RESUME CLASSIFICATION AND SCORING USING NLP AND MACHINE LEARNING

PARIMALA MUDIMELA

INTRODUCTION

REVA UNIVERSITY

PROBLEMS WITH MANUAL SCREENING OF RESUMES

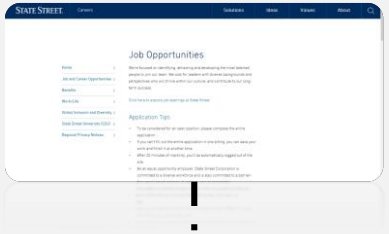
- Recruiters scan resumes by keeping people for manually resume scanning. Or if they are dealing with very few resumes then they scan themselves. The ways mostly depend on the number of resumes the State street is getting on a **day to day basis**.

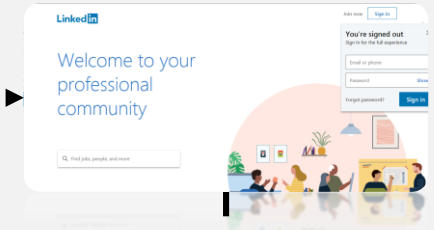
- There are lots of pain points and disadvantages by doing manual screening (Job skills, Experience, Company, College), When dealing with recruitment process more than 50% of the time is spent on screening the resumes. Even after spending time the chances of error and missing a correct candidate is still high.

AS-IS Process

REVA UNIVERSITY

JOB DESCRIPTION
Crafting a compelling job description is essential to helping us attract the most qualified candidates for any job.





ONLINE ADVERTISEMENT
HR creates a job advertisement, which is posted on our website or on Social Media websites like LinkedIn.

APPLY ONLINE
Candidates apply online that confirms receipt of application.

SCREENING
HR evaluates all applications and also reviews candidates who have previously registered with State street. Profiles matching the requirements of the advertised position are selected to proceed to the next step.

...Continued AS-IS Process

- CHALLENGES IN SCREENING PROCESS

SCREENING PROCESS:
Screening resumes usually involves a three-step process based on the role's minimum and preferred qualifications. Both types of qualifications should be related to on-the-job performance and are ideally captured in the job description.

Step 1: Screening resumes based on minimum qualifications
Step 2: Screening resumes based on preferred qualifications
Step 3: Shortlisting candidates based on minimum and preferred qualifications

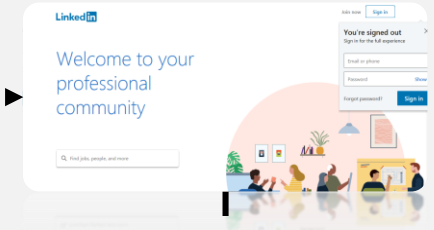
Every 100 candidates screened, we need to shortlist 12 of them to interview, two of them will receive an offer, and one candidate will accept to result in one successful hire.

The important thing to remember is that screening process is applied consistently and objectively across all resumes.

This is where the recruitment process can get a little challenging. Almost all candidates who have applied for the job are well qualified. So how can we zero-in on the right candidate(s)?

To-BE Process

JOB DESCRIPTION
Crafting a compelling job description is essential to helping us attract the most qualified candidates for any job.

