```
#Mounting Drive:
from google.colab import drive
drive.mount("/content/drive")
```

```
Mounted at /content/drive
```

```
drive.flush_and_unmount()
```

```
#PROJECT VARIABLES TO COPY TO DRIVE: DO NOT MODIFY!

#DATASET GLOBAL PATH CONFIGURATION:
PROJECT_ROOT_DIR = '/content/drive/MyDrive/testProj/'
DATASETS_ROOT_DIR = PROJECT_ROOT_DIR + 'datasets/'
TAMIL_DATASET_HOME = DATASETS_ROOT_DIR + 'ta/'
SPEECH_CORPORA_ZIPS_DIR = DATASETS_ROOT_DIR + 'zips/speech_corpora/'
CV_DATASET_DIR = TAMIL_DATASET_HOME + 'commonvoice/'
OPENSLR_DATASET_DIR = TAMIL_DATASET_HOME + 'openslr/'
ULCA_DATASET_DIR = TAMIL_DATASET_HOME + 'ulca_corpus/'
  #UTILS VARS:
UTILS_HOME = PROJECT_ROOT_DIR+'utils/'
DEEPSPEECH_HOME = UTILS_HOME + 'deepspeech/'
KENLM_HOME=UTILS_HOME + 'kenlm/'

  #MODEL VARS:
LM_PATH = PROJECT_ROOT_DIR + 'lm/'
MODEL_HOME = PROJECT_ROOT_DIR + 'model_without_ulca/'
FEATURE_CACHE_PATH = MODEL_HOME + 'feature_cache/'
CHECKPOINTS_DIR = MODEL_HOME + 'checkpoints/'
EXPORTS_DIR = MODEL_HOME + 'exports/'
SUMMARY_DIR = MODEL_HOME + 'summary/'
TEST_OUTPUT_JSON_PATH = MODEL_HOME + 'output.json'
```

Analysis of ULCA Dataset

Generating the required data

```
#Testing DFs
import pandas as pd
ulca_df = pd.read_csv(TAMIL_DATASET_HOME + 'csv/complete_datasets/ulca_dataset.csv')

# #Test results DFs
# ulca_test_result_df = pd.read_csv(MODEL_HOME + 'ulca_test_results_modified.csv')

# #Training DF = Train + Dev
# train_df = pd.concat([pd.read_csv(TAMIL_DATASET_HOME + 'train-without-ulca.csv'), pd.rea
```

```
st = train_words_set.union(oov_train) #266313
ulca_word_set # 201072
len(st - ulca_word_set)
```

```
65241
```

```
#LM Set and Train words set
train_words_set = set()
lm_words_set = set()

with open(LM_PATH + 'intermediates/old/vocab-5000000.txt') as f:
  for line in f:
    line = line.strip()
    lm_words_set.add(line)

def addTrainWords(wordList):
  global train_words_set
  for x in wordList:
    if len(x) > 0:
      train_words_set.add(x)

train_df.apply(lambda x : addTrainWords(str(x['transcript']).strip().split(' ')), axis = 1

print('No. of unique words in speech train dataset:', len(train_words_set))
print('No. of unique words in Language Model:', len(lm_words_set))
```

```
No. of unique words in speech train dataset: 116107
No. of unique words in Language Model: 5000000
```

```
#No. of words in datasets:
ulca_word_set = set()

def addWords(wordList):
  global ulca_word_set

  for x in wordList:
      ulca_word_set.add(x)

ulca_test_results_df.apply(lambda x : addWords(str(x['src']).strip().split(' ')), axis = 1

print('Total no. of unique words in ULCA dataset:', len(ulca_word_set))
```

```
Total no. of unique words in ULCA dataset: 201072
```

```
word_domain_dict = dict()

def findDomain(transcript, domain):
  wordList = transcript.split(' ')
  for word in wordList:
    if len(word) > 0:
      if not(word in word_domain_dict):
        word_domain_dict[word] = set()
```

```
      word_domain_dict[word].add(domain)

ulca_test_results_df.apply(lambda x : findDomain(x['transcript'], x['domain']), axis = 1)

intersection_set = set(word_domain_dict.keys()).intersection(train_words_set)
print('No. of words in intersection set(AM & ULCA):', len(intersection_set))

intersection_dict = dict()
for w in intersection_set:
  intersection_dict[w] = word_domain_dict[w]
```

```
No. of words in intersection set(AM & ULCA): 50866
```

```
word_domain_dict = dict()

def findDomain(transcript, domain):
  wordList = transcript.split(' ')
  for word in wordList:
    if len(word) > 0:
      if not(word in word_domain_dict):
        word_domain_dict[word] = set()

      word_domain_dict[word].add(domain)

ulca_df.apply(lambda x : findDomain(x['transcript'], x['domain']), axis = 1)

intersection_set = set(word_domain_dict.keys()).intersection(lm_words_set)
print('No. of words in intersection set(LM & ULCA):', len(intersection_set))

intersection_dict = dict()
for w in intersection_set:
  intersection_dict[w] = word_domain_dict[w]
```

```
No. of words in intersection set(LM & ULCA): 200996
```

```
df = pd.DataFrame.from_dict(list(intersection_dict.keys()))
df = df.rename(columns={0 : 'train_word'})
df['domains'] = df.apply(lambda x : intersection_dict[x['train_word']], axis = 1)
df.head(10000)
```

	train_word	domains
0	அவரிடமிருந்த	{news}
1	டிரைவரையும்	{religion}
2	விடைபெறும்	{news, religion}
3	கிழமைகளிலும்	{news}
4	விலைமாதர்	{religion}

```
domainFreqCounts = dict()
def countDomains(domain_set):
  for domain in domain_set:
    if domain in domainFreqCounts:
      domainFreqCounts[domain] += 1
    else:
      domainFreqCounts[domain] = 1

df.apply(lambda x : countDomains(intersection_dict[x['train_word']]), axis = 1)
print(domainFreqCounts)
```

```
{'news': 133974, 'religion': 75663, 'education': 57863, 'technology': 31541, 'general
```

```
#OOV Training Words
to_append = []
print('No. of unique training words:', len(train_words_set))

def addOOVWords(wavFileName, wordList, duration, old_loc):
  for w in wordList:
    if len(w) > 0 and not(w in train_words_set):
      to_append.append([w, wavFileName, old_loc, duration])

ulca_df.apply(lambda x : addOOVWords(x['wav_filename'], str(x['transcript']).strip().split

oov_train_df = pd.DataFrame(to_append, columns=['unknown_word', 'wav_filename', 'old_loc',
del to_append

oov_train_df.drop_duplicates(inplace=True)

oov_train_df = pd.merge(oov_train_df, ulca_test_result_df, on='wav_filename', how='inner')

oov_train_df.drop(columns=['loss', 'char_distance', 'char_length', 'word_distance', 'trans
oov_train_df.drop_duplicates(inplace=True)

oov_train_df['cer'].round(decimals=5)
oov_train_df['wer'].round(decimals=5)

print('Total CSV length:', len(oov_train_df.index))
oov_train_df.to_csv(TAMIL_DATASET_HOME + 'unknown_training_words.csv', index = False)
```

```
No. of unique training words: 116107
Total CSV length: 1501312
```

```
a = oov_train_df['unknown_word'].value_counts()
a = pd.DataFrame(a)
a
```

	unknown_word
அப்டின்னு	6059
அப்படின்னு	5998
மாநிலச்	5685
எடப்பாடி	5353
செய்தியாளர்களிடம்	4983
...	...
யுனெஸ்கோவில்	1
ஒருகூட்டம்	1
காணிக்கையாகக்	1
பேசிக்கொண்டிருந்தபோது	1
டிக்கெட்டுகளில்	1

