M.S. in Applied Data Science Portfolio

La Monte Henry Piggy Yarroll

Syracuse University School of Information Studies

Table of Contents

1. Introduction: Purpose and Motivation

Why I Pursued the MS in Applied Data Science

I began the Synoptic Key of Life (SKOL) project in 2019 as a personal research endeavor, driven by my experience as an amateur mycologist struggling to identify fungal specimens using the scattered taxonomic literature. While I had a clear vision of what I wanted to build—a semantic search system that could match specimen descriptions to published species accounts—I lacked the systematic training in machine learning, natural language processing, and big data systems needed to implement it.

I enrolled in the MS in Applied Data Science program at Syracuse University specifically to acquire these skills. Each course I selected advanced a specific component of SKOL, transforming what began as an aspirational hobby project into a deployed, functional system now accessible at https://synoptickeyof.life.

What is SKOL?

SKOL is a Synoptic Key for mycological taxonomy. Unlike dichotomous keys that force users through a rigid sequence of binary choices, a synoptic key allows users to begin their search from any observable characteristic. As described on the SKOL website:

> "SKOL aims to be a synoptic key for all fungi described in the open access mycological taxonomic literature. We want the machines to do the hard work of reading and understanding the literature, so that human users can simply describe what they see and find matching species descriptions." (Yarroll et al., 2026)

The project applies machine learning and natural language processing techniques to make the vast corpus of fungal taxonomy publications accessible to researchers, students, and enthusiasts through semantic search capabilities. The core research question SKOL addresses is: "Which literature has descriptions similar to my collection?" (Yarroll, 2025a).

The Challenge

Mycological taxonomy literature spans centuries of scientific publications across numerous journals and books. Traditional keyword-based searches fail to connect researchers with relevant descriptions because:

- Taxonomic terminology has evolved over time
- Different authors describe similar features using different vocabulary
- Species descriptions are scattered across thousands of publications
- Finding morphologically similar species requires reading and comparing numerous descriptions
- Until 2011, every new species publication required a Latin description, creating multilingual challenges (Yarroll et al., 2024a)

Course Integration

The following courses each contributed essential capabilities to SKOL:

Course	Contribution	SKOL Related
IST664 (Natural Language Processing)	SBERT embeddings, semantic search, Taxon class design	*
IST718 (Big Data Analytics)	PySpark classification pipeline, scalable processing	*
IST736 (Text Mining)	Novel Glycemic Increment metric for personalized diabetic nutrition	
IST769 (Advanced Big Data Management)	CouchDB document store, Redis caching, data pipelines	*
IST691 (Deep Learning)	LLM fine-tuning for feature extraction	*

Course	Contribution	SKOL Related
IST690 (Independent Study)	Django/React website deployment	*

---

2. Program Learning Goals

Goal 1: Collect, Store, and Access Data

Requirement: Collect, store, and access data by identifying and leveraging applicable technologies.

Technologies Deployed The SKOL project required a comprehensive data management strategy spanning the full data lifecycle from acquisition through serving:

Data Collection: - Web scraping with shell scripts using `wget` to extract journal issues and articles from archives (Yarroll, 2025a) - RSS feed ingestion from Ingenta Connect for new article notifications - OCR processing with Tesseract for older page-scan PDFs - YEDDA annotation tool for manual corpus labeling (Yarroll et al., 2024a)

Data Storage: - **CouchDB**: Document store for taxa records, article metadata, and JSON structures. CouchDB's native JSON support made it ideal for storing the nested feature structures generated by LLMs (Yarroll, 2025a) - **Redis**: In-memory cache for SBERT embeddings, enabling fast similarity searches without recomputing vectors - **Parquet**: Columnar format for intermediate processing stages in the PySpark pipeline

Data Access: - REST API (Django REST Framework) for web application queries - PySpark DataFrame operations for batch processing - Direct CouchDB queries for taxa retrieval

Evidence from Projects **IST769 (Big Data Management):** The project proposal outlined a comprehensive metadata tracking system:

