

Report on Task 2

I created a UI where you can upload the two files and an analysis can be run. I first displayed the top 5 items of both the files so you can get a reference. I first created word clouds for the answers of the two files. This is because it is the easiest way to understand if there has been a significant change in the answer. According to our files, the word cloud is almost the same suggesting minuet changes. Then I run a Similarity score distribution. I took the two answers, matched by original id and encoded it with a SentenceTransformer model 'all-MiniLM-L6-v2' which is generally used for clustering or semantic search. I then computed the cosine similarity of these embeddings and found that most have a similarity score of 1 meaning there is no change. The mean similarity score is 0.98 and the median was 1. The standard deviation was of 0.14. Most answers showed high similarity meaning minor changes. Some question may have been completely rewritten or added indicated by low similarity of some answers.

I then categorized these answers into three categories based on their similarity score. I found that most were mostly similar with just a few being completely different and no answers with somewhat similar score.

The thresholds were

If score<0.3: Completely Different

Else If score<0.7: Somewhat Similar

Else If score<1: Very Similar

Then I checked the type of the change. I found that majority were mainly grammar/format changes, which included the non changes as well, then 10 items where both were empty so no change, very few where there was a content change and some which were removed in final.

The threshold was as:

If no answer in both: Both Empty

If answer in v1 and no in final: Removed in Final

If answer not in v1 and in final: Added in Final

If the common words in the answers in v1 and v2 divided by their length is greater than 0.9:
Grammar Change or No Change

Else: Some content has been changed

The statistics summary is as:

Total questions: 119

Average similarity score: 0.98

Change Similarity Distribution:

Similarity_Category

Very Similar 0.97479

Completely Different 0.02521

Name: proportion, dtype: float64

Change Type Distribution:

Change_Type

Mainly Format/Grammar 0.890756

Both Empty 0.084034

Content Change 0.016807

Removed in Final 0.008403

We see majority were very similar and just some format and grammar changes were done.