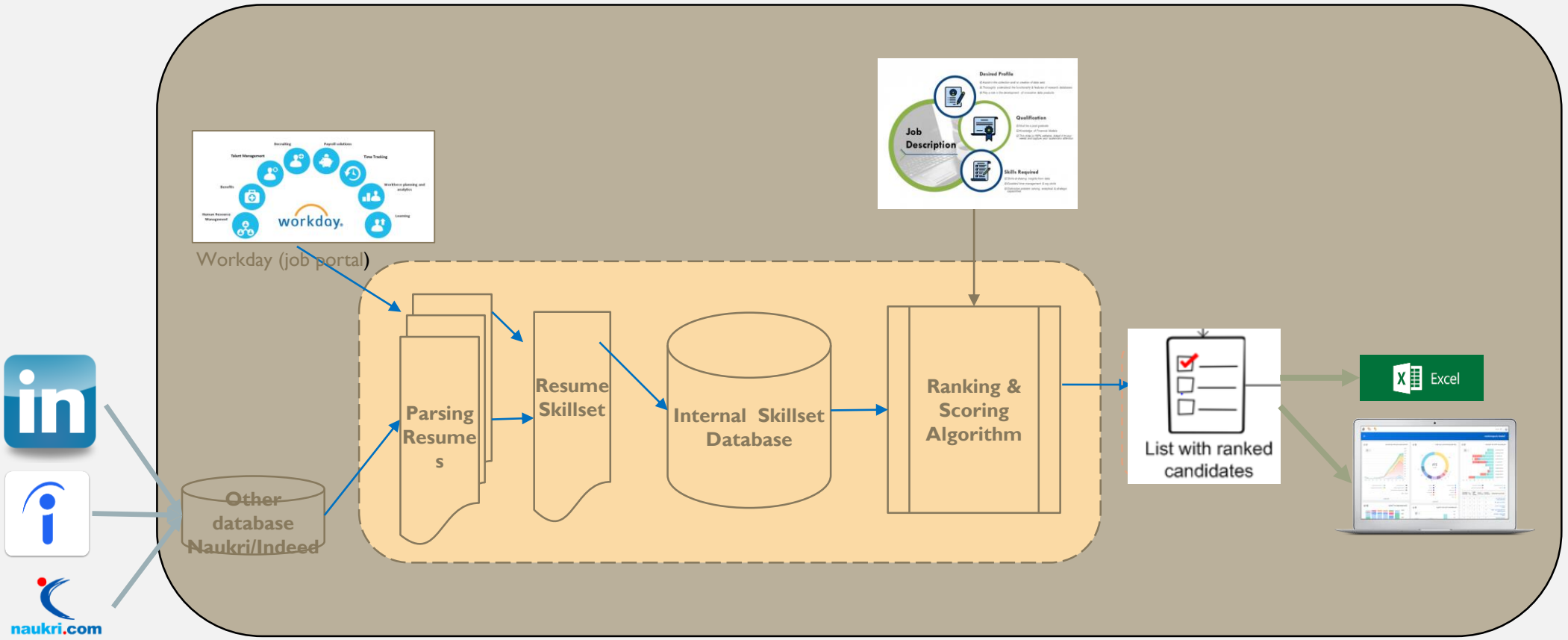


ONLINE ADVERTISEMENT
HR creates a job advertisement, which is posted on our website or on Social Media websites like LinkedIN.

APPLY ONLINE
Candidates apply online that confirms receipt of application.

WEB APPLICATION
By using a Excel/Web APP we can easily make decision according to Job preferences with less or no manual intervention. We can sort and filter only necessary resumes.

Technical Architecture



Project Requirements

The software requirements in this project include:

- Python
- NLTK
- ML
- PostgreSQL
- Flask
- HTML
- CSS

BUSINESS OBJECTIVE

REVA UNIVERSITY

The business objective is to develop an automatic system to screen the resumes and rank them

DATA UNDERSTANDING

- Data comprises of unstructured resumes which are in the form of .doc or .docx or pdf format
- Resumes is collected from various job portals like Workday, LinkedIn ,Naukri etc.
- Job Description which is in the format of .doc or .docx or .pdf.
- Unstructured resumes are converted in to structured text format.

DATA COLLECTION

REVA UNIVERSITY

- Resumes are collected from Workday using API(application Programmable Interface).A python script is written to communicate with the Workday API and extract the resume data.
- Data are collected from LinkedIn using Selenium(mimics the human behavior)