> "It is necessary to track provenance for data carefully. Items in the metadata include the name of the journal, Bibtex entries, human and download URLs, any internal keys, and the date of the scrape." (Yarroll, 2025a)

The Jupyter notebook demonstrates ingestion from multiple sources: - Mycotaxon at Ingenta Connect - Studies in Mycology at Ingenta Connect - Historical works from MycoWeb archives

IST664 (NLP): The team collected and organized a comprehensive dataset:

> "For our final project, we worked with a comprehensive dataset of over 7,000 scientific articles focused on mycelia. This dataset, gathered through extensive collaboration between the Imperial Institute of Agricultural Research, the Mycological Society of America, and other respected institutions, includes detailed classifications of fungal species." (Yarroll et al., 2024a)

The labeled corpus comprises: - 190 journal issues manually annotated - 60,754 paragraphs categorized into three classes - Over 2.1 million words across 300,000+ lines

IST718 (Big Data Analytics): Scaled data ingestion to production volumes: - 1,021 journal issues in the unlabeled corpus - Approximately 4.3 million lines and 25 million words processed via PySpark

Key Deliverables

- IST769 Final Project Proposal and Jupyter notebook
- IST664 data preprocessing pipeline
- Django application with CouchDB integration

Goal 2: Create Actionable Insight

Requirement: Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.

Insight Generated SKOL creates actionable insight for the mycological research community by transforming unstructured scientific literature into a searchable semantic space:

Primary Insight: Researchers can now find relevant species descriptions using natural language queries rather than exact keyword matches. A user describing their specimen as having "pileus campanulate, lamellae yellow rust-brown" receives ranked matches from the literature based on semantic similarity, not string matching.

Secondary Insights: - Identification of potential cryptic synonyms (descriptions that overlap significantly despite different names) - Morphological clustering enabling cladogram construction - Gap analysis revealing understudied taxonomic groups

Scientific Context The IST664 report articulates the societal value:

> "The end solution provides a highly accurate search matching capability on a domain-specific corpus of mycology article content. The intent is for a user's query to be a formal description of a particular

> specimen, though partial descriptions seem to work well." (Yarroll et al., 2024a)

The IST691 paper emphasizes accessibility:

> "It opens the taxonomic literature to the enthusiastic amateur… By combining natural language processing, machine learning, and distributed computing, SKOL supports rapid consistent species identification and enhances research productivity in taxonomy." (Yarroll et al., 2025)

Full Data Science Lifecycle

Phase	SKOL Implementation
Problem Definition	Enable semantic search over taxonomic literature
Data Collection	Web scraping, RSS feeds, OCR
Data Preparation	Paragraph segmentation, annotation, Taxon class construction
Modeling	Text classification, SBERT embeddings, LLM feature extraction
Evaluation	Classifier accuracy (94%), Jaccard distance for JSON extraction
Deployment	synoptickeyof.life website
Monitoring	User feedback system, error logging

Personalized Medicine: Glycemic Increment for Diabetics (IST736) Beyond mycological research, I applied data science to create actionable insight in a second distinct context: personalized medicine for diabetics.

The Challenge: Diabetes affects 38.4 million Americans (11.6% of the population), with an additional 97.6 million having prediabetes (Balasi & Yarroll, 2025). Dietary management is a cornerstone of diabetes treatment, but existing metrics like Glycemic Index (GI) and Glycemic Load (GL) have significant limitations:

- GI requires fasting conditions and standardized 50g carbohydrate portions
- GL calculations require precise food weighing and extensive lookup tables
- Both metrics are population averages, not individualized predictions
- Neither accounts for how foods interact within composite meals

Novel Contribution—Glycemic Increment (G+): The IST736 project introduced a truly novel metric:

> "Glycemic Increment (G+) directly estimates the blood sugar consequences of particular foods and composite meals (glycemic response). It provides an individualized model of each patient's response to particular foods. The inputs to the model are just a list of foods in a particular meal. The output is a number that directly estimates an

> increase or decrease in blood sugar over a reasonable period." (Balasi & Yarroll, 2025)

Unlike GI/GL, Glycemic Increment: - Uses actual Continuous Glucose Monitor (CGM) data from the individual patient - Does not require fasting states or precise food measurements - Accounts for the patient's "typical serving" of each food - Models the combined effects of foods eaten together

Methodology: Using two years of CGM data combined with free-form meal logs, we built linear regression models where coefficients directly represent each food's contribution to blood sugar change. The unit (mg-hours/dL) allows direct comparison of glycemic effects across different measurement intervals.

