Original Article *Available online at www.bpasjournals.com*

Natural Language Processing and Machine Learning for Semantic Subject Indexing of Google Books Using Annif

Dr Sukumar Mandal
Assistant Professor, Department of Library and Information Science, The University of Burdwan
Email: sukumar.mandal5@gmail.com
ORCID: 0000-0003-1415-402X

How to cite this article: Dr Sukumar Mandal. (2026). Natural Language Processing and Machine Learning for Semantic Subject Indexing of Google Books Using Annif. *Library Progress International,* 46(1), 1-12

Abstract

This contemplate investigates the application of the Annif Toolkit for automatic (semantic) subject indexing in large-scale collections of digitized books with Google Books as a sample test collection. This study addresses the issues related to manual subject indexing in the context of large-scale collections of heterogeneous content that are characterized by their volume, heterogeneity and lack of scalability and consistency. In this interpretation thoroughly explored the use of a machine learning approach based on a vocabulary driven approach using a curated set of 5000 Google Book records analyzed through an automated indexing pipeline utilizing a TF-IDF backend and a controlled subject vocabulary. It has evaluated system performance using common Information Retrieval metrics (precision, recall, F1-Score, etc.) as well as ranking-based metrics (Normalized Discounted Cumulative Gain). Therefore, consequences indicate that the system is capable of identifying highly relevant subject concepts from the contents of a wide range of different types of books at very high rates of recall and that it provides high-quality rankings; however, our results also indicate that the system has moderate precision due to its design, which was optimized for high recall, making it suitable for large-scale digital libraries. Finally, the integrated system has highlighted the potential of the Annif Toolkit to provide real-time subject recommendations, to index content in search systems automatically, and to provide decision-support tools to enable efficient, consistent and semantically-enriched organization of knowledge.

Keywords: Semantic subject indexing, Natural language processing, Machine learning, Annif, Digital libraries, Automated knowledge organisation

1. Introduction

Indexing is one of the important concept in automated and digital library system. Semantic level subject indexing has systematically affected the modern information retrieval system. Information and knowledge have never been organized and shared in such an extensive manner as massive growth of digital information in recent years. Natural language processing and machine learning are also playing an important role to creating the automated subject indexing for libraries (Golub, 2021). The emergence of large-scale digital libraries (for instance Google Books) has created an unprecedented challenge - to deal with the exponential increase of textual content found within them (Kasprzik, 2024). Subject Indexing has traditionally been the systematic process of describing a particular item with terms from a predetermined vocabulary (or 'controlled vocabulary'). Manual subject indexing has proven to be far too labor-intensive and costly to perform at the scale necessary for discovery purposes in the current era; furthermore, it has also demonstrated inconsistent application in supporting effective searching of digital materials (Suominen et al., 2022). Therefore, the need for the development and utilization of new methods for automated subject indexing is a pressing necessity so that continued scholarly and public access to digital materials will remain both meaningful and accessible (Kasprzik, 2023).

Semantic Subject Indexing, particularly when utilizing Artificial Intelligence (AI) and Machine Learning (ML), is an attractive model to address many issues of indexing and classification. This is because AI and ML enable the design of automated systems capable of classifying and assigning subject headings to documents; and generating additional metadata quickly and at scale enabling improved discovery and enhanced quality of access to information (Golub, 2019). One of the best known Open Source tools utilized to facilitate automatic classification and subject heading assignments of documents is the Annif AI Tool Kit developed by the National Library of Finland. Annif has gained popularity due to its ability to classify documents across a variety of document types and languages regardless of the presence of specific indexing vocabularies and/or ML algorithms used in other tools, thereby offering a flexible tool for libraries with a wide range of needs and specifications (Suominen, 2019). Annif is continually being improved upon, including its inclusion of advanced Natural Language Processing (NLP) and Large Language Models (LLMs) to improve its performance and reduce the processing time for documents (Suominen et al., 2025).

Annif AI Tool Kit is currently being studied as an example of the utilization of Semantic Subject Indexing (SSI) at a large scale via the vast Google Books Collection. This study will explore how Annif may be configured and trained so that it will generate similar, accurate, and consistent subject headings for enormous collections of digital books. The work will also evaluate the quality of Annif's subject headings through established standards for Information Retrieval (IR); and compare the advantages and disadvantages of SSI automatically generated across multiple subjects. Additionally, this research will discuss the possible implications of SSI for scalable digital libraries and knowledge organization systems. Finally, by providing evidence of Annif's implementation in a "real world" environment, this research will increase our understanding of automatic indexing and future IR methodologies in numerous international academic settings.

2. Background and Related Work

2.1 Digital Knowledge Growth

The rapid expansion of digital collections is creating a need for efficient and scalable methods of organizing content. Manual subject indexing still has theoretical benefits, however, manual subject indexing has become impractical as the number of items in a collection grows. This context has provided a basis for the study of semantic subject indexing (using Artificial Intelligence) by exploring whether an automated method that specifically using the Annif AI Toolkit which could be used to support access to large-scale resource collections from Google Books (Massa et al., 2023).

2.2 Semantic Indexing

Semantic level subject indexing uses a controlled vocabulary to provide consistency in the application of subject terms to enable information retrieval; this is achieved by utilizing pre-identified controlled terms which are typically defined by an established knowledge organization system (KOS) that could be based on a thesaurus enabling subject heading lists, classification schemes etc. These KOS's have traditionally been applied within a digital environment, however with advancements in automation and semantic technologies they are also now being applied within the semantic web environment with a high degree of precision (Golub, 2021)**.**

2.3 Indexing Challenges

Advances in indexing technology have highlighted several challenges in both manual and automated indexing methods. Challenges include the quality control necessary to ensure that terms assigned by machines (and other computer generated) are accurate enough to support end-user needs and the costly time consuming nature of manually indexing documents as well. Therefore, developing new indexing systems could potentially offer an alternative method to facilitate end-users in achieving their desired search results while minimizing the indexing errors inherent in using machine-generated facets and terms (Dylag, Zlatev & Boniface, 2025)**.**

