Inferences:

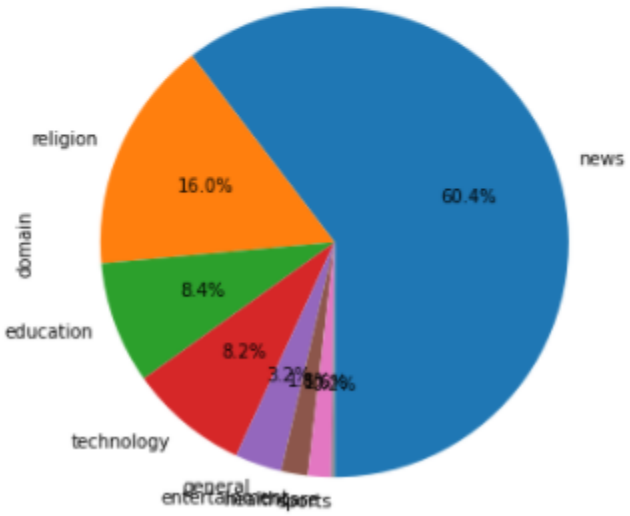
1. Ulca test analysis overview:

1 to 8 of 8 entries Filter

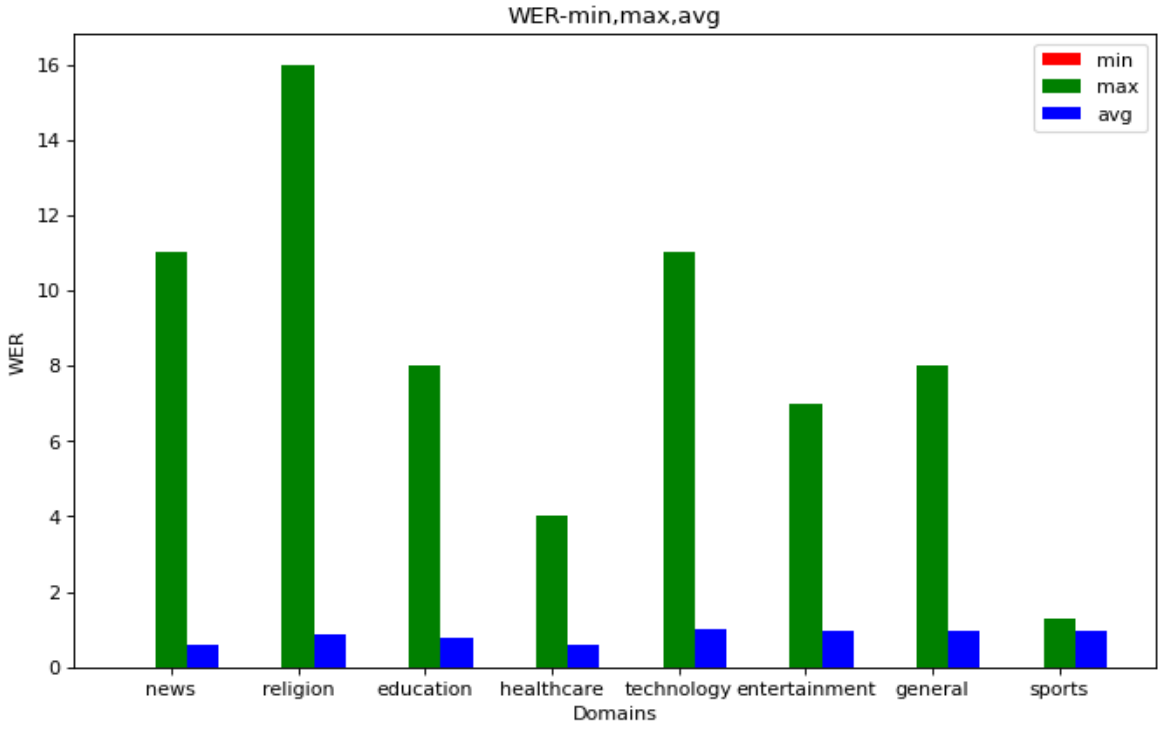
index	loss	char_distance	char_length	word_distance	word_length	cer	wer	durationSec
count	643377.0	643377.0	643377.0	643377.0	643377.0	643377.0	643377.0	643377.0
mean	127.5869000752709	30.956440780444435	92.27594085582793	7.03462355663942	10.540673664119172	0.4027229548600489	0.710506194342079	5.766138655873461
std	101.09941415918064	26.486428128940336	61.54566162082664	5.133924377306195	6.680129910716798	0.34515103894429017	0.3151855820185174	3.335173057696041
min	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.51
25%	57.53913497924805	12.0	42.0	3.0	5.0	0.17307692307692302	0.5	3.09
50%	101.8063735961914	24.0	80.0	6.0	9.0	0.35849056603773505	0.7647058823529409	5.04
75%	167.9835205078125	42.0	130.0	10.0	15.0	0.6101694915254231	1.0	7.83
max	1378.768798828125	269.0	354.0	47.0	47.0	31.33333333333333	16.0	15.0