Approach	Audience	Actionable Output
Population GI/GL	General public	Generic dietary guidelines
Glycemic Increment	**Individual diabetic**	**Personalized food-effect predictions**

Results: Models using regular meal types with trigram tokenization showed moderate predictive power (correlation ~0.25, meaningful slopes). The food coefficients validated expectations—konjac noodles (GI=0) showed minimal effect while foods like "masoor beef" (always eaten with rice) showed high G+ values.

This work demonstrates that patients can use their own sensor data to produce personalized guidance for controlling blood sugar through diet—a concrete contribution to the emerging field of individualized medicine.

Key Deliverables

- IST664 MycoSearch semantic search system
- IST736 Glycemic Increment paper and analysis code (https://github.com/piggyatbaqaqi/sugarbowl)
- IST691 feature extraction pipeline
- Live website at synoptickeyof.life

Goal 3: Apply Visualization and Predictive Models

Requirement: Apply visualization and predictive models to help generate actionable insight.

Predictive Models Text Classification (IST718):

The PySpark pipeline classifies paragraphs into three categories: Nomenclature, Description, and Miscellaneous Exposition. Multiple classifiers were evaluated:

> "The notebook conducts a comprehensive evaluation of various machine learning classifiers, including BernoulliNB, AdaBoostClassifier, RandomForestClassifier, SGDClassifier, RidgeClassifier, and OneVsRestClassifier." (Yarroll et al., 2024a)

Feature engineering techniques included: - TF-IDF weighting for term importance - Suffix-based features (e.g., -aceae, -spore, -mycetes) for taxonomic vocabulary

Best Result: Logistic Regression with combined TF-IDF and suffix features achieved **94% accuracy** (Yarroll et al., 2024b).

Semantic Embeddings (IST664):

SBERT (Sentence-BERT) generates 768-dimensional vector representations:

> "The SBERT pretrained SentenceTransformer creates a text embedding of our Taxon object. The SBERT model is based on Transformer Architecture and the Attention Mechanism. The multi-headed attention embeds the text into query, key, and value vectors which each help represent the data and explain the meaning, context, and position." (Yarroll et al., 2024a)

The embedding enables cosine similarity search where values approaching 1.0 indicate high semantic similarity.

LLM Feature Extraction (IST691):

Multiple language models were evaluated for converting prose descriptions to structured JSON:

Model	Outcome
ChatGPT 4.0	Accurate but too slow/expensive for batch processing
Llama 3.3 70B	Hardware limitations, no Latin support
Gemma3 27B	Good quality but out-of-memory failures
Gemma3 12B	Acceptable performance
Mistral 7B Instruct	Selected for fine-tuning

Fine-tuning results showed overfitting with the small (16-sample) training set: - Training loss dropped to near 0 - Evaluation loss increased to ~1.3 - Base model Jaccard score: 0.762 - Fine-tuned model Jaccard score: 0.799 (worse due to overfitting)

Visualizations IST664 Paragraph Distance Analysis:

The team created visualizations to determine the maximum distance between Nomenclature and Description paragraphs:

> "The distribution diagram shows the count of articles (y-axis) and the number of paragraphs that exist between the species name and the

> species description (x-axis)… This information was used to validate the design of the taxon object.” (Yarroll et al., 2024a)

Key finding: Most descriptions occur within 5-6 paragraphs of the nomenclature.

IST691 Training Curves:

Loss curves during fine-tuning revealed the overfitting problem, informing the decision to prioritize larger training sets in future work.