2.4 Annif Toolkit

Annif is a freely available, multi-lingual tool for subject indexing that was created by the National Library of Finland to aid in the use of automated subject indexing. Annif is capable of using multiple forms of machine learning and is neutral with respect to vocabularies and thus, it can be used with a variety of combinations of vocabularies (e.g., LCSH, DDC). The addition of large language models has provided Annif with flexibility and scalability **(Suominen, 2019).**

2.5 Google Books Corpus

Google Books Database is among the larger databases that illustrate the necessity for Automated Semantic Indexing Systems (ASIS). Due to the large amount of books in the Google Books collection and also due to the variety of types of books included in the collection; it is necessary to utilize Artificial Intelligence (AI) in order to utilize the organization and availability of the books at Google Books to its fullest potential. Annif, utilizes AI to address the complex problem of organizing the vast amounts of books available through the Google Books collection, with the goal of improving the subject access to the Google Books collection for library patrons and other users of Information Centers (Yang, et al., 2023).

3. Materials and Methods

3.1 Data Collection and Preparation

Approximately 5,000 bibliographic records exist in this system that include references for books as well as links to each reference located within the Google Books repository (https://books.google.co.in/) where the bibliographic data was retrieved. A Python-based environment was then developed and configured so as to support a machine learning environment. This was accomplished through the installation of Python (https://www.python.org/) and the software foundation's documentation, followed by the implementation of Annif, an open-source toolkit for developing artificial intelligence applications (https://annif.org/); and then finally utilizing Ubuntu Long Term Support (LTS) as the operating system platform for Annif. The next stage involved cleaning and normalizing the bibliographic data to ensure consistency across all relevant bibliographic fields (e.g., book title and link) in order to be able to automatically apply subject terms to the documents.

3.2 Annif Indexing Pipeline for Google Books

Step 1: Activate the Annif Virtual Environment

The command is used to activate the Python virtual environment that contains Annif, and the other necessary dependencies for Annif to run in a controlled and consistent environment.

```
source annif-venv/bin/activate
```

Step 2: Navigate to the Annif Directory

The task relating to navigate that takes libraries into the Annif working directory is for access to configuration files and executing essential commands to execute semantic indexing.

```
cd annif-venv/annif
```

Step 3: Configure the Annif Project

At this stage, the Annif project is configured by editing the projects.cfg file to specify the project identifier, language parameters, linguistic normalisation, TF–IDF backend, and linked controlled vocabulary, thereby establishing a structured framework for subject indexing of Google Books content.

```
[GoogleBKS]
name=Google Books Indexing
language=en
analyzer=snowball(english)
backend=tfidf
vocab=GoogleBKS
```

Step 4: Verify the Project

This process lists all registered Annif projects to verify that the GoogleBKS project has been correctly added and configured.

```
annif list-projects
```

Step 5: Load Vocabulary

This task imports the controlled subject vocabulary from a TSV file into the Annif system, enabling the authorised terms to be used during training and subject assignment.

```
annif load-vocab GoogleBKS --language en data-sets/BKS/GoogleBKSVOC.tsv
```

Step 6: Train the Model

This phase trains the TF–IDF–based Annif model using labelled data, enabling it to learn associations between textual content and subject terms for effective indexing.

```
annif train GoogleBKS data-sets/BKS/GoogleBKS.tsv
```

Step 7: Evaluate the Model

This evaluation explores the performance of the trained Annif model by comparing the generated subject assignments with the reference subjects in the dataset, thereby determining the accuracy and effectiveness of the indexing process.

annif eval GoogleBKS data-sets/BKS/GoogleBKS.tsv

Step 8: Generate Subject Suggestions

This procedure applies the trained Annif model to a sample text input to automatically generate relevant subject terms, producing subject suggestions for the given input.
echo "Machine Learning System for Libraries" | annif suggest GoogleBKS

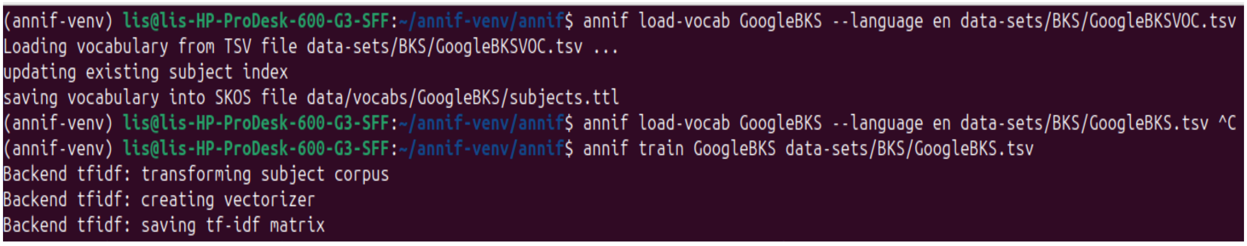
Step 9: Run Annif Web Service

The Annif web service can be launched via the annif run command to provide an interactive web-based interface for both testing and subject indexing on a local machine. The service can be launched with --host 0.0.0.0 and a defined port number, allowing users to remotely access the Annif interface by entering the web address http://127.0.0.1:5000 in their web browsers. Once the service has been launched, users can enter text into the Annif web interface in order to receive real-time automated suggestions of subjects based upon that input.

4. Results

4.1 Load Vocabulary and Train Model

Figure 1 outlines the systematic workflow used to ingest a controlled subject vocabulary, and to train the Annif model using a curated collection of 5,000 Google Book records. The first step in the workflow is to execute the annif load-vocab command to create a TSV-formatted vocabulary file, process each subject term and convert the subject terms to an SKOS representation of structured data, which can be persisted for later reuse as it will normalize subject concepts, index subject concepts uniformly, and align the subject concepts with the predefined project parameters. The second step in the workflow is to train the model by executing the annif train command to utilize the TF-IDF backend. In this step, the Google Books corpus is converted into a vectorized representation of the text; a TF-IDF matrix is created and saved; and the model learns statistical relationships between book titles and controlled subject concepts, creating a stable, transparent, and reproducible pipeline for semantic indexing of collections.

