Similarity analysis by Word TF-IDF List

Fu-yin Cherng, Mrini Khalil

March 2017

1 Constituting Word TF-IDF List for Pairs of Documents

The TF-IDF formula that suggested during the last meeting involved a document frequencies of the current document and the document to which it is compared. We therefore had to issue word frequency lists by pairs of documents. These pairs were: Stack Training Set and Stack Testing Set, Stack Training Set and MOOC DSP, Stack Training Set and MOOC Reactive, MOOC DSP and MOOC Reactive. The pairs containing Stack were issued 5 times as the training and testing sets are randomly split at 80% train and 20% test.

For a pair of documents, we: constituted a word list that is the union of both documents word lists counted the inverse document frequency using (+1) on both sides to smooth the results and avoiding divisions by zero computed the tf.idf weights per word, and normalized the weights such that the sum of all tf.idf weights for a document is equal to 1

The TF-IDF formula for word w before normalizing was as follows:

$$TF_{current-doc}(w) \times \log \frac{1}{DF_{current-doc}(w)DF_{comparison-doc}(w)}$$

2 Comparing the Divergence between Documents

By applying Cross Entropy, JS divergent, KL divergent, L1 and L2 distance, we obtained the value of divergence between the aforementioned pairs of documents. Here are the terms of the result value based on their corresponding pairs of documents.

- train_test = divergence(Stack Training Set, Stack Test Set)
- train_dsp = divergence(Stack Training Set, MOOC DSP)
- train_reactive = divergence(Stack Training Set, MOOC Reactive)
- dsp_reactive = divergence(MOOC DSP, MOOC Reactive)

Figure 1 to 5 illustrate the results for each method. We can see that the divergence value of train_dsp is closer to value of train_reactive in the result of Cross Entropy, JS divergence, KL divergence and L1 distance. The only exception is L2 distance (Figure 5). The Δ between train_dsp and train_reactive is larger than the Δ between train_dsp and train_test.

To sum up, our result suggests that the document of MOOC DSP is not homogeneous with the document of Stack Exchange. Hence, applying the classifiers built by document of Stack Exchange to data of MOOC may not be appropriate and feasible.