2. Ulca dataset consists majorly news.

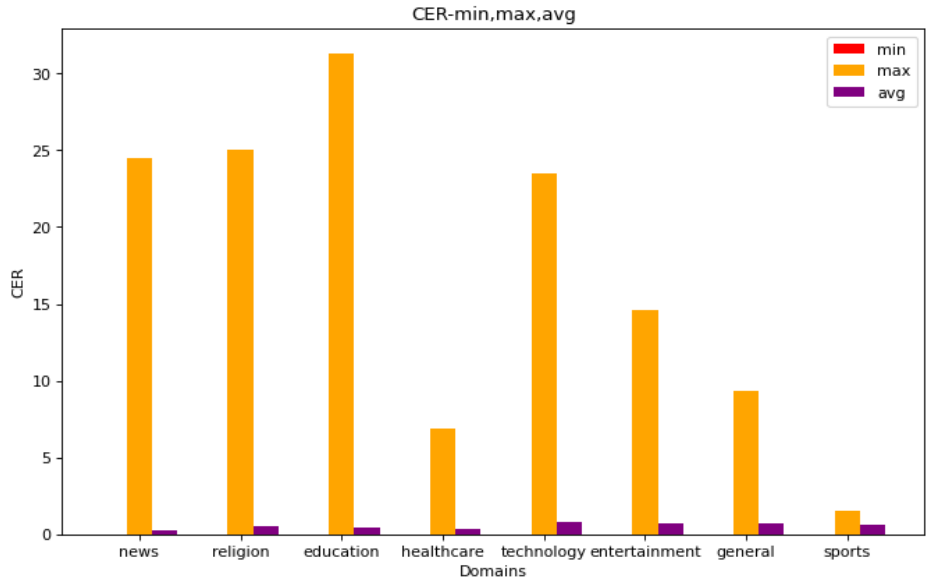
```
news             388747
religion         103085
education         54082
technology        53047
general           20836
entertainment     11809
healthcare        10346
sports             1425
Name: domain, dtype: int64
<matplotlib.axes._subplots.AxesSubplot at 0x7f8ed756c290>
```



3. WER analysis (comparison across all domains)



4. CER analysis (comparison across all domains)



5. 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}

6. LM Data consists majorly news related words.

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

7. 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

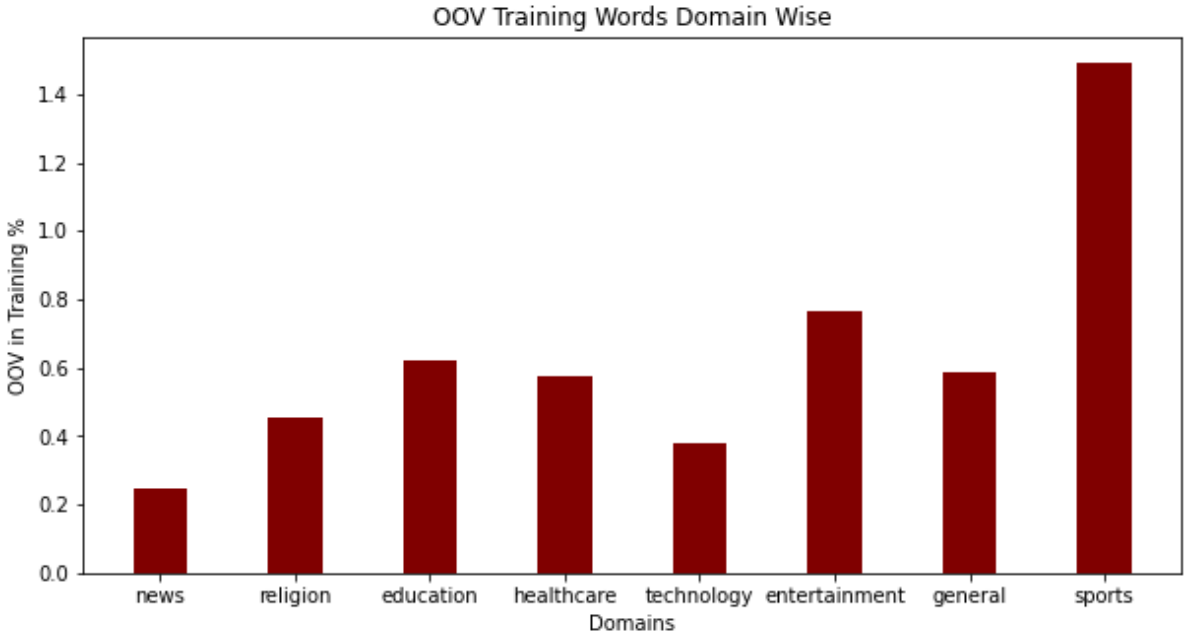
ஃபுல்லாம்	9
ஓகேய்வா	9
பண்டேன்	8
நுப்பத்தி	8
அப்டீன்னேன்	7
அட்வைசிங்	6
இதுக்காண்டி	5
பர்பஸ்க்காக	5
திருத்தேர்	5
ஆயரூபா	4
ஃபர்தரா	4
இஷட்டு	3
தொளசண்ட்	3
அப்டீன்னே	3
கேக்குதுங்களா	3
பார்க்கவேண்டியிருக்க	2
ரெக்கொயர்மெண்ட்	2
செரிறி	2
தொகுத்திருக்கு	2
செலவழித்திருக்கிறே	2
அதலெல்லாம்	2
திருப்புல்லா	1
பால்பண்	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

