Partnership Analysis & Level Classification

By:
Abraham Serrano
Prasanna Rajendra

Supervisor:
Robin Donatello
Kelsey Kyllonen

Outline

1. Abstract
2. Quarto
3. Proposed methodology
4. Results
5. Next Step

Introduction

- CHC(Center for Healthy Communities) has subcontractors who partner with local organizations for calfresh outreach
  - These subcontractors report to CHC on activities

- Wanted to see if we could determine a subcontractor's level of participation in the contract based on the descriptions they listed for their partnership activities

QuartoDoc

- Main usage for language flexibility
- Reticulate package

Importing package

```r
{r}
library(reticulate)
```

List in python :

```python
{python}
pr_list_object=[1,2,3]
print("Python list")
print(pr_list_object)
```

Python and R Communication

Reticulate package in R can convert python container to R

Branches Diverge

1. In R,
   - We did Text Preprocessing
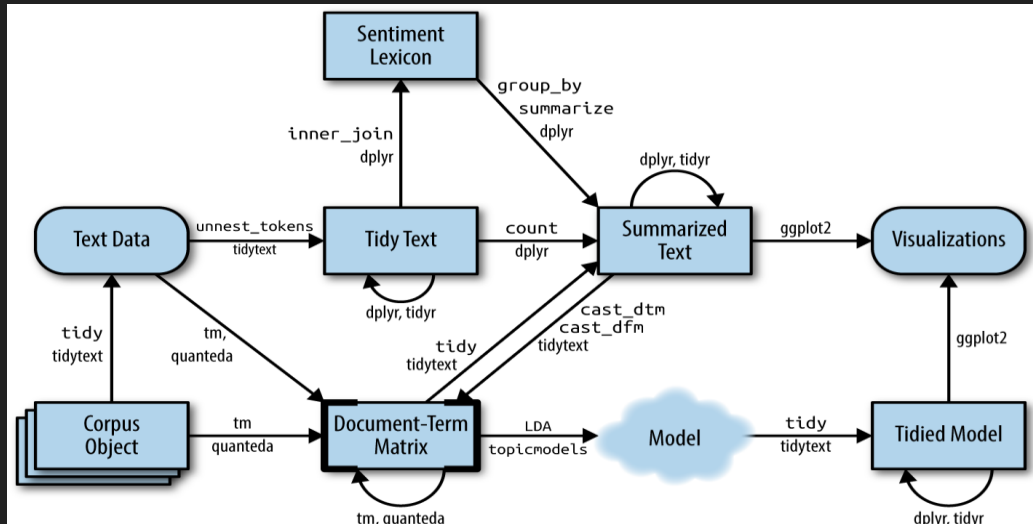   - Created DTM
   - Performed LDA

1. In Python ,
   - Multi Class Model Decision Tree
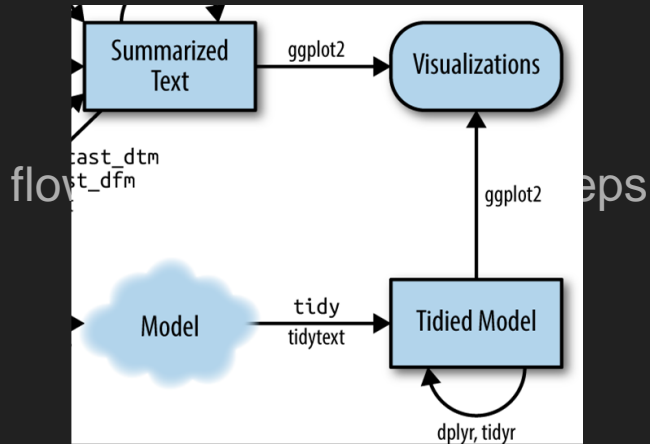   - Data Preprocessing on Numerical and Categorical Features