150206 rows × 1 columns

```
#OOV LM Words
to_append = []

def addOOVWords(wavFileName, wordList, duration, old_loc):
  for w in wordList:
    if len(w) > 0 and not(w in lm_words_set):
      to_append.append([w, wavFileName, old_loc, duration])

ulca_df.apply(lambda x : addOOVWords(x['wav_filename'], str(x['transcript']).split(' '), x
oov_lm_df = pd.DataFrame(to_append, columns=['unknown_word', 'wav_filename', 'old_loc', 'd
oov_lm_df.drop_duplicates(inplace=True)

oov_lm_df = pd.merge(oov_lm_df, ulca_test_result_df, on='wav_filename', how='inner')
oov_lm_df.drop(columns=['loss', 'char_distance', 'char_length', 'word_distance', 'transcri
oov_lm_df.drop_duplicates(inplace=True)

oov_lm_df['cer'].round(decimals = 5)
oov_lm_df['wer'].round(decimals = 5)

print('Total CSV length:', len(oov_lm_df.index))
#oov_lm_df.to_csv(TAMIL_DATASET_HOME + 'unknown_lm_words.csv', index = False)

del to_append
oov_lm_df
```

```
Total CSV length: 3496
```

	unknown_word	wav_filename	
0	எட்நூத்தி	split_dataset/part1/a/48_7_237file-idkgTgcdjaG...	ulca_corpus/resampled/corpu
1	கொஷ்டீன்	split_dataset/part1/a/48_47_1038file-idl-tBAnF...	ulca_corpus/resampled/corpu
2	ஒன்பதாயிரத்தி	split_dataset/part1/a/48_11_177file-idDCDWXDSP...	ulca_corpus/resampled/corp
3	டுவெண்ட்டி	split_dataset/part1/a/48_67_646file-idw7_7ep58...	ulca_corpus/resampled/corpu
4	எட்நூத்தி	split_dataset/part1/b/48_56_1141file-id6uW09fF...	ulca_corpus/resampled/corpu
...	...	...	
3491	ஆறாயிரத்தி	split_dataset/part153/c/48_23_NSD-Tamil-Tamil-...	ulca_corpus/resampled
3492	ஓகேங்களா	split_dataset/part153/c/48_11_611file-id3XrDzx...	ulca_corpus/resampled/corp
3493	ஆறாயிரத்தி	split_dataset/part153/d/48_0_9_Regional-Chenna...	ulca_corpus/resampled/corpus
3494	அறநூத்தி	split_dataset/part153/e/48_43_448file-idMNnDV1...	ulca_corpus/resampled/corpu
3495	எட்நூத்தி	split_dataset/part153/e/48_6_241file-idY6smQhb...	ulca_corpus/resampled/corpu

3496 rows × 11 columns

```
#OOV to training and LM:
to_append = []

def addOOVWords(wavFileName, wordList, duration, collectionSource, domain):
  for w in wordList:
    if len(w) > 0 and not(w in train_words_set) and not(w in lm_words_set):
      to_append.append([w, wavFileName, duration, collectionSource, domain])
```

```
ulca_df.apply(lambda x : addOOVWords(x['wav_filename'], str(x['transcript']).split(' '), x
output_df = pd.DataFrame(to_append, columns=['unknown_word', 'wav_filename', 'duration_sec
output_df = output_df.drop_duplicates()

output_df = pd.merge(output_df, ulca_test_result_df, on='wav_filename', how='inner')
output_df.drop(columns=['loss', 'char_distance', 'char_length', 'word_distance', 'word_len
output_df.drop_duplicates(inplace=True)

print('Total CSV length:', len(output_df.index))
#output_df.to_csv(TAMIL_DATASET_HOME + 'unknown_words.csv', index = False)

del to_append

output_df['cer'].round(decimals = 6)
output_df['wer'].round(decimals = 6)

print('Unknown LM DataFrame is same as this dataframe!' if output_df.drop(columns=['durati
```

```
Total CSV length: 3496
Unknown LM DataFrame is same as this dataframe!
```

```
#Number of unique OOV Words in LM
unique_to_lm_words = oov_lm_df['unknown_word'].unique()
print('No. of unique words:', len(unique_to_lm_words))
```

```
No. of unique words: 76
```

```
print('LM Set size:', len(lm_words_set))
print('Train Set size:', len(train_words_set))
print('No. of common elements to LM and train set:', len(train_words_set.intersection(lm_w
print('Union Set size:', len(lm_words_set.union(train_words_set)))
```

```
LM Set size: 5000000
Train Set size: 116107
No. of common elements to LM and train set: 98590
Union Set size: 5017517
```

Generating ulca_test_result_df

```
#Reading JSON Files
import json

data_dict = dict()
params_dict = dict()
for i in range(1, 61):
  print('Reading part #', i, ': ', end='')
  print('data JSON...', end='')
  try:
    df = pd.read_json(ULCA_DATASET_DIR + 'corpus' + str(i) + '/data.json')
    data_dict[i] = df
    print('done. params JSON...', end='')
```

```
    except:
      print('not available!')
      continue

    try:
      f = open(ULCA_DATASET_DIR + 'corpus' + str(i) + '/params.json')
      params = json.load(f)
      params_df = pd.json_normalize(params)
      params_dict[i] = params_df
      f.close()
      print('done.')
    except:
      print('not available!')
```

```
 Reading part # 1 : data JSON...done. params JSON...done.
 Reading part # 2 : data JSON...done. params JSON...done.
 Reading part # 3 : data JSON...done. params JSON...done.
 Reading part # 4 : data JSON...done. params JSON...done.
 Reading part # 5 : data JSON...done. params JSON...done.
 Reading part # 6 : data JSON...done. params JSON...done.
 Reading part # 7 : data JSON...done. params JSON...done.
 Reading part # 8 : data JSON...done. params JSON...done.
 Reading part # 9 : data JSON...done. params JSON...done.
 Reading part # 10 : data JSON...done. params JSON...done.
 Reading part # 11 : data JSON...done. params JSON...done.
 Reading part # 12 : data JSON...done. params JSON...done.
 Reading part # 13 : data JSON...done. params JSON...done.
 Reading part # 14 : data JSON...done. params JSON...done.
 Reading part # 15 : data JSON...done. params JSON...done.
 Reading part # 16 : data JSON...done. params JSON...done.
 Reading part # 17 : data JSON...done. params JSON...done.
 Reading part # 18 : data JSON...done. params JSON...done.
 Reading part # 19 : data JSON...done. params JSON...done.
 Reading part # 20 : data JSON...done. params JSON...done.
 Reading part # 21 : data JSON...done. params JSON...done.
 Reading part # 22 : data JSON...done. params JSON...done.
 Reading part # 23 : data JSON...done. params JSON...done.
 Reading part # 24 : data JSON...done. params JSON...done.
 Reading part # 25 : data JSON...done. params JSON...done.
 Reading part # 26 : data JSON...not available!
 Reading part # 27 : data JSON...done. params JSON...done.
 Reading part # 28 : data JSON...done. params JSON...done.
 Reading part # 29 : data JSON...done. params JSON...done.
 Reading part # 30 : data JSON...not available!
 Reading part # 31 : data JSON...done. params JSON...done.
 Reading part # 32 : data JSON...done. params JSON...done.
 Reading part # 33 : data JSON...done. params JSON...done.
 Reading part # 34 : data JSON...done. params JSON...done.
 Reading part # 35 : data JSON...done. params JSON...done.
 Reading part # 36 : data JSON...done. params JSON...done.
 Reading part # 37 : data JSON...done. params JSON...done.
 Reading part # 38 : data JSON...done. params JSON...done.
 Reading part # 39 : data JSON...done. params JSON...done.
 Reading part # 40 : data JSON...done. params JSON...done.
 Reading part # 41 : data JSON...done. params JSON...done.
 Reading part # 42 : data JSON...done. params JSON...done.
 Reading part # 43 : data JSON...done. params JSON...done.
 Reading part # 44 : data JSON...done. params JSON...done.
```

```
Reading part # 45 : data JSON...done. params JSON...done.
Reading part # 46 : data JSON...done. params JSON...done.
Reading part # 47 : data JSON...done. params JSON...done.
Reading part # 48 : data JSON...done. params JSON...done.
Reading part # 49 : data JSON...done. params JSON...done.
Reading part # 50 : data JSON...done. params JSON...done.
Reading part # 51 : data JSON...done. params JSON...done.
Reading part # 52 : data JSON...done. params JSON...done.
Reading part # 53 : data JSON...done. params JSON...done.
Reading part # 54 : data JSON...done. params JSON...done.
Reading part # 55 : data JSON...not available!
Reading part # 56 : data JSON...done. params JSON...done.
Reading part # 57 : data JSON...done. params JSON...done.
```

```
for i in data_dict:
  data_dict[i].to_csv(ULCA_DATASET_DIR + 'metadata_part_' + str(i) + '.csv', index=False)

for i in params_dict:
  params_dict[i].to_csv(ULCA_DATASET_DIR + 'params_part_' + str(i) + '.csv', index=False)


ulca_metadata_df = pd.read_csv(ULCA_DATASET_DIR + 'metadata_part_1.csv')
for i in range(2, 61):
  try:
    d = pd.read_csv(ULCA_DATASET_DIR + 'metadata_part_' + str(i) + '.csv')
    ulca_metadata_df = pd.concat([ulca_metadata_df, d])
  except:
    print('Skipping part', i, '...')

ulca_metadata_df.head(1000)
```

```
Skipping part 26 ...
Skipping part 30 ...
Skipping part 55 ...
```