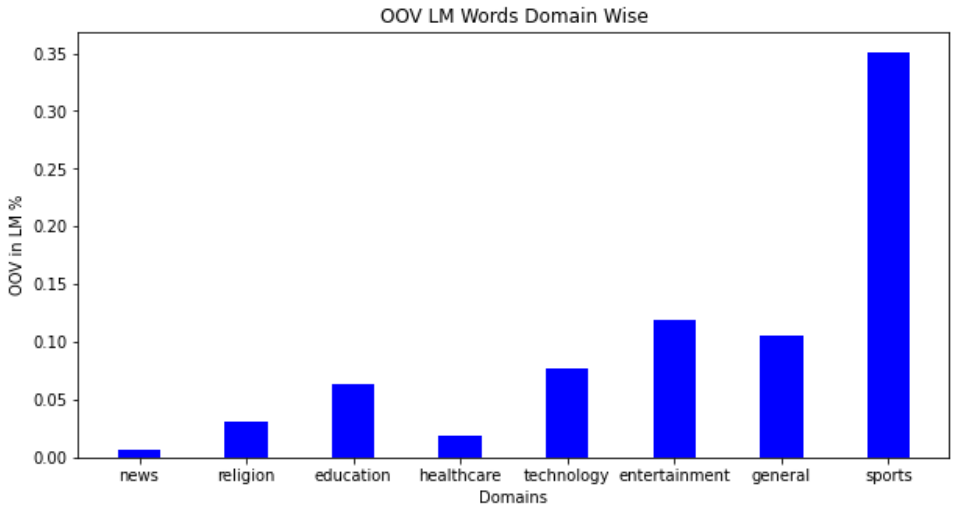
8. Total OOV Words not known to AM is 116107

9. Most common OOV words wrt LM across all domains: {'பெர்ஸன்ட்', 'அப்டின்றான்', 'அப்டின்ட்டு', 'டுவெண்ட்டி', 'டுவல்வ்', 'பண்ட்டு'}

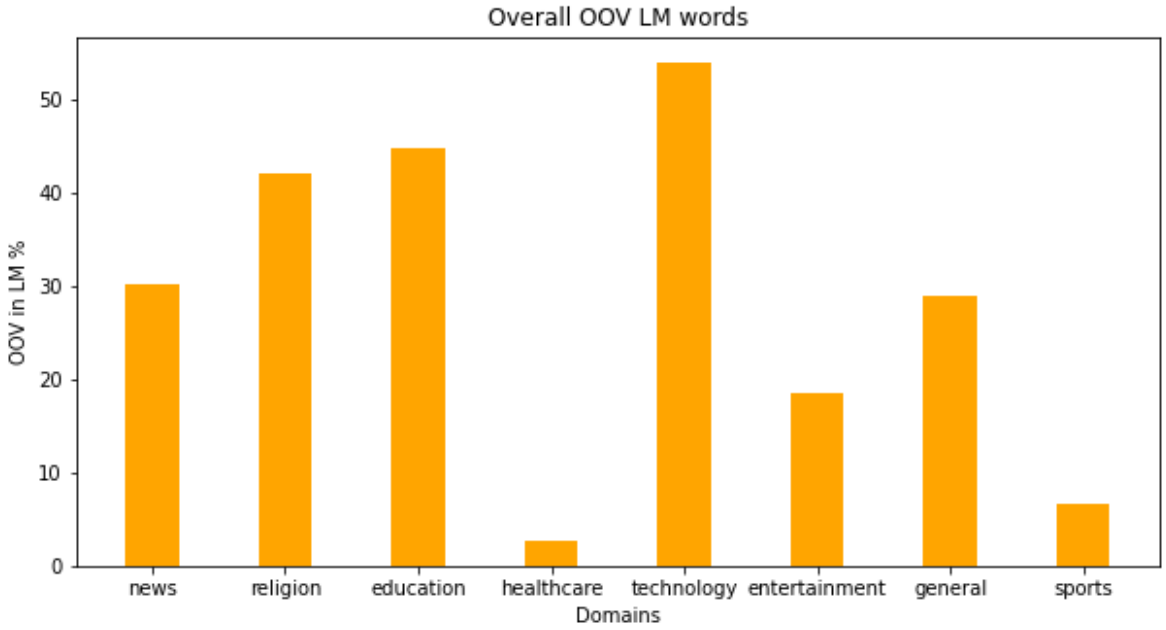
10. Numbers (in Tamil) constitute a major part of OOV words across all domains

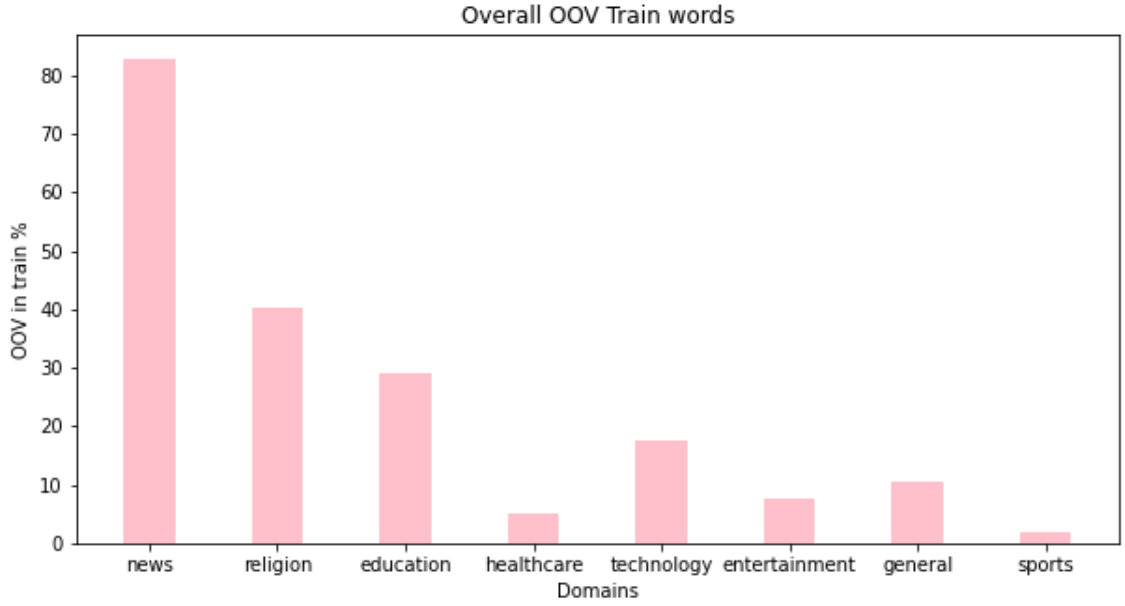
11. OOV words for both AM and LM is majorly from sports followed by entertainment. News has least OOV words.