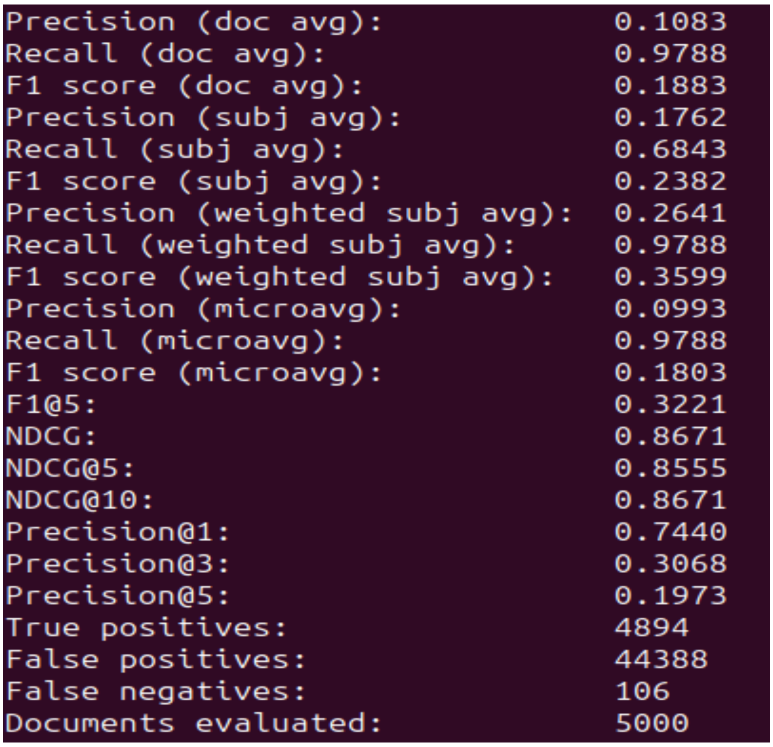


```
(annif-venv) lis@lis-HP-ProDesk-600-G3-SFF:~/annif-venv/annif$ annif load-vocab GoogleBKS --language en data-sets/BKS/GoogleBKSVOC.tsv
Loading vocabulary from TSV file data-sets/BKS/GoogleBKSVOC.tsv ...
updating existing subject index
saving vocabulary into SKOS file data/vocabs/GoogleBKS/subjects.ttl
(annif-venv) lis@lis-HP-ProDesk-600-G3-SFF:~/annif-venv/annif$ annif load-vocab GoogleBKS --language en data-sets/BKS/GoogleBKS.tsv ^C
(annif-venv) lis@lis-HP-ProDesk-600-G3-SFF:~/annif-venv/annif$ annif train GoogleBKS data-sets/BKS/GoogleBKS.tsv
Backend tfidf: transforming subject corpus
Backend tfidf: creating vectorizer
Backend tfidf: saving tf-idf matrix
```

Figure 1: Load vocabulary and train model for datasets

4.2 Model Evaluation

Annif's model has been tested on a test set of 5,000 Google Books records, with clearly interpretable results. The model's document-level recall was 0.9788; it shows a very good ability to get relevant subject concepts, but the precision is 0.1083 and its F1 score is modest. Similarly, the same tendencies can be seen at the subject level as well; high recall and moderate precision show the model prefers broad coverage in automatic subject indexing. When considering subject frequency, there is better balance among weighted precision and F1 scores and therefore better reliability for the most frequently occurring subjects. In addition, micro-averaging also confirmed the model's recall oriented behavior. Other additional information provided by ranking-based metrics, the model exhibits excellent NDCG values; this indicates that the model places the most relevant subjects at the top of the list most often. Additionally, the model shows strong early relevance in subject recommendation, based on Precision@1 and Precision@3 scores. Overall, the testing of the Annif model has demonstrated that the model is very successful in finding relevant subject concepts for Google Books documents and is very good in terms of ranking quality and recall. However, lower precision may indicate some room for improvement through such methods as adjusting thresholds, refining vocabularies, or developing hybrid back-end models; however, the testing demonstrated a solid base for large scale semantic indexing. These results are illustrated in **Figure 2** which graphically illustrate the model's recall oriented behavior, excellent ranking ability, and consistent performance across all document, subject, and ranking based measures.

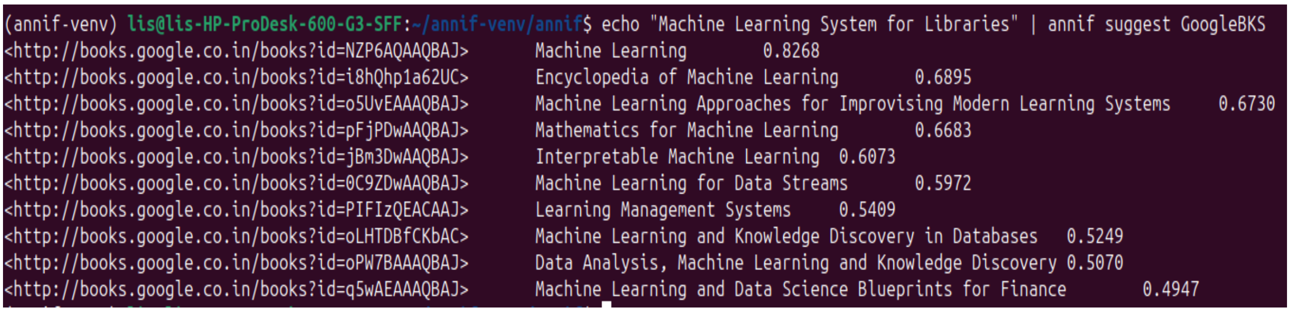


```
Precision (doc avg):            0.1083
Recall (doc avg):               0.9788
F1 score (doc avg):             0.1883
Precision (subj avg):           0.1762
Recall (subj avg):              0.6843
F1 score (subj avg):            0.2382
Precision (weighted subj avg):  0.2641
Recall (weighted subj avg):     0.9788
F1 score (weighted subj avg):   0.3599
Precision (microavg):           0.0993
Recall (microavg):              0.9788
F1 score (microavg):            0.1803
F1@5:                           0.3221
NDCG:                           0.8671
NDCG@5:                         0.8555
NDCG@10:                        0.8671
Precision@1:                    0.7440
Precision@3:                    0.3068
Precision@5:                    0.1973
True positives:                 4894
False positives:                44388
False negatives:                106
Documents evaluated:            5000
```