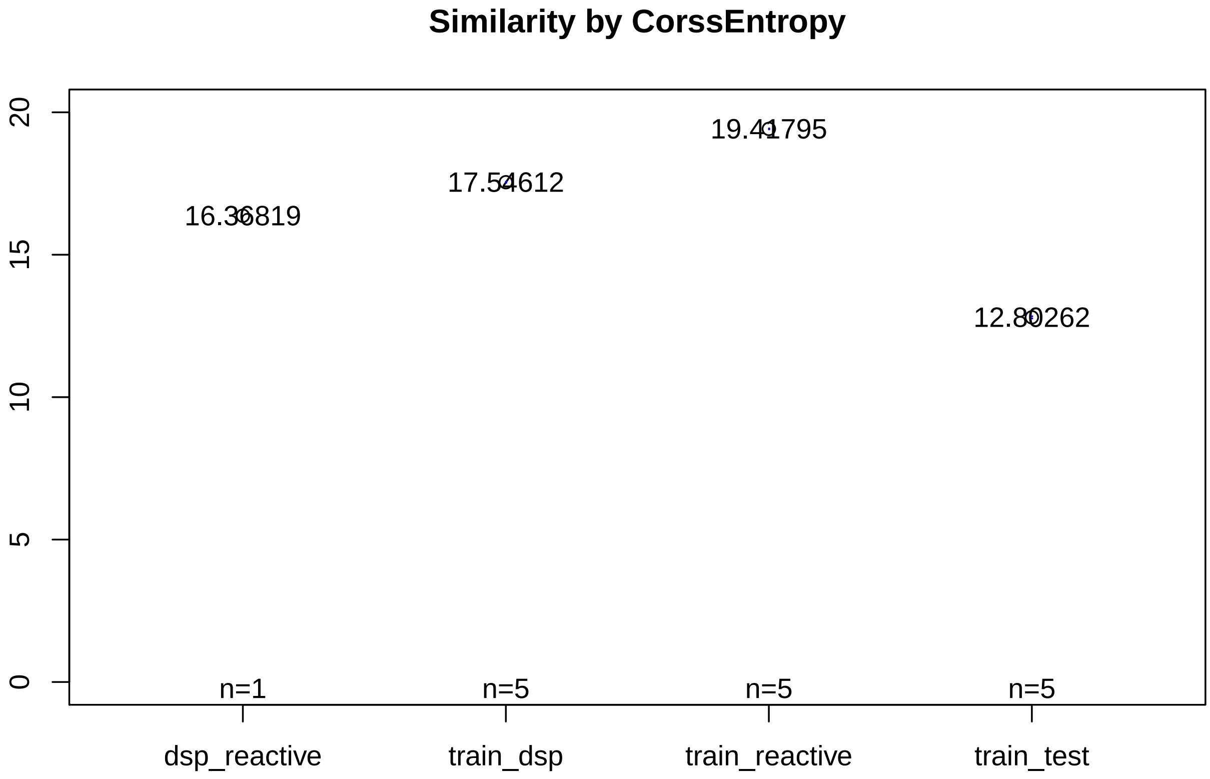

Figure 1: The result of cross entropy.

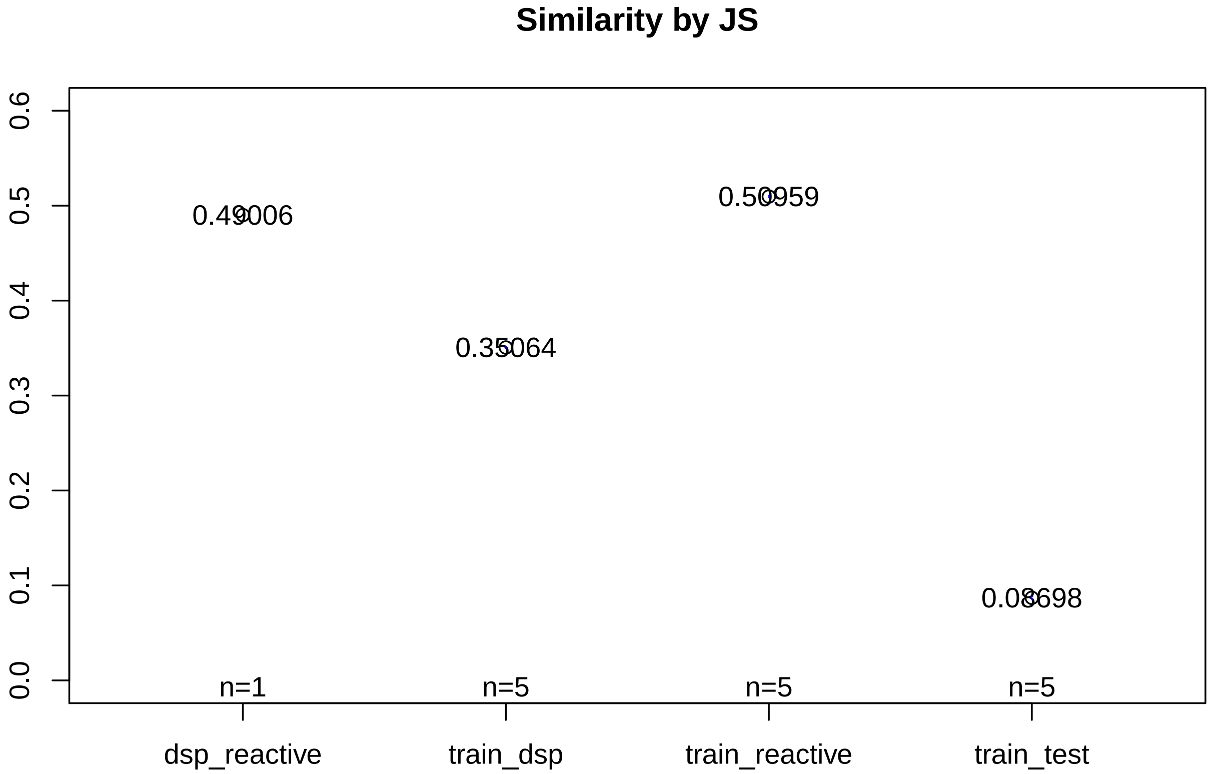

Figure 2: The result of JS divergence.