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





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

12. wav file duration and WER are negatively correlated: (value: -0.1730614399471933)

Sl. No.	Data Source	Observation
1.	OOV for Religion wrt LM	a. 8/32 are numbers. b. 7/32 are Tanglish words. c. 17/32 are modern tamil words. d. WER: i. Min-0.00 ii. Max-16.00 iii. Avg-0.856 e. CER: i. Min-0.00 ii. Max-25.0 iii. Avg-0.524
2.	OOV for News wrt LM	a. 9/23 are numbers b. 6/23 are Tanglish words (accented) c. 8/23 are modern tamil d. WER: i. Min-0.00 ii. Max-11.00 iii. Avg-0.605 e. CER: i. Min-0.00 ii. Max-24.5 iii. Avg-0.289
3.	OOV for Education wrt LM	a. 10/34 are numbers b. 11/34 are Tanglish words (accented) c. 13/34 are modern tamil d. WER: i. Min-0.00 ii. Max-8.00

		iii. Avg-0.766 e. CER: i. Min-0.00 ii. Max-31.33 iii. Avg-0.43
4.	OOV for Healthcare wrt LM	a. 2/2 are modern tamil b. WER: i. Min-0.00 ii. Max-4.00 iii. Avg-0.611 c. CER: i. Min-0.00 ii. Max-6.857 iii. Avg-0.315
5.	OOV for technology wrt LM	a. 9/41 are numbers b. 12/41 are Tanglish words (accented) c. 20/41 are modern tamil d. WER: i. Min-0.00 ii. Max-11.00 iii. Avg-1.00 e. CER: i. Min-0.00 ii. Max-23.50 iii. Avg-0.798
6.	OOV for entertainment wrt LM	a. 1/14 are numbers b. 4/14 are Tanglish words (accented) c. 9/14 are modern tamil d. WER: ii. Min-0.00 iii. Max-7.00 iv. Avg-0.954 e. CER: i. Min-0.00 ii. Max-14.60 iii. Avg-0.694

7.	OOV for general wrt LM	a. 6/22 are numbers b. 6/22 are Tanglish words (accented) c. 10/22 are modern tamil
8.	OOV for sports wrt LM	a. 3/5 are Tanglish words (accented) b. 2/5 are modern tamil

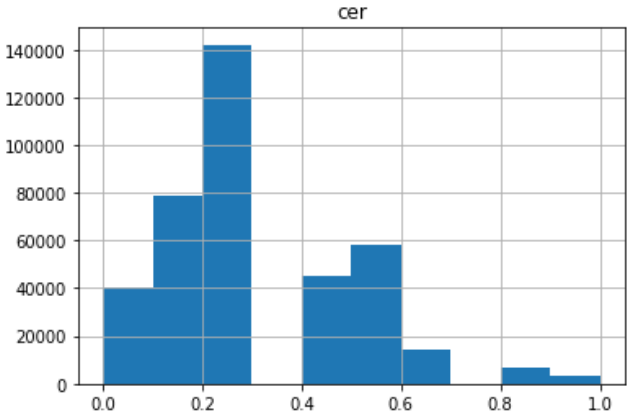
13. In depth domain analysis:

