Final Project for Data620
December 2019

Authors:
Scott Karr
Vikas Sinha

Proposal

Our project will use Topic Modeling algorithms to determine the top trends in one or more academic disciplines (e.g, Machine Learning, Economics), based on analysis of research papers published in that category on the website Arxiv.org.

Objective

Our objective is to determine the most common topics pursued by researchers within selected disciplines, in order to show how the interest of the research community in those disciplines has evolved over a period of time. We aim to implement our analysis using topic modeling and natural language processing algorithms.

Dataset

The dataset will consist of a selection of papers published in the selected categories on Arxiv.org in the last few years.

Libraries Used

1. **gensim:** open-source library for unsupervised topic modeling and natural language processing
2. **nltk:** a suite of libraries and programs for symbolic and statistical natural language processing (NLP)
3. **spaCy:** an openz-source software library for advanced natural language processing
4. **pyLDAvis:** Python library for interactive topic model visualization
5. **matplotlib:** Matplotlib is a plotting library for the Python programming language

References

1. honhttps://www.machinelearningplus.com/nlp/topic-modeling-gensim-pyt/
2. https://arxiv.org/search/advanced