

# Fathom Assignment Documentation

## 1)Extraction of audio :

For the extraction of audio from those video files I've used a project named "ffmpeg", this worked pretty good along with python. A library called "ffmpeg" in this project is commonly used to handle the multimedia stuff with python, so basically a simple function with a for loop and an if condition along with a couple of imports is gonna handle the the task of extracting the audio(in the wave format). I've extracted and stored the audio in the wave file format.

Main Library's used : ffmpeg

## 2)Audio to Text conversion :

Now to convert the a wave file to text, I've made use of Google Speech-to-Text translation API, but I had encountered a problem with that, it is only working for a audio file of length approximately 15 sec but the files I had are of length 1-1.5hrs, so what I did is, I trimmed a single video to 15 sec and sent that trimmed 15 sec file for translation and then I appended the contents returned by Google to a text file, I repeated this process for all the audio files I had, but then I had encountered another problem that is, upto what length I had to trim a file since all the audio files are of different lengths, so then what I did is initially I've found the duration of an audio file and then trimmed that file until duration end-point is reached, and when the end-point is reached conversion for that audio file is completed, than I had switched to another audio file in the same directory(the script I had written takes all the files in the current directory and filters the wave format files from the, we can change that by simply changing the path for listing the files).

Main library's used :ffmpeg, SpeechRecognition, contextlib

### 3)Cluster/topic mining :

Now after all the audio files are converted to text files then I appended all the text files into one and named it the master file, then I had created a lexicon using the master file, then from the master file I had read 100 bytes of data at a time and then tokenized, lemmatized and vectorized the data in that 100 bytes and added that vector to a list, I repeated this process until the end of the file and when end is reached I broke out of the loop. At the end of this process the whole file is converted to a list of vectors, then I fitted this list to a “Mean Shift” classifier, which in turn classified the the text file to total number of topics(I had used some common sense in choosing the bandwidth, so it may be varied according to the content we are dealing with).

### 4)Assigning Human-readable labels:

To assign human-readable labels to the mined clusters I made use of “Term Frequency - Inverse Document Frequency(TF-IDF)”. It’s a way to score the importance of words in a document based on how frequently they appear across multiple documents.

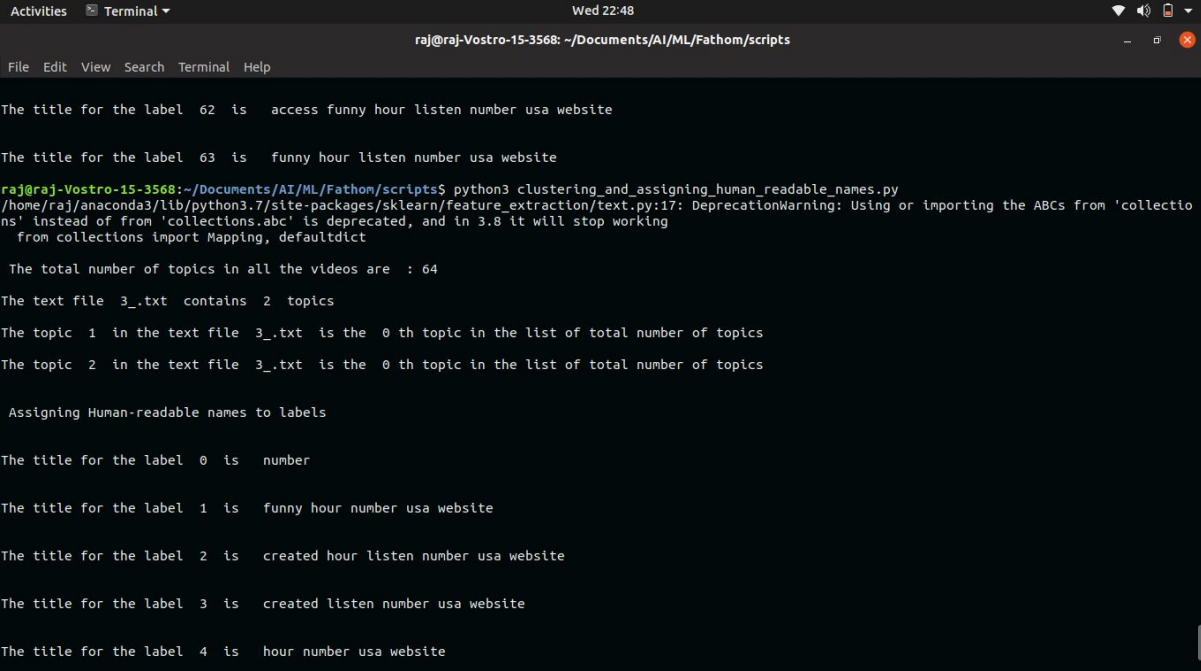
Initially I found term frequency, which is how often does a word appears in a document. Then I found the Inverse Document Frequency, which is how often the word is found in other documents.

Then I’ve multiplied them both to get the TF-IDF score of each vector in a cluster, then I found the vector with maximum TF-IDF score in each cluster. Then the selected vector in each cluster is transferred into a sentence by mapping the vector with lexicon through which it was vectorized. The results are not so satisfying, since the bandwidth, removal of stop words has to be optimized according to the size of the datasets being used.

# Demonstrating the results

In this section I'm gonna share some of the screenshots of the results I've obtained.