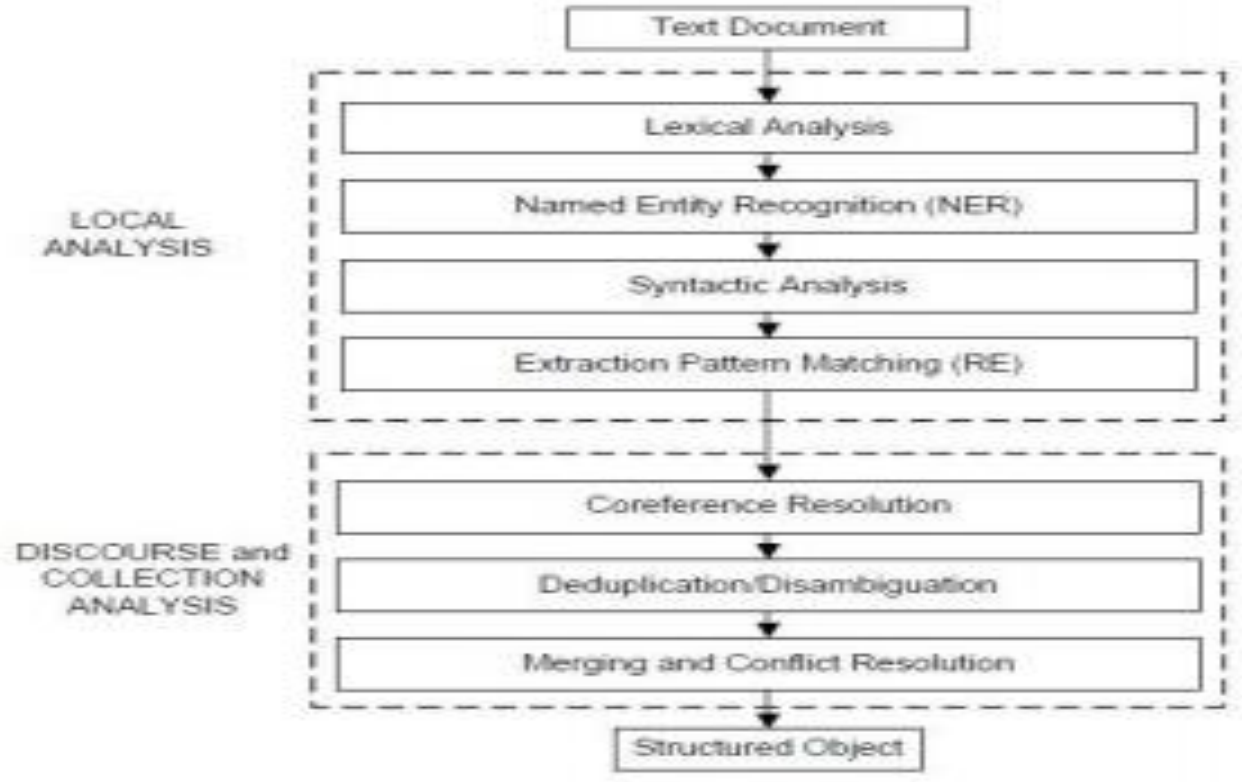
DATA PREPARATION

The various attributes from the resumes like

- Experience
- Technical Skills
- Non Technical skills
- Email
- Phone Number
- Education

are extracted from the resume documents and stored in text format.

RESUME PARSER WORKFLOW



INFORMATION EXTRACTION PROCESS

REVA UNIVERSITY

The designed Information Extraction System consists of 4 phases:

- Text Segmentation
- Named Entity Recognition
- Named Entity Clustering
- Text Normalization

TEXT SEGMENTATION

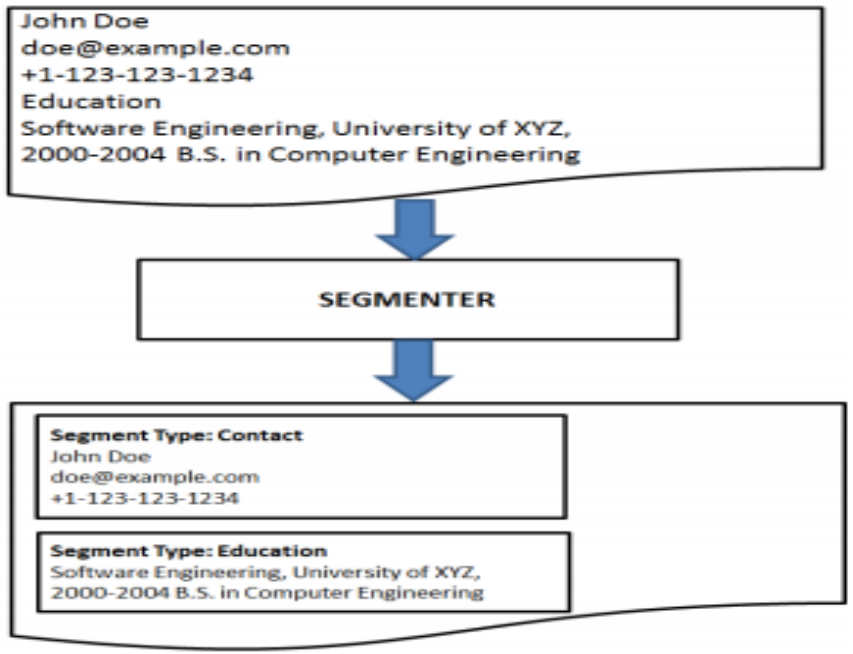


Figure2. Segmentation in the Resume by Segmenter

Segment Type	Related Info under the Segment
Contact	Name
	Phone
	Email
	Web
Education	Degree
	Program
	Institution
Experience	Position
	Company
	Date Range

