- The 2 notebooks have the updated code which i will describe in a bit.
- The 2 text files have more keywords for drought, earthquake, flood and disaster which we have extracted from the chinese room editor and used to increase our precision and improve our modelling.

Updates:

- We have added a performance metric to our modelling. The code generates the confusion matrix for the models. This is used to calculate the precision, recall and thereafter the F score.

- This performance measure was used to calculate the F score for Oromo with the following seeded words: **Lines 30-37**

```
In [130]: # Anchor words
          #["chocho'a", "lafaa","socho'a","lafa"],

          #sochoʹʹaa sochoʹʹan sochoʹʹuun  sochoʹʹuu
          anchor_words = [["hoongee","hongee"],["galaana","galaanaa"],["balaa"]]

          anchored_topic_model = Corex(n_hidden=25, max_iter=1500,seed=3192)
          anchored_topic_model.fit(doc_word_mat, words=words, anchors=anchor_words, anchor_strength=6);
```

and it returned the following:

```
           Predicted++  Predicted--
Actual++           161         1645
Actual--            20          144
precision 0.8895027624309392
recall 0.08914728682170543
F1_score 0.16205334675390037
```

- Now we used the words extracted from the chinese room editor and added them to the seeded words as follow: **Lines 38-49**

```
In [141]: # Anchor words
          #["chocho'a", "lafaa","socho'a","lafa"],

          #sochoʹʹaa sochoʹʹan sochoʹʹuun  sochoʹʹuu
          anchor_words = [["hoongee","hongee","oolaa"],["galaana","galaanaa","guutuu","lolaa"],
                          ["balaa","dhibee","irraa"],
                          ["sochii","lafaa"]]

          anchored_topic_model = Corex(n_hidden=25, max_iter=1500,seed=3192)
          anchored_topic_model.fit(doc_word_mat, words=words, anchors=anchor_words, anchor_strength=6);
```

and it returned an improved precision and recall score of:

```
After adding more words from chinease room translator

          Predicted++  Predicted--
Actual++          248         1558
Actual--           27          137

precision 0.9018181818181819

recall 0.13732004429678848

F1_score 0.2383469485824123
```

- This shows that once we retrieve the topic model with limited seeded keywords and then follow it up by searching for more keywords we are bound to get a better score. A more precise model.
- Following this, **Lines 50 to 54** have the Top 10/20 documents extracted according to the Occurrence Count of the seeded word for a given topic. This can also be used to extract top 20-25 documents for the same but the count reduces to 1 after the top 10 documents.

Issues:

- We were planning to run the same thing for Tigrinya but the chinese room editor returned the romanized version of the Tigrinya text. Therefore we tried to trans-literate it online but we were not able to get the correct text.
- We are meeting Ulf on Friday to correct this issue and also understand more about the CRE so that we can try and incorporate some of the methods automatically in the system.

Looking Forward:

- Given that we have a performance measure : precision, recall, f score, top 10 occurrence, specificity and sensitivity. We are going to expand this to Guided LDA seeding and compare the results.
- We are also looking forward to creating our own topic modelling system and compare the scores for the same.

Please review the document along with the notebook and give us your feedback on the same.