Search Results Visualization:

> “A dynamic color-coded ranking system visually represents the relevance of results based on cosine similarity, making it easier to understand performance metrics at a glance.” (Yarroll et al., 2024a)

Results display with magenta indicating highest similarity and light grey indicating lowest.

Key Deliverables

- IST718 classifier Jupyter notebook with model comparisons
- IST664 embedding implementation and visualizations
- IST691 fine-tuning code with training curves

Goal 4: Use Programming Languages

Requirement: Use programming languages such as R and Python to support the generation of actionable insight.

Python-Centric Implementation All SKOL components are implemented in Python, leveraging its rich ecosystem for data science and machine learning:

Core Libraries:

Library	Application
pandas	Data manipulation and DataFrame operations
numpy	Numerical computing
scikit-learn	Text classification, TF-IDF vectorization
sentence-transformers	SBERT embeddings
transformers (Hugging Face)	LLM fine-tuning
PySpark	Distributed processing
Django	Web application framework
redis-py	Embedding cache management
couchdb	Document store integration

Web Technologies:

Technology	Application
Django REST Framework	API endpoints
React JS	Frontend interactivity
HTML5/CSS3	Responsive design
JavaScript	Client-side logic

Key Classes Developed **Taxon Class (skol/taxon.py):**

> "The Taxon classes encapsulate key components of a mycology article, such as: Nomenclature (the scientific naming conventions), Descriptions (detailed textual information about the species), and Metadata (supplementary details such as filename, paragraph number, page number)." (Yarroll et al., 2024a)

The Taxon class: - Links structured metadata with descriptive content - Standardizes data format for embedding compatibility - Merges Description paragraphs into cohesive text strings

SKOL Class (dr-drafts-mycosearch/src/data.py):

Prepares Taxon objects for embedding in the MycoSearch framework, handling the interface between SKOL's data structures and Dr. Draft's embedding pipeline.

SKOL_TAXA Class:

Loads taxa from CouchDB and provides them to the search interface, including metadata fields for pdf_page, pdf_label, line_number, and paragraph_number.

Code Repositories All source code is publicly available: - https://github.com/piggyatbaqaqi/skol - https://github.com/piggyatbaqaqi/dr-drafts-mycosearch

Key Deliverables

- Complete Python codebase across both repositories
- Jupyter notebooks demonstrating data processing
- Django application with React frontend

Goal 5: Communicate Insights

Requirement: Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).

Communication Artifacts **Academic Papers:**

Document	Audience	Format
IST664 Team Report	Academic/Technical	16-page paper with code, diagrams, results
IST691 Conference Paper	Academic	IEEE-style format with figures and references
IST718 Final Report	Academic	Technical report with methodology and results

Presentations:

Presentation	Audience	Format
IST718 Final Presentation	Course faculty, peers	PowerPoint with video recording
MASMC 2025 Poster	Mycology research community	Conference poster
IST690 Final Presentation	Faculty supervisor	Live demonstration

Public-Facing Communication:

The SKOL website (synoptickeyof.life) includes an About page explaining the project to general audiences:

> "SKOL guides you through creating a technical description of the organism you have in front of you. Your description becomes part of the synoptic key, accessible to other users searching for similar organisms. Along the way, SKOL connects you to relevant taxonomic literature, helping you learn more about the species you are studying." (Yarroll et al., 2026)

Technical Documentation:

- README files in GitHub repositories
- Code comments explaining implementation decisions
- API documentation for REST endpoints

Visual Communication System architecture diagrams effectively communicate complex data flows:

> "The flow diagram depicts the process of transferring the annotated mycology articles from SKOL to MycoSearch, converting them to Taxon format, and embedding the content into MycoSearch." (Yarroll et al., 2024a)

The IST664 report includes: - Figure D: SKOL-to-MycoSearch data flow - Figure E: Vector-based search process - Figures B, C: Paragraph distance distributions

Key Deliverables

- All written reports and presentations
- Website About page and documentation
- Conference poster presentation

Goal 6: Apply Ethics

Requirement: Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

Ethical Principles Applied **Transparency:**

All SKOL code is released under the GNU General Public License v3 (GPLv3):

> "SKOL is open source software released under the GNU General Public License v3 (GPLv3). The source code is available on GitHub." (Yarroll et al., 2026)

This ensures that anyone can inspect, modify, and build upon the work.