Figure 2: Performance of model evaluation

4.3 Subject Suggestions

Annif's ability to suggest subjects based on the training model is illustrated using the example of an Annif-generated subject recommendation for a user-entered search question ("Machine Learning System for Libraries") as shown in **Figure 3**. The model has generated a ranked list of semantically-relevant subjects from the Google Books corpus, along with a confidence score associated with each of the subjects; the confidence scores are indicative of how strongly the subject is associated with the entered search question. The subject that the model suggests as most relevant, "Machine Learning" (confidence score = 0.8268) is consistent with the thematic core of the search question. In addition, other subjects suggested by the model that are also very relevant to the theme of the search question include "Encyclopedia of Machine Learning", "Machine Learning Approaches for Improving Modern Learning Systems", and "Mathematics for Machine Learning" which indicate that the model is sensitive to context. Subjects suggested by the model, in addition to those identified above that are relevant to libraries and education, also include "Interpretability in Machine Learning", "Machine Learning for Data Streams", and "Learning Management Systems". Overall, the results illustrate the success of the trained Annif model in providing users with a coherent, ranked list of subject recommendations, which provide context to enhance discovery and automate knowledge organization in library systems.

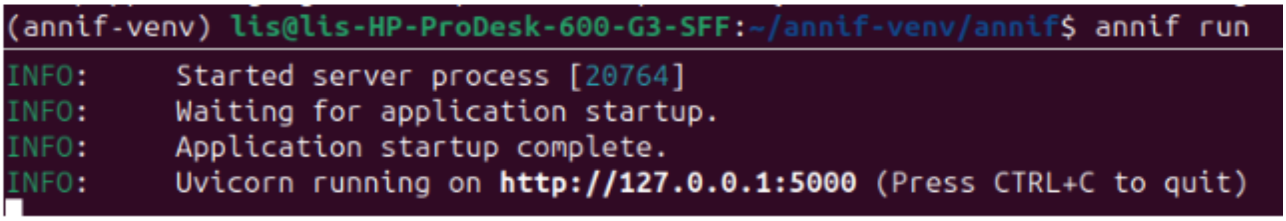


```
(annif-venv) lis@lis-HP-ProDesk-600-G3-SFF:~/annif-venv/annif$ echo "Machine Learning System for Libraries" | annif suggest GoogleBKS
<http://books.google.co.in/books?id=NZP6AQAAQBAJ>       Machine Learning        0.8268
<http://books.google.co.in/books?id=i8hQhp1a62UC>       Encyclopedia of Machine Learning        0.6895
<http://books.google.co.in/books?id=o5UvEAAAQBAJ>       Machine Learning Approaches for Improvising Modern Learning Systems     0.6730
<http://books.google.co.in/books?id=pFjPDwAAQBAJ>       Mathematics for Machine Learning        0.6683
<http://books.google.co.in/books?id=jBm3DwAAQBAJ>       Interpretable Machine Learning  0.6073
<http://books.google.co.in/books?id=0C9ZDwAAQBAJ>       Machine Learning for Data Streams       0.5972
<http://books.google.co.in/books?id=PIFIzQEACAAJ>       Learning Management Systems     0.5409
<http://books.google.co.in/books?id=oLHTDBfCKbAC>       Machine Learning and Knowledge Discovery in Databases   0.5249
<http://books.google.co.in/books?id=oPW7BAAAQBAJ>       Data Analysis, Machine Learning and Knowledge Discovery 0.5070
<http://books.google.co.in/books?id=q5wAEAAAQBAJ>       Machine Learning and Data Science Blueprints for Finance        0.4947
```

Figure 3: Command oriented semantic subject suggestions

4.4 Annif Web Server

Successful operation of the Annif Web Server provides an environment that supports testing and validation of the automated subject indexing by providing the ability to interactively test the indexing (as depicted in **Figure 4**). The execution of the 'annif run' command will start up the backend services and launch the uvicorn web server which is used to confirm that the application has been successfully loaded and is able to receive/accept incoming request. It is accessible locally via http://127.0.0.1:5000; therefore users are able to use a web browser to interactively enter search requests, view suggested subjects based on their search requests, observe how the model behaves and provide support for evaluating, demonstrating and iteratively refining the semantic indexing workflow processes.