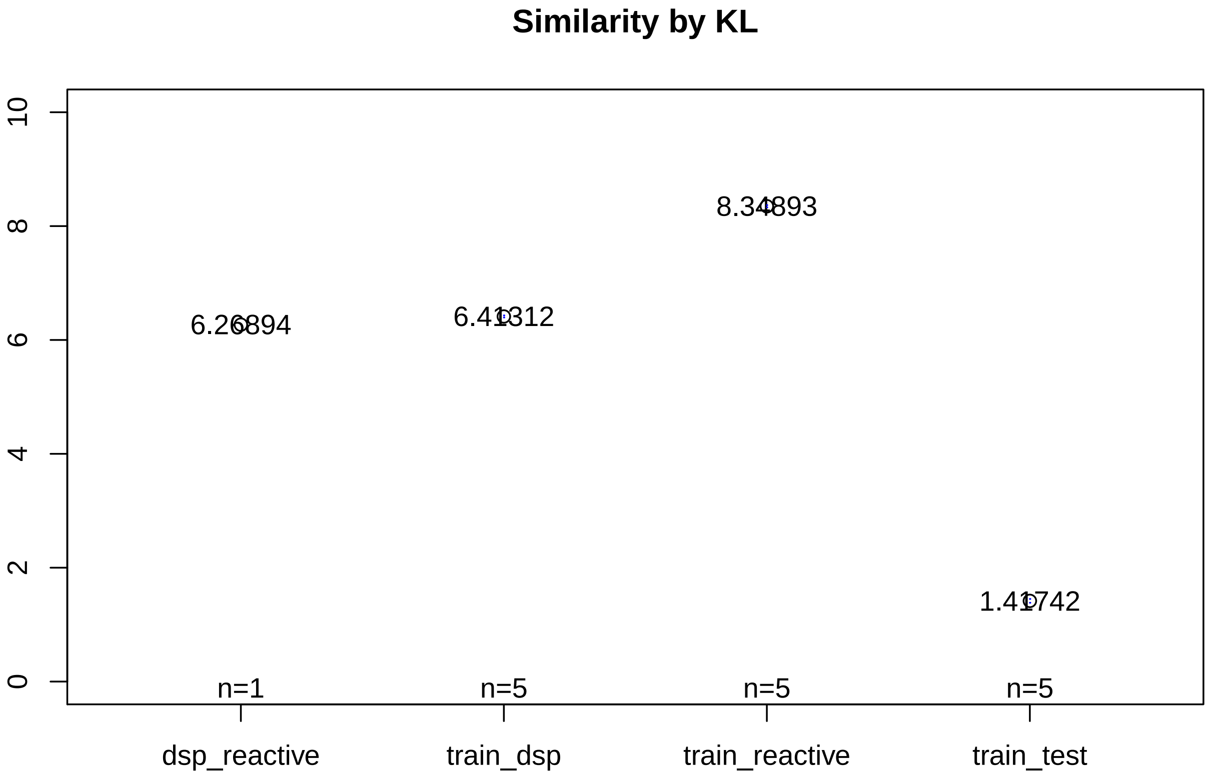

Figure 3: The result of KL divergence

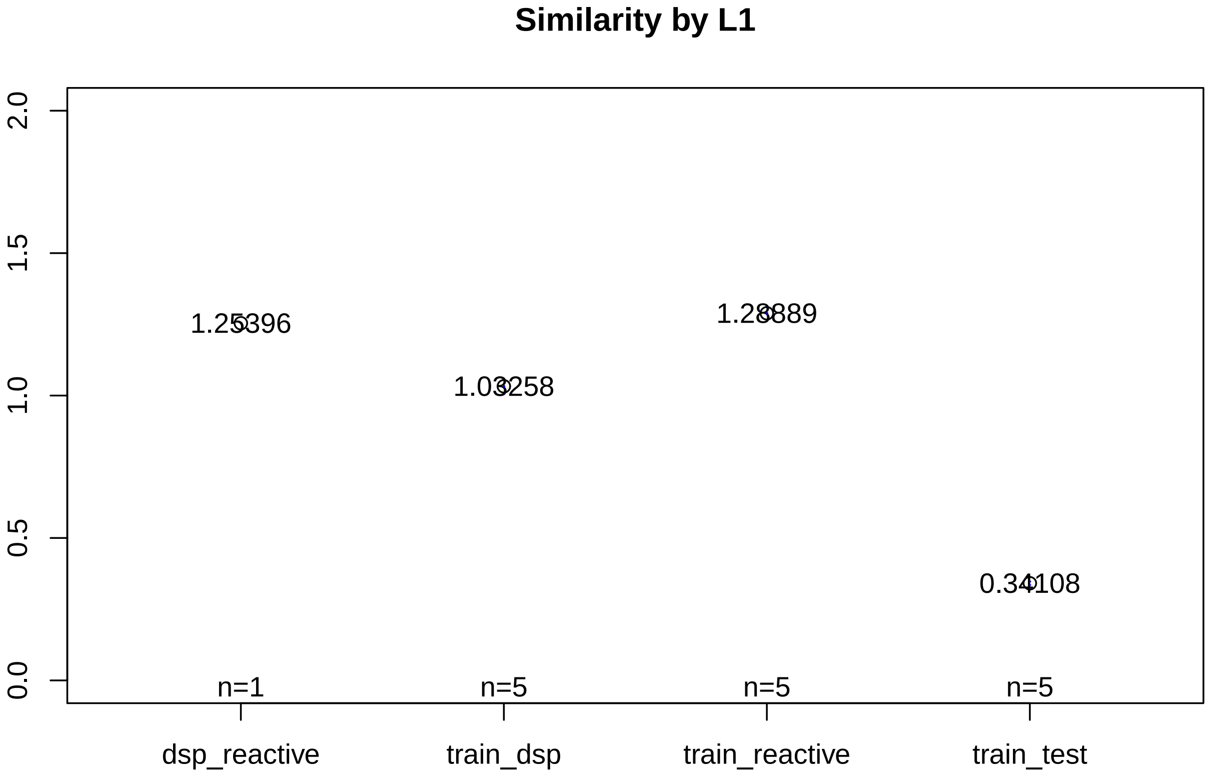

Figure 4: The result of L1 distance.