	audioFilename	collectionSource	snr	duration	gender	
0	26_Regional-Chennai-Tamil-0645-20204671655.wav	['newsonair.nic.in', 'http://newsonair.nic.in/...	{'methodType': 'WadaSnr', 'methodDetails': {'s...	12.30	non-specified	
1	35_Regional-Chennai-Tamil-	['newsonair.nic.in',	{'methodType': 'WadaSnr',	7.89	non-	ஆலை

```
def oov_words(src,res):
  src_mod=src.split(' ')
  res_mod=res.split(' ')
  for word in src_mod:
    if word in res_mod and training_set:

ulca_test_results_df.apply(lambda x:oov_words(x['src'],x['res']))
```

3	Chennai-Tamil-	[newsonair.nic.in',	WadaSnr',	3.84	non-	ஒகே

```
ulca_params_df = pd.read_csv(ULCA_DATASET_DIR + 'params_part_1.csv')
for i in range(2, 61):
  try:
    d = pd.read_csv(ULCA_DATASET_DIR + 'params_part_' + str(i) + '.csv')
    ulca_params_df = pd.concat([ulca_params_df, d])
  except:
    print('Skipping part', i, '...')

l = [1]
def getAndIncCounter():
  d = l[0]
  l[0] += 1
  if l[0] == 26 or l[0] == 30 or l[0] == 55:
    l[0] += 1
  return d

ulca_params_df = ulca_params_df[['collectionSource', 'domain']].reset_index()
ulca_params_df['part'] = ulca_params_df.apply(lambda x : 'corpus' + str(getAndIncCounter()
ulca_params_df = ulca_params_df.reindex(columns=['part', 'collectionSource', 'domain'])

ulca_params_df['domain'] = ulca_params_df['domain'].apply(eval).apply(lambda x : x[0])
ulca_params_df['collectionSource'] = ulca_params_df['collectionSource'].apply(eval).apply(

ulca_params_df
```

```
Skipping part 26 ...
Skipping part 30 ...
Skipping part 55 ...
```

	part	collectionSource	domain
0	corpus1	newsonair.nic.in	news
1	corpus2	newsonair.nic.in	news
2	corpus3	newsonair.nic.in	news
3	corpus4	newsonair.nic.in	news
4	corpus5	newsonair.nic.in	news
5	corpus6	newsonair.nic.in	news
6	corpus7	newsonair.nic.in	news
7	corpus8	newsonair.nic.in	news
8	corpus9	newsonair.nic.in	news
9	corpus10	newsonair.nic.in	news
10	corpus11	newsonair.nic.in	news
11	corpus12	newsonair.nic.in	news
12	corpus13	newsonair.nic.in	news
13	corpus14	newsonair.nic.in	news
14	corpus15	newsonair.nic.in	news
15	corpus16	newsonair.nic.in	news
16	corpus17	newsonair.nic.in	news
17	corpus18	newsonair.nic.in	news
18	corpus19	newsonair.nic.in	news
19	corpus20	Dawah_Team_Tamil_Islamic_Bayans	religion
20	corpus21	Dawah_Team_Tamil_Islamic_Bayans	religion
21	corpus22	Dawah_Team_Tamil_Islamic_Bayans	religion
22	corpus23	Dawah_Team_Tamil_Islamic_Bayans	religion
23	corpus24	Ekam_Channel_Tamil	healthcare

```
ulca_params_df.to_csv(ULCA_DATASET_DIR + 'complete_metadata.csv', index=False)
```

```
params_dict = ulca_params_df.set_index('part').to_dict()

def getCollectionSource(partNo):
  return params_dict['collectionSource'][partNo]

def getDomain(partNo):
  return params_dict['domain'][partNo]
```

```
ulca_df['collection_source'] = ulca_df.apply(lambda x : getCollectionSource(str(x['old_loc
ulca_df['domain'] = ulca_df.apply(lambda x : getDomain(str(x['old_loc']).split('/')[2]), a
ulca_df.to_csv(TAMIL_DATASET_HOME + 'csv/complete_datasets/ulca_dataset.csv', index = Fals
print('CSV Length:', len(ulca_df.index))
ulca_df.head(10000)
```

```
CSV Length: 643377
```

	wav_filename	wav_filesize	transcript	
0	split_dataset/part1/a/48_7_237file-idkgTgcdjaG...	673964	ஆயிரத்தி எட்நூத்தி ஆறாம் ஆண்டு நடைபெற்ற வேலூர்...	ulca_co
1	split_dataset/part1/a/48_47_1038file-idl-tBAnF...	195884	அடுத்த கொஷ்டீன் வந்து நிவாஸ் பாபு	ulca_cor
2	split_dataset/part1/a/48_0_21_Regional-Chennai...	745964	தில்லிக்கு முழு மாநில அந்தஸ்து வழங்க கோரி பொது...	ulca_corp
3	split_dataset/part1/a/48_33_969file-id2-YxRL7M...	299564	சரி இது வந்து பிச்ச இல்லாம	ulca_co
4	split_dataset/part1/a/48_20_6_495file-idaN531X...	377324	மூணாவது வந்து கிராஸ் பப்ளிகேஷன்ஸ் வந்து எகனாமி...	ulca_cor
...	...	...	...	
9995	split_dataset/part3/b/48_12_Regional-Chennai-T...	244844	முதலமைச்சர் திரு எடப்பாடி பழனிசாமி கூறியுள்ளார்	ulca_corp
9996	split_dataset/part3/b/48_88_5376file-idgKsWAZe...	288044	குழந்தைகள் எப்படி வழங்கப்பட வேண்டும்	ulca_cor
9997	split_dataset/part3/b/48_6_125file-idr04m13ZxK...	187244	சென்னை மற்றும் புறநகரை பொறுத்தவரையில்	ulca_co
9998	split_dataset/part3/b/48_16_658file-idDzTLDRdh...	959084	அப்படி நடக்கும்போது இதோட மாட்டாள் லெவல் வந்து ...	ulca_co
9999	split_dataset/part3/b/48_1_59_NSD-Tamil-Tamil-...	276524	காவல்துறையைச் சேர்ந்த ஒருவர் வீர மரணம் அடைந்தார்	ulca_