## 1. Firstly the total number of topics present in all the videos

A terminal window titled 'Terminal' with a dark background. The window shows the output of a Python script. The output includes several lines of text: 'The title for the label 62 is access funny hour listen number usa website', 'The title for the label 63 is funny hour listen number usa website', a deprecation warning about 'collections.abc', 'The total number of topics in all the videos are : 64', 'The text file 3\_.txt contains 2 topics', 'The topic 1 in the text file 3\_.txt is the 0 th topic in the list of total number of topics', 'The topic 2 in the text file 3\_.txt is the 0 th topic in the list of total number of topics', 'Assigning Human-readable names to labels', 'The title for the label 0 is number', 'The title for the label 1 is funny hour number usa website', 'The title for the label 2 is created hour listen number usa website', 'The title for the label 3 is created listen number usa website', and 'The title for the label 4 is hour number usa website'.

```
Activities Terminal Wed 22:48
raj@raj-Vostro-15-3568: ~/Documents/AI/ML/Fathom/scripts

File Edit View Search Terminal Help

The title for the label 62 is access funny hour listen number usa website

The title for the label 63 is funny hour listen number usa website

raj@raj-Vostro-15-3568:~/Documents/AI/ML/Fathom/scripts$ python3 clustering_and_assigning_human_readable_names.py
/home/raj/anaconda3/lib/python3.7/site-packages/sklearn/feature_extraction/text.py:17: DeprecationWarning: Using or importing the ABCs from 'collections' instead of from 'collections.abc' is deprecated, and in 3.8 it will stop working
  from collections import Mapping, defaultdict

The total number of topics in all the videos are : 64

The text file 3_.txt contains 2 topics

The topic 1 in the text file 3_.txt is the 0 th topic in the list of total number of topics
The topic 2 in the text file 3_.txt is the 0 th topic in the list of total number of topics

Assigning Human-readable names to labels

The title for the label 0 is number

The title for the label 1 is funny hour number usa website

The title for the label 2 is created hour listen number usa website

The title for the label 3 is created listen number usa website

The title for the label 4 is hour number usa website
```

## 2. Secondly, the number of topics present in a single video(converted to text file) or text file.

```
Activities Terminal Wed 22:48
raj@raj-Vostro-15-3568: ~/Documents/AI/ML/Fathom/scripts

The title for the label 62 is access funny hour listen number usa website

The title for the label 63 is funny hour listen number usa website

raj@raj-Vostro-15-3568:~/Documents/AI/ML/Fathom/scripts$ python3 clustering_and_assigning_human_readable_names.py
/home/raj/anaconda3/lib/python3.7/site-packages/sklearn/feature_extraction/text.py:17: DeprecationWarning: Using or importing the ABCs from 'collections' instead of from 'collections.abc' is deprecated, and in 3.8 it will stop working
  from collections import Mapping, defaultdict

The total number of topics in all the videos are : 64

The text file 3_.txt contains 2 topics

The topic 1 in the text file 3_.txt is the 0 th topic in the list of total number of topics
The topic 2 in the text file 3_.txt is the 0 th topic in the list of total number of topics

Assigning Human-readable names to labels

The title for the label 0 is number

The title for the label 1 is funny hour number usa website

The title for the label 2 is created hour listen number usa website

The title for the label 3 is created listen number usa website

The title for the label 4 is hour number usa website
```

3.out of total number of topics, which topic belongs to which video

```
Activities Terminal Wed 22:48
raj@raj-Vostro-15-3568: ~/Documents/AI/ML/Fathom/scripts

The title for the label 62 is access funny hour listen number usa website

The title for the label 63 is funny hour listen number usa website

raj@raj-Vostro-15-3568:~/Documents/AI/ML/Fathom/scripts$ python3 clustering_and_assigning_human_readable_names.py
/home/raj/anaconda3/lib/python3.7/site-packages/sklearn/feature_extraction/text.py:17: DeprecationWarning: Using or importing the ABCs from 'collections' instead of from 'collections.abc' is deprecated, and in 3.8 it will stop working
  from collections import Mapping, defaultdict

The total number of topics in all the videos are : 64

The text file 3_.txt contains 2 topics

The topic 1 in the text file 3_.txt is the 0 th topic in the list of total number of topics
The topic 2 in the text file 3_.txt is the 0 th topic in the list of total number of topics

Assigning Human-readable names to labels

The title for the label 0 is number

The title for the label 1 is funny hour number usa website

The title for the label 2 is created hour listen number usa website

The title for the label 3 is created listen number usa website

The title for the label 4 is hour number usa website
```

4. Assigning human-readable names to the clusters

```
Activities Terminal Wed 22:52
raj@raj-Vostro-15-3568: ~/Documents/AI/ML/Fathom/scripts
File Edit View Search Terminal Help
The title for the label 0 is number
The title for the label 1 is funny hour number usa website
The title for the label 2 is created hour listen number usa website
The title for the label 3 is created listen number usa website
The title for the label 4 is hour number usa website
The title for the label 5 is hour listen number picture usa website
The title for the label 6 is famous funny hour number usa website
The title for the label 7 is created hour listen number usa website
The title for the label 8 is camera hour listen number usa website
The title for the label 9 is created funny hour listen number usa website
The title for the label 10 is hour number usa website
The title for the label 11 is funny hour number usa website
The title for the label 12 is created hour listen number usa website
```

Report by:

K.Rajkamal  
N.I.T Raipur  
CSE - 5th sem  
Mobile :9494315656  
Mail: rajkamalk99@gmail.com