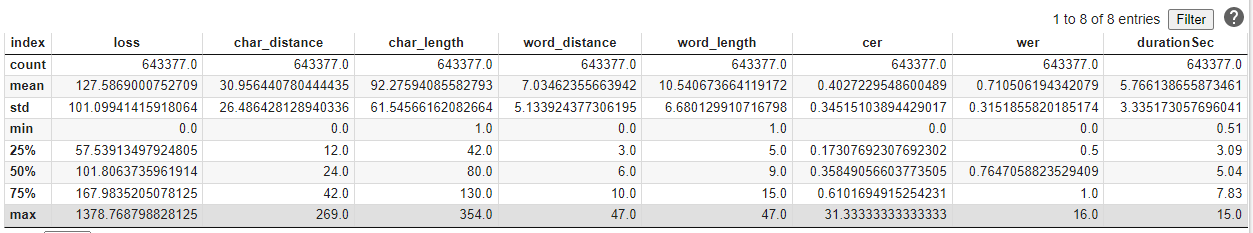
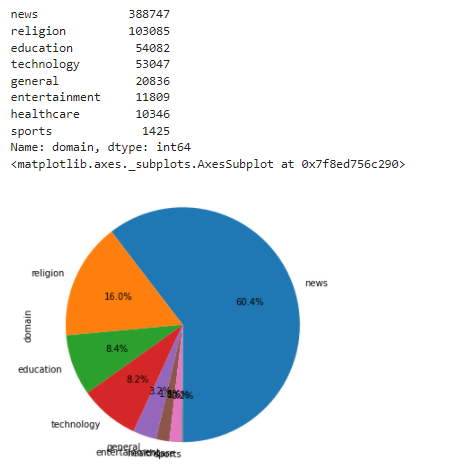
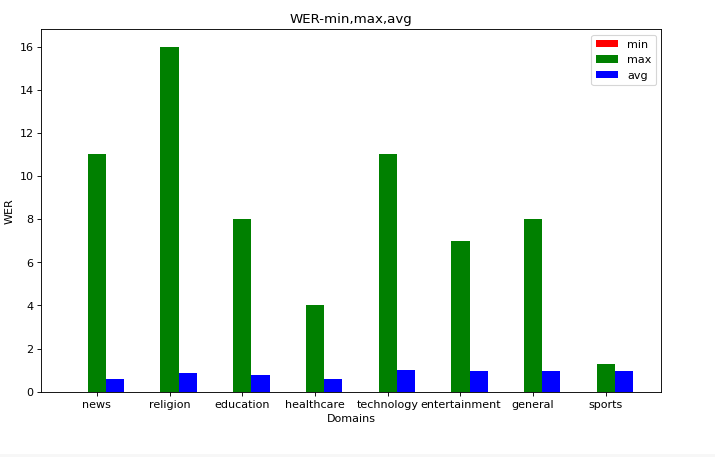
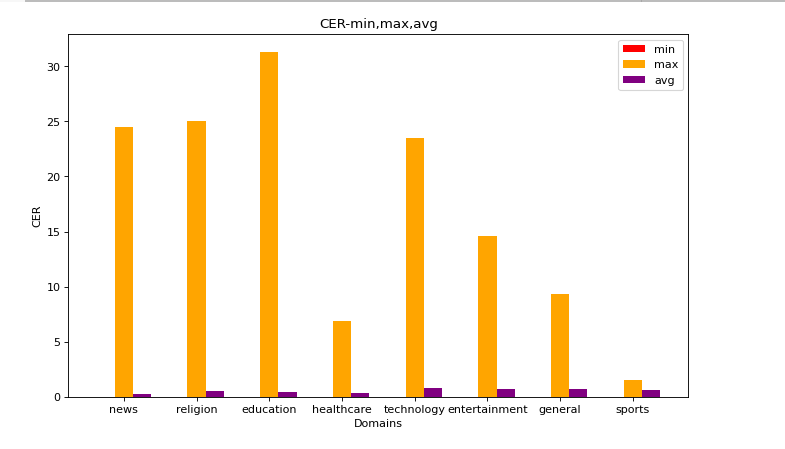
**Inferences:**

* + 1. Ulca test analysis overview:
    2. Ulca dataset consists majorly news.



