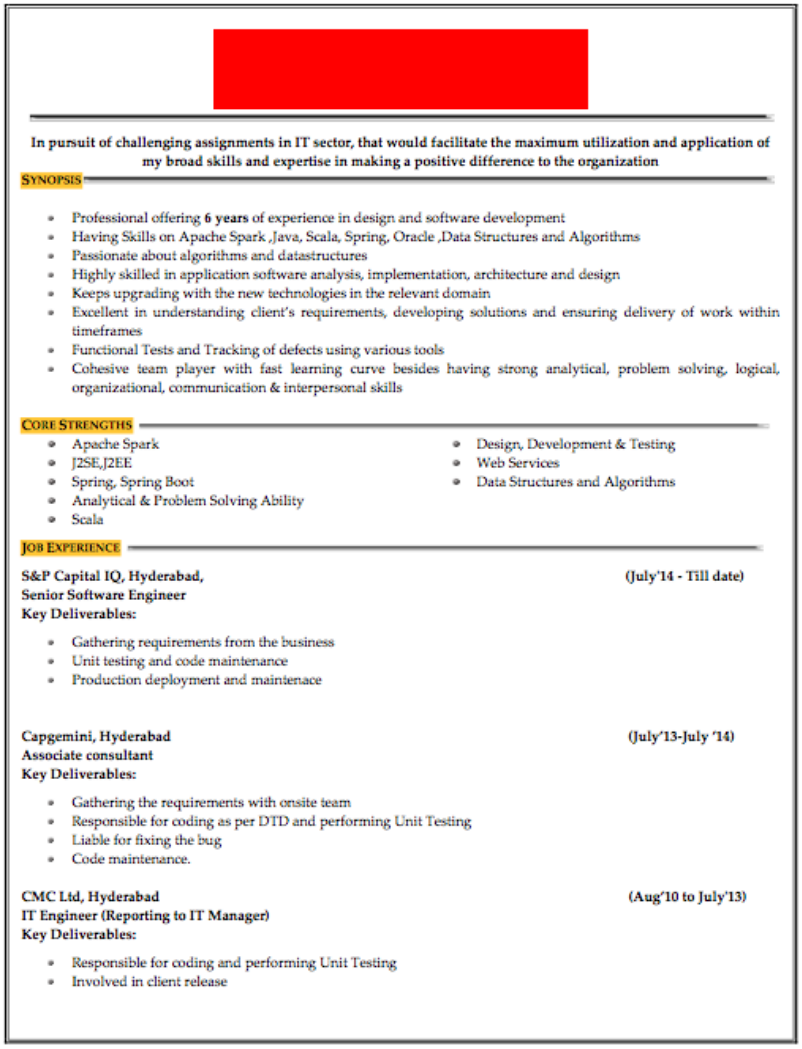
Problem Statement

You are provided a dataset of 83 documents. Each document is a CV/Resume of a person. Most resumes have a hierarchical structure consisting of sections and sub-sections. The task assigned to you is to detect starting words of all top-level sections in a resume.

To illustrate, shown below is a screenshot of a sample resume. There are a bunch of words highlighted in yellow colour. These are the starting words of each section. We will call such words as ‘section marker words’.



[redacted]

In pursuit of challenging assignments in IT sector, that would facilitate the maximum utilization and application of my broad skills and expertise in making a positive difference to the organization

SYNOPSIS

- Professional offering **6 years** of experience in design and software development
- Having Skills on Apache Spark ,Java, Scala, Spring, Oracle ,Data Structures and Algorithms
- Passionate about algorithms and datastructures
- Highly skilled in application software analysis, implementation, architecture and design
- Keeps upgrading with the new technologies in the relevant domain
- Excellent in understanding client's requirements, developing solutions and ensuring delivery of work within timeframes
- Functional Tests and Tracking of defects using various tools
- Cohesive team player with fast learning curve besides having strong analytical, problem solving, logical, organizational, communication & interpersonal skills