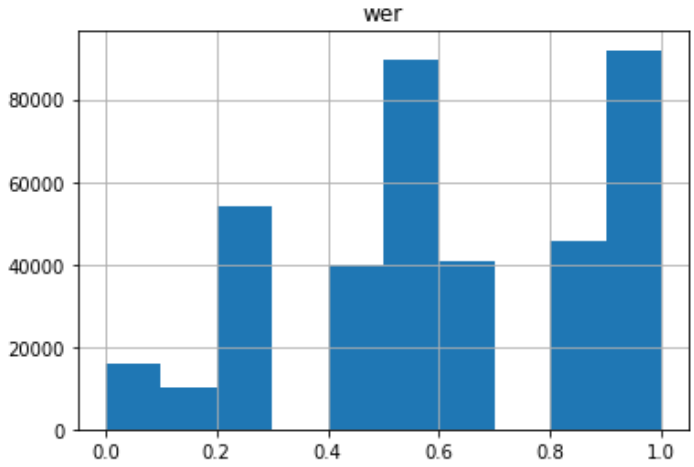
1. News:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	388747.000000	388747.000000	388747.00000	388747.000000	388747.000000	388747.000000	388747.000000
mean	123.126505	30.337160	113.63342	7.358251	12.469184	0.289545	0.605358
std	87.148827	26.217196	61.79334	5.168112	6.707807	0.256556	0.294998
min	0.000000	0.000000	1.00000	0.000000	1.000000	0.000000	0.000000
25%	61.280657	11.000000	67.00000	4.000000	7.000000	0.125000	0.400000
50%	103.616951	24.000000	105.00000	6.000000	12.000000	0.243697	0.615385
75%	163.108940	42.000000	154.00000	10.000000	17.000000	0.415760	0.833333
max	1286.996948	232.000000	349.00000	37.000000	39.000000	24.500000	11.000000

```
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%



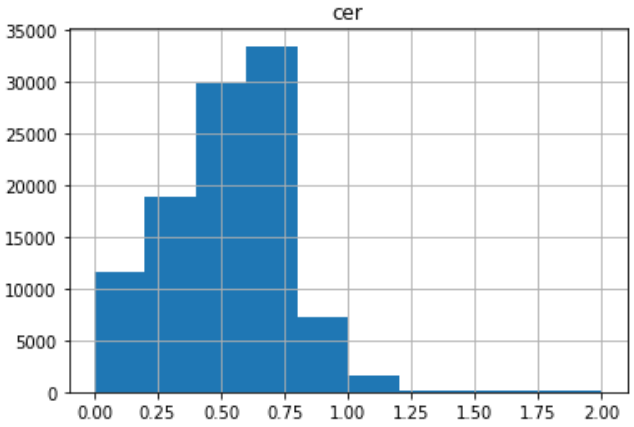


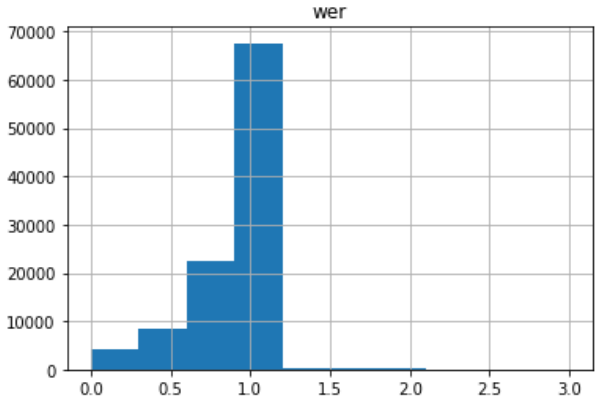
AM data consists mainly of news related words. So, for most of the wav files CER lies between 0 and 0.3. Many files have relatively faster speech rates, so AM might have partially transcribed the syllables. The LM on receiving this input yields the nearest match, which may or may not be the most accurate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

2. Religion:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	103085.000000	103085.000000	103085.000000	103085.000000	103085.000000	103085.000000	103085.000000
mean	101.043972	25.364505	51.263297	5.430237	6.421439	0.524132	0.856779
std	83.780615	20.199738	35.923676	3.927646	4.353611	0.306157	0.262819
min	0.000000	0.000000	2.000000	0.000000	1.000000	0.000000	0.000000
25%	44.400000	12.000000	26.000000	3.000000	3.000000	0.400000	0.800000
50%	78.000000	20.000000	41.000000	4.000000	5.000000	0.500000	1.000000
75%	131.000000	33.000000	66.000000	7.000000	8.000000	0.700000	1.000000
max	1255.500000	269.000000	354.000000	47.000000	47.000000	25.000000	16.000000





Total Records=103085

No. of 0-0.5 WER records=12115

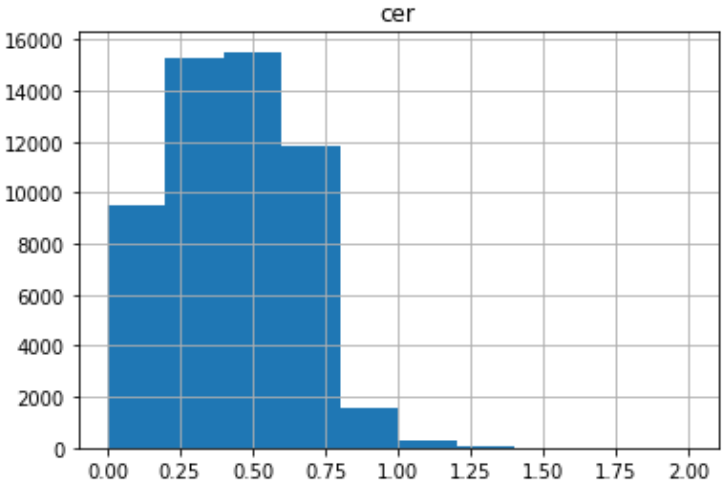
%records with 0-0.5 WER= 11.75243730901683%

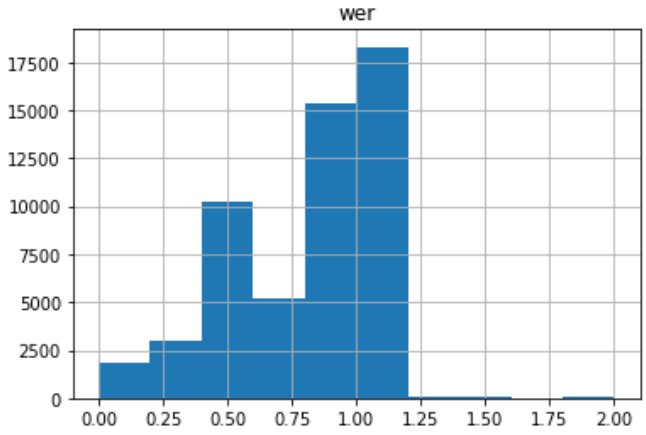
No. of 0-0.25 CER records=16007
Total Records=103085
%records with 0-0.25 CER=15.527962361158268%

Total religion OOV words in AM data: 46720
Total religion OOV words in LM data: 32

3. Education Analysis:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	54082.000000	54082.000000	54082.000000	54082.000000	54082.000000	54082.000000	54082.000000
mean	134.870579	33.779021	80.409841	7.613513	9.862061	0.430252	0.766521
std	116.766810	28.711413	53.776569	5.583064	6.256023	0.285652	0.272546
min	0.043945	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000
25%	49.228918	13.000000	38.000000	3.000000	5.000000	0.256881	0.615385
50%	100.988815	26.000000	68.000000	6.000000	8.000000	0.424619	0.833333
75%	184.787163	47.000000	110.000000	10.000000	13.000000	0.600000	1.000000
max	1218.222168	249.000000	339.000000	42.000000	45.000000	31.333333	8.000000



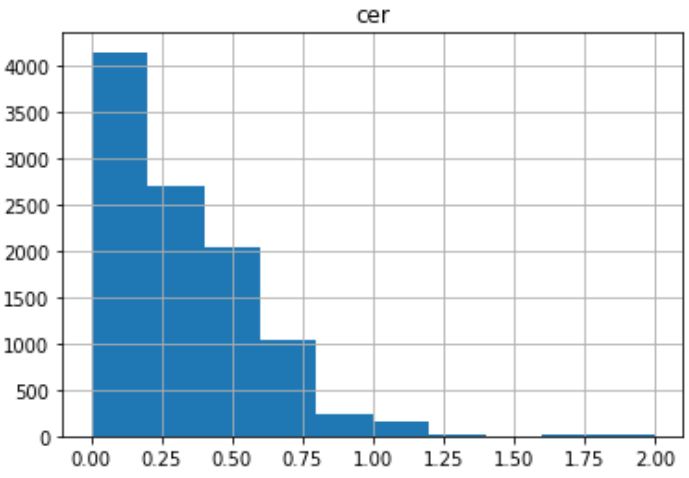


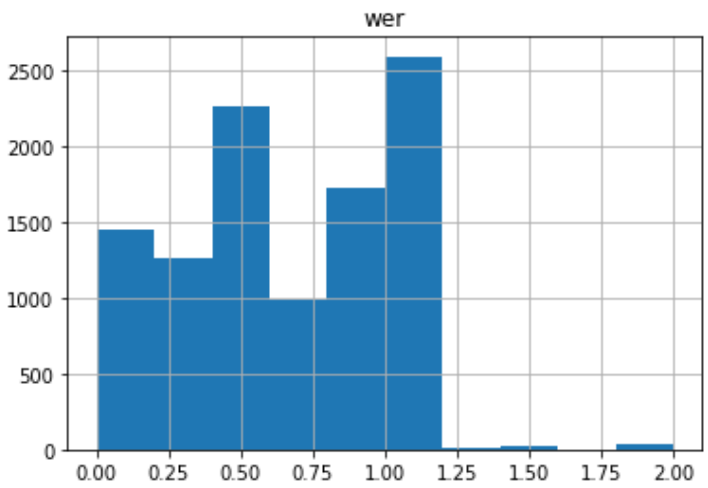

