Project Title

Topic database creation over The Rice Thresher

Project Mentor(s)

Edgar Avalos-Gauna

Project Summary

Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. It is often used to extract the key ideas or phrases from full text documents. The goal is to analyse the shift on key topics over the years on the Rice Thresher

Project Description

Surveying newspapers are among the richest sources of information available to scholars studying peoples and cultures. Due to the formal structure within a newspaper, it is possible to analyze it following a systematic approach. Additionally, with the implementation of novel IT technologies, this process can be done faster and more precise. Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. It involves counting words and grouping similar word patterns to infer topics within unstructured data. Topic modeling can link words with the same context and differentiate across the uses of words with different meanings. Topic modelling will help to identify the most important and potentially interesting topics over a given period of time. It will provide a general picture of the society at that time. By studying the Rice Thresher, this project aims to provide a timeline of the major events that happened through Rice University's history. This might also provide some information in changes happening in Houston during the past century. In order to carry out this project, as a first step, only headlines will be studied and then a full text analysis will be made. Additionally, different machine learning techniques will be assessed. Some of these techniques are Vector Space Model (VSM), Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). These techniques will be implemented using Python as main coding language.

Key Tasks for Fellow(s)

- Data collection: web scraping, text document transformation, and data consolidation
- Exploratory Data Analysis: Bag of words, statistical metrics, visualizations
- Unsupervised learning: Topic modeling, clustering analysis
- Conference presentation: Most relevant results will be used to present a paper on a conference
- Journal paper: Most relevant findings will be used to wirte a journal paper

Qualifications

Basic kowledge on statistics and python programming

Learning Outcomes

Natural language processing, webscraping, and unsupervised machine learning analysis