Open Access:

SKOL exclusively uses open access mycological literature:

> "SKOL started by indexing taxonomic literature from major mycological journals: older issues of Mycologia, Mycotaxon, and Persoonia." (Yarroll et al., 2026)

The project respects robots.txt and scraping guidelines:

> "We want to be a well-behaved web scraper. Respect robots.txt, a standardized file that tells us what parts of a web site a scraper is allowed to access." (Yarroll, 2025a)

Reproducibility:

- All data processing pipelines are documented and version-controlled
- Model parameters and training procedures are recorded
- Provenance metadata tracks the source of every document

Bias Awareness and Limitations:

The project openly acknowledges its limitations:

1. **Language Bias:** Initial focus on English-language literature, with Latin translation challenges documented: > "The fine-tuned model has failed to

translate the Latin description, and is reporting Latin and English results as if they were separate features.” (Yarroll et al., 2025)

2. **Model Limitations:** Overfitting issues with small training sets are clearly reported rather than hidden.
3. **Coverage Bias:** The corpus currently emphasizes certain journals and time periods.

Privacy:

SKOL collects no personal data. The project focuses entirely on published scientific literature in the public domain or under open access licenses.

Attribution:

All source materials are properly cited, and search results link back to original publications:

> “Along with the text and similarity score, the journal which it came from along with the URL where it can be found is returned to the user.” (Yarroll et al., 2024a)

Collaborative Credit:

All contributors are acknowledged: - Christopher Murphy, Jennifer Balasi, Shintaro Osuga (IST664) - David Caspers, Christopher Murphy (IST718) - Padmaja Kurumaddali, Patrick Le (IST691) - Dr. Gregory Block (IST690 advisor)

Key Deliverables

- GPLv3 license in repositories
- Documented limitations in all reports
- Attribution and citation practices throughout

---

3. Synthesis: Integrated Learning Through SKOL

The SKOL project demonstrates how the diverse courses in the MS Applied Data Science program integrate into a coherent professional capability. Rather than isolated assignments, each course contributed essential components to a functioning system.

Additionally, the IST736 Glycemic Increment project demonstrates breadth—applying the same data science lifecycle to a completely different domain (personalized medicine for diabetics). This shows that the cognitive strategies developed through this program transfer across contexts, from scientific research tools to individual health optimization.

```
IST664 (NLP)          -> Semantic Search Foundation
       |                  (Taxon class, SBERT embeddings, MycoSearch)
```

```
        v
IST718 (Big Data)      -> Scalable Classification Pipeline
        |                  (PySpark, 94% accuracy classifier)
        v
IST769 (Data Mgmt)     -> Production Data Architecture
        |                  (CouchDB, Redis, metadata tracking)
        v
IST691 (Deep Learning) -> Feature Extraction for UI
        |                  (LLM fine-tuning, JSON generation)
        v
IST690 (Independent)   -> Public Website Deployment
        |                  (Django, React, production hosting)
        v
    synoptickeyof.life  -> Live Production System
```

Integration Points

The Taxon Class: Developed incrementally across projects, this class encapsulates the core data structure linking nomenclature to descriptions with full metadata. It serves as the bridge between raw article content and the embedding space.

MycoSearch: Forked from Dr. Draft's SOTA Literature Search and extended with domain-specific adaptations. The SKOL class in data.py demonstrates how academic tools can be adapted for specialized applications.

Data Flow: Documents flow from web scrapers through text extraction, classification, Taxon construction, embedding, and finally to the search interface—a complete pipeline touching every skill area in the program.