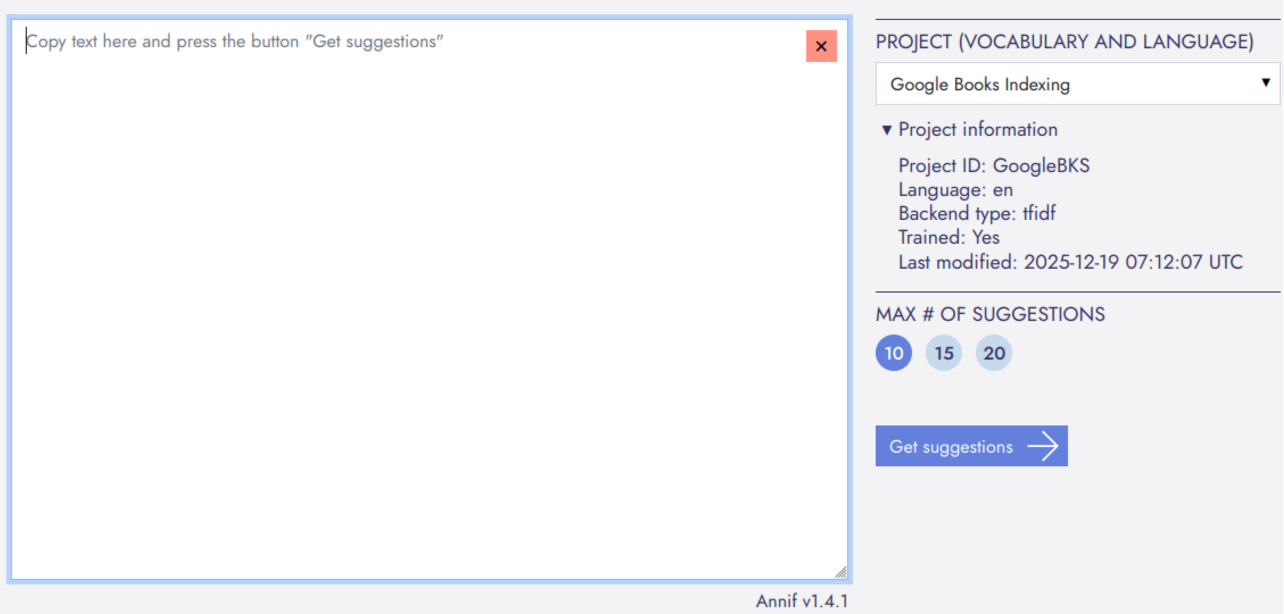


```
(annif-venv) lis@lis-HP-ProDesk-600-G3-SFF:~/annif-venv/annif$ annif run
INFO:     Started server process [20764]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://127.0.0.1:5000 (Press CTRL+C to quit)
```

Figure 4: Annif web server running by Terminal

4.5 Semantic Subject Interface for Google Books

The Web-Based Semantic Subject Interface gives users an opportunity to assess the quality of automated subject indexes developed on a sample of Google Books content through an interactive validation process that can be run on-line. Users can enter descriptive text about a book or just its title to see a ranked list of the top suggestions from the TF-IDF model that was used to train this index, thus enabling users to evaluate the semantic relevance directly. Additionally, the Web-Based Semantic Subject Interface allows access to key metadata about the project (e.g., the unique identifier GoogleBKS; the language(s) of the books in the sample; the type of back-end system being used; whether or not the model has been trained), which increases both the transparency and the ability to reproduce results. A search window of the interface is illustrated in **Figure 5** and shows how users can input their query in a text area at the top of the window; select the number of possible subjects they would like to see using slider controls; and review detailed metadata about the project in a clearly organized panel. The three features work together to provide users with the capability to examine the behavior of the models as it occurs in real time; relate the output from large scale automated subject indexing to well-informed judgments made by humans; and illustrate how Annif can be practically applied for developing indexes.



PROJECT (VOCABULARY AND LANGUAGE)

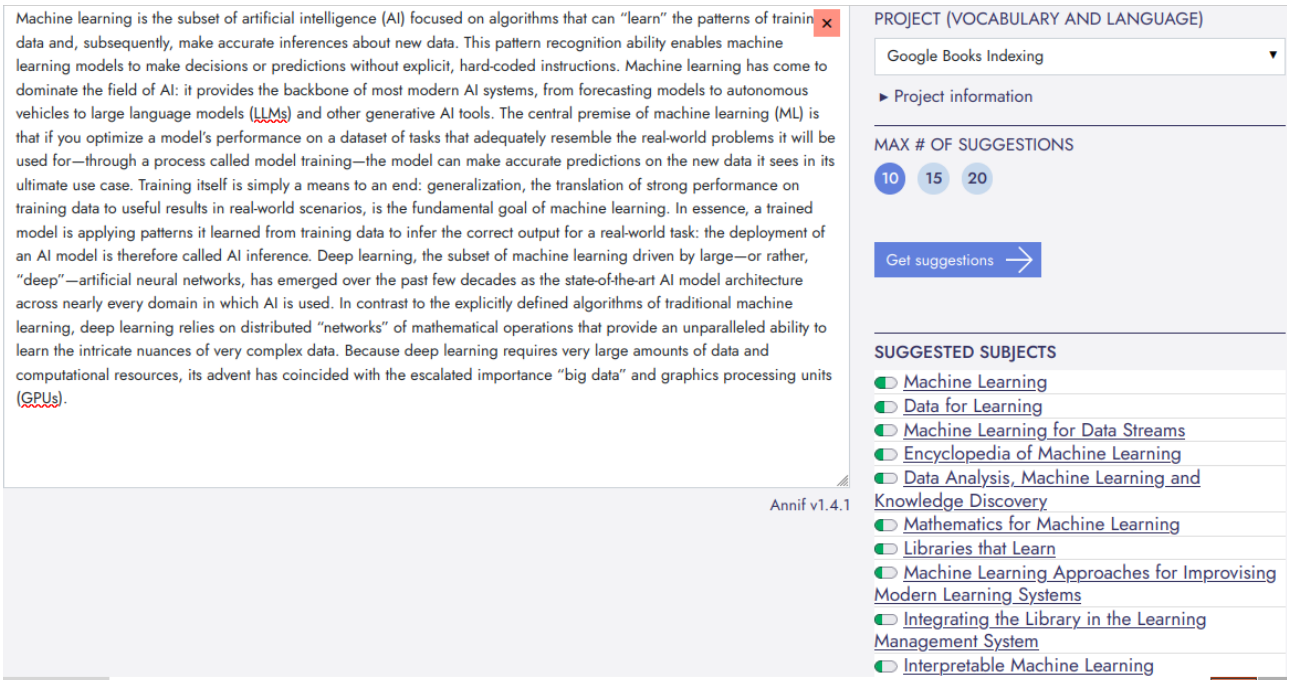
Google Books Indexing

Project ID: GoogleBKS
Language: en
Backend type: tfidf
Trained: Yes
Last modified: 2025-12-19 07:12:07 UTC

