Wikipedia Article Analysis:
Final Report

Harlin W. Sanders
Springboard Data Science Career Track
Capstone Project #2

Problem Statement/Challenge

- Given a Wikipedia article, build a model to:
  - identify the primary topic of a single 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
  - cluster the articles based on their content and recommend similar articles based on those primary topics.

Data & Data Wrangling Steps for Part I

- obtain article from the Wikipedia library
- access and analyze the various components of the page (i.e., title, URL(s), images, links, summary, and full content)
- tokenize each word contained in the article and create a bag of words
- To refine the results:
  - remove non-alphabetic characters as well as English stop words
  - use lemmatization to produce a much better and more concise bag of words based on the stems of the key words in the article

Results of Part I

- Random article was "Status of First Nations treaties in British Columbia”
- Bag of Words produced terms such as 'treaty', 'nation', 'process', 'government', etc.

Data & Data Wrangling Steps for Part II

- obtain 100 articles from the Wikipedia library
- tokenize each word contained in each article and create a bag of words
- remove non-alphabetic characters as well as English stop words
- use lemmatization to produce a much better and more concise gensim bag of words based on the stems of the key words in the article
- TF-idf
- K-Means
  - Recommendation based on k-means clustering

Results of Part II

- Gensim bag of words for these 100 random articles revealed that top terms:
  - “first”
  - “also”
  - “system”
  - “writing”
  - “team”
- TF-idf (using the 2nd article in the corpus) revealed topics/scores:
  - “goapele” with a score of 0.48
  - “dawn” with a score of 0.21

Results of Part II
continued

- K-means
  - K=5
  - Majority of articles fell into clusters 0 and 1

- Recommendation

Conclusions

- This model worked quite well for analyzing both one and multiple articles.
- Tokenization and the initial bag of words highlighted topics found in the article.
- The gensim bag of words and Tf-idf performed well and provided a concise list of topics.
- The k-means clustering could easily be integrated into an app or website that provides recommendations to users/readers based on their interests.