NAMED ENTITY RECOGNITION

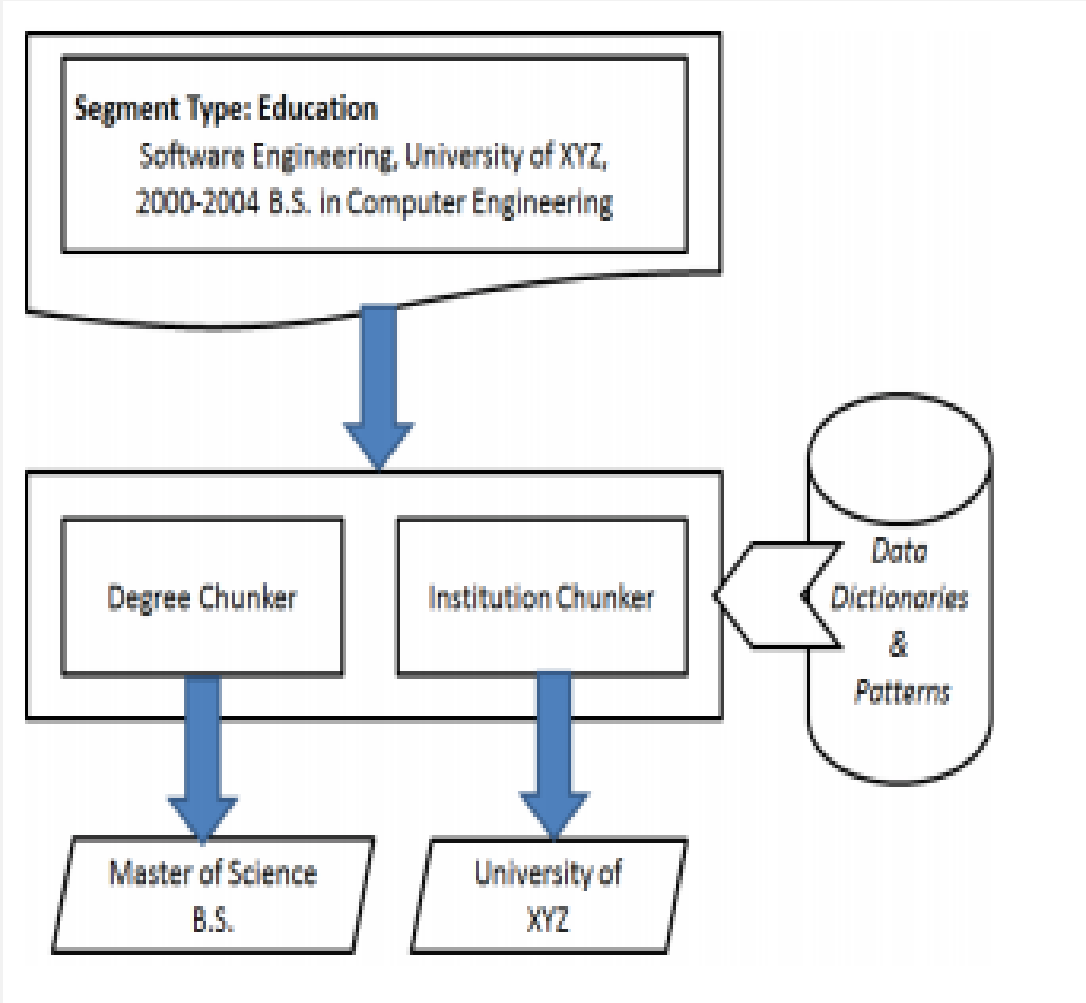
REVA UNIVERSITY

- The tokenized text documents are fed to a Named Entity Recognizer.
- The term Named Entity refers to noun phrases (type of token) within a text document that have some predefined categories like Person, Location, Organization ,expressions of times, quantities, monetary values, percentages, etc. Numeric values like phone numbers and dates are also Named Entities.
- Resumes consist of mostly named entities and some full sentences. Because of this nature of the resumes, the most important task is to recognize the named entities. For each type of information, there is specially designed chunker.