* + 1. WER analysis (comparison across all domains)
    2. CER analysis (comparison across all domains)



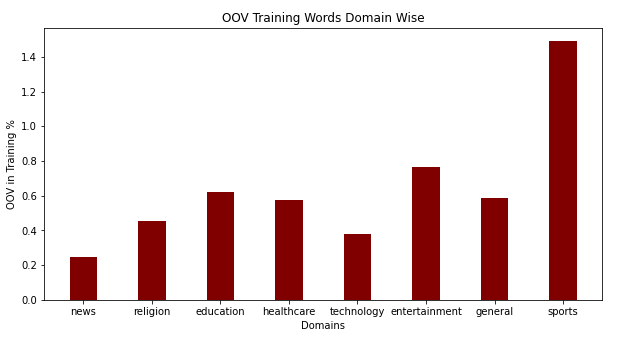
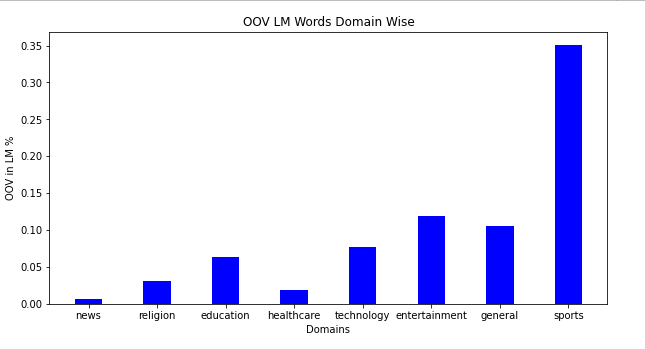
* + 1. Words in Training dataset (AM data) consists news related words in majority   
         
       {'news': 37838, 'religion': 28975, 'education': 24312, 'general': 9489, 'technology': 11348, 'healthcare': 6680, 'sports': 1759, 'entertainment': 7979}
    2. LM Data consists majorly news related words.

{'news': 133974, 'religion': 75663, 'education': 57863, 'technology': 31541, 'general': 21714, 'entertainment': 16991, 'sports': 3880, 'healthcare': 12617}

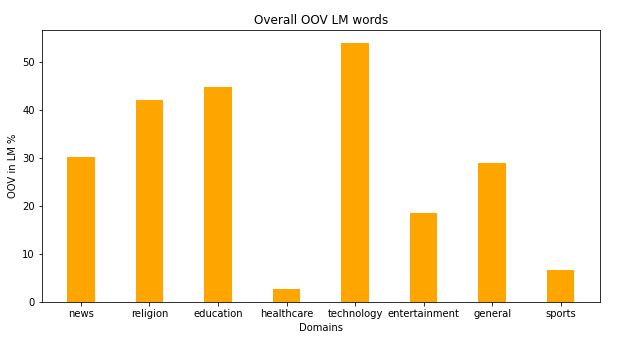
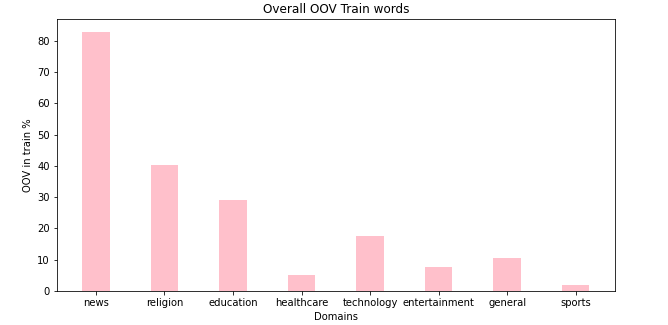
* + 1. Total OOV Words not known to LM is 76.