Cognitive Strategy Development

The portfolio demonstrates not just technical skills but the cognitive strategy for approaching new data science problems:

1. **Problem Decomposition:** Breaking the SKOL vision into tractable components (collection, classification, embedding, search, interface)
2. **Tool Selection:** Choosing appropriate technologies for each component (PySpark for scale, SBERT for semantics, CouchDB for documents)
3. **Iterative Refinement:** Starting with hand-annotated data, training classifiers, correcting errors, expanding the corpus
4. **Integration Testing:** Ensuring components work together through the complete pipeline

---

4. Strengths and Challenges

Areas of Strength

Natural Language Processing: The IST664 project built a strong foundation in text processing, embeddings, and semantic search. I am comfortable with transformer architectures, vectorization techniques, and similarity metrics.

System Integration: Successfully connecting multiple technologies (PySpark, CouchDB, Redis, Django, React) into a functioning pipeline demonstrates practical engineering skills beyond individual algorithms.

Domain Expertise: My background in amateur mycology enabled meaningful feature engineering decisions, such as recognizing suffix patterns (-aceae, -spore) and understanding the structure of taxonomic descriptions.

Areas of Challenge

LLM Fine-tuning: The IST691 project revealed the difficulty of fine-tuning with small training sets. The overfitting observed (training loss near 0, evaluation loss increasing) indicates the need for larger annotated corpora. Future work requires either: - Substantially more hand-labeled training examples - Alternative approaches such as few-shot learning or retrieval-augmented generation

Latin Language Support: Many taxonomic descriptions include Latin text, particularly older publications. Current models struggle with Latin translation:

> "Perhaps the prompt needs to specify the order of operations" (Yarroll et al., 2025)

Scaling: While the system handles 170 annotated journal issues well, expanding to the full corpus of 1,000+ issues requires ongoing engineering effort and computational resources.

---

5. Lifelong Learning Plan

Immediate Next Steps

1. **Expand Training Corpus:** Create more hand-labeled JSON examples for LLM fine-tuning to address overfitting
2. **Web Crawlers:** Implement automated ingestion of new publications from open-access journals
3. **Decision Tree UI:** Build the interactive description builder using extracted features

Continuing Education

- Follow developments in transformer architectures and domain-specific fine-tuning

- Participate in mycology community events (Western PA Mushroom Club, MASMC conferences)
- Contribute to open source ML/NLP tools in the scientific domain

Career Application

The skills developed through this program apply broadly to scientific research domains requiring: - Text mining of domain-specific literature - Semantic search over unstructured documents - Integration of machine learning into research workflows

I intend to continue developing SKOL as an open-source contribution to the mycology community while applying these skills professionally.

---

6. References

Foundational Works

Ait, A., Cánovas Izquierdo, J.L., & Cabot, J. (2023). On the Suitability of Hugging Face Hub for Empirical Studies. *ESEM 2023*. https://doi.org/10.48550/arXiv.2307.14841

Gisolfi, N. (2024). Dr Draft's state-of-the-art (SOTA) Literature Search. Auton Lab. https://github.com/autonlab/dr-drafts-sota-literature-search

Jang, A. et al. (2023). Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP 2019*. https://doi.org/10.48550/arXiv.1908.10084

Data Sources

Hennebert, G.L., & Korf, R.P. (Eds.). (1974-2010). *Mycotaxon: A New Journal on Taxonomy and Nomenclature of Fungi and Lichens*. Ithaca, NY.

Murrill, W.A. (Ed.). (1909-1961). *Mycologia*. New York Botanical Garden, Mycological Society of America.

Nauta, M.M., & Noordeloos, M.E. (Eds.). (1959-1998). *Persoonia: A Mycological Journal*. Riksherbarium, Leiden, The Netherlands.

