense because by clustering, redundant phrases are removed and more phrases are included in the summary)
    - Both the recall and precision is better than bigram
  + The precision is lower than LexRank, but recall is higher and the f-measure is better.

## Weighted Rouge Scores

Not finished yet.