**Total OOV words unknown to both LM and AM:**

|  |  |
| --- | --- |
| OOV Word | Count |
| எட்நூத்தி | 755 |
| ஒன்பதாயிரத்தி | 394 |
| கொஷ்டீன் | 390 |
| ஆறாயிரத்தி | 353 |
| டுவெண்ட்டி | 289 |
| டுவல்வ் | 275 |
| அறநூத்தி | 187 |
| அப்டின்றான் | 179 |
| ஓகேங்களா | 120 |
| அப்டின்ட்டு | 100 |
| பெர்ஸன்ட் | 100 |
| மூனாயிரத்தி | 87 |
| இரநூத்தி | 41 |
| பதிமூன் | 24 |
| பண்ட்டு | 24 |
| அப்டின்னுச்சு | 15 |
| ஐநூறுவா | 12 |
| அப்டின்டு | 11 |
| எப்பியுமே | 10 |
| ஃபுல்லாம் | 9 |
| ஓகேய்வா | 9 |
| பண்டேன் | 8 |
| நுப்பத்தி | 8 |
| அப்டீன்னேன் | 7 |
| அட்வைசிங் | 6 |
| இதுக்காண்டி | 5 |
| பர்பஸ்க்காக | 5 |
| திருத்தோ் | 5 |
| ஆயரூபா | 4 |
| ஃபர்தரா | 4 |
| இஷட்டு | 3 |
| தொளசண்ட் | 3 |
| அப்டீன்னே | 3 |
| கேக்குதுங்களா | 3 |
| பார்க்கவேண்டியிருக்க | 2 |
| ரெக்கொயர்மெண்ட் | 2 |
| செரிறி | 2 |
| தொகுத்திருக்கு | 2 |
| செலவழித்திருக்கிறே | 2 |
| அதலெல்லாம் | 2 |
| திருப்புல்லா | 1 |
| பால்பண் | 1 |
| வாங்கப்பட்டிருக் | 1 |
| கவலைப்படுவீர்க | 1 |
| உன்டே | 1 |
| திருவாங்க | 1 |
| உக்காந்துக்கிட் | 1 |
| தடுப்புச்சு | 1 |
| மொழிப்பு | 1 |
| நினைச்சுகிட் | 1 |
| குறிசொல்லு | 1 |
| கண்டறியப்பட்டிருக் | 1 |
| கதைத்திருக்கிற | 1 |
| தமிழ்சின | 1 |
| சட்டப்போர | 1 |

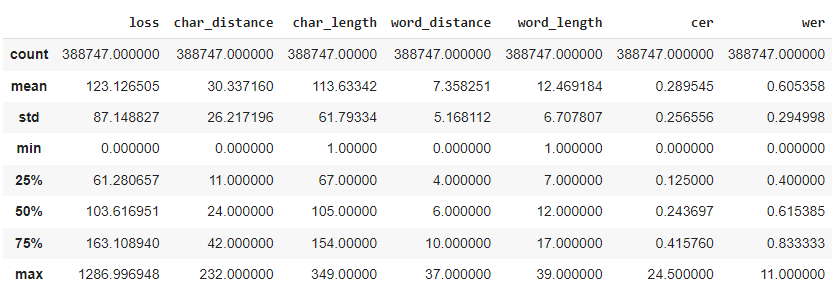
* + 1. Total OOV Words not known to AM is 116107
    2. Most common OOV words wrt LM across all domains: {'பெர்ஸன்ட்', 'அப்டின்றான்', 'அப்டின்ட்டு', 'டுவெண்ட்டி', 'டுவல்வ்', 'பண்ட்டு'}
    3. Numbers (in Tamil) constitute a major part of OOV words across all domains
    4. OOV words for both AM and LM is majorly from sports followed by entertainment. News has least OOV words.

(OOV words in each domain/Total words in that domain)\*100

(OOV words in each domain/Total OOV Words) \*100

* + 1. wav file duration and WER are negatively correlated: (value: -0.1730614399471933)

|  |  |  |
| --- | --- | --- |
| Sl. No. | Data Source | Observation |
| 1. | OOV for Religion wrt LM | 1. 8/32 are numbers. 2. 7/32 are Tanglish words. 3. 17/32 are modern tamil words. 4. WER:    1. Min-0.00    2. Max-16.00    3. Avg-0.856 5. CER: 6. Min-0.00 7. Max-25.0 8. Avg-0.524 |
| 2. | OOV for News wrt LM | 1. 9/23 are numbers 2. 6/23 are Tanglish words (accented) 3. 8/23 are modern tamil 4. WER:    1. Min-0.00    2. Max-11.00    3. Avg-0.605 5. CER: 6. Min-0.00 7. Max-24.5 8. Avg-0.289 |
| 3. | OOV for Education wrt LM | 1. 10/34 are numbers 2. 11/34 are Tanglish words (accented) 3. 13/34 are modern tamil 4. WER: 5. Min-0.00 6. Max-8.00 7. Avg-0.766 8. CER: 9. Min-0.00 10. Max-31.33 11. Avg-0.43 |
| 4. | OOV for Healthcare wrt LM | 1. 2/2 are modern tamil 2. WER:    1. Min-0.00    2. Max-4.00    3. Avg-0.611 3. CER: 4. Min-0.00 5. Max-6.857 6. Avg-0.315 |
| 5. | OOV for technology wrt LM | 1. 9/41 are numbers 2. 12/41 are Tanglish words (accented) 3. 20/41 are modern tamil 4. WER:    1. Min-0.00    2. Max-11.00    3. Avg-1.00 5. CER: 6. Min-0.00 7. Max-23.50 8. Avg-0.798 |
| 6. | OOV for entertainment wrt LM | 1. 1/14 are numbers 2. 4/14 are Tanglish words (accented) 3. 9/14 are modern tamil 4. WER:    1. Min-0.00    2. Max-7.00    3. Avg-0.954 5. CER: 6. Min-0.00 7. Max-14.60 8. Avg-0.694 |
| 7. | OOV for general wrt LM | 1. 6/22 are numbers 2. 6/22 are Tanglish words (accented) 3. 10/22 are modern tamil |
| 8. | OOV for sports wrt LM | 1. 3/5 are Tanglish words (accented) 2. 2/5 are modern tamil |