10000 rows × 7 columns

```
df1 = ulca_test_result_df.copy()
df1 = df1.drop(columns=['transcript', 'collection_source','domain'])
df2 = pd.merge(df1, ulca_df, on='wav_filename', how='inner')
df2 = df2.drop(columns=['wav_filesize', 'old_loc'])
df2
```

	wav_filename	src	
0	split_dataset/part2/e/48_12_NSD-Tamil-Tamil-07...	ஆசியாவை மையப்படுத்தி பிராந்திய பாதுகாப்பு கட்ட...	மை பிராந்திய
1	split_dataset/part2/b/48_0_5_Regional-Chennai-...	திருநெல்வேலி தூத்துக்குடி தென்காசி மாவட்டங்களி...	திரு தூத்துக்குடி மாவ
2	split_dataset/part1/d/48_32_Regional-Chennai-T...	பள்ளிகள் இதோடு இணைக்கப்படுகின்றன குழந்தைகளுக்க...	பள்ளிக இணைக்கப்ட குழந்ை
3	split_dataset/part3/b/48_4_NSD-Tamil-Tamil-124...	கர்நாடகாவில் மூன்று மக்களவை மற்றும் இரண்டு சட்...	கர்நாடகா மக்களவ இ
4	split_dataset/part1/b/48_11_NSD-Tamil-Tamil-07...	இதேபோல மாநிலங்களவையில் ஆள்கடத்தல் தடுப்பு மசோத...	மாநிலங்க ஆள்கடத்
...	...	...	
643372	split_dataset/part151/a/48_52_1007file-idKcR_F...	சைக்கிள்	வாசக சைலன்ச
643373	split_dataset/part152/a/48_4_105_1854file-idbC...	மிஞ்சினால்	காரர்களாக
643374	split_dataset/part153/e/48_182_4765file-id4ei1...	தமிழ்ல பண்றாங்க	கட்சிய அவர்கள் க ப
643375	split_dataset/part151/e/48_86_666file-id7wVOqX...	இஸ்லாம்	எம்பி இஸ்லாமிக்
643376	split_dataset/part150/c/48_80_1475file-id0CSwl...	ஏற்கனவே	சட்டமாக்கப்ட புதுப்ப

643377 rows × 14 columns

```
ulca_test_result_df.to_csv("/content/drive/MyDrive/testProj/model_without_ulca/ulca_test_r
df2.to_csv("/content/drive/MyDrive/testProj/model_without_ulca/ulca_test_results_modified.
```

```
import pandas as pd
#Testing DFs
ulca_df = pd.read_csv(TAMIL_DATASET_HOME + 'csv/complete_datasets/ulca_dataset.csv')


#Test results DFs
ulca_test_result_df = pd.read_csv(MODEL_HOME + 'ulca_test_results.csv')

#Training DF = Train + Dev
train_df = pd.concat([pd.read_csv(TAMIL_DATASET_HOME + 'train-without-ulca.csv'), pd.read_


ulca_test_result_df['wav_filename'] = ulca_test_result_df.apply(lambda x : '/'.join(str(x[
ulca_test_result_df
```

wav_filename src

Importing Required CSV's

```
#importing required csvs
import pandas as pd
# oov_train_df = pd.read_csv(TAMIL_DATASET_HOME + 'unknown_training_words.csv')
# oov_lm_df = pd.read_csv(TAMIL_DATASET_HOME + 'unknown_lm_words.csv')
ulca_test_results_df = pd.read_csv('/content/drive/MyDrive/testProj/tests/ulca_tests/model
```

Chennai-T...

```
ulca_test_results_df.describe()
```

	loss	char_distance	char_length	word_distance	word_length	
count	643377.000000	643377.000000	643377.000000	643377.000000	643377.000000	643
mean	127.586900	30.956441	92.275941	7.034624	10.540674	
std	101.099414	26.486428	61.545662	5.133924	6.680130	
min	0.000000	0.000000	1.000000	0.000000	1.000000	
25%	57.539135	12.000000	42.000000	3.000000	5.000000	
50%	101.806374	24.000000	80.000000	6.000000	9.000000	
75%	167.983521	42.000000	130.000000	10.000000	15.000000	
max	1378.768799	269.000000	354.000000	47.000000	47.000000	

id7WVOqX...

```
ulca_test_results_df['modifiedRes'] = ulca_test_result_df.apply(lambda x : str(x['res']).r
ulca_test_results_df['modifiedSrc'] = ulca_test_result_df.apply(lambda x : str(x['src']).r
ulca_test_results_df['modifiedWER'] = ulca_test_result_df.apply(lambda x : 0 if str(x['mod
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-10-c1a0f06ccda3> in <module>()
----> 1 ulca_test_results_df['modifiedRes'] = ulca_test_result_df.apply(lambda x :
str(x['res']).replace(" ", ""), axis = 1)
      2 ulca_test_results_df['modifiedSrc'] = ulca_test_result_df.apply(lambda x :
str(x['src']).replace(" ", ""), axis = 1)
      3 ulca_test_results_df['modifiedWER'] = ulca_test_result_df.apply(lambda x : 0
if str(x['modifiedRes']) == str(x['modifiedSrc']) else x['wer'], axis = 1)

NameError: name 'ulca_test_result_df' is not defined
```

```
ulca_test_results_df
```

	wav_filename	src	
0	split_dataset/part2/e/48_12_NSD-Tamil-Tamil-07...	ஆசியாவை மையப்படுத்தி பிராந்திய பாதுகாப்பு கட்ட...	மை பிராந்திய
1	split_dataset/part2/b/48_0_5_Regional-Chennai-...	திருநெல்வேலி தூத்துக்குடி தென்காசி மாவட்டங்களி...	திரு தூத்துக்குடி மாவ
2	split_dataset/part1/d/48_32_Regional-Chennai-T...	பள்ளிகள் இதோடு இணைக்கப்படுகின்றன குழந்தைகளுக்க...	பள்ளிக இணைக்கப்ட குழந்த
3	split_dataset/part3/b/48_4_NSD-Tamil-Tamil-124...	கர்நாடகாவில் மூன்று மக்களவை மற்றும் இரண்டு சட்...	கர்நாடகா மக்களவ இ
4	split_dataset/part1/b/48_11_NSD-Tamil-Tamil-07...	இதேபோல மாநிலங்களவையில் ஆள்கடத்தல் தடுப்பு மசோத...	மாநிலங்க ஆள்கடத்
...	...	...	
643372	split_dataset/part151/a/48_52_1007file-idKcR_F...	சைக்கிள்	வாசக சைலன்ச
643373	split_dataset/part152/a/48_4_105_1854file-idbC...	மிஞ்சினால்	காரர்களாக
643374	split_dataset/part153/e/48_182_4765file-id4ei1...	தமிழ்ல பண்றாங்க	கட்சிய அவர்கள் க ப
643375	split_dataset/part151/e/48_86_666file-id7wVOqX...	இஸ்லாம்	எம்பி இஸ்லாமிக்
643376	split_dataset/part150/c/48_80_1475file-id0CSwl...	ஏற்கனவே	சட்டமாக்கப்ட புதுப்ப

