Name: Alexis Collier

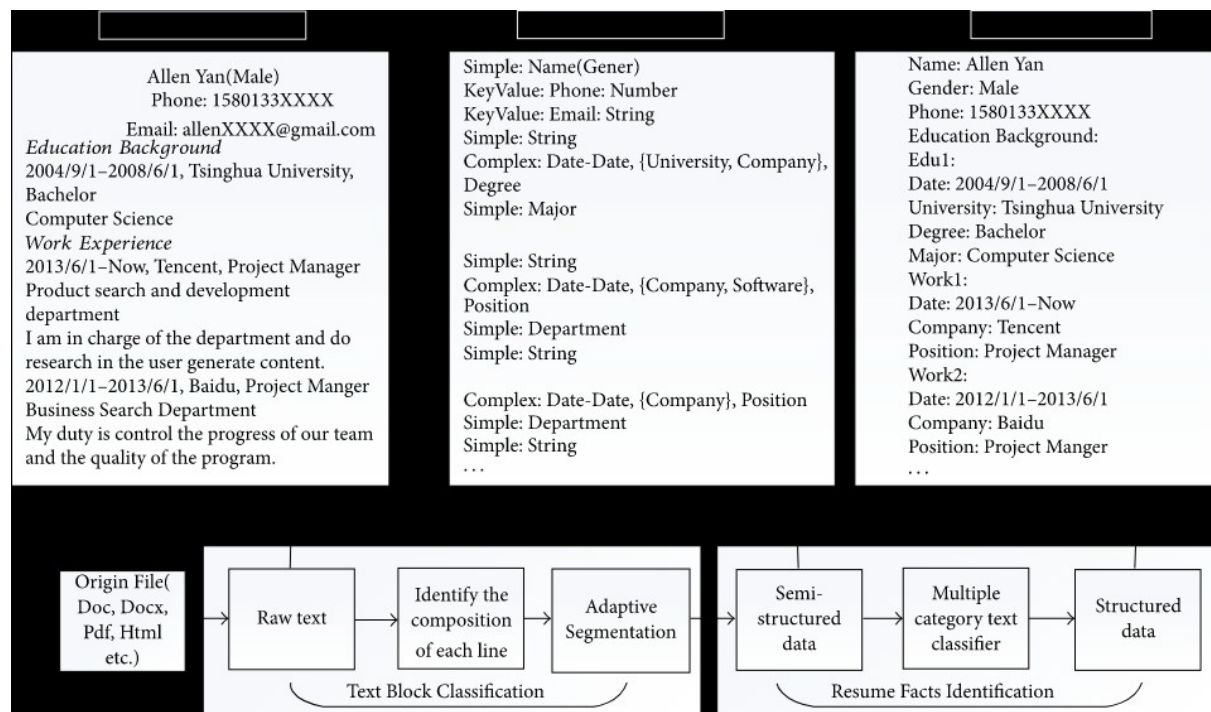**Email:** info@alexiscollier.com

Country: United States

College: FullStack Academy

Specialization: NLP

Problem description:

The resumes contain information irrelevant to the HR/authority and must manually be processed to shortlist the promising candidates, making the shortlisting task a herculean task for HR. Using the NER (Named Entity Recognition) model of NLP, this problem can be solved by finding and classifying the entities present in each resume into predefined classes such as person name, college name, academic information, relevant experiences, skill set, etc.

Data cleaning:

I will first be putting the JSON data into a more appropriate format, getting rid at the same time of the unnecessary spaces on the right and the left and only keeping the start point and end point of each label (I will not need the text itself).

