Using Text Based Models to Predict Priority of Activities

Rica Rebusit

Department of Mathematics and Statistics, California State University, Chico

Introduction

The Center for Healthy Communities (CHC) promotes food security, nutrition education and overall health in the community. Federal food assistance is available through the Supplemental Nutrition Assistance Program (SNAP), known in California as CalFresh Outreach (CFO) which CHC coordinates on 50+ CA college campuses. CHC helps students apply and receive CFO benefits and wants to understand the impact of CFO activities such as meetings, events, and trainings with campus partners have on the amount of applications received. The activities are reported as written descriptions to CHC and they want to predict which activity should be prioritized.

Methodology

Collect Data

Data of partnership activities were collected. This included campus partners, year, quarter, and a description of the activity.

Cleaning and Preprocessing

Remove unnecessary variables, whitespace, and splits text descriptions into a list of words.

Tokenization

The process of breaking down sentences into words and removing stopwords such as “and”, “the”, “we”, “to”

sentence
We started to plan for the trip tomorrow
Bob went to go eat a burger

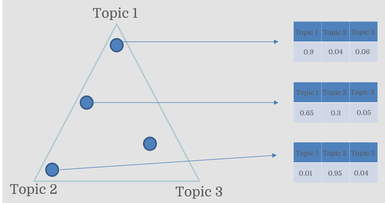
word
started
plan
trip
tomorrow
bob
eat
burger

Methodology - LDA

Latent Dirichlet Allocation (LDA) is a topic modeling method which clusters rows of “documents” to find topics that show how frequent they appear.

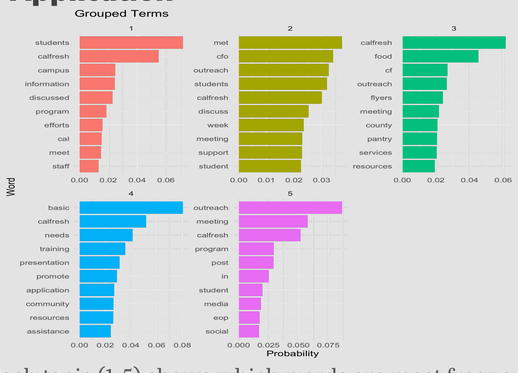
Partner	Year	Quarter	Activity	Description
CSUC	2023	1	Training	Trained new workers
BUTTE	2023	2	Meeting	Discussed financial distributions
SAC	2023	3	Event	Took pictures and posted on Instagram

Here, the description are the documents to be analyzed.

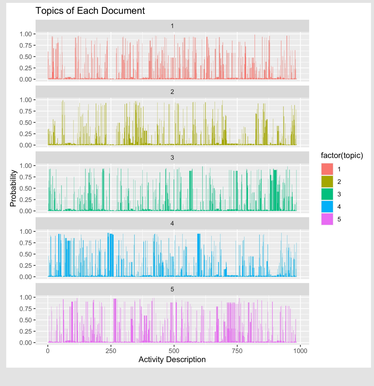


Once key topics in the documents are identified and sorted, the dots represent the probability the word is in the topic.

Application



Each topic (1-5) shows which words are most frequent.

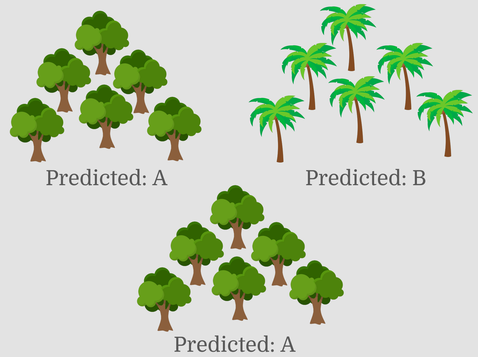


Descriptions of activities are in chronological order of quarter.

The probability of the topics, show which activity description is more frequent in.

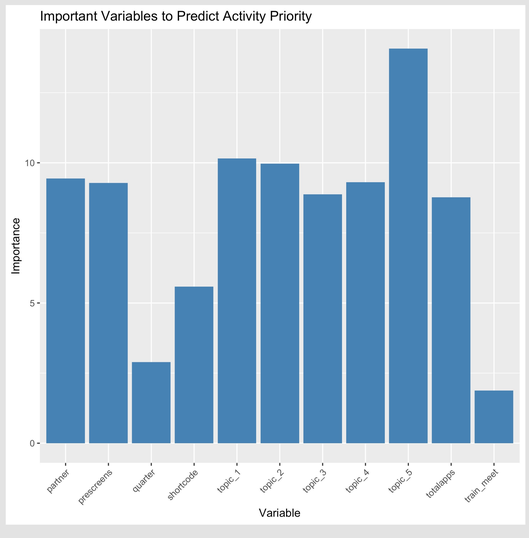
Methodology – Random Forest

Random Forest is a machine learning method for classification. It builds a series of decision trees that output a prediction that merge together as a majority vote to output a final prediction.



Final Prediction: A

Application



The importance for each variable is used to rank which variable in the Random Forest model has more significance in predicting the target. In this case, the target is the priority of the activity.

Prediction

Using the Random Forest model, we can predict the priority of activities. The model looks at the predictor variables and uses them to make a prediction for the target. For example, to predict the priority, new data about the campus partners, activity description, quarter, and such, the random forest model predicts the probability of the priority based on the predictors. This will then provide insight on how CHC will distribute money to campus partners.

Future Work

This is still an ongoing project, and it can be easily picked up from where it was left off. Ideally, we would want more labeled data to automate the priority of the activity. This will benefit CHC as they would be able to understand the impact of CFO activities.

Acknowledgements

Supported and guided by Dr. Robin Donatello. Center for Healthy Communities CalFresh Outreach

California State University Chico
College of Natural Sciences
MATHEMATICS AND STATISTICS DEPARTMENT

DATA SCIENCE INITIATIVE CHICO STATE