Figure 5: Semantic Subject Interface for Google Books (http://127.0.0.1:5000/)

4.6 Intelligent Search Indexing for Google Books

A user's ability to access an automated intelligent search index of Google Books via an Internet interface and generate semantically related subject suggestion from the inputted text are illustrated by this example. A user can choose the number of subject suggestions to be displayed; ten, fifteen or twenty. This is intended to allow for the selection of the level of indexing detail that the user prefers. For example, if a user wants to have a very specific thematic focus, they could select ten subject suggestions. However, if a user wants a larger semantic scope, they could select twenty subject suggestions. In addition, **Figure 6** illustrates how a trained project identifies a representative passage about machine learning and links the passage to relevant subjects from a controlled vocabulary, organized in terms of rank. The suggestions will include both the central theme(s) of the passage and related ideas, demonstrating Annif's capability to find the optimal balance between precision and recall given the desired output quantity. This flexible indexing method helps promote subject-based retrieval, as well as uniformity in the application of indexing practices and supports the discoverability of Google Books beyond search capabilities.



Machine learning is the subset of artificial intelligence (AI) focused on algorithms that can "learn" the patterns of trainin data and, subsequently, make accurate inferences about new data. This pattern recognition ability enables machine learning models to make decisions or predictions without explicit, hard-coded instructions. Machine learning has come to dominate the field of AI: it provides the backbone of most modern AI systems, from forecasting models to autonomous vehicles to large language models (LLMs) and other generative AI tools. The central premise of machine learning (ML) is that if you optimize a model's performance on a dataset of tasks that adequately resemble the real-world problems it will be used for—through a process called model training—the model can make accurate predictions on the new data it sees in its ultimate use case. Training itself is simply a means to an end: generalization, the translation of strong performance on training data to useful results in real-world scenarios, is the fundamental goal of machine learning. In essence, a trained model is applying patterns it learned from training data to infer the correct output for a real-world task: the deployment of an AI model is therefore called AI inference. Deep learning, the subset of machine learning driven by large—or rather, "deep"—artificial neural networks, has emerged over the past few decades as the state-of-the-art AI model architecture across nearly every domain in which AI is used. In contrast to the explicitly defined algorithms of traditional machine learning, deep learning relies on distributed "networks" of mathematical operations that provide an unparalleled ability to learn the intricate nuances of very complex data. Because deep learning requires very large amounts of data and computational resources, its advent has coincided with the escalated importance "big data" and graphics processing units (GPUs).

Annif v1.4.1

PROJECT (VOCABULARY AND LANGUAGE)

Google Books Indexing ▾

▸ Project information

MAX # OF SUGGESTIONS

10 15 20

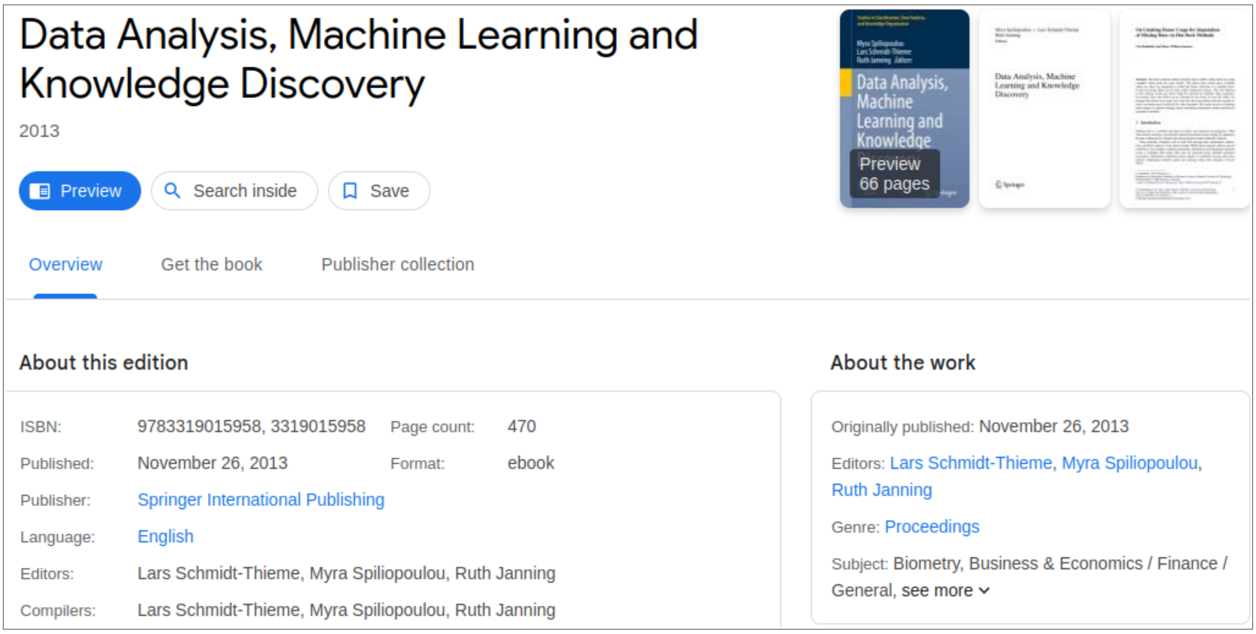
Get suggestions →

SUGGESTED SUBJECTS

- Machine Learning
- Data for Learning
- Machine Learning for Data Streams
- Encyclopedia of Machine Learning
- Data Analysis, Machine Learning and Knowledge Discovery
- Mathematics for Machine Learning
- Libraries that Learn
- Machine Learning Approaches for Improving Modern Learning Systems
- Integrating the Library in the Learning Management System
- Interpretable Machine Learning

Figure 6: Intelligent Search Indexing for Google Books (http://127.0.0.1:5000/)

The ability to apply semantic subject indexes directly in the Google Books environment allows for an instant and natural move between conceptual subjects (as identified by a user) and available bibliographic resources. Once a user identifies a suggested subject, the search system immediately associates the selected subject with relevant Google Book Titles and makes accessible detailed book level data for the user to browse. The process is illustrated in **Figure 7**, where the selection of a subject opens the record of a selected title, allowing the user to view the title's publication date, authors/editors, publisher, and preview options. In addition to improving the users experience of exploratory searching, it also helps increase user engagement and demonstrates the practical uses of the described approach.



