# Folklore Document Classification
## Curtis Mann

The goal of this project was to attempt to create a program that identifies the region of origin for a given folklore tale. Unfortunately, it was not wildly successful. This document will explain the contents of this folder, the process and methods of the project, as well as some evaluation of the project.

...

Contents

Inside this folder you will find:

Three Python files: (1)HTML_Parser.py, (2)Query_Location_Parser.py,
(3)Document_Classification.py
One HTML file: thompson.htm
Two .txt files: Test.txt, Key.txt

The files listed above are necessary for the program to run. All other files in this folder are generated by one of the above Python programs.

Methods and Process

I used Stith Thompson's *Motif-Index of Folk Literature* (thompson.htm, http://www.ruthenia.ru/folklore/thompson) as a source for folk literature motifs, with a motif being a short summary of a folktale followed by it region(s) of origin.

The first Python program ((1)HTML_Parser.py) uses the Beautiful Soup HTML parsing library together with regular expressions to remove all of the HTML tags and to extract the motifs from the source file. This takes 10-15 minutes to run on my computer. The output of this first program is QueriesOut.txt.

The second Python program ((2)Query_Location_Parser.py) reads in QueriesOut.txt and uses regular expressions and NLTK generate two more files. Regular expressions are used to further remove unnecessary information such as motif ID's and deal with unexpected anomalies in the data. The NLTK stopword list is used to remove meaningless words. The output is two .txt files, CleanQueries.txt and Locations.txt. Every line in Queries.txt contains a folklore motif . Every line in Locations.txt contains the region(s) of origin for its corresponding motif in CleanQueries.txt.

The third Python program ((3)Document_Classification.py) reads in both Locations.txt and CleanQueries.txt. Using these files, the program creates one large string for each location (China, Ireland, Africa, India) composed of that location's queries concatenated together. A "query" object is created for each location. A TF-IDF vector is created for each query object.

Next, Test.txt is read in. This file contains 49 folktales from varying regions, all sourced from http://www.pitt.edu/~dash/folktexts.html. A “document” object is created for each tale and a TF-IDF vector is created for each folktale. To evaluate the similarity between the queries and the documents, the cosine similarity is calculated for all region's queries against each folktale. The region with the highest cosine similarity score is considered the “winner.” The results are compared to the information in Key.txt, which is composed of the correct region of origin for each folktale.

Output is written to ClassificationOutput.txt and scores.txt. ClassificationOutput.txt contains an ID for each tale, the region of origin, the program's “winning” region of origin, and the “winning” cosine similarity score. At the bottom of ClassificationOutput.txt are some statistics which will be covered in the next section. scores.txt contains the cosine similarity scores for all region's queries against each folktale.

## Evaluation

The data at the end of ClassificationOutput.txt can be seen below:

```
irish:     4/17,  0.23529411764705882
indian:    11/20, 0.55
chinese:   0/3,   0.0
african:   4/9,   0.4444444444444444
```

As you can tell, the results were not what I had hoped for. In retrospect, I believe the project was a little too ambitious. I should have perhaps limited it to just Irish folklore and written a program that included more features (utilizing my knowledge of the Irish lanuage (which often appears in tales), creating a corpus of Irish names, writing a stemming algorithm based off the grammar of the Irish Language, etc...) to determine whether or not it was Irish in origin.

Originally I intended for the program to analyze as many as 10 different regions of origin. However, as I began to work on this I realized I had to scale back to achieve somewhat satisfactory results in the given time period. I chose Irish, Indian, Chinese, and African folktales because there were many entries in Stith Thompson's index. However, not many Chinese or African tales were found on the folktale database.