IST736 Text Mining

Professor Bei Yu

Spring 2025

HW3 Exploring Common Topics in Health Research News

The Health Research News Headlines (HW3\_headlines\_10000.csv) is a data set that sampled 10,000 headlines from health research press releases posted on the EurekAlert! website. You are going to explore the common themes of health research news by applying kMeans, LDA, and BERTopic algorithms.

For each algorithm, describe your choice of number of clusters, label the clusters, and explain their meanings. For the clusters that are hard to interpret, you can label them as “unsure”.

[Extra points] This data set also contains timestamp for each headline. Analyze whether common health topic change over time. Correct solution upgrades a previous grade, e.g. from A- to A.

If the dataset is too large for your computer to process, sample it down.

Submit your .pynb file with your explanations included as comments.