Data Analysis, Machine Learning and Knowledge Discovery

2013

About this edition

ISBN:	9783319015958, 3319015958	Page count:	470
Published:	November 26, 2013	Format:	ebook
Publisher:	Springer International Publishing		
Language:	English		
Editors:	Lars Schmidt-Thieme, Myra Spiliopoulou, Ruth Janning		
Compilers:	Lars Schmidt-Thieme, Myra Spiliopoulou, Ruth Janning		

About the work

Originally published: November 26, 2013

Editors: Lars Schmidt-Thieme, Myra Spiliopoulou, Ruth Janning

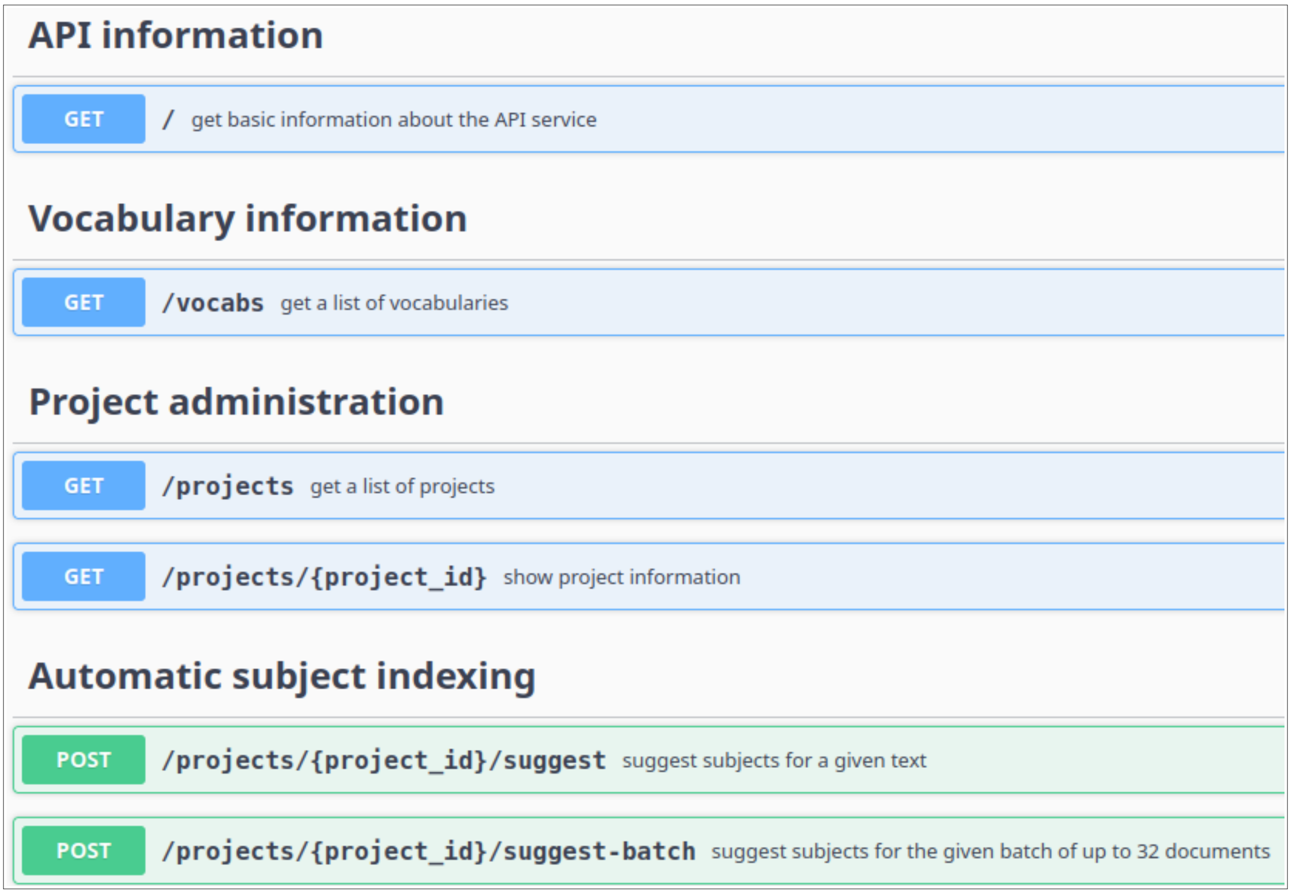
Genre: Proceedings

Subject: Biometry, Business & Economics / Finance / General, see more ⌄

Figure 7: Details of Google Books (https://tinyurl.com/2j6au883)

4.7 REST API Interface

The Annif interface presents an automated, programmatically defined, and structured method for automatic subject indexing of Google Books providing a scalable and efficient way to search through information. The Annif interface defines all endpoints used in accessing the service information, searching controlled vocabularies, managing indexing projects etc., which allows users to have a clear understanding of the process and enables compatibility between systems. An additional benefit of this tool is that it has dedicated semantic processing endpoints, allowing users to submit individual pieces of content, or multiple pieces of content in bulk for subject suggestions. Therefore Annif is ideal for large-scale automation (see **Figure 8**). Annif's use of an API based architecture ensures that there is less opportunity for variation when implementing automated processes.



