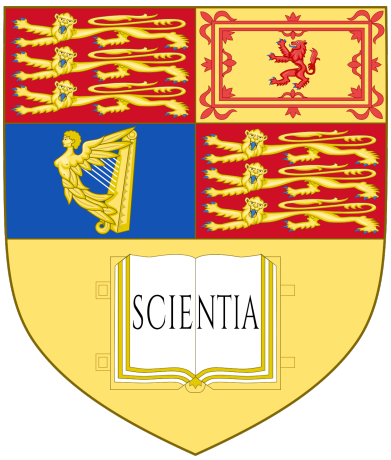
Imperial College London

Project Proposal: Comparing Latent Dirichlet Allocation (LDA) and Supervised Learning for Topic Modeling the STEM Research Funding Landscape in UK

Student: Zitong Zhao

supervisor: Samraat Pawar



Department of Life Sciences, Imperial College London,
Silwood Park Campus, Ascot SL5 7PY, UK

April 2023

Keywords

machine learning, latent Dirichlet allocation (LDA), supervised learning, Label-LDA, STEM funding landscape

Introduction

The project focuses on comparing the utility of unsupervised learning and supervised learning in natural language processing. The dataset used is the UKRI research funding situation, and topic analysis is used to classify the distribution of STEM subjects, which is then analyzed.
Some topic modelling methods based on the latent Dirichlet allocation (LDA) proposed by Blei et al. were applied as a new document classification algorithm([3],[2]). Standard LDA is an unsupervised method that automatically generates potential topics based on the discrete probability distribution of words.Classification accuracy can be significantly improved if labeled the data and added prior information. Therefore, extending LDA to supervised algorithms is a hot research topic. Labeled-LDA is a type of topic model in supervised learning that constrains LDA by defining a one-to-one correspondence between the latent topics of LDA and user labels ([4]). Labeled-LDA can directly learn the correspondence between topics and labels. Support vector machines (SVM) is also a popular supervised learning method for natural language processing because of the nature of SVM as discriminative classifiers that can effectively handle textual data ([1]).
UK Research and Innovation (UKRI) is a government-established organization that promotes the UK's international leadership in science, technology, and socio-economic development through supporting and funding activities such as research, innovation, and higher education. It contains many research proposals with approved amounts. In this article, we will conduct a topic analysis based on natural language processing to categorize all proposals.
The distribution of UKRI funding can reveal the research directions and importance of different STEM disciplines, enabling us to understand which disciplines are at the forefront of development. This can encourage cooperation and communication between different disciplines, facilitate interdisciplinary research, and help ensure the optimal use of funding, promoting fair distribution.

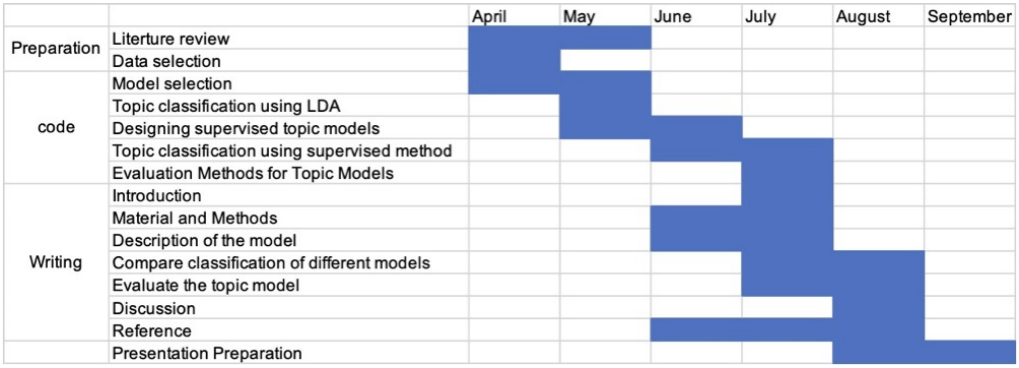
Proposed Methods

Using Mallet Tool for LDA topic modelling (https://mimno.github.io/Mallet/) Using GitHub code for label-LDA (https://github.com/JoeZJH/Labeled-LDA-Python) Then evaluation Methods for Topic Models by calculating recall, precision, and estimating the probability of unseen held-out documents given some training documents([5]).

Anticipated outcomes

Design of a topic model that can identity the STEM topic of UK funding landscape using LDA and supervised learning methods.
Comparison of different methods for classification of datasets in STEM.
Evaluation the efficiency and posterior probability between unsupervised and supervised methods.
Analysing the impact of the distribution of STEM funding on research development in the UK.

Timeline



Budget

ortable hard disk for data storage and transmission - £88 Commuting cost of train fare to the lab twice a week - £410

References

[1] Roberto Basili. Book Review. *Computational Linguistics*, 29(4):655–661, 12 2003.

[2] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256, 2009.

[5] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.