Instant Resume Evaluation Engine

Jinesh Dhruv (jad6566@g.rit.edu)
Advisor: Prof. Christopher Homan

Introduction

- The job seeker prefers having the resume in PDF format as it can be opened on any operating system, formatting does not get messed up and it provides a better appearance
- Existing resume parsing system still has trouble parsing and extracting the content from the PDF due to the different layout-structures

Contribution

- Created a hybrid content-based & segmentation-based technique for resume parsing
- Developed a tool where user can upload the resume for information extraction

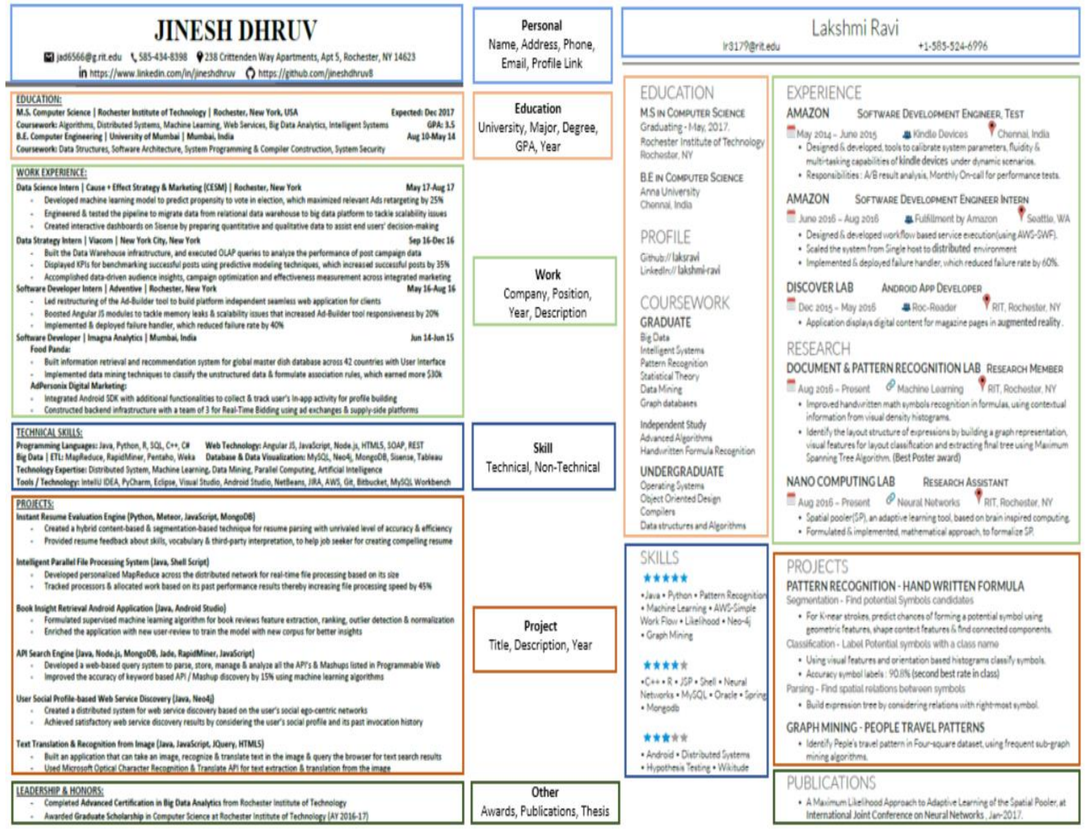
Tools and Technology

- Software & Python packages used: Meteor, MongoDB, Python 2.7, pdfminer, NLTK, Fuzzy

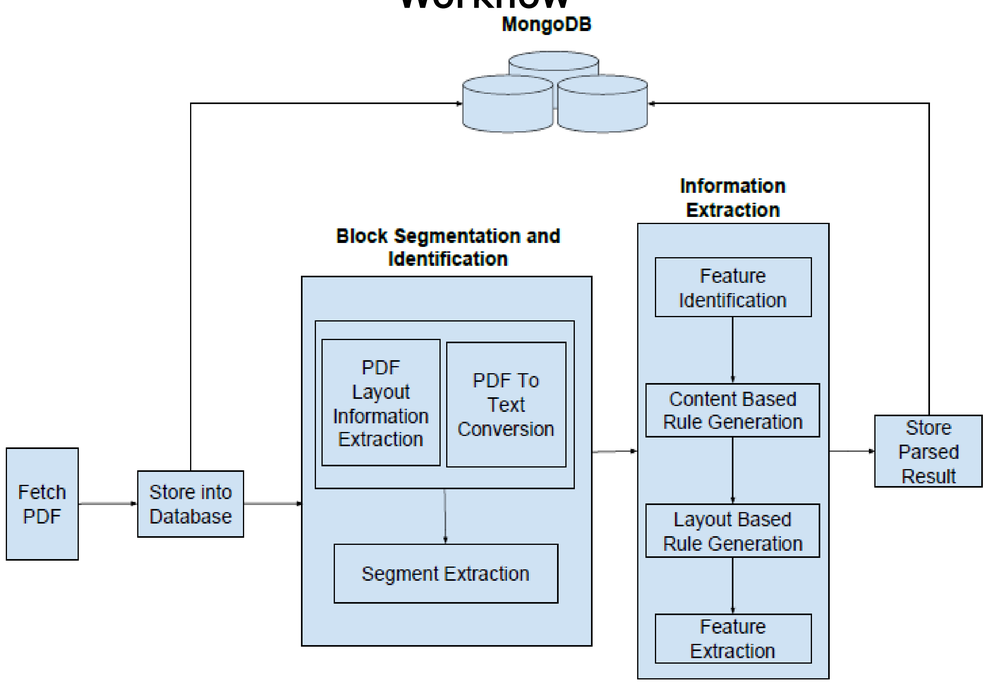
Corpus Word-List

Domain List	Definition
Segment keyword list	List of all possible segment titles like education, qualification, university, project, personal project, academic project, etc.
University list	A university list in the US
US location list	A list of US Zip Codes, State and City
Degree list	A list of aliases & abbreviations of Masters, M.S., MS, Bachelor's, etc.
Major list	A list of majors in United States
Company list	A list of all company in US

Hierarchical Layout



Workflow



Implementation

Block Segmentation and Identification

- Stored document layout information
- Converted PDF to text
- Search segment keywords using fuzzy logic
- Created bounding box near the keyword
- Stored each sentence into respective segment

Feature Extraction

- Created segment specific feature list
- Searched main feature in the respective segment
  - Used keyword-based technique
  - Used layout information like font-size, font-name and position

Segmentation Result

Segment Detection	Precision	Recall	F1-score
Personal	97%	97%	97%
Education	76%	72%	74%
Work Experience	90%	75%	81%
Project	53%	47%	50%
Skills	75%	70%	72%
Other	41%	40%	40%

Feature Extraction Result

Attributes	Precision	Recall	F1-score
Name	92%	90%	91%
Email	92%	90%	91%
Phone	92%	90%	91%
Profile Link	81%	77%	78%
Address	65%	57%	60%
University	88%	75%	80%
Major	70%	65%	67%
Date	55%	35%	42%
Degree	65%	60%	62%
GPA	60%	40%	48%

Conclusion

- The proposed system produces better results for resume parsing due to the use of content-based and layout-based techniques