Table Reconstrcution

1 Introduction

PDF documents have become the mainstream document format because of its unique cross-platform convenience advantage. PDF documents contain a large amount of valuable data information, and the table is one of the important carriers of these data. However, the structure of PDF documents is complex, and it is difficult for us to obtain accurate table information directly from the document format. Therefore, for PDF tables, we need to reconstruct the structure of the table, so as to achieve the extraction of the table.

2 Purpose

This Lab focuses on the table line reconstruction for the tables without frame lines. In this Lab, you will use python to complete table line drawing of a specific table without frame lines and show me the result in the lab report.

You need to find and cut any one table without frame lines from any PDF document and draw the table line for it. After programming, you need to answer two questions. In this Lab, we have prepared a bonus question, if you can answer this question well, you will get extra points from the teaching assistant.

Table 2
$^{40}Ar/^{39}Ar$ closure temperature calculations for samples mentioned in the text[a]

Sample	Location	Mineral	Composition	Diff radius, a (μm)	Act. energy, E (cal/mol)	Do/a^2	dT/dt (°C/Ma)	EdT/dt ($\times 10^{-9}$)	Approximate T_c (°C)	Ages, Ma (+/−)		%^{39}Ar (plateau age)
										Integrated	Plateau	
W. San. Felsic Dyke	Santoy	Biotite	Ann65	60	45 000	2139	30	42.81	290	1705 (7)	1757 (7)	82.4
9088	Santoy	Biotite	Ann80	1000	42 000	7.7	30	39.95	335	1711 (17)	1713 (6)	58.8
9222-20	Santoy	Biotite	Ann55	100	47 000	770	30	44.71	330	1709 (6)	1732 (6)	88.4
9222-62	Santoy	Biotite	Ann40	90	50 000	951	30	47.56	365	1733 (8)	1733 (8)	100.0
9222-73	Santoy	Hornblende	Ferroan pargasite	90	64 100	296	2.5	4.065	480	1715 (11)	1716 (11)	99.2
9222-9	Santoy	Hornblende	Ferroan pargasite	90	64 100	295	2.5	4.065	480	1713 (9)	1716 (9)	98.7
9222-56	Santoy	Hornblende	Ferroan pargasite	140	64 100	122	2.5	4.065	495	1711 (18)	1717 (19)	95.6
9222-41	Santoy	Hornblende	Ferroan pargasite	170	64 100	83	2.5	4.065	505	1737 (19)	1741 (19)	97.8
8822-1099	Brownell	Biotite	Ann55	90	47 000	951	4	4.471	300	1713 (8)	1727 (8)	91.9
kbl-1	Brownell	Biotite	Ann55	140	47 000	393	4	4.471	310	1827 (15)	1759 (6)	92.9
kbl-9	Brownell	Biotite	Ann55	125	47 000	493	4	4.471	310	1719 (7)	1745 (7)	73.5
kbl-2	Brownell	Biotite	Ann45	80	49 000	1203	4	4.661	320	1738 (7)	1743 (7)	96.7
kbl-8	Brownell	Biotite	Ann45	70	49 000	1571	4	4.661	315	1724 (8)	1756 (7)	90.4

Figure 1: Table line reconstruction

3 Question

1. How to automatically locate the tables in a PDF?
2. What do you think is the most difficult step to extract the table from the PDF? why?

4 Bonus question

1. How to accurately identify the header of the table, and use natural language processing (NLP) or other methods to understand the information in the table, and then extract the entities and relationships from table to construct a specific knowledge graph?