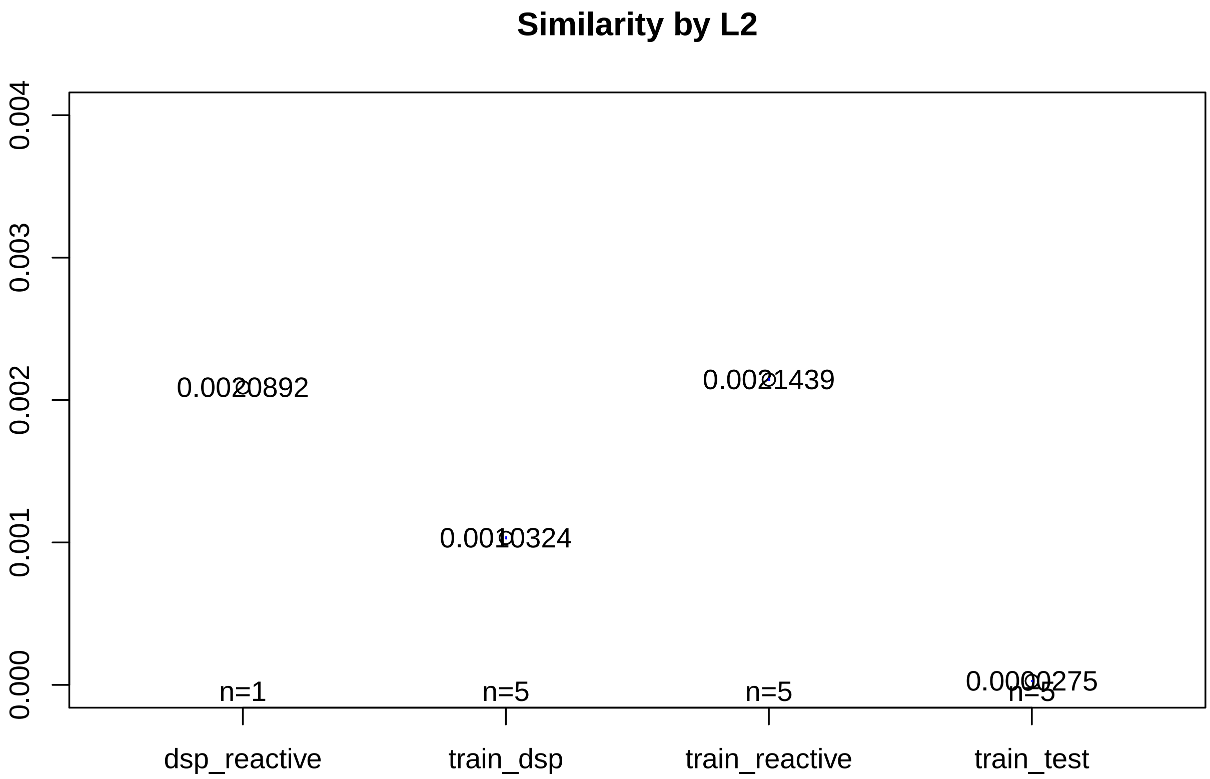

Figure 5: The result of L2 distance. Δ of train_dsp to train_reactive is 0.0011114. Δ of train_dsp to train_test is 0.0010049.