In R

Create LDA Model

- Python df -> R df
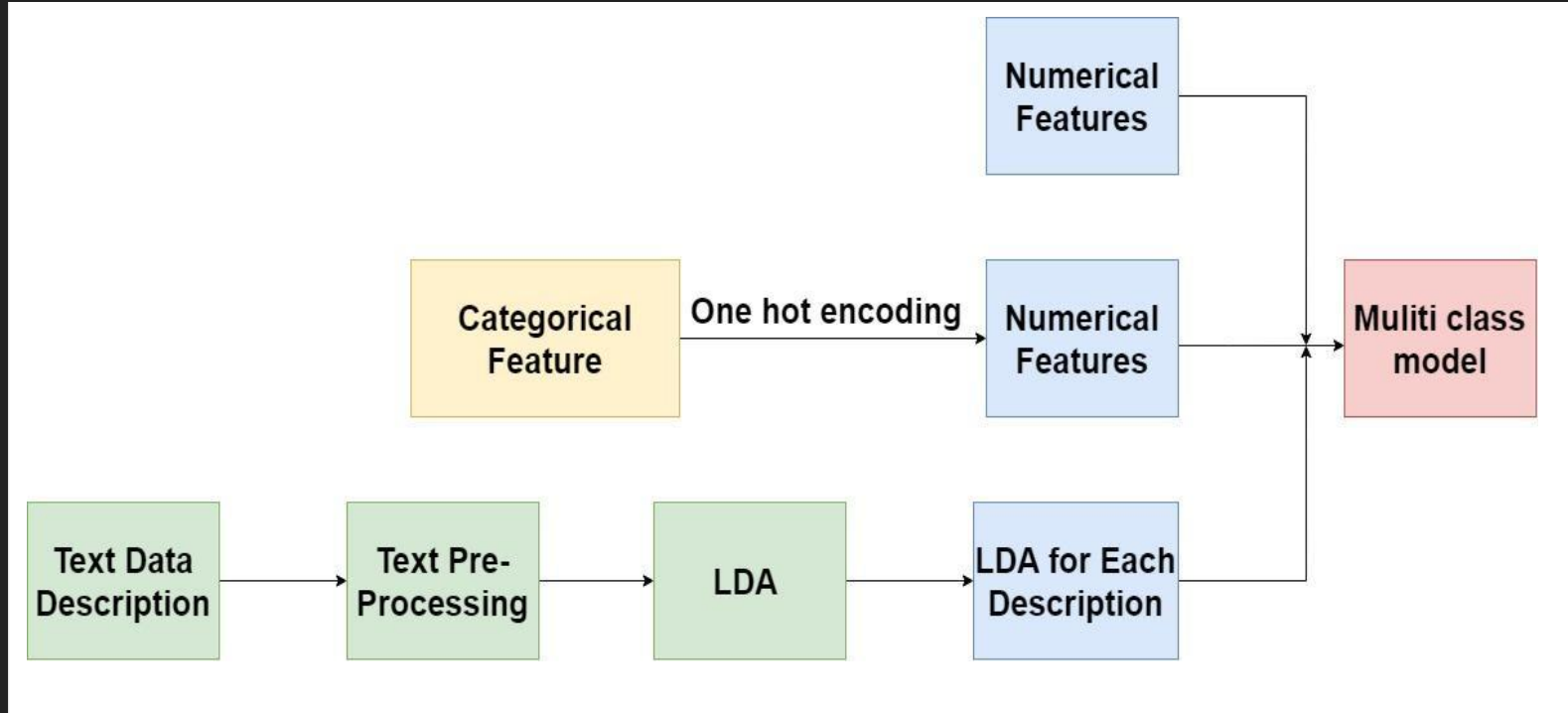- R df -> Document Term Matrix
- Document Term Matrix -> LDA Model

Next Steps:

- Get topic probabilities for number of topics I (the user) want
- Visualizing the model

- This flow...st_dfm ...eps

Proposed methodology

Processing Data

- Numeric data:
  - Numeric data will be directly passed to the multi-class model
- Categorical data:
  - All the categorical data will be converted into each new dummy column by one hot encoding.
- Text Data:
  - Will perform LDA on the text data will divide into 4 clusters , will use the gamma value of each cluster as a new feature for out model

After all the data is converted into numeric, using Decision Tree to perform Multi-Label Classification

Results

LDA Values for each Topic

```
[(0,
  '0.091*"student" + 0.073*"outreach" + 0.057*"meeting" + 0.051*"meet" + '
  '0.042*"staff" + 0.038*"presentation" + 0.031*"update" + 0.017*"application
  '+ 0.016*"community" + 0.015*"weekly"'),
 (1,
  '0.152*"basic" + 0.082*"resource" + 0.049*"discuss" + 0.045*"promote" + '
  '0.043*"social" + 0.033*"fair" + 0.033*"quarterly" + 0.032*"great" + '
  '0.029*"medium" + 0.024*"event"'),
 (2,
  '0.074*"program" + 0.050*"student" + 0.043*"outreach" + 0.034*"information"
  '+ 0.034*"service" + 0.033*"food" + 0.027*"campus" + 0.026*"provide" + '
  '0.026*"meeting" + 0.025*"training"'),
 (3,
  '0.074*"collaboration" + 0.031*"action" + 0.023*"day" + 0.022*"summer" + '
  '0.022*"network" + 0.020*"train" + 0.020*"training" + 0.019*"income" + '
  '0.015*"life" + 0.014*"quarter"')]
```

Word Cloud

Next Step

- We can do some research on Partner feature since the distribution is not normal.
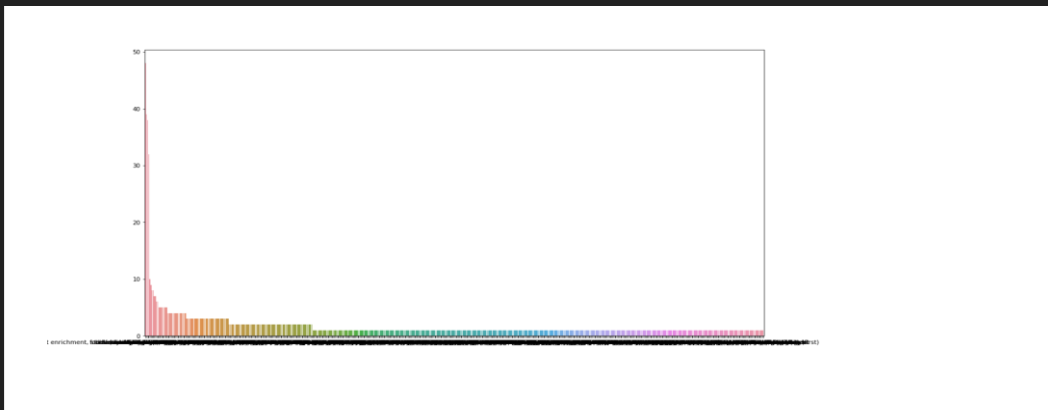
- Once we get the Label data ready we can Grid CV over different parameters to find best fit algorithm

Thank you !!!