Predict Your Paper's Fate

Author: Zhe Guan
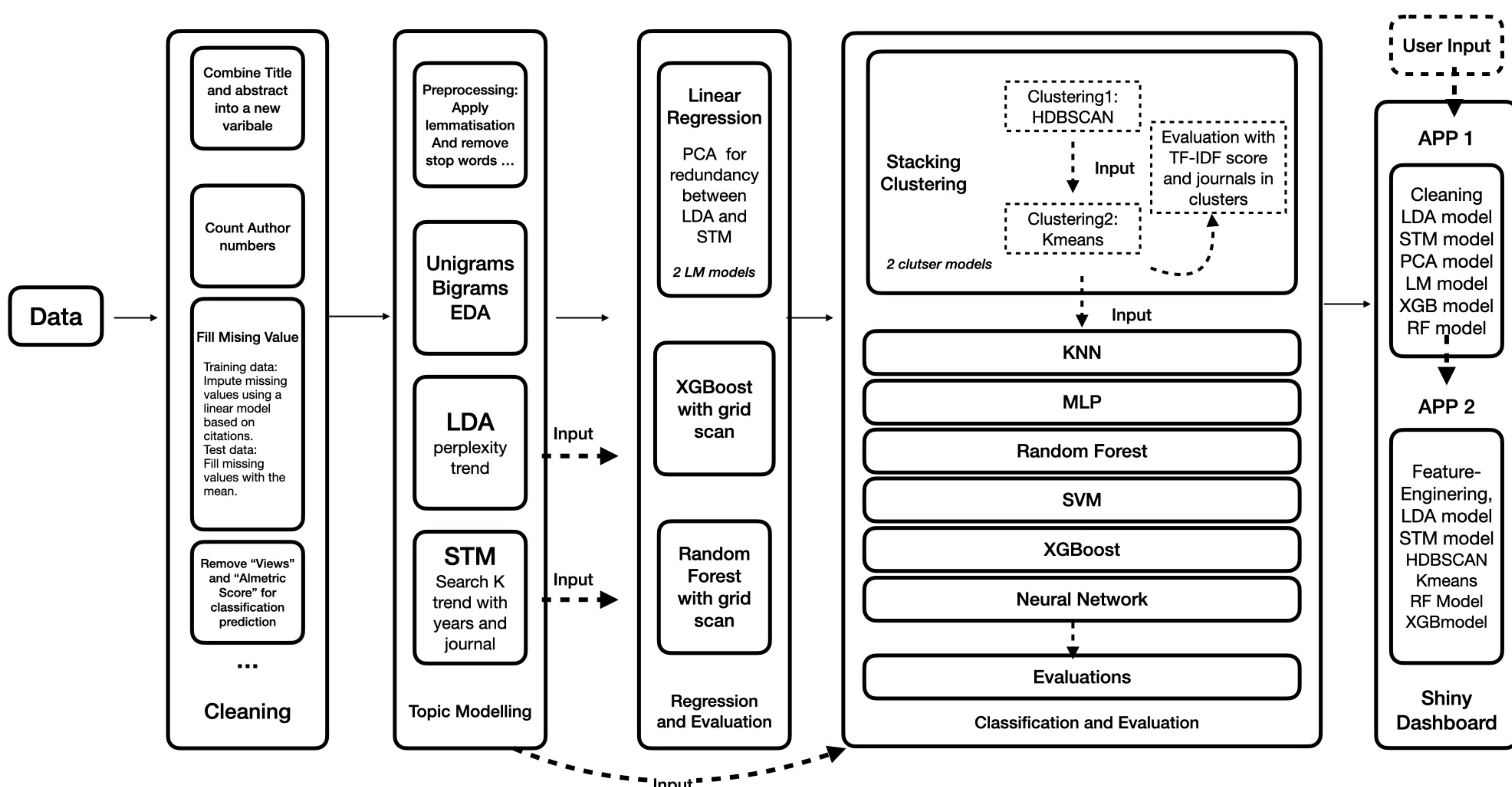
I designed two dashboards that can advise the target journal and citation counts of your paper based on abstract topic modelling.

Really? Let's try...

...

Where does it suggest to submit?

Trash Bin. Accurate prediction!

Made by author based on *pixton* app

Highlight:

✳ Bigrams are used to provide more informative insights into the data.

✳ Various topic modelling techniques are used to provide more data for model training.

✳ Stacking clustering is used as a feature engineering technique.

✳ Two dashboards are developed to predict target journal and citation counts for users.

Work Flow:

Cleaning
- Combine Title and abstract into a new varibale
- Count Author numbers
- Fill Mising Value
  Training data: Impute missing values using a linear model based on citations. Test data: Fill missing values with the mean.
- Remove "Views" and "Almetric Score" for classification prediction
- ...

Topic Modelling
- Preprocessing: Apply lemmatisation And remove stop words ...
- Unigrams Bigrams EDA
- LDA perplexity trend
- STM Search K trend with years and journal

Regression and Evaluation
- Linear Regression
  PCA for redundancy between LDA and STM
  2 LM models
- XGBoost with grid scan
- Random Forest with grid scan

Classification and Evaluation
- Stacking Clustering
  Clustering1: HDBSCAN → Input → Clustering2: Kmeans
  2 cluster models
  Evaluation with TF-IDF score and journals in clusters
- Input → KNN
- MLP
- Random Forest
- SVM
- XGBoost
- Neural Network
- Evaluations

Shiny Dashboard
- User Input
- APP 1
  Cleaning
  LDA model
  STM model
  PCA model
  LM model
  XGB model
  RF model
- APP 2
  Feature-Enginering,
  LDA model
  STM model
  HDBSCAN
  Kmeans
  RF Model
  XGBmodel

Data → Input

Predict your target journal: https://yhj0vi-zhe-guan.shinyapps.io/target_journal_dashboard/
Predict your citation count: https://yhj0vi-zhe-guan.shinyapps.io/Citation_dashboard/
Repository: https://github.com/phy-guanzh/Journal_Abstracts_Analysis