NAMED ENTITY RECOGNITION..CONT

Chunkers use four types of information to find named entities:

- Clue words: like prepositions (e.g. in the work experience information segment the word after "at" most probably a company name)
- Well Known or Famous names: Through data-dictionaries of well-known institutions, well known places, companies or organization, academic degrees, etc.
- From prefixes and suffixes of word: For institutions (e.g. University of, College etc.) and companies (e.g. Corp., Associates, etc.)
- Style of Writing Name of person: Generally the name of the person is written as First Letter capitalize then we will guess that this word possibly name of person.



NAMED ENTITY CLUSTERING

REVA UNIVERSITY

- Each segment (e.g. education information) contains a block of related information. For example, an education segment will have a number of blocks of information about educational institutions that a person attended. For example, an education information block can contain institution name, degree, major, and date information
- The named entity are in the block of information are need to be grouped together to do the more process on it.
- Named entities (chunks) are grouped according to their proximity and type.

TEXT NORMALIZATION

REVA UNIVERSITY

- In text normalization, some of the named entities are transformed to make it consistent.
- In normalization phase, we expand some of abbreviations using dictionaries

Input	Output	Type
B.S.	Bachelor of Science	Degree
JOHN DOE	John Doe	Name
University of ABC	ABC University	Institution

Term (Full Form/word)	Abbreviation
Bachelor of Science	B.S.,BS ,BSc
Master of Science	M.S.,MS ,MSc
Bachelor of Arts	B.A., BA
Doctor of Philosophy	Ph.D., PhD
Doctor of Medicine	Medicine Doctor, M.D.
Bachelor of Computer Application	BCA,B.C.A

CALCULATION OF SCORE

REVA UNIVERSITY

$$S = \frac{|\{Sr\}|}{|\{RSj\}|} * 50\% + \frac{|\{Er\}|}{|\{REj\}|} * 20\% + \frac{|\{Xr\}|}{|\{RXj\}|} * 20\% + \frac{|\sum Yw|}{|\sum Cw|} * 10\% \quad (1)$$

Where:

- **S:** is the relevance score result.
- **Sr:** is the set of applicant's skills.
- **RSj:** the required skills in the job post.
- **Er:** is the set of concepts that describe applicant educational information.
- **REj:** is the set of concepts from the required educational information in job post.
- **Xr:** set of concepts that describe applicant experience information.
- **RXj:** concepts that represent the required experience information in the job post.
- **Yw:** the total number of employment years.
- **Cw:** number of companies that the applicant worked in.

As shown in the formula, we have set the following weighting values:

Skills weight = 50%, Educational level weight = 20%, Job experience weight = 20% and Loyalty level weight = 10%. The results of using the scoring formula are detailed in the next section

MODEL EVALUATION

REVA UNIVERSITY

- Used precision, recall, and F-measure metrics for performance evaluation.
- Precision measures the amount of relevant items retrieved as a percentage of the entire number of things retrieved.
  Precision= #(relevant items retrieved)/#(retrieved items)
- Recall measures the amount of relevant items retrieved as a percentage of the amount of relevant items within the collection.
  Recall= #(relevant items retrieved)/#(relevant items)

Recognized Entity	Precision	Recall	F-Score
College Name	100.0%	100.0%	100%.0
Location	100.0%	97.78%	98.88%
Designation	100.0%	100.0%	100.0%
Email Address	95.83%	100.0%	97.87%
Name	100.0%	100.0%	100%.0
Skills	96.36%	96.36%	96.36%
Years of Experience	100.0%	100.0%	100.0%
Graduation Year	96.55%	87.50%	91.80%
Degree	100.0%	100.0%	100.0%
Companies worked at	98.08%	100.0%	99.03%

CONCLUSION

REVA UNIVERSITY

- In this project, we came with a web application that helps to shortlist the resumes of the applicants for a job posting.
- In piloting this project with screeners we estimated that using the tool roughly sped up the screening process by a factor of 20 as compared to manual screening.
- This can result in great savings in time and reduce the cost of hiring.
- Recruiters can spend more time in other important tasks


Thank
you