```
No. of 0-0.25 WER records=3113

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

4. Healthcare

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	10346.000000	10346.000000	10346.000000	10346.000000	10346.000000	10346.000000	10346.000000
mean	60.816806	15.027257	52.179200	3.780785	6.347477	0.314965	0.611292
std	56.667055	15.145642	34.667719	3.203744	4.087229	0.289941	0.357722
min	0.057791	0.000000	2.000000	0.000000	1.000000	0.000000	0.000000
25%	20.891195	4.000000	26.000000	2.000000	3.000000	0.093750	0.333333
50%	44.157618	11.000000	44.000000	3.000000	5.000000	0.274510	0.666667
75%	82.496147	22.000000	69.000000	5.000000	8.000000	0.476190	1.000000
max	735.703674	176.000000	292.000000	35.000000	36.000000	6.857143	4.000000





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

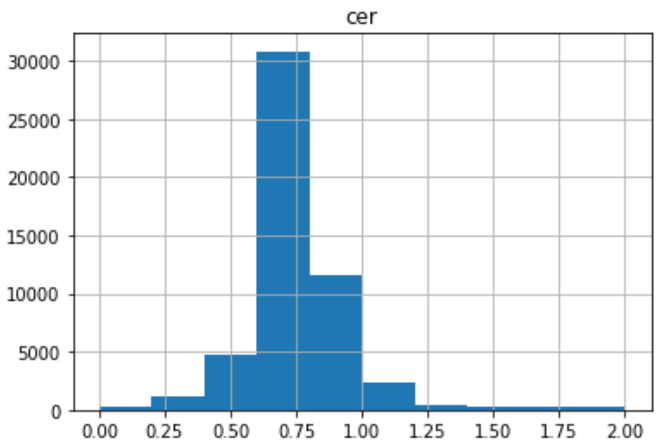
transcripts. The number of such files is very less, explaining the decreasing trend in both the graphs.

