Resume Parser Using Natural Language Processing Final Report

Abstract

The purpose of this project is to use the power of Natural Language Processing to parse information from a resume pdf files.

Design

In order to parse resumes information, data was downloaded from different data resources. Then, EDA, topic modeling and multiple classification models were implemented to get the best resluts.

Data

- ### PDF Files

  The First dataset is **1,388** resumes pdf files comes from Kaggle and GitHub , and after text extraction and intial cleaning, the resume texts and thier path converted into dataframe with **1388** rows and **3** columns, which are: Resume, cagtegory.,id

  - Resume Columns contains resumes (one resume in each row).
  - cagtegory Columns contains the category of each resume.

- ### Ready Dataset

  These datasets comes from Kaggle - resumes and Kaggle - ResumeDataSet and after combining them **1,388** rows and **3**columns are result, but all columns are dropped except **TEXT** column to convert them into an unsupervised learning dataset, and after cleaning the number of rows becomes **1297**

- **Topic Modeling**:

after trying many models we chose SVM with TF-IDF.

Cleaning:

- Drop unneeded columns.
- Handle nulls.
- Handle Duplicates.

Classification:

MODEL	Training Accuracy
Naif Bayes	**0.63**
Random Forst	**0.053**

Tools

- **Data manipulation and cleaning:** Pandas and Numpy, Matplotlib
- **Text precossing:** NLTK, , gensim, scikit-learn ,
- **Topic Modeling:** LAS,LDA,NMF
- **Visualization:** , matplotlib, seaborn