643377 rows × 14 columns

Python Code for Calculating WER

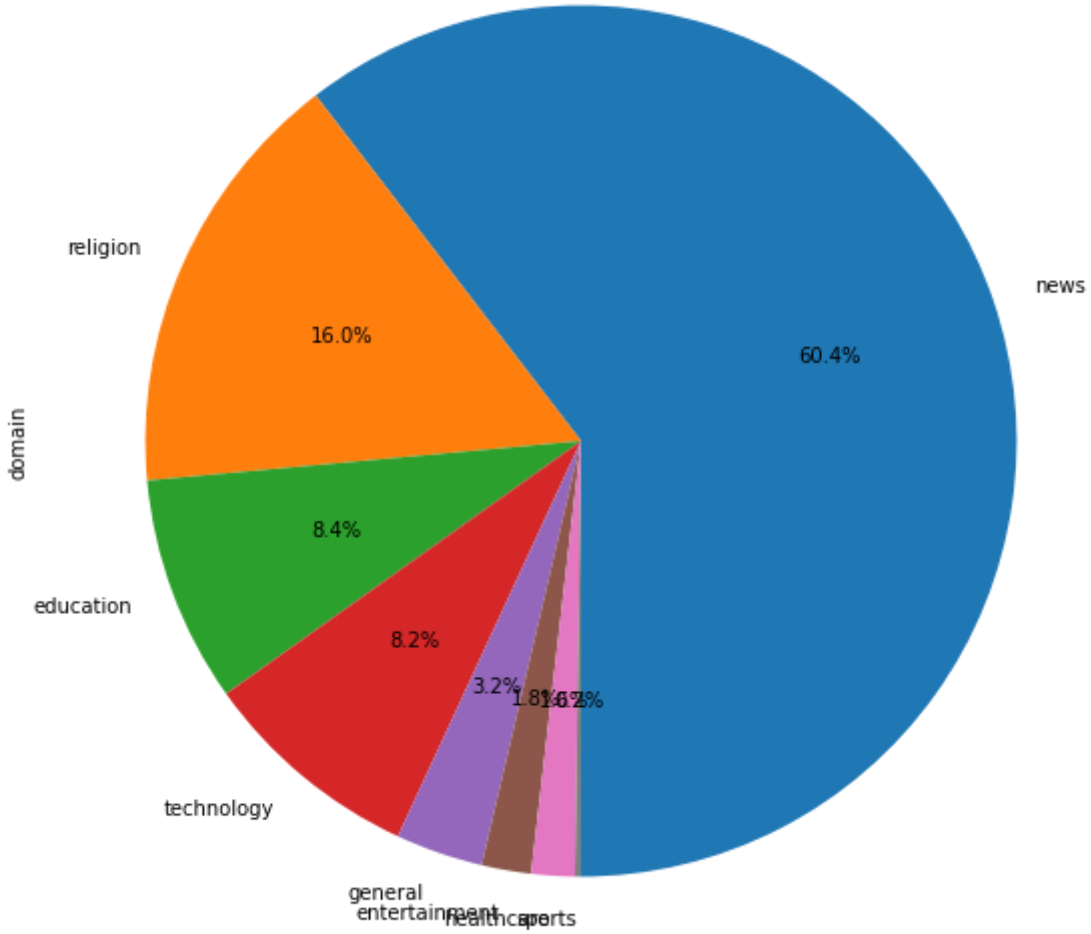
[ ] ↳ *1 cell hidden*

Analyzing the generated data

ULCA Data Analysis (WER, CER Pie charts)

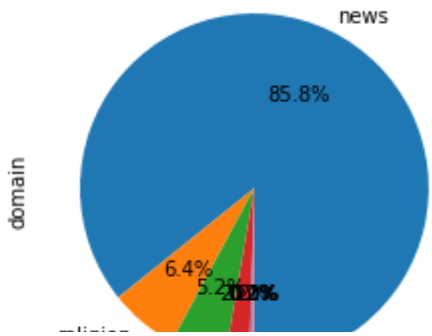
```
#Understanding how much of our dataset is of what type data
df = ulca_test_results_df
print(df['domain'].value_counts(normalize=True))
df['domain'].value_counts().plot(kind='pie', autopct='%1.1f%%', startangle=270, figsize=(1
```

```
news             0.604229
religion         0.160225
education        0.084060
technology       0.082451
general          0.032385
entertainment    0.018355
healthcare       0.016081
sports           0.002215
Name: domain, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7fd398d97490>
```



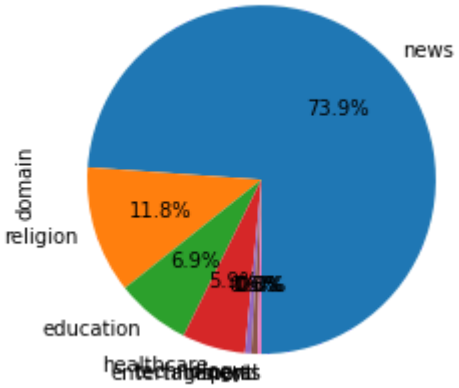

```
#Pie chart to show CER
df = ulca_test_results_df[ulca_test_results_df['cer'] <=0.2]
print(df['domain'].value_counts())
df['domain'].value_counts().plot(kind='pie', autopct='%1.1f%%', startangle=270)
```

```
news             161764
religion          12016
education          9731
healthcare         4215
technology          356
entertainment       355
general             203
sports                6
Name: domain, dtype: int64
<matplotlib.axes._subplots.AxesSubplot at 0x7f8ec6e77f90>
```




```
df = ulca_test_results_df[ulca_test_results_df['cer'] == 0]
print(df['domain'].value_counts())
df['domain'].value_counts().plot(kind='pie', autopct='%1.1f%%', startangle=270)
```

```
news             15475
religion          2474
education         1455
healthcare        1227
technology         122
entertainment      122
general             73
sports               2
Name: domain, dtype: int64
<matplotlib.axes._subplots.AxesSubplot at 0x7fef0ffa8e90>
```




```
ulca_test_results_df['durationSec'].corr(ulca_test_results_df['wer'])
```

```
-0.1730614399471933
```

Overall WER and CER Analysis

[ ] ↳ *6 cells hidden*

News Analysis

[ ] ↳ *16 cells hidden*

Religion Analysis

[ ] ↳ *16 cells hidden*

Education Analysis

[ ] ↳ *15 cells hidden*

Healthcare Analysis

[ ] ↳ *15 cells hidden*

Technology Analysis

[ ] ↳ *15 cells hidden*

Entertainment Analysis

[ ] ↳ *15 cells hidden*

General Domain Analysis

[ ] ↳ *15 cells hidden*

Sports Analysis

[ ] ↳ *15 cells hidden*

OOV Words Analysis

Bar Graphs

[ ] ↳ *6 cells hidden*

%OOV across all domains in AM(Training data)

```
#Formula used: (#OOV Words in training data for one domain)/(Total #OOV Words across all d
arr = sports_training_output_df['unknown_word'].unique()
arr1= news_training_output_df['unknown_word'].unique()
arr2 = religion_training_output_df['unknown_word'].unique()
arr3 = edu_training_output_df['unknown_word'].unique()
arr4 = tech_training_output_df['unknown_word'].unique()
arr5 = health_training_output_df['unknown_word'].unique()
arr6 = general_training_output_df['unknown_word'].unique()
arr7 = entertainment_training_output_df['unknown_word'].unique()
sports_oov_cnt=len(arr)/(1161.07)
news_oov_cnt=len(arr1)/(1161.07)
religion_oov_cnt=len(arr2)/(1161.07)
education_oov_cnt=len(arr3)/(1161.07)
tech_oov_cnt=len(arr4)/(1161.07)
health_oov_cnt=len(arr5)/(1161.07)
general_oov_cnt=len(arr6)/(1161.07)
entertainment_oov_cnt=len(arr7)/(1161.07)
print(news_oov_cnt,religion_oov_cnt,education_oov_cnt,general_oov_cnt,tech_oov_cnt,health_
```

```
82.81929599421224 40.23874529528797 28.925904553558357 10.548028973274652 17.42702851
```

%OOV across all domains in LM

```
#Formula used: (#OOV Words in LM data for one domain)/(Total #OOV Words across all domains
arr = sports_lm_output_df['unknown_word'].unique()
arr1= news_lm_output_df['unknown_word'].unique()
arr2 = religion_lm_output_df['unknown_word'].unique()
arr3 = edu_lm_output_df['unknown_word'].unique()
arr4 = tech_lm_output_df['unknown_word'].unique()
arr5 = health_lm_output_df['unknown_word'].unique()
arr6 = general_lm_output_df['unknown_word'].unique()
arr7 = entertainment_lm_output_df['unknown_word'].unique()
sports_oov_cnt=len(arr)/(0.76)
news_oov_cnt=len(arr1)/(0.76)
religion_oov_cnt=len(arr2)/(0.76)
education_oov_cnt=len(arr3)/(0.76)
tech_oov_cnt=len(arr4)/(0.76)
health_oov_cnt=len(arr5)/(0.76)
general_oov_cnt=len(arr6)/(0.76)
entertainment_oov_cnt=len(arr7)/(0.76)
print(news_oov_cnt,religion_oov_cnt,education_oov_cnt,general_oov_cnt,tech_oov_cnt,health_
```

```
30.263157894736842 42.10526315789474 44.73684210526316 28.94736842105263 53.947368421
```