* + 1. In depth domain analysis:
       1. **News:**

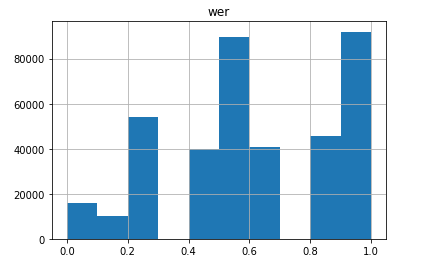
No. of 0-0.2WER records=37304

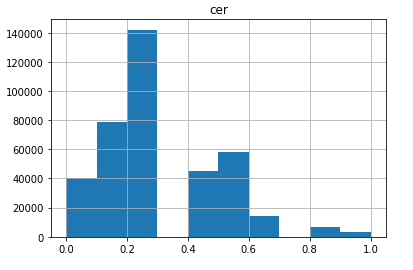
Total Records=388747

%records with 0-0.2 WER =9.595958296784284%

No. of 0-0.2CER records=161764

Total Records=388747

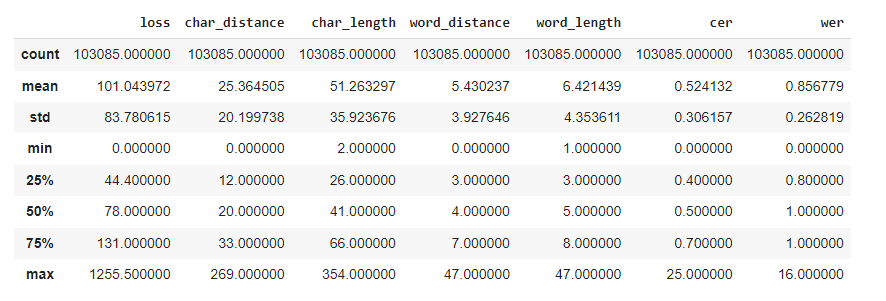
%records with 0-0.2 CER=41.61163944673528%

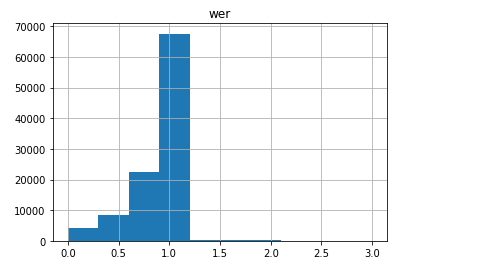
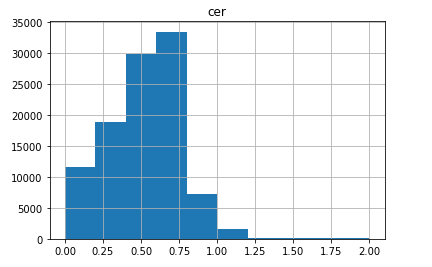


AM data consists mainly of news related words. So, for most of the wav files CER lies between 0 and 0.3. Many files have relatively faster speech rates, so AM might have partially transcribed the syllables. The LM on receiving this input yields the nearest match, which may or may not be the most accurate transcription. This results in a fluctuating WER with majority records in range between 0 and 1.