API information

GET / get basic information about the API service

Vocabulary information

GET /vocabs get a list of vocabularies

Project administration

GET /projects get a list of projects

GET /projects/{project_id} show project information

Automatic subject indexing

POST /projects/{project_id}/suggest suggest subjects for a given text

POST /projects/{project_id}/suggest-batch suggest subjects for the given batch of up to 32 documents

Figure 8: REST API interface of Annif for Google Books indexing (http://127.0.0.1:5000/v1/ui/)

5. Discussions

5.1 Impact on Digital Libraries

As shown in this study, the Annif AI Toolkit allows digital libraries to be enhanced through scalable, consistent, and efficient subject indexing of large collections (e.g., Google Books) using automated subject assignment. The automation of subject assignment provides reduced manual labor while also providing increased coverage across rapidly growing collections. Annif is used as an aid in a librarian's semi-automated workflow process; librarians can view and edit suggested subjects, thus enhancing metadata quality, increasing efficiency, and aiding users with more successful subject-based searching.

5.2 Influence on Library Cataloguing

The system offers a consistent method of representing subjects in library catalogue through alignment of its computer generated recommendations to controlled vocabulary resources, thereby reducing variability based on individual judgement and/or local practice which can be especially beneficial for large collection management. Findings also show that Annif operates as a successful decision support system and supports faster creation of records, authority control, quality assurance, and semantic validation.

5.3 Enhancement of Semantic Retrieval Systems

Annif's inclusion in the system provides a more semantically rich way to retrieve information via moving the process of finding related documents from being based on keywords to based on conceptual relationships. Using controlled vocabulary lists to link the content of each document will allow users to find documents that are more relevant to their needs; especially when the user is looking for something in an area of study that can include multiple disciplines. The adjustable subject limit (as seen in the other modules) allows the user to adjust the amount of specificity desired when retrieving documents and thus to achieve both precision and recall. Finally, at very large scales (e.g., Google Books), adding semantic enrichment to the retrieval process makes exploratory search easier and makes it possible to browse through themes or categories more easily.

5.4 Insights from Findings

The evaluation of the AI toolkit demonstrates that it is capable of achieving a good level of recall and ranking performance when utilized with large-scale digital libraries which require broad subject coverage to be used effectively. The trained Annif model achieves relatively moderate precision, but is able to provide strong levels of both early precision and NDCG indicating that Annif will continue to prioritize the most relevant subjects. Overall, the results confirm Annif as an effective solution for addressing the challenges of scalability, consistency and coverage in discovery-oriented digital collections.

5.5 Challenges, Limitations, and Future Directions

These studies have identified several issues with respect to the quality of training data, the degree of granularity of the vocabulary, and challenges associated with indexing information that is complex, interdisciplinary or has nuances. Concept drift indicates a continued need for periodic retraining of the system and updating the vocabulary. Future research could investigate advanced machine learning techniques (and potentially large language models), specialized ontologies, etc. as ways to improve both the semantic understanding and the ability to generate high-quality data in order to develop the next generation of systems for indexing in digital libraries.

6. Conclusion

The outcome of this integrated framework demonstrate the real-world application feasibility of the Annif AI Toolkit as a viable option for providing scalability in performing semantic subject indexing in very large-scale digital collections (using Google Books as a challenging example of such a collection). Additionally, a TF-IDF based Annif model was implemented and evaluated using 5000 records to provide evidence that high levels of recall are achievable, that good ranking quality is achieved, and that the automatically generated subjects generated by

Annif consistently align with controlled vocabulary sources, thus making it suitable for environments where there exists a need for complete coverage and semantic coherence. While the primary measure of Annif's effectiveness is its ability to function technically well, the research also provides evidence that Annif serves as an effective decision support tool for assisting professional cataloguers in improving the efficiency of their work, while maintaining intellectual control through authority validation and quality assurance processes. By providing concept driven search capabilities as opposed to surface level keyword searching, semantic indexing has the potential to improve discovery and thematic navigation of vast digital repositories. On a larger scale, the research highlights the importance of developing automated indexing systems to allow for sustainable, transparent, and reproducible subject access to digital information, and indicates that the use of advanced machine learning methods and more robust semantic resources in the future will serve to further enhance the organization and accessability of large-scale digital collections.

References

Golub, K. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, *59*(8), 702–719. https://doi.org/10.1080/01639374.2021.2012311

Kasprzik, A. (2024). The automation of subject indexing at zbw and the role of metadata in times of large language models. *Procedia Computer Science*, *249*, 160–166. https://doi.org/10.1016/j.procs.2024.11.059

Suominen, O., Lehtinen, M., & Inkinen, J. (2022). Annif and finto ai: Developing and implementing automated subject indexing. *JLIS*, *1*. https://doi.org/10.4403/jlis.it-12740

Kasprzik, A. (2023). Automating subject indexing at ZBW: Making research results stick in practice. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *33*(1). https://doi.org/10.53377/lq.13579

Golub, K. (2019). Automatic subject indexing of text. *KNOWLEDGE ORGANIZATION*, *46*(2), 104–121. https://doi.org/10.5771/0943-7444-2019-2-104

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *29*(1), 1–25. https://doi.org/10.18352/lq.10285

Suominen, O., Inkinen, J., & Lehtinen, M. (2025). *Annif at the germeval-2025 llms4subjects task: Traditional xmtc augmented by efficient llms*. arXiv. https://doi.org/10.48550/ARXIV.2508.15877

Massa, S., Annosi, M. C., Marchegiani, L., & Messeni Petruzzelli, A. (2023). Digital technologies and knowledge processes: New emerging strategies in international business. A systematic literature review. *Journal of Knowledge Management*, *27*(11), 330–387. https://doi.org/10.1108/JKM-12-2022-0993

Dylag, J. J., Zlatev, Z., & Boniface, M. (2025). Pretrained language models for semantics-aware data harmonisation of observational clinical studies in the era of big data. *BMC Medical Informatics and Decision Making*, *25*(1), 400. https://doi.org/10.1186/s12911-025-03055-y

Yang, H., Wang, N., Yang, L., Liu, W., & Wang, S. (2023). Research on the automatic subject-indexing method of academic papers based on climate change domain ontology. *Sustainability*, *15*(5), 3919. https://doi.org/10.3390/su15053919