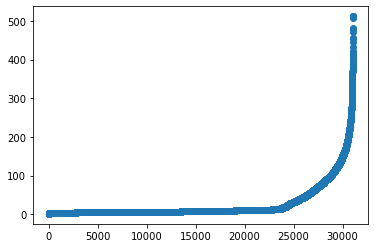
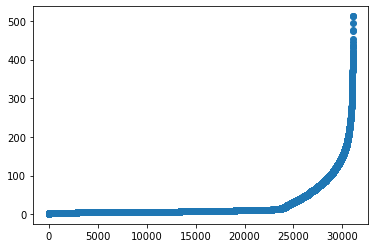
**25/05:**

About the sentence length: There are 3 sentences in each fold that were longer than 512 characters. I truncated them. I drew some crude plots with the sentence lengths to discuss if truncating any further was worth it. This is for fold 1:



And this for fold 2:



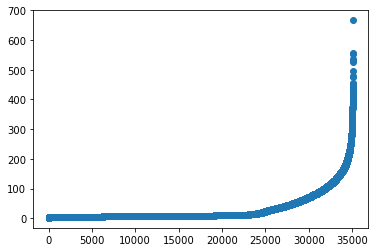
These tables show how many samples are lost for different length thresholds, in the two folds

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **400** | **350** | **300** | **250** | **225** | **200** | **175** | **150** | **125** | **100** |
| **Fold1** | 24 | 54 | 98 | 216 | 319 | 480 | 694 | 1052 | 1537 | 2246 |
| **Fold2** | 25 | 57 | 108 | 221 | 338 | 478 | 709 | 1064 | 1616 | 2326 |

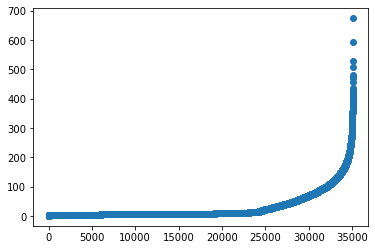
**14/06:**

I added the IEMOCAP datset, with 8068 extra rows, and re-did the same calculations as before to inform my sentence truncation threshold decisions.

Fold1:



Fold2:



How much we lose:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **400** | **350** | **300** | **250** | **225** | **200** | **175** | **150** | **125** | **100** |
| **Fold1** | 32 | 71 | 134 | 266 | 405 | 604 | 909 | 1438 | 2169 | 3235 |
| **Fold2** | 24 | 49 | 97 | 240 | 352 | 558 | 848 | 1324 | 2050 | 3082 |