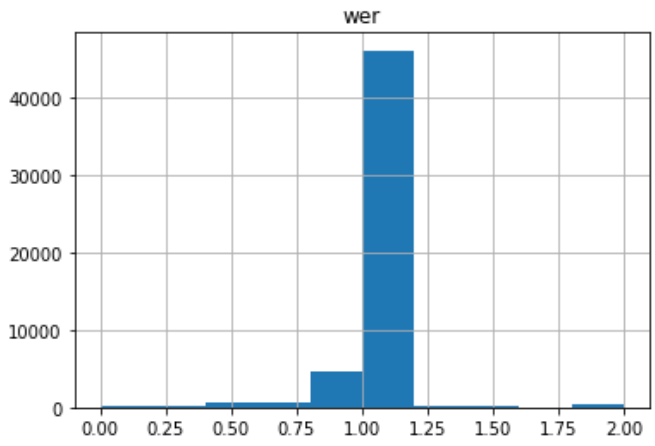
```
No. of 0-0.25 WER records=2077

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

5. Technology:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	53047.000000	53047.000000	53047.000000	53047.000000	53047.000000	53047.000000	53047.000000
mean	188.417629	41.085113	57.232586	7.564066	7.727770	0.798721	1.002002
std	143.665302	30.727321	45.080685	5.721644	5.894071	0.560034	0.254271
min	0.488861	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000
25%	82.848755	19.000000	25.000000	3.000000	3.000000	0.674419	1.000000
50%	149.564285	33.000000	45.000000	6.000000	6.000000	0.743590	1.000000
75%	254.222343	55.000000	77.000000	10.000000	10.000000	0.814815	1.000000
max	1378.768799	249.000000	341.000000	47.000000	47.000000	23.500000	11.000000



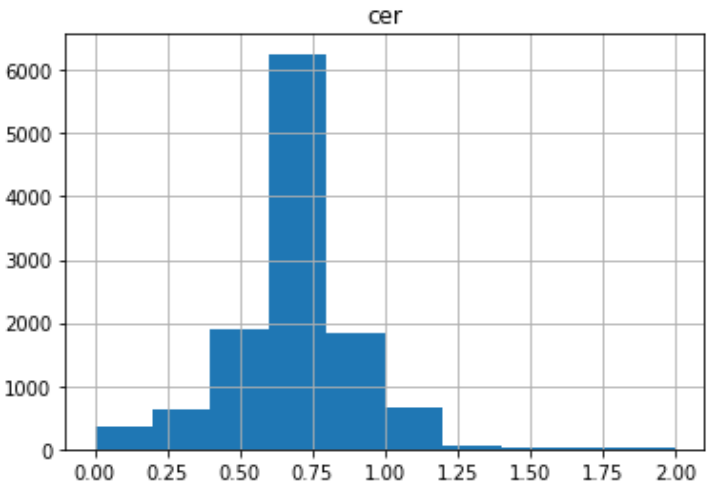


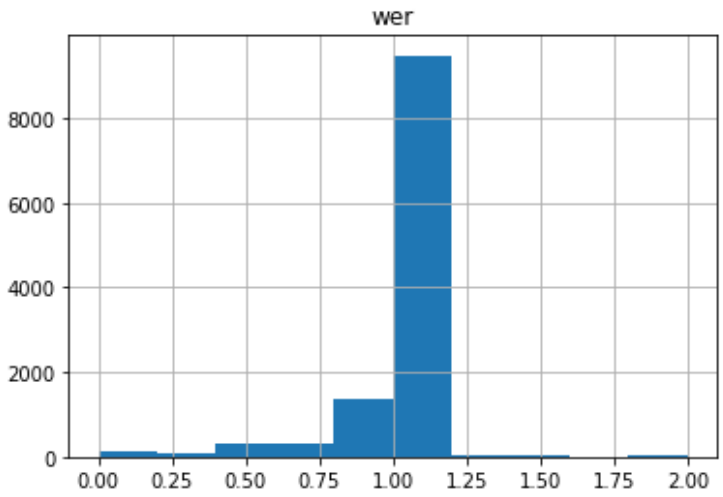
```
No. of 0-0.25 WER records=162
Total Records =53047

%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

6. Entertainment:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	11809.000000	11809.000000	11809.000000	11809.000000	11809.000000	11809.000000	11809.000000
mean	146.620959	32.317216	48.986028	6.272335	6.594462	0.694326	0.954933
std	120.623872	24.784651	37.065261	4.612088	4.758754	0.293577	0.194312
min	0.600000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000
25%	59.300000	15.000000	23.000000	3.000000	3.000000	0.600000	1.000000
50%	110.100000	26.000000	38.000000	5.000000	5.000000	0.700000	1.000000
75%	199.200000	44.000000	65.000000	8.000000	9.000000	0.800000	1.000000
max	812.000000	194.000000	285.000000	33.000000	34.000000	14.600000	7.000000



