Information Extraction From Profile Using Machine Learning

Pavithra C P

November 22, 2018

Abstract

Corporate firms and enlisting agencies process varied profiles daily. This is often no task for humans. An automatic intelligent system is needed which may put off all the very important info from the unstructured profile and rework all of them to a standard structured format. To realize additional attention from the recruiters, most profiles are written in various formats, together with varied font size, font color, and table cells. However, the variety of format is harmful to data processing, like profile data extraction, automatic job matching, and candidates ranking. This work proposes a three step approach for information extraction from profile by, [1] Text block identification, [2] Vectorization of data, [3] Multi-class classifier to predict different segments of a profile. Bernoulli Naive Bayes Classifier is used as classification algorithm for prediction.

1. Methodology

The methodology introduces a machine learning based approach for tackling the problem of information extraction from profile. The solution involves a three step process :

- Block Identification
- Vectorization
- Bernoulli Naive Bayes Classifier to predict different profile segments.

The work proposed is on resume segmentation. There are various resume dataset are available for this work. Using this dataset the two step process of the information extraction is performed.

1.1. Block Identification

In Block Identification step, it segments the resume into different sections on basis of headings and its contents. The headings can be identified by different font size with respect to the content or by keyword based extraction. Synonyms of the keyword are also considered in case of keyword based extraction.

1.2. Vectorization

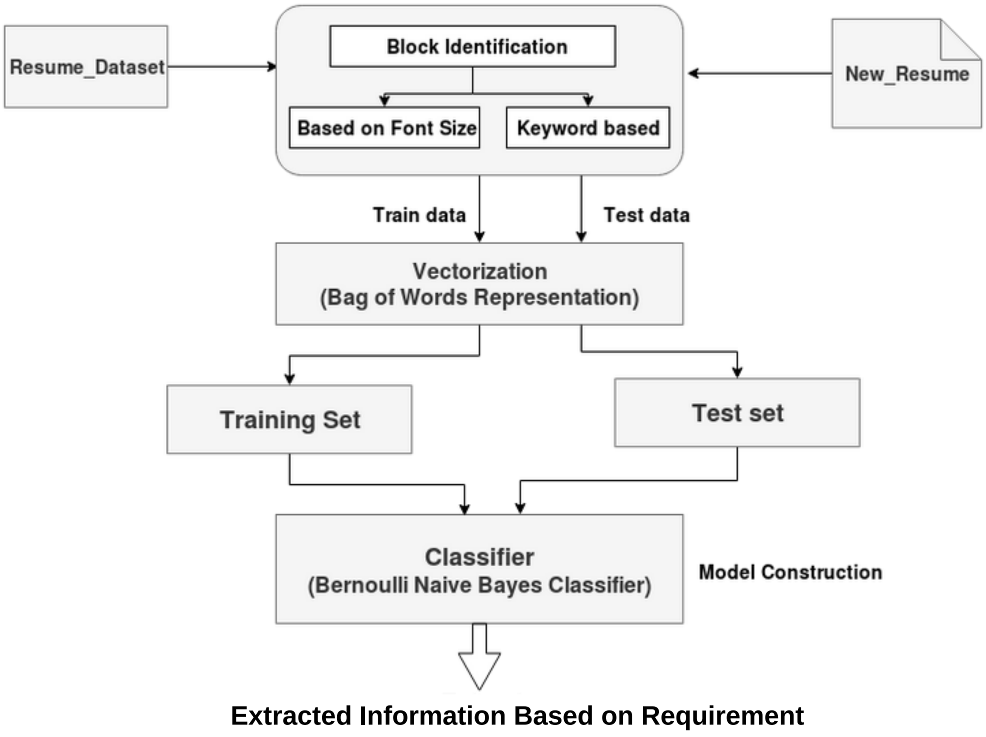
Vectorization is a process by which the system takes an assortment of documents text and applies numerical feature vectors to them. This is called the Bag of Words representation. Bag of Words is a data science method that falls under the text feature extraction category. So, using BOW approach, texts are converted into vectors.

1.3. Prediction using Bernoulli NB Classifier

The input to this step will be vectors of different blocks. Selected vectors according to the requirement (eg :- Educational background, Skills, Experience etc..) will be provided for training. A model for resume segmentation is created with this training data using Bernoulli Naive Bayes Classifier.

When a new resume is given, it follows (1) Block identification, (2) Vectorization. Then vectors are passed into the model, for predicting the vectors corresponding to learnt model. The output will be extracted information based on requirement.

2. System Architecture


Resume_Dataset
Block Identification
Based on Font Size
Keyword based
New_Resume
Train data
Test data
Vectorization
(Bag of Words Representation)
Training Set
Test set
Classifier
(Bernoulli Naive Bayes Classifier)
Model Construction
Extracted Information Based on Requirement