OOV Words correctly transcribed

```
#List of OOV words correctly transcribed.
oov_train = set(pd.read_csv(TAMIL_DATASET_HOME + 'unknown_training_words.csv')['unknown_wo
oov_lm = set(pd.read_csv(TAMIL_DATASET_HOME + 'unknown_lm_words.csv')['unknown_word'].uniq
```

```
correctlyTranscribedAMWords = dict()
correctlyTranscribedLMWords = dict()
totalWordsCount = 0

def checkCorrectOOVWord(src, res):
  global correctlyTranscribedAMWords, correctlyTranscribedLMWords, totalWordsCount

  srcList = src.split(' ')
  res = res.replace(" ", "")

  for word in srcList:
    if len(word) <= 0:
      continue

    totalWordsCount += 1

    if word in oov_train and word in res: #OOV word, there in res also
      if word in correctlyTranscribedAMWords:
        correctlyTranscribedAMWords[word] += 1
      else:
        correctlyTranscribedAMWords[word] = 1

    if word in oov_lm and word in res:
      if word in correctlyTranscribedLMWords:
        correctlyTranscribedLMWords[word] += 1
      else:
        correctlyTranscribedLMWords[word] = 1

ulca_test_results_df.apply(lambda x : checkCorrectOOVWord(str(x['src']), str(x['res'])), a
print('Correctly Transcribed OOV Words count:')
print('AM: ', len(correctlyTranscribedAMWords), '/', len(oov_train), ' OOV words (', round
print('LM: ', len(correctlyTranscribedLMWords), '/', len(oov_lm), ' OOV words (', round((l
print('Correctly Transcribed AM Words({word : # of instances of correct transcription}):\n
print('\nCorrectly Transcribed LM Words({word : # of instances of correct transcription}):
print('Total Words scanned:', totalWordsCount)
```

```
Correctly Transcribed OOV Words count:
AM: 64850/150206 OOV words (43.17%)
LM: 4/76 OOV words (5.26%).
Correctly Transcribed AM Words({word : # of instances of correct transcription}):
 {'ஆசியாவை': 2, 'சுஷ்மா': 253, 'சுவராஜ்': 65, 'தென்காசி': 164, 'பாதிப்புகள

Correctly Transcribed LM Words({word : # of instances of correct transcription}):
 {'கன்சர்வேட்': 1, 'சட்டப்போர': 1, 'ஆறாயிரத்தி': 1, 'ஈடுபடுத்தப்படுகிற': 1}
Total Words scanned: 6781627
```

	Unnamed: 0	wav_filename	transcript	wav_filesize	
0	0	split_dataset/part154/a/TA0421-TA0423_1-A.022.wav	ஆனா	30444	/c
1	1	split_dataset/part154/a/TA0727-TA0728_2-A.110.wav	உண்மைதான் அது உண்மைத்தன்	60460	/c
2	2	split_dataset/part154/a/TA0615-TA0616_2-B.022.wav	அய்யயோ என்ன என்ன வயசு இருக்கும் உங்களுக்கு	75628	/c
3	3	split_dataset/part154/a/TA0625-TA0626_1-A.401.wav	இல்லை	19244	/c
4	4	split_dataset/part154/a/TA0574-TA0576_2-B.193.wav	சும்மா நம்ப எடுத்து	42828	/c
...	...	...	...	...	
42206	42207	split_dataset/part161/e/000080039.wav	காங்கேசன் துறையில் ஆரம்பித்து காலி வரை இலங்கை ...	253804	/c
42207	42208	split_dataset/part161/e/000020172.wav	தமிழன் உயிரை விட்டும் அவனுக்கு கிடைக்காத நியாய...	228204	/c

```
#Rank of Correctly transcribed AM OOV Words in LM Text
rankdict=dict()
cnt=1
with open(LM_PATH + 'intermediates/old/vocab-5000000.txt') as f:
  for word in f:
    word=word.strip()
    if word in correctlyTranscribedAMWords:
      rankdict[word]=cnt



    cnt+=1
print(rankdict)


#print(len(correctlyTranscribedAMWordsSet))
```

```
{'இ': 423, 'இதுகுறித்து': 429, 'ஜ': 519, 'தொடர்பில்': 553, 'பங்கேற்றனர்': 760,
```

```
c=0
for val in rankdict.values():
  if(val<=500000):
    c+=1
print(c)
```

```
63026
```

```
rankdictset=set()
for word in rankdict:
  rankdictset.add(word)
```

```
correctlyTranscribedAMWordsSet=set()
for word in correctlyTranscribedAMWords:
  correctlyTranscribedAMWordsSet.add(word)
for word in correctlyTranscribedAMWordsSet:
  if not(word in rankdictset):
    print(word)
```

```
ஆறாயிரத்தி
ஈடுபடுத்தப்படுகிற
சட்டப்போர
கன்சர்வேட்
```

```
print(len(rankdictset))
```

```
64846
```

```
result_set=lm_words_set.difference(rankdictset)
print(len(result_set))
```

```
4935154
```

```
print(len(rankdict))
```

```
64846
```

```
#List of OOV words correctly transcribed in Intersection-set(AM and ulca)
correctlyTranscribedWords = dict()

totalWordsCount = 0

def checkCorrectOOVWord(src, res):
  global correctlyTranscribedWords, totalWordsCount

  srcList = src.split(' ')
  res = res.replace(" ", "")
```

```
  for word in srcList:
    if len(word) <= 0:
      continue

    totalWordsCount += 1

    if word in intersection_set and word in res: #OOV word, there in res also
      if word in correctlyTranscribedWords:
        correctlyTranscribedWords[word] += 1
      else:
        correctlyTranscribedWords[word] = 1



ulca_test_results_df.apply(lambda x : checkCorrectOOVWord(str(x['src']), str(x['res'])), a
print('Correctly Transcribed Words in intersection of ULCA and AM count:')
print('AM: ', len(correctlyTranscribedWords), '/', len(intersection_set), ' Words (', roun
print('Correctly Transcribed in intersection of ULCA and AM count({word : # of instances o
print('Total Words scanned:', totalWordsCount)
```

```
Correctly Transcribed Words in intersection of ULCA and AM count:
AM: 35770/50866 Words (70.32%)
Correctly Transcribed in intersection of ULCA and AM count({word : # of instances of
 {'மையப்படுத்தி': 15, 'பிராந்திய': 182, 'பாதுகாப்பு': 5558, 'கட்டமைப்பை': 111
Total Words scanned: 6781627
```