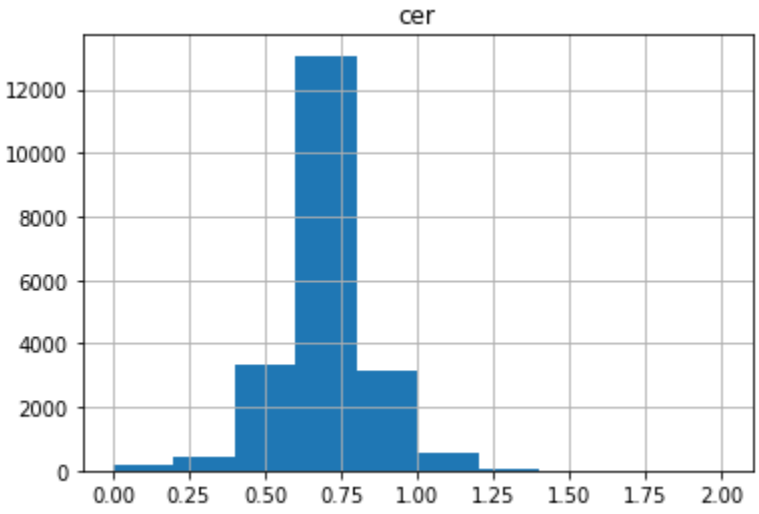


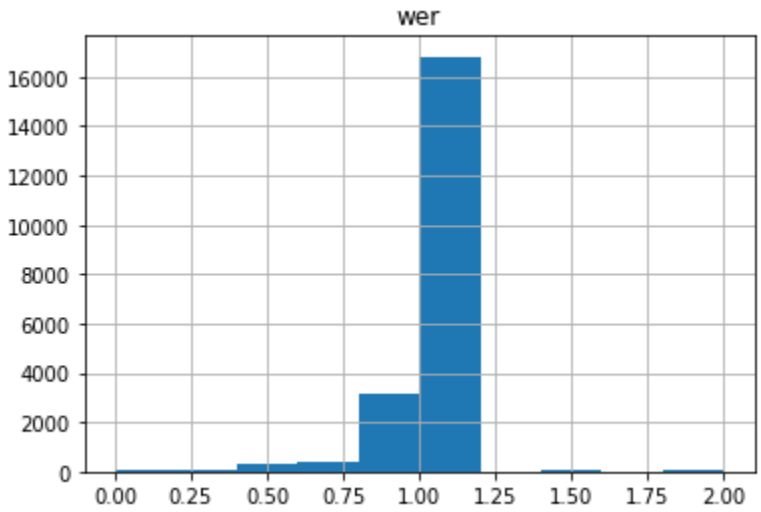

```
No. of 0-0.25 WER records=171
Total Records=11809
%records with 0-0.25 WER =1.4480480989076128%


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

7. General:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	20836.000000	20836.000000	20836.000000	20836.000000	20836.000000	20836.000000	20836.000000
mean	179.803609	42.248848	61.287243	7.840756	8.107650	0.696871	0.965876
std	136.951261	31.303169	44.247893	5.601440	5.699808	0.192426	0.138371
min	1.500000	0.000000	2.000000	0.000000	1.000000	0.000000	0.000000
25%	81.800000	20.000000	29.000000	4.000000	4.000000	0.600000	1.000000
50%	141.600000	34.000000	49.000000	6.000000	7.000000	0.700000	1.000000
75%	238.600000	55.000000	81.000000	10.000000	11.000000	0.800000	1.000000
max	1369.100000	252.000000	327.000000	42.000000	43.000000	9.300000	8.000000



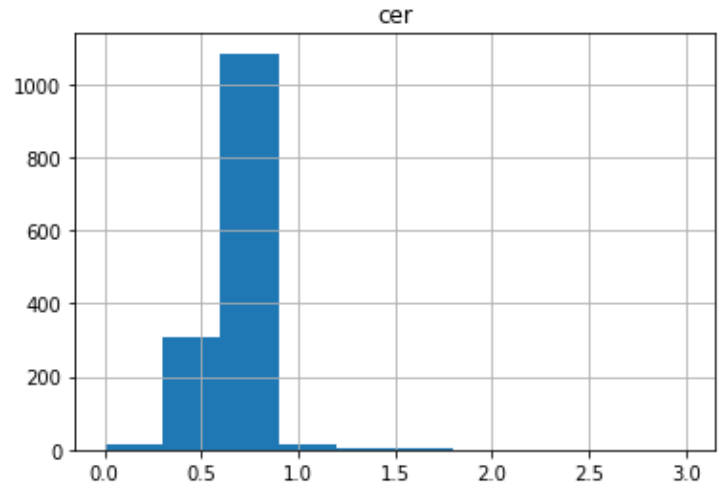


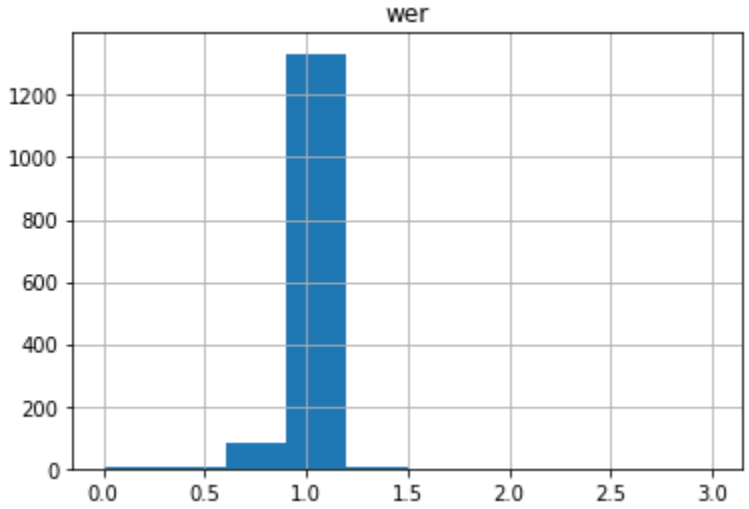

```
No. of 0-0.25 WER records=97

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

8. Sports:

	loss	char_distance	char_length	word_distance	word_length	cer	wer
count	1425.000000	1425.000000	1425.000000	1425.000000	1425.000000	1425.000000	1425.000000
mean	287.147653	59.508772	90.557895	11.284211	11.609825	0.656961	0.968539
std	185.323280	34.206475	50.951863	6.275392	6.372841	0.111813	0.079705
min	8.995679	0.000000	6.000000	0.000000	1.000000	0.000000	0.000000
25%	146.414078	33.000000	51.000000	6.000000	7.000000	0.603604	0.969697
50%	248.073380	53.000000	80.000000	10.000000	10.000000	0.661290	1.000000
75%	386.734955	80.000000	120.000000	15.000000	15.000000	0.714286	1.000000
max	1137.396240	235.000000	327.000000	37.000000	37.000000	1.555556	1.285714






```
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)