CORE STRENGTHS

- Apache Spark
- J2SE,J2EE
- Spring, Spring Boot
- Analytical & Problem Solving Ability
- Scala
- Design, Development & Testing
- Web Services
- Data Structures and Algorithms

JOB EXPERIENCE

S&P Capital IQ, Hyderabad, **(July'14 - Till date)**
Senior Software Engineer
Key Deliverables:

- Gathering requirements from the business
- Unit testing and code maintenance
- Production deployment and maintenace

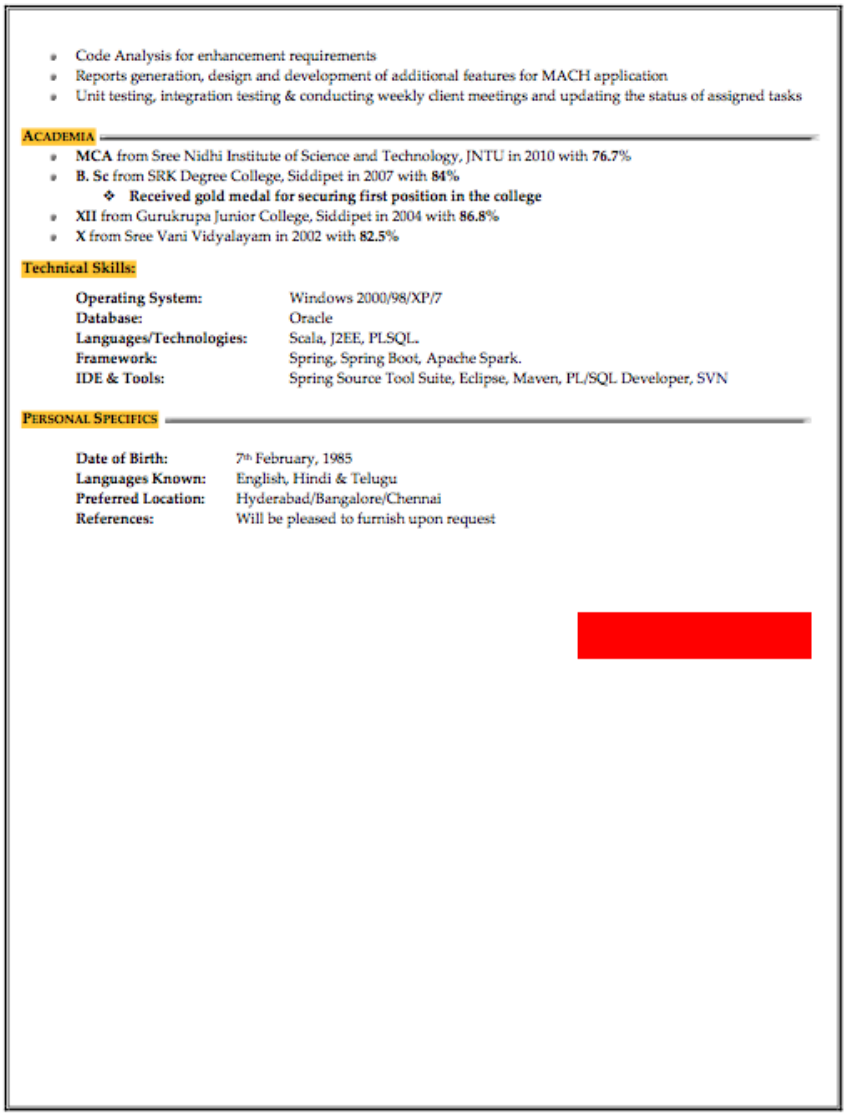
Capgemini, Hyderabad **(July'13-July '14)**
Associate consultant
Key Deliverables:

- Gathering the requirements with onsite team
- Responsible for coding as per DTD and performing Unit Testing
- Liable for fixing the bug
- Code maintenance.

CMC Ltd, Hyderabad **(Aug'10 to July'13)**
IT Engineer (Reporting to IT Manager)
Key Deliverables:

- Responsible for coding and performing Unit Testing
- Involved in client release

Confidential and Proprietary
Copyright(c) 2017. Aganitha cognitive solutions private limited. All rights reserved.