Total news OOV words in AM data: 96159

Total news OOV words in LM data: 23

* + - 1. **Religion:**



Total Records=103085

No. of 0-0.5 WER records=12115

%records with 0-0.5 WER= 11.75243730901683%

No. of 0-0.25 CER records=16007

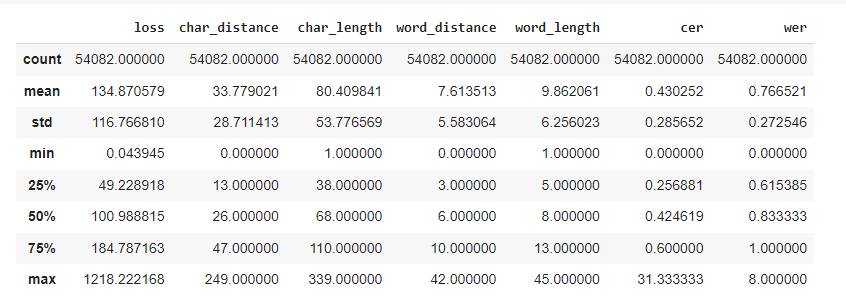
Total Records=103085

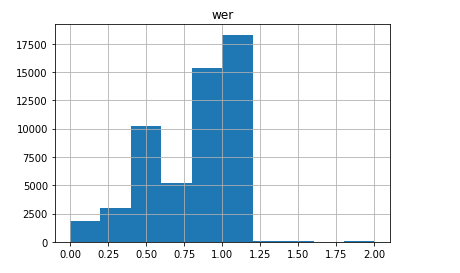
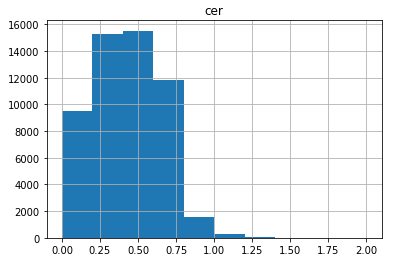
%records with 0-0.25 CER=15.527962361158268%

Total religion OOV words in AM data: 46720

Total religion OOV words in LM data: 32

* + - 1. **Education Analysis:**

****



No. of 0-0.25 WER records=3113

Total Records=54082

%records with 0-0.25 WER=5.756074109685293%

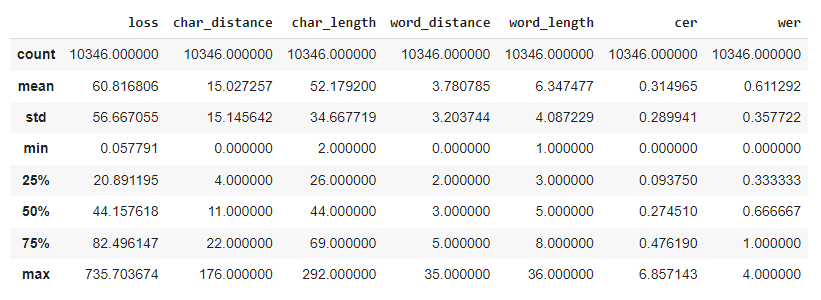
No. of 0-0.25 CER records=13179

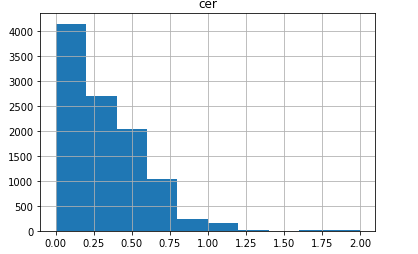
Total Records=54082

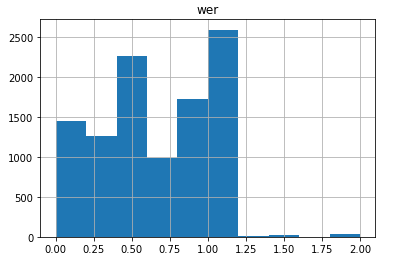
%records with 0-0.25 CER=24.36855145889575%

Total education OOV words in AM data: 33585

Total education OOV words in LM data: 34

* + - 1. **Healthcare**



For 0 CER transcripts, the WAV files are of the right pace as in training set, resulting in correct performance. As the CER increases to 0.2, some noise (like muffled recordings, other background noise, etc.) in the WAV files, which causes the AM to fail. As CER further increases to 0.3, the speech rates are faster, coupled with noisy recordings, making the WER also jump up. Very high CER and WER files are entirely not in Tamil, and both the source model and our model disagree on the transcripts. The number of such files is very less, explaining the decreasing trend in both the graphs.