This is what the annotated text looks like:

```
{'entities': [(1749, 1755, 'Companies worked at'),
  (1696, 1702, 'Companies worked at'),
  (1417, 1423, 'Companies worked at'),
  (1356, 1793, 'Skills'),
  (1209, 1215, 'Companies worked at'),
  (1136, 1247, 'Skills'),
  (928, 932, 'Graduation Year'),
  (858, 889, 'College Name'),
  (821, 856, 'Degree'),
  (787, 791, 'Graduation Year'),
  (744, 750, 'Companies worked at'),
  (722, 742, 'Designation'),
  (658, 664, 'Companies worked at'),
  (640, 656, 'Designation'),
  (574, 580, 'Companies worked at'),
  (555, 572, 'Designation'),
  (470, 493, 'Companies worked at'),
  (444, 468, 'Designation'),
  (308, 314, 'Companies worked at'),
  (234, 240, 'Companies worked at'),
  (175, 198, 'Companies worked at'),
  (93, 136, 'Email Address'),
  (39, 48, 'Location'),
  (13, 37, 'Designation'),
  (0, 12, 'Name')]}
```

I will now be replacing the "\n" with simple spaces:

This is how the plain text looks like now:

```
dt[0][0]

'Govardhana K Senior Software Engineer  Bengaluru, Karnataka, Karnataka - Email me on Indeed: indeed.com/r/
Salesforce Developer Oracle 5 Years 2 Month • Core Java Developer Languages Core Java, Go Lang Oracle PL-SC
WORK EXPERIENCE  Senior Software Engineer  Cloud Lending Solutions -  Bangalore, Karnataka -  January 2018
Staff Consultant  Oracle -  Bangalore, Karnataka -  January 2014 to October 2016  Associate Consultant  Ora
ing  Adithya Institute of Technology -  Tamil Nadu  September 2008 to June 2012  https://www.indeed.com/r/C
```

Next, using regex and simple Python code, I will remove leading and trailing white spaces from entity spans. This is the final output:

```
data[0][0]

'Govardhana K Senior Software Engineer  Bengaluru, Karnataka, Karnataka - Email m
Salesforce Developer Oracle 5 Years 2 Month • Core Java Developer Languages Core
WORK EXPERIENCE  Senior Software Engineer  Cloud Lending Solutions -  Bangalore,
Staff Consultant  Oracle -  Bangalore, Karnataka -  January 2014 to October 2016
ing  Adithya Institute of Technology -  Tamil Nadu  September 2008 to June 2012
K/b2de315d95905b68?isid=rex-download&ikw=download-top&co=IN  SKILLS  APEX. (Less
m/in/govardhana-k-61024944/  ADDITIONAL INFORMATION  Technical Proficiency:  Lang
r, NetBeans, Eclipse, SQL developer, PL/SQL Developer, WinSCP, Putty Web Technolc
ddleware: Web logic, OC4J Product FLEXCUBE: Oracle FLEXCUBE Versions 10.x, 11.x a
```

```
data[0][1]

{'entities': [[1749, 1755, 'Companies worked at'],
  [1696, 1702, 'Companies worked at'],
  [1417, 1423, 'Companies worked at'],
  [1356, 1793, 'Skills'],
  [1209, 1215, 'Companies worked at'],
  [1136, 1247, 'Skills'],
  [928, 932, 'Graduation Year'],
  [858, 889, 'College Name'],
  [821, 856, 'Degree'],
  [787, 791, 'Graduation Year'],
  [744, 750, 'Companies worked at'],
  [722, 742, 'Designation'],
  [658, 664, 'Companies worked at'],
  [640, 656, 'Designation'],
  [574, 580, 'Companies worked at'],
  [555, 572, 'Designation'],
  [470, 493, 'Companies worked at'],
  [444, 468, 'Designation'],
  [308, 314, 'Companies worked at'],
  [234, 240, 'Companies worked at'],
  [175, 198, 'Companies worked at'],
  [93, 136, 'Email Address'],
  [39, 48, 'Location'],
```

The entities of the data are now in lists, which makes them easier to iterate and use.

I will be transforming data to a more suitable shape depending on my choice of model for this project.

GitHub Repo link: https://github.com/colla00/NLP-Resume-Extraction-Project