Training Words Analysis

```
#List of training words not correctly transcribed.

incorrectlyTranscribedWords = dict()
uniqueWordSet = set()
totalWordsCount = 0

def checkIncorrectTrainTranscription(src, res):
  global incorrectlyTranscribedWords, totalWordsCount

  srcList = src.split(' ')
  res = res.replace(" ", "")

  totalWordsCount += len(srcList)

  for word in srcList:
    uniqueWordSet.add(word)

    if word in train_words_set and not(word in res):
      if word in correctlyTranscribedAMWords:
        incorrectlyTranscribedWords[word] += 1
      else:
        incorrectlyTranscribedWords[word] = 1

ulca_test_results_df.apply(lambda x : checkIncorrectTrainTranscription(str(x['src']), str(
```

```
print('Incorrectly Transcribed Training Words count: ', len(incorrectlyTranscribedWords),
print('List ({word : # of instances of incorrect transcription}):\n', incorrectlyTranscrib
print('Total Words scanned: ', totalWordsCount, '(', len(uniqueWordSet), ' unique).', sep=
```

```
Incorrectly Transcribed Training Words count: 46806/116107 (40.31%)
List ({word : # of instances of incorrect transcription}):
 {'இந்திய': 1, 'இரண்டாயிரத்தி': 1, 'ரிசர்வ்': 1, 'பிறகே': 1, 'ஒருங்கிணைப்பு':
Total Words scanned: 6781627(201072 unique).
```

```
df1
```

	wav_filename	src	
0	split_dataset/part2/e/48_12_NSD-Tamil-Tamil-07...	ஆசியாவை மையப்படுத்தி பிராந்திய பாதுகாப்பு கட்ட...	அ மை பிராந்திய
1	split_dataset/part2/b/48_0_5_Regional-Chennai-...	திருநெல்வேலி தூத்துக்குடி தென்காசி மாவட்டங்களி...	திரு தூத்துக்குடி மாவ
2	split_dataset/part1/d/48_32_Regional-Chennai-T...	பள்ளிகள் இதோடு இணைக்கப்படுகின்றன குழந்தைகளுக்க...	பள்ளிக இணைக்கப்ப குழந்த
3	split_dataset/part3/b/48_4_NSD-Tamil-Tamil-124...	கர்நாடகாவில் மூன்று மக்களவை மற்றும் இரண்டு சட்...	கர்நாடகா மக்களை இர
4	split_dataset/part1/b/48_11_NSD-Tamil-Tamil-07...	இதேபோல மாநிலங்களவையில் ஆள்கடத்தல் தடுப்பு மசோத...	மாநிலங்க ஆள்கடத்
...	...	...	
630918	split_dataset/part152/b/48_11_197file-idA0_9NN...	அவர்கள் வந்து	அவ
630919	split_dataset/part153/c/48_185_2038file-idiVZ1...	அதனால என்னை	அதனா
630920	split_dataset/part152/b/48_1_355file-iduhOX5JQ...	தியானம் என்பது	தியான
630921	split_dataset/part150/e/48_0_21_Regional-Chenn...	இசையும்	
630922	split_dataset/part151/a/48_3_11_1348file-idl6g...	எல்லா மனிதர்களும்	எல்லா ம

21066 rows × 14 columns

```
df2[(df2['wer'] > 0) & (df2['durationSec'] >= 3) & (df2['durationSec'] <= 12)]
```

	wav_filename	src	
414	split_dataset/part2/a/48_15_NSD-Tamil-Tamil-12...	இரண்டாயிரத்தி பதினாறாம் ஆண்டு நாடாளுமன்றத்தில்...	இரண்டா பதினாறாம் நாடாளுமன்ற
415	split_dataset/part2/a/48_47_1959file-idbpZW8yR...	கடந்த ஐந்து நாட்களாக பாதிக்கப்பட்ட இரண்டு லட்ச...	கடந்த ஐந்து நா பாதிக்கப்பட்ட
416	split_dataset/part1/c/48_5_NSD-Tamil-Tamil-124...	மாநிலங்களை கலந்தாலோசித்து ஒருமித்த கருத்து ஏற்...	மாநில கலந்தாலே ஒருமித்த கருத்
417	split_dataset/part2/d/48_0_20_Regional-Chennai...	சபரிமலை ஐயப்பன் கோவிலில் அனைத்து வயது பெண்களைய...	சபரிமலை ஐ கோவிலில் அ வயது பெண்க
418	split_dataset/part1/e/48_5_20_Regional-Chennai...	தொழில்துறை அமைப்பான அசோசெம் நூற்றாண்டு விழா கொ...	தொழி அமை அசோசெம் நூற் விழ
...	...	...	
643371	split_dataset/part153/b/48_23_1694file-idfdUlF...	ஆரம்பிக்க	கரைத்துக்குடித்த தரவின்ப
643372	split_dataset/part151/a/48_52_1007file-idKcR_F...	சைக்கிள்	வாசகசாலை சைலன்சர்கள்
643373	split_dataset/part152/a/48_4_105_1854file-idbC...	மிஞ்சினால்	ட காரர்களாகவு
643374	split_dataset/part153/e/48_182_4765file-id4ei1...	தமிழ்ல பண்றாங்க	கட்சியின அவர்கள் கடப்ப பவர்
643375	split_dataset/part151/e/48_86_666file-id7wVOqX...	இஸ்லாம்	எம்பிலிட் இஸ்லாமிக் கல் நடி

441154 rows × 14 columns

```
df1
```

	wav_filename	src	
0	split_dataset/part2/e/48_12_NSD-Tamil-Tamil-07...	ஆசியாவை மையப்படுத்தி பிராந்திய பாதுகாப்பு கட்ட...	அ மை பிராந்திய
1	split_dataset/part2/b/48_0_5_Regional-Chennai-...	திருநெல்வேலி தூத்துக்குடி தென்காசி மாவட்டங்களி...	திரு தூத்துக்குடி மாவ
2	split_dataset/part1/d/48_32_Regional-Chennai-T...	பள்ளிகள் இதோடு இணைக்கப்படுகின்றன குழந்தைகளுக்க...	பள்ளிக இணைக்கப்ப குழந்தை
3	split_dataset/part3/b/48_4_NSD-Tamil-Tamil-124...	கர்நாடகாவில் மூன்று மக்களவை மற்றும் இரண்டு சட்...	கர்நாடகா மக்களன இர
4	split_dataset/part1/b/48_11_NSD-Tamil-Tamil-07...	இதேபோல மாநிலங்களவையில் ஆள்கடத்தல் தடுப்பு மசோத...	இ மாநிலங்க ஆள்கடத்த
...	...	...	
630918	split_dataset/part152/b/48_11_197file-idA0_9NN...	அவர்கள் வந்து	அவ
630919	split_dataset/part153/c/48_185_2038file-idiVZ1...	அதனால என்னை	அதனா
630920	split_dataset/part152/b/48_1_355file-iduhOX5JQ...	தியானம் என்பது	தியான
630921	split_dataset/part150/e/48_0_21_Regional-Chenn...	இசையும்	
630922	split_dataset/part151/a/48_3_11_1348file-idl6g...	எல்லா மனிதர்களும்	எல்லா ம

21066 rows × 14 columns

```
df1 = ulca_test_results_df[ulca_test_results_df['wer'] == 0]
df2 = ulca_test_results_df
df2[~df2['wav_filename'].isin(df1['wav_filename'])]
```

	wav_filename	src	
413	split_dataset/part2/e/48_1_8_NSD-Tamil-Tamil-1...	மக்களுக்கு ஆரோக்கியம் மற்றும் சுகாதாரம் பற்றிய...	மக் ஆரோக்கியம் சுகாதாரம் ப
414	split_dataset/part2/a/48_15_NSD-Tamil-Tamil-12...	இரண்டாயிரத்தி பதினாறாம் ஆண்டு நாடாளுமன்றத்தில்...	இரண்ட பதினாறாம் நாடாளுமன்
415	split_dataset/part2/a/48_47_1959file-idbpZW8yR...	கடந்த ஐந்து நாட்களாக பாதிக்கப்பட்ட இரண்டு லட்ச...	கடந்த ஐந்து நா பாதிக்கப்பட்ட
416	split_dataset/part1/c/48_5_NSD-Tamil-Tamil-124...	மாநிலங்களை கலந்தாலோசித்து ஒருமித்த கருத்து ஏற்...	மாநில கலந்தா ஒருமித்த கரு
417	split_dataset/part2/d/48_0_20_Regional-Chennai...	சபரிமலை ஐயப்பன் கோவிலில் அனைத்து வயது பெண்களைய...	சபரிமலை ஐ கோவிலில் அ வயது பெண்க
...	...	...	
643372	split_dataset/part151/a/48_52_1007file-idKcR_F...	சைக்கிள்	வாசகசா சைலன்சர்க
643373	split_dataset/part152/a/48_4_105_1854file-idbC...	மிஞ்சினால்	கார்களாகவு
643374	split_dataset/part153/e/48_182_4765file-id4ei1...	தமிழ்ல பண்றாங்க	கட்சியின அவர்கள் கடப்

```
df2['durationSec'].mean()
```

```
5.766138655873461
```