Wood, M. (1995-2025). MykoWeb Journals. https://www.mykoweb.com/systematics/journals.html

Course Deliverables

Balasi, J. & Yarroll, L.H.P. (2025). Glycemic Increment: Individualized Medicine for Diabetics. Syracuse University. IST736 Final Project. https://github.com/piggyatbaqaqi/sugarbowl

Yarroll, L.H.P., Balasi, J., Murphy, C., & Osuga, S. (2024a). Mycology Literature Search. Syracuse University. IST664 Final Project. https://github.com/piggyatbaqaqi/skol/blob/main/IST664/IST664_Team3_Balasi_Murphy_Osuga_Yarroll.p

Yarroll, L.H.P., Caspers, D., & Murphy, C. (2024b). Synoptic Key of Life II, Final Project Report. Syracuse University. IST718 Final Project. https://github.com/piggyatbaqaqi/skol/blob/main/IST718/IST718_Final_Report_FINAL.pdf

Yarroll, L.H.P. (2025a). Synoptic Key of Life Pipelines. Syracuse University. IST769 Final Project Proposal.

Yarroll, L.H.P., Kurumaddali, P., & Le, P. (2025). Synoptic Key of Life: Feature Extraction for Fungal Taxonomy. Syracuse University. IST691 Final Project.

Yarroll, L.H.P. (2025b). Synoptic Key of Life Initial Website. Syracuse University. IST690 Independent Study Proposal.

Yarroll, L.H.P. (2025c). Synoptic Key of Life. Mid Atlantic Mycology Conference 2025 (MASMC2025). https://github.com/piggyatbaqaqi/skol/blob/main/MASMC2025/Synoptic%20Key%20of%

Yarroll, L.H.P. et al. (2026). About SKOL - Synoptic Key of Life. https://synoptickeyof.life/about/

Technical References

Adithya, S.K. (2023). A Beginner's Guide to Fine-Tuning Mistral 7B Instruct Model. *Medium.* https://adithyask.medium.com/a-beginners-guide-to-fine-tuning-mistral-7b-instruct-model-0f39647b20fe

Khan, H.K. (2023). A Step-by-Step Guide to Fine-Tuning the Mistral 7B LLM. *E2E Cloud.* https://www.e2enetworks.com/blog/a-step-by-step-guide-to-fine-tuning-the-mistral-7b-llm

---

7. Appendices

Folder Structure

```
Portfolio/
|-- 01_Overview/
|   |-- overview.pdf           # This document converted to PDF
|   |-- resume.pdf             # Current professional resume
|
|-- 02_IST664_NLP/
|   |-- README.txt             # Software requirements, how to review
|   |-- IST664_Team3_Balasi_Murphy_Osuga_Yarroll.pdf
|   |-- SKOL_presentation3.pptx
|
|-- 03_IST736_Text_Mining/
```

```
|   |-- README.txt
|   |-- IST736_Balasi_Yarroll_Final_Project.pdf
|   |-- glycemic_increment.ipynb
|
|-- 04_IST691_Deep_Learning/
|   |-- README.txt
|   |-- IST691_final_project_paper.pdf
|   |-- mistral_transfer_learning.ipynb
|
|-- 05_IST718_Big_Data/
|   |-- README.txt
|   |-- IST718_Final_Report_FINAL.docx
|   |-- IST_718_Final_Project_Classifier.ipynb
|
|-- 06_IST769_Data_Management/
|   |-- README.txt
|   |-- IST769_Final_Project_Proposal.pdf
|   |-- ist769_skol.ipynb
|
|-- 07_IST690_Independent_Study/
|   |-- README.txt
|   |-- IST_690_SKOL_Website.pdf
|
|-- 08_Video_Presentation/
    |-- portfolio_presentation.mp4  # 10-minute summary
```

Software Requirements for Reviewing Deliverables

File Type	Software Required
.pdf	Any PDF reader (Adobe Reader, browser)
.docx	Microsoft Word or LibreOffice
.pptx	Microsoft PowerPoint or LibreOffice
.ipynb	Jupyter Notebook, JupyterLab, or VS Code
.mp4	Any video player

Repository Links

- **SKOL Main Repository:** https://github.com/piggyatbaqaqi/skol
- **MycoSearch Fork:** https://github.com/piggyatbaqaqi/dr-drafts-mycosearch
- **Live Website:** https://synoptickeyof.life

Portfolio prepared for IST782 - Applied Data Science Portfolio Syracuse University School of Information Studies Spring 2026