# Second Capstone Project
# Milestone Report #2

## Problem Statement/Challenge

Given a set of Wikipedia articles, the challenge is to build a model to identify the primary topic of each article and analyze the content and structure of that article. The second part of the project will focus on conducting similar analyses of a group of Wikipedia articles. After topic modeling is complete for the corpus, I will cluster the articles based on their content and recommend similar articles based on those primary topics.

A model such as this will be vital to those interested in clustering, transforming, visualizing, and extracting insights from any unlabeled and unknown dataset. This model would be useful for a number of clients/customers. For example, law enforcement or intelligence services might find this model useful as they investigate documents discovered in a raid or review of forensics information.

## Project Details and Description of the Data

For this challenge, I obtained articles from the Wikipedia package. I then accessed and analyzed the various components of the page. Then, I tokenized each word contained in the document and created a bag of words. To refine the results, I removed non-alphabetic characters as well as English stop words. I then used lemmatization to produce a much better and more concise bag of words based on the stems of the key words in the article.

For the next portion of the project, I obtained several articles using the Wikipedia library. I tokenized each article in a similar fashion as mentioned above and created a gensim corpus. I then

created a gensim bag of words. I also used Tf-idf to find topics from a particular article with results based on the entire corpus.

## Initial Findings

My initial findings suggested that my model worked quite well for analyzing only one article. The article was processed cleanly and efficiently. Tokenization and the initial bag of words highlighted topics found in the article. I was then able to verify the results by revisiting the title and summary of the article.

I experienced similar results for the group of articles. The gensim bag of words and Tf-idf performed quite well and provided a concise list of topics. I had to add extra code to catch Disambiguation errors as several phrases from my randomly-selected group were associated with multiple Wikipedia pages. I then began clustering the articles based on key words and topics. Initial results proved promising but require additional work.

## Next Steps

Moving forward, I will continue to explore these results and compare them as I increase the number of Wikipedia articles. Additionally, I plan to refine and tweak the clustering (currently using k-means) to be able to provide recommendations to the reader/user for additional articles with similar topics.