- Code Analysis for enhancement requirements
- Reports generation, design and development of additional features for MACH application
- Unit testing, integration testing & conducting weekly client meetings and updating the status of assigned tasks

ACADEMIA

- **MCA** from Sree Nidhi Institute of Science and Technology, JNTU in 2010 with **76.7%**
- **B. Sc** from SRK Degree College, Siddipet in 2007 with **84%**
  - ❖ **Received gold medal for securing first position in the college**
- **XII** from Gurukrupa Junior College, Siddipet in 2004 with **86.8%**
- **X** from Sree Vani Vidyalayam in 2002 with **82.5%**

Technical Skills:

Operating System:	Windows 2000/98/XP/7
Database:	Oracle
Languages/Technologies:	Scala, J2EE, PLSQL.
Framework:	Spring, Spring Boot, Apache Spark.
IDE & Tools:	Spring Source Tool Suite, Eclipse, Maven, PL/SQL Developer, SVN

PERSONAL SPECIFICS

Date of Birth:	7th February, 1985
Languages Known:	English, Hindi & Telugu
Preferred Location:	Hyderabad/Bangalore/Chennai
References:	Will be pleased to furnish upon request

We have made the job a little simpler for you by providing you pre-processed data from resumes. You have access to each word of the document. You are also given a few features (discussed in next section) related to each word. You have to predict whether the word is a section marker word or not.

Data Format

Each document is represented in the pickle file format (pickled using Python 2). To know more about how to load pickle files, refer to https://docs.python.org/2/library/pickle.html.

Confidential and Proprietary
Copyright(c) 2017. Aganitha cognitive solutions private limited. All rights reserved.

Each pickle file contains two lists:

1. Word List
2. Output List

Input Format

Each element of a word list denotes one word of the document and they are ordered in left to right reading order. Each element of the word list is again a list of size 16 consisting of following word features:

1. Word ID - We have used NLTK's english word corpus to convert the words to unique ID's. Any word which is not present in the english vocabulary has been given an ID corresponding to <UNK> token.
2. X Coordinate - Float value denoting the starting x(top left) co-ordinate of the word.
3. Y Coordinate - Float value denoting the starting y(top left) co-ordinate of the word.
4. Page ID - String denoting the page where the word occurs.
5. Font Style - String denoting the Style of the font used for that word.
6. Font Color - String denoting the color of the font used for that word.
7. Font Size - Float value denoting the font size for that word.
8. Font Bold - 1 indicating bold formatting used for that word, else 0.
9. Boolean variable denoting if the word starts with an uppercase character.
10. Boolean variable denoting if the word consists of all uppercase characters.
11. Boolean variable denoting if the word is a single digit.
12. Boolean variable denoting if the word is a number (multiple digits).
13. Boolean variable denoting if the word is a number but with special characters like ',' and '.'.
14. Boolean variable denoting if the word is an email address.
15. Boolean variable denoting if the word is a hyperlink.
16. Boolean variable denoting if the word is a calendar month.

Output Format

Each output list is of size equal to the number of words in the document. Each element is a label which indicates if the corresponding word is a section marker word(1) or not(0) .

NLTK Word Vocabulary format

You are also given a file named `nltk_words.vocab`, providing word to ID map as a json. The keys are the words (in lowercase) and corresponding values are the IDs assigned to them. You can use this map to reconstruct the words in original document if you see the necessity to do so. Any out of vocabulary word has been given an ID equal to 'size of vocabulary + 1'.

Confidential and Proprietary
Copyright(c) 2017. Aganitha cognitive solutions private limited. All rights reserved.

To retrieve the mapping as a python `dict`, you can use the following python code.

```
import codecs
import json

with codecs.open(vocabPath, 'r', encoding="UTF-8") as f:
    word2id = json.load(f)
```

Platform instructions

Pickling has been done using python 2 and hence please stick to python 2. Python 3 will fail in unpickling.

Confidential and Proprietary
Copyright(c) 2017. Aganitha cognitive solutions private limited. All rights reserved.