With ULCA Scorer

```
msr_output = pd.read_json("/content/drive/MyDrive/testProj/model_without_ulca/msr_results_
msr_output_orig = pd.read_json("/content/drive/MyDrive/testProj/model_without_ulca/msr_tes
```

```
msr_output
```

	wav_filename	src	res	loss
0	/content/drive/MyDrive/testProj/datasets/ta/sp...	அந்த கடிதத்துக்கு அன்றே அவர் பதில் கடிதம் எழுத...	அந்த கடிதத்துக்கு அன்றே அவர் பதில் கடிதம் எழுத...	79.786972
1	/content/drive/MyDrive/testProj/datasets/ta/sp...	இந்த	இந்த	8.690358
2	/content/drive/MyDrive/testProj/datasets/ta/sp...	அந்த	அந்த	7.919152
3	/content/drive/MyDrive/testProj/datasets/ta/sp...	அந்த	அந்த	5.868518
4	/content/drive/MyDrive/testProj/datasets/ta/sp...	அந்த	அந்த	5.838010
...	...	...	...	...
42206	/content/drive/MyDrive/testProj/datasets/ta/sp...	கண்டிப்பா	அந்த பா	24.781662
42207	/content/drive/MyDrive/testProj/datasets/ta/sp...	அமான்டா	ஆனால் தான்	22.945990
42208	/content/drive/MyDrive/testProj/datasets/ta/sp...	கண்டிப்பா	அங்கு பா	22.524199
42209	/content/drive/MyDrive/testProj/datasets/ta/sp...	அடடா	வா வா	21.889162

```
print(msr_output['wer'].value_counts().sort_index())
```

```
0.000000       10
0.125000        1
0.142857        2
0.222222        1
0.230769        1
              ...
0.952381        1
1.000000    39351
1.111111        1
1.500000        1
2.000000       11
Name: wer, Length: 68, dtype: int64
```

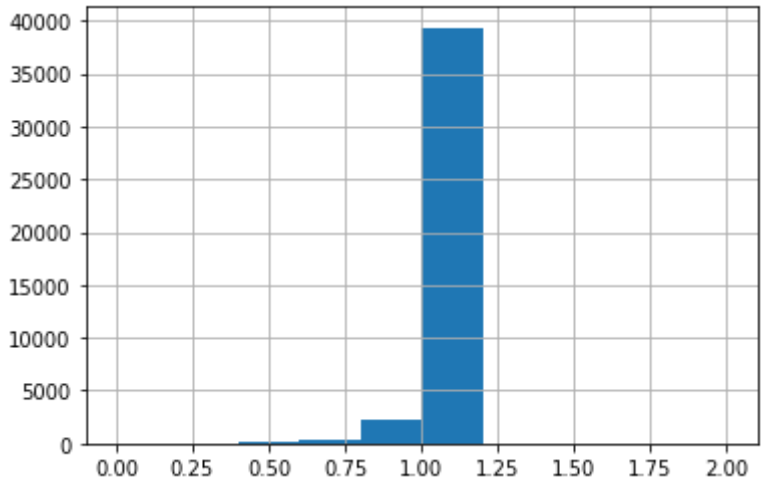
```
print(msr_output_orig['wer'].value_counts().sort_index())
```

```
0.000000       11
0.272727        1
0.285714        1
0.375000        1
0.384615        1
0.428571        5
0.444444        1
0.461538        1
0.500000       26
0.555556        2
0.571429       10
0.583333        2
0.600000        1
0.625000        9
0.636364        2
```

```
0.666667       28
0.692308        3
0.700000       10
0.714286       40
0.727273        8
0.750000       72
0.764706        1
0.769231        3
0.777778       26
0.785714        2
0.789474        1
0.800000       53
0.812500        3
0.818182       31
0.833333       50
0.846154       14
0.850000        1
0.857143      234
0.866667       10
0.875000      251
0.882353        3
0.888889      184
0.894737        3
0.900000      164
0.909091      126
0.916667      112
0.923077       63
0.928571       54
0.933333       29
0.937500       28
0.941176       23
0.944444       11
0.947368        9
0.950000       11
0.952381        3
1.000000    40463
1.111111        1
2.000000        9
Name: wer, dtype: int64
```

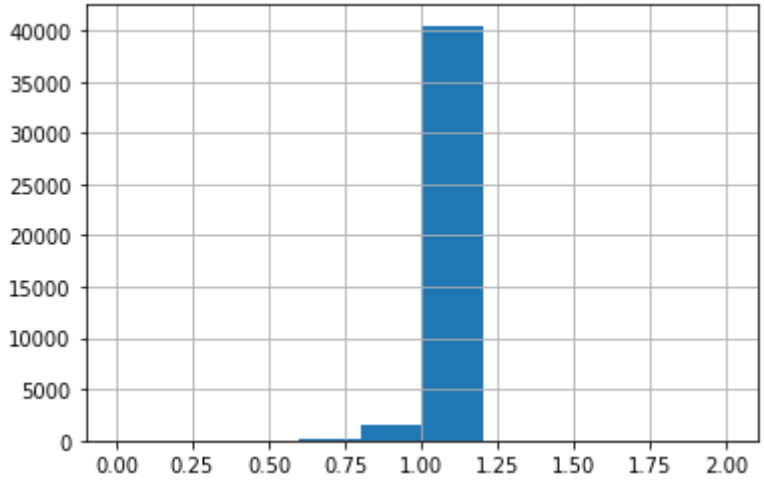
```
msr_output['wer'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f231b29a8d0>
```



```
msr_output_orig['wer'].hist()
```

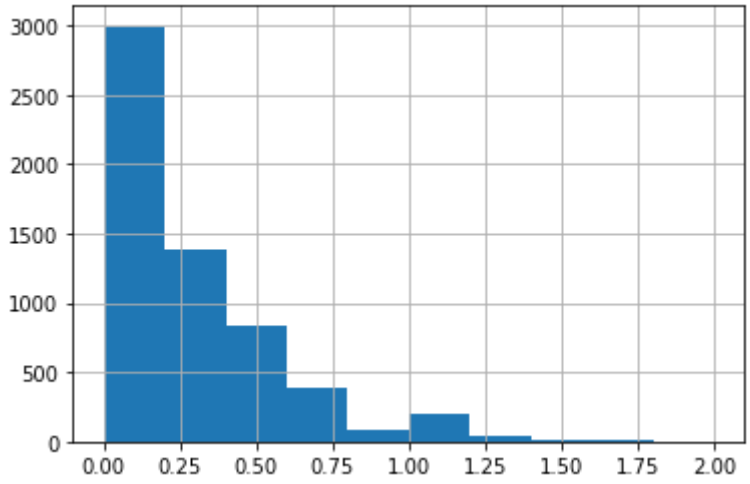
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f23209a4410>
```




```
test_df_orig = pd.read_json("/content/drive/MyDrive/testProj/model_without_ulca/openslr+cv
test_df = pd.read_json("/content/drive/MyDrive/testProj/model_without_ulca/openslr+cv_resu
```

```
test_df_orig['wer'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f232068fc10>
```




```
test_df['wer'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f231b82f550>
```

```
orig_words = set()
ulca_words = set()

with open("/content/drive/MyDrive/testProj/lm/ulca-tamil-vocab-cleaned.txt") as f:
  for line in f:
    ulca_words.add(line.strip())

print('Len(ULCA):', len(ulca_words))

with open("/content/drive/MyDrive/testProj/lm/intermediates/old/vocab-5000000.txt") as f:
  for line in f:
    orig_words.add(line.strip())

print('Len(Orig):', len(orig_words))
```

```
Len(ULCA): 500000
Len(Orig): 5000000
```

```
print(len(orig_words - ulca_words))
print(len(ulca_words - orig_words))
```

```
4503089
3089
```