No. of 0-0.25 WER records=2077

Total Records=10346

%records with 0-0.25 WER=20.075391455635028%

No. of 0-0.25 CER records=4904

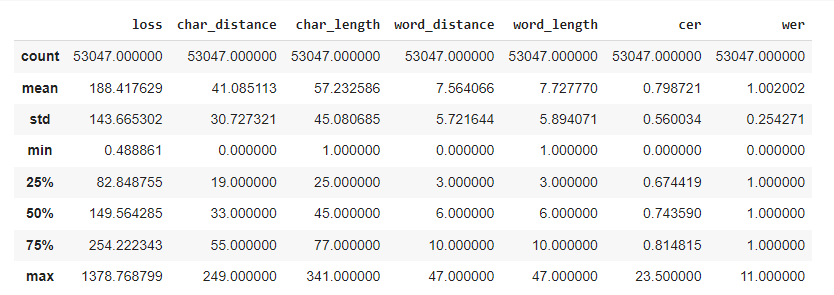
Total Records=10346

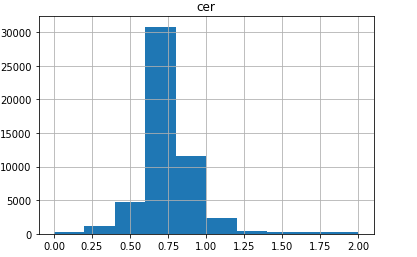
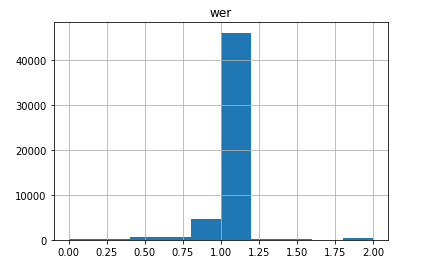
%records with 0-0.25 CER=47.399961337715055%

Total OOV words in AM data: 5939

Total OOV words in LM data: 2

* + - 1. **Technology:**





No. of 0-0.25 WER records=162

Total Records =53047

%records with 0-0.25 WER =0.3053895602013309

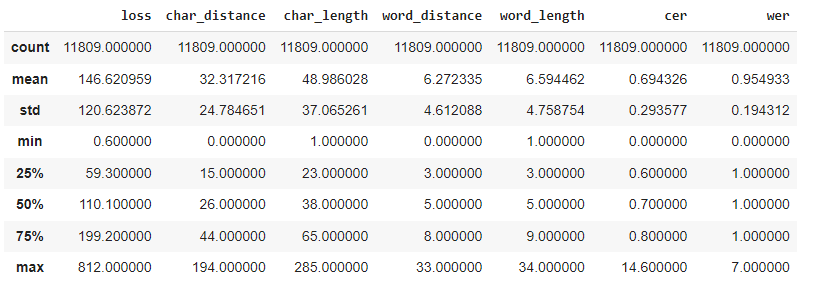
No. of 0-0.25 CER records=535

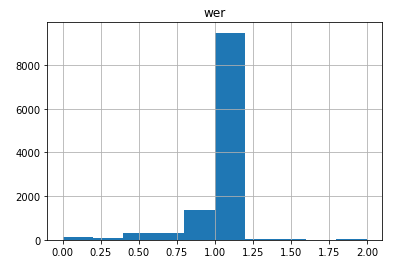
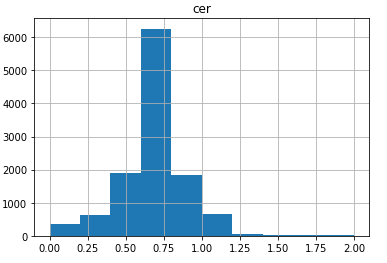
Total Records=53047

