Physics Dataset Practice Problems

Week of July 27, 2020

3 Medical Physics Abstracts

As researchers we must understand the trends in our fields. This can be a little daunting for those just entering a field. Even for experienced researchers it can be difficult to have a quantitative understanding of the trends. Machine learning techniques can be adapted to quantify these trends. This dataset contains the last 5 years of abstracts for manuscripts published in Medical Physics. The goal is to apply natural language processing and clustering techniques on this dataset to show the most common subfields in medical physics.

3.1 Data Import

This dataset is actually quite easy to extract. You can export citations, with abstracts, to all the papers from a journal for a given time period from `https://pubmed.ncbi.nlm.nih.gov/`. To minimize the manual efforts with this dataset, an extracted `.nbib` file was converted to a `.csv` file (`papers.csv`). Import the abstracts with pandas into a list.

3.2 Preprocessing

This example directly follows the natural language processing example in the Udemy class. Go through the abstracts and remove the stop words and apply stemming (import re, nltk).

3.3 Bag of Words Model

This example utilizes the bag of word model. Use `CountVectorizer` from `sklearn.feature_extraction.text` to convert the abstracts to vectors.

3.4 K-means Clustering

This example uses KMeans clustering to identify clusters as subfields. Use KMeans (from `sklean.cluster`) along with the elbow method to decide how many subfields you think there are in these abstracts. Then, apply this number of clusters to classify all the abstracts. From this, you can calculate the cluster centers (`kmeans.cluster_centers_`).

3.5 Making Some Sense of the Clusters

There are many words that are common with a lot of the abstracts (purpose, methods, results...). To find what identifies the clusters, you cannot simply print the words with the highest occurrences in the clusters. Instead, calculate the average cluster center. Then, calculate the difference between each cluster array and this average center. The words with the highest difference from the average center are a better representation of a cluster. For each cluster, find the 8 words (or 10, or 20, or however many you like) with the highest difference relative to the average and print these words. See if you think these groupings of words resemble a subfield, such as machine learning.