TEXTURE: Extracting Data from Text Highlights

Maeda Hanafi (maeda.hanafi@nyu.edu), Azza Abouzied (azza@nyu.edu)

THE PROBLEM

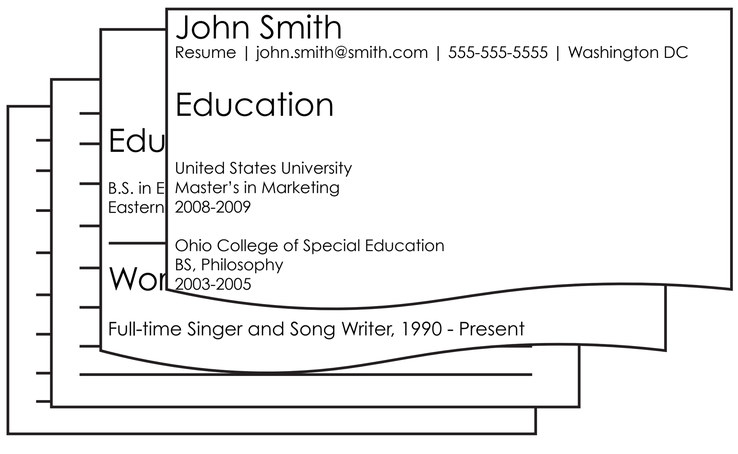
- Users want to organize structured information from large collections of documents without manually extracting this information from each document.
- Examples:
  - An Islamic scholar trying to extract narrator chains from a large collection of prophetic narrations known as Hadiths.
  - An employer with hundreds of resumes trying to find each candidate's school, major, and year.
  - A researcher trying to extract titles, abstracts, and images from a large collection of research papers.

THE SOLUTION: TEXTURE

- Texture is a tool for synthesizing data extraction scripts from user examples.
- Users highlight areas of the text they wish to extract.
- Texture learns a script that best describes the patterns of the highlighted areas.

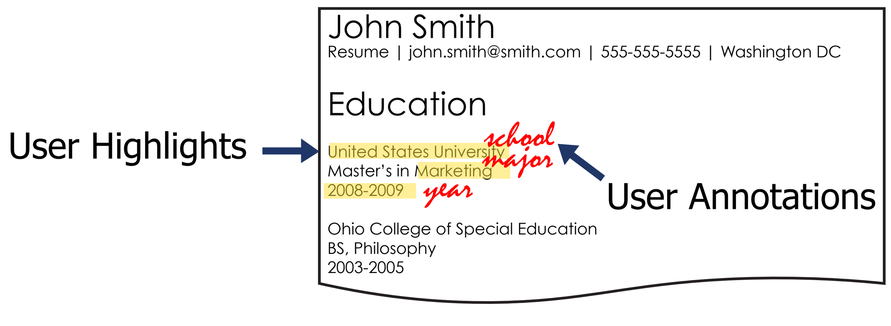
  Texture then applies the script to the entire collection and users can visually examine the correctness of the inferred script.
- Users can provide more highlighting examples to refine Texture's learned data extraction scripts.

1 UPLOAD DOCUMENT COLLECTION



2 HIGHLIGHT AND ANNOTATE

Users highlight areas of text they wish to extract such as an applicant's school, major, and graduation year.



3 TEXTURE LEARNS TO EXTRACT

Concepts are the building blocks of extraction scripts.

- Regular expressions that describe string patterns.

```
Digit Digit Digit Digit - Digit Digit Digit Digit
    1996 - 1997

1 Capital Letter Lower-case Letters +
    Master's in Computer Science
```

- Elements of a labelled dictionary:

```
University Dictionary
  New York University Abu Dhabi
  University of Connecticut
  University of Michigan at Dearborn
```

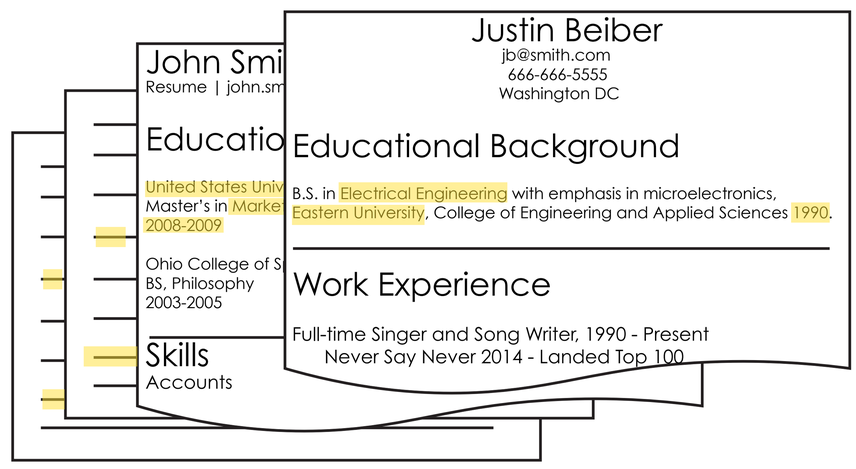
Learning means
1) Synthesizing every valid regular expression
2) Then, picking out the smallest expression or determining which dictionary the highlighted text belongs to.

Extraction scripts are simple production rules:

- Relationships of concepts.

```
school := in (university dictionary)
                word+ "university"
education := permutation (school, year, degree)
year := digit digit digit digit
```

4 TEXTURE EXTRACTS DATA



NAME	SCHOOL	MAJOR	YEAR
John	US Univ.	Marketing	2009
Justin	Eastern U	Electrical Eng.	1990
Maeda	SCSU	Biology	2012
Ayesha	CMU	Philosophy	2009
Juan	NYUAD	Science	2003

Red-colored rows indicate the script's confidence of the results. The user can accept the results or provide more annotations to refine the extraction script.