%records with 0-0.25 CER=1.0085395969611852%

Total OOV words in AM data: 20234

Total OOV words in LM data: 41

* + - 1. **Entertainment:**



No. of 0-0.25 WER records=171

Total Records=11809

%records with 0-0.25 WER =1.4480480989076128%

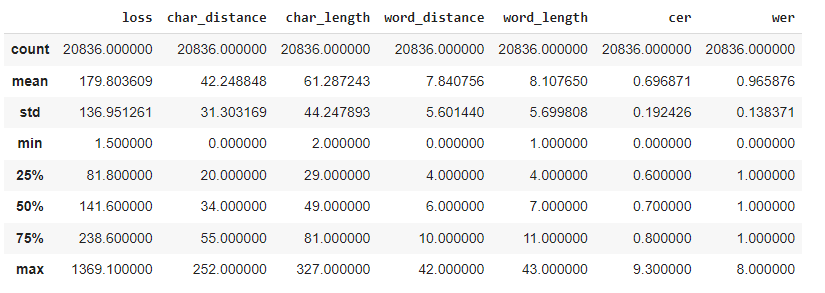
No. of 0-0.25 CER records=472

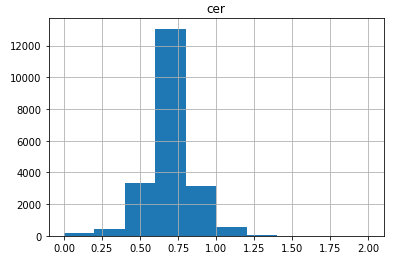
Total Records=11809

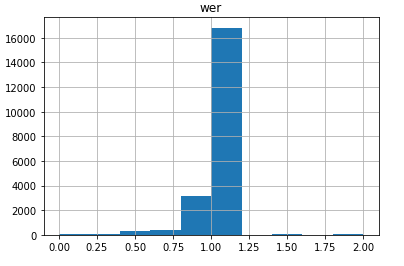
%records with 0-0.25 CER=3.9969514776865105%

Total OOV words in AM data: 9026

Total OOV words in LM data: 14

* + - 1. **General:**





No. of 0-0.25 WER records=97

Total Records=20836

%records with 0-0.25 WER =0.46554041082741404%

No. of 0-0.25 CER records=263

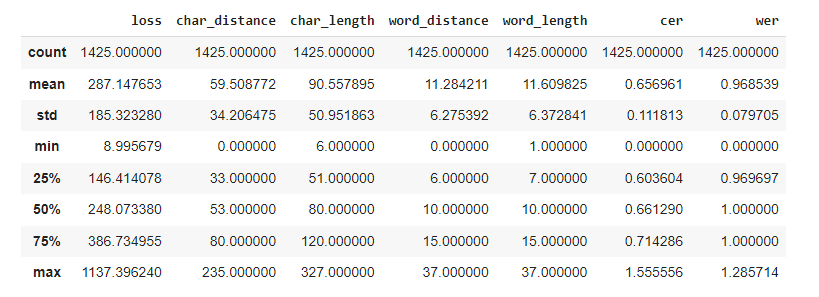
Total Records=20836

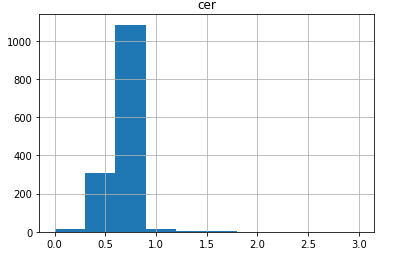
%records with 0-0.25 CER=1.2622384334805146%

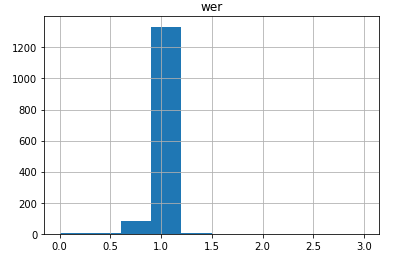
Total OOV words in AM data: 12247

Total OOV words in LM data: 22

* + - 1. **Sports:**







No. of 0-0.5 WER records=2

Total Records=1425

%records with 0-0.5 WER =0.14035087719298245%

No. of 0-0.5 CER records=84

Total Records=1425

%records with 0-0.5 CER=5.894736842105263%

Total OOV words in AM data: 2126

Total OOV words in LM data: 5

More training is required targeting domains, rather than in general. on all these domains.

In LM more data needs to be added for religion, education, sports, healthcare, entertainment and technology related text corpuses.

':

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | % of total files | %0 WER  (within each done) | %0 WER  (across all domains) | %OOV  (within each one in lm and am) | %OOV  (across all in lm and am) | Avg WER |
| news |  |  |  |  |  |  |
| religion |  |  |  |  |  |  |
| education |  |  |  |  |  |  |
| general |  |  |  |  |  |  |
| technology |  |  |  |  |  |  |
| healthcare |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| sports |  |  |  |  |  |  |
| entertainment |  |  |  |  |  |  |

* + 1. List of OOV Words correctly transcribed.(list, cnt, %)(AM, LM)
    2. List of training words not performing well (not crctly transcribed)