# CS767 Extra Project

**Pragya Sen,** BU ID**: U97864862**

**OBJECTIVE:** Develop a system that extracts affiliation and contact details of the first and last authors from the enclosed PDF file containing academic paper references and outputs the data into structured Excel files.

**DATASET:** Used the dataset given with 27 pages worth of references in a PDF file.

**METHODOLOGY:**

Step 1: PDF OCR Extraction

1. Extracted scanned text using Tesseract OCR.
2. Cleaned text by normalizing quotes and merging broken lines.

Step 2: LLM-Based Reference Parsing

1. Passed reference blocks to LLaMA 3 via Ollama to extract: Reference title, First author, Last author
2. Returned structured JSON for reliability.

Step 3: Multi-Source Online Enrichment

| **Source** | **Purpose** |
|---|---|
| Google Scholar | Academic affiliation & verified profile |
| CrossRef API | DOI inference + affiliation by metadata |
| DuckDuckGo | Email + affiliation snippets from the web |
| LLaMA Fallback | Infer missing affiliation/email |
| Local Cache | Avoid repeated lookups |

Step 4: Final Export

Output 2 Excel files:

1. **OCR_output.xlsx** – raw extracted references
2. **Verified_output.xlsx** – enriched author data

**Results:**

1. **OCR_output.xlsx**:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | reference | title | first_author | last_author |
| 2 | Wiener '48 | Time, communication, and the ne | N. Wiener | |
| 3 | [Fuegi '03] Fuegi, J., & Francis, J. | Lovelace & Babbage and the crea | Fuegi | Francis |
| 4 | Kagermann '11] Kagermann, H., | Industrie 4.0: Mit dem Internet d | Kagermann | Wahlster |
| 5 | Hey '09] Hey, T., Tansley, S., & T | The Fourth Paradigm: Data-Inten | Hey, T. | Tolle, K. M. |
| 6 | Mitchell '97 | Machine Learning | T.M. Mitchell | |
| 7 | [Géron '19] Géron, A. (2019). Ha | Hands-on Machine Learning with | A. Géron | |
| 8 | [Bayes '63] Bayes, T. (1763). LIT. | An Essay Towards Solving a Probl | Bayes, T. | |
| 9 | Russell '21] Russell, S., Norvig, P. | Artificial Intelligence: A Modern A | S. Russell | P. Norvig |
| 10 | [Chapmann '17] Chapmann, J. (2 | Machine Learning: Fundamental / | Chapmann | J. |
| 11 | [Knuth '97] Knuth, D. E. (1997). T | The Art of Computer Programmin | D. E. Knuth | |
| 12 | [Bhargava '16] Bhargava, A. (201 | Grokking Algorithms | A. Bhargava | |
| 13 | [Han '11] Han, J., Pei, J., & Kamb | Data Mining: Concepts and Techn | J. Han | M. Kamber |
| 14 | [Allen '83] Allen, J. F. (1983). Ma | Maintaining knowledge about ter | J. F. | Allen |
| 15 | [Wolpert '97] Wolpert, D.H., Ma | No Free Lunch Theorems for Opti | Wolpert | Macready |
| 16 | [Topol '19] Topol, E. (2019). Dee | Deep medicine: how artificial inte | E. Topol | Topol |
| 17 | [Zheng '18] Zheng, A., & Casari, A | Feature Engineering for Machine | Zheng, A. | Casari, A. |
| 18 | [Deisenroth '19] Deisenroth, M.I | Mathematics for Machine Learnir | Deisenroth, M.P | Ong C.S |

2. **Verified_output.xlsx**:

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | reference | title | first_author | last_author | first_author_affiliation | first_author_email | last_author_affiliation | last_author_email | |
| 2 | Wiener '48 | Time, communication, and the nervous sys | N. Wiener | | Massachusetts Institute of T | Not publicly available (or unknown) | | | |
| 3 | [Fuegi '03] Fu | Lovelace & Babbage and the creation of the | Fuegi | Francis | Flare/MITH | fuegi@uwm.edu | Flare/MITH | | |
| 4 | Kagermann ' | Industrie 4.0: Mit dem Internet der Dinge a | Kagermann | Wahlster | Germany | kagermann@acatec | Germany | wahler@forschungsgesellschaft-auto | |
| 5 | Hey '09] Hey | The Fourth Paradigm: Data-Intensive Scien | Hey, T. | Tolle, K. M. | University of Washington | | Microsoft Research | | |
| 6 | Mitchell '97 | Machine Learning | T.M. Mitchell | | | | Machine Learning | | |
| 7 | [Géron '19] ( | Hands-on Machine Learning with Scikit-Lea | A. Géron | | | | | | |
| 8 | [Bayes '63] B | An Essay Towards Solving a Problem in the | Bayes, T. | | Fellow of the Royal Society | | | | |
| 9 | Russell '21] F | Artificial Intelligence: A Modern Approach | S. Russell | P. Norvig | Stanford University | stuart@cs.stanford | Google Research (formerly | pnorvig@norvignon.com | |
| 10 | [Chapmann | Machine Learning: Fundamental Algorithm | Chapmann | J. | | | | | |
| 11 | [Knuth '97] K | The Art of Computer Programming | D. E. Knuth | | Stanford University (emeriti | knuth@cs.stanford.edu | | | |
| 12 | [Bhargava '1 | Grokking Algorithms | A. Bhargava | | University of California, Ber | abhart@eecs.berkeley.edu | | | |
| 13 | [Han '11] Ha | Data Mining: Concepts and Techniques | J. Han | M. Kamber | University of Wisconsin-Ma | han@cs.wisc.edu | Monash University | michael.kamber@monash.edu | |
| 14 | [Allen '83] Al | Maintaining knowledge about temporal int | J. F. | Allen | | | | | |
| 15 | [Wolpert '97 | No Free Lunch Theorems for Optimization | Wolpert | Macready | University of Texas at Austi | wolpert@cs.utexas. | Sarnoff Corporation | macready@sarnoff.com | |
| 16 | [Topol '19] T | Deep medicine: how artificial intelligence c | E. Topol | Topol | Scripps Research Translatio | etopol@scripps.edu | | | |
| 17 | [Zheng '18] Z | Feature Engineering for Machine Learning: | Zheng, A. | Casari, A. | Carnegie Mellon University | azheng@andrew.cr | University of Bologna | | |
| 18 | [Deisenroth | Mathematics for Machine Learning | Deisenroth, M.P | Ong C.S | University of Cambridge | mpd31@cam.ac.uk | National University of Sing | csong@nus.edu.sg | |
| 19 | Saul '00] Sau | An Introduction to Locally Linear Embeddin | Saul, L. K. | Roweis, S. T. | Massachusetts Institute of T | lksaul@mit.edu | University of Texas at Austi | stroeweis@utexas.edu | |
| 20 | [Van der Ma | Visualizing Data Using t-SNE | Van der Maaten | Hinton | University of Amsterdam | lvdmaaten@science | Stanford University | geoffrey.hinton@stanford.edu | |
| 21 | [McInnes '18 | Umap: Uniform manifold approximation ar | McInnes | Melville | The University of Melbourn | lachlan.mcinn [at] s | The University of Melbourr | jmelville [at] unimelb.edu.au | |