

Historical News Question Answering: Course Project

Assignment created by: Georgios Peikos & David La Barbera

1 Goal

Develop a search engine that retrieves and ranks historical newspaper passages from digitized archives in response to users' natural language questions about events, people, and facts from 1800 to 1920. The system should cope with challenges such as noisy OCR text, archaic language, and temporal expressions that require normalization to historical publication dates. It should identify and link relevant entities across time and geography, correct transcription errors to improve retrieval accuracy, and extract concise answers supported by original sources. The goal is to provide transparent, contextually accurate access to factual information drawn from historical news content.

The benchmark used in this project was built from the Chronicling America collection, which includes over twenty-one million digitized newspaper pages (1756–1963) curated by the Library of Congress and the National Endowment for the Humanities. It comprises 39,330 newspaper pages published between 1800 and 1920 across fifty-three US states, ensuring broad geographic and temporal coverage.

2 Project Description and Material

This project enables students to explore and apply approaches that combine Information Retrieval and Natural Language Processing techniques in practice. Students can leverage knowledge acquired from previous courses, material presented during this course, and external resources to develop their own ideas.

The project aims to help students gain a solid understanding of Information Retrieval systems for historical archives and the specific challenges of Question Answering over historical newspaper text, for which dedicated resources will be provided. They are expected to analyze the provided document collection, conduct baseline experiments using the lab materials, and then design and implement their own approach to address the task. Finally, students will effectively communicate their investigation by clearly presenting their methodology, findings, and insights in both a written report and an oral presentation.

In detail, students should first read the provided material to understand the main challenges that news-related IR systems face. Students are encouraged to explore the dataset by analyzing term distributions and entity occurrences. They can experiment with different indices built on the original OCR text and the cleaned version, allowing them to compare retrieval effectiveness across both. By indexing the raw and corrected contexts separately, they can gain insights into how text quality affects search performance and answer accuracy. Possible project directions include developing and evaluating retrieval methods for natural factual question answering, experimenting with neural, lexicon-based, or hybrid retrieval models, and assessing their performance using standard ranking metrics. Moreover, they can experiment with another relevance dimension, i.e. novelty, by leveraging the publication dates of newspapers or by applying advanced query understanding techniques that integrate external knowledge to enhance temporal and semantic relevance. Students may also focus on approaches to identify and link entities, and other aspects, found in the newspaper passages.

As large language models are increasingly used to explore historical events, figures, and social contexts, it has become essential to ensure that these systems rely on Retrieval-Augmented Generation approaches to enhance factual grounding. Such methods first retrieve relevant passages from historical newspapers and then generate accurate, evidence-based answers supported by explicit references to the sources. This project represents a first step toward that goal, as students will focus on developing effective retrieval systems for historical news content, which can be easily extended to generation and RAG frameworks to evaluate how integrating external evidence improves the reliability of LLM-produced answers.

This project aims to strengthen both technical and analytical skills in NLP and Information Retrieval while promoting the ability to apply them to real-world challenges. Students are strongly encouraged to think creatively and innovatively when defining their approach to this search task. The project is also designed to foster collaboration, helping students develop key soft skills such as teamwork, effective communication, and an understanding of team dynamics.

Students are encouraged to build upon the discussions and materials presented in the labs when conceptualizing and developing their projects. They should also make use of the provided resources, including carefully selected academic papers and online tutorials, to support their exploration of NLP, Information Retrieval, and financial data processing.

Resources related to the task:

- ◊ **Lab Material:** The material covered during the lab sessions serves as the primary resource for students.
- ◊ **Insights on Historical Data Retrieval** [Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey](#)
- ◊ **Original Paper that introduced the Dataset:** [ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages](#)
- ◊ **Research on news-related Question Answering in 2025:** [Google search of related studies](#). For inspiration!
- ◊ **GitHub Resource** [Link to the dataset used in this project](#).
- ◊ **PyTerrier Documentation and Experiments:** Comprehensive documentation and guides for conducting experiments using PyTerrier, an IR research platform.
- ◊ Additional material will be provided as teams progress with their chosen project ideas (e.g. keyword extraction, usage of LLMs, etc.)..

3 Participation Guidelines

- ◊ **Eligibility:** Open to all students enrolled in the course.
- ◊ **Team Formation:** Teams are required to be formed with a composition of **3 members**.
- ◊ **Communication:** Each team must appoint one member as the **communication leader**. This person is responsible for registering the team and its members with the instructor and handling all communication between the team and the course instructor.
- ◊ **Participation in Exams:** You are required to complete and submit one project to be eligible to participate in the exam! Please remember that, to be eligible to take part in an exam session, you must **first** submit the project.
 - For example, if you submit the project by 15 January, you can participate in all subsequent exam sessions.
 - If you decide to complete a project in May, you will be eligible for all sessions following that month, such as those in June, July, and September.
 - Your project grade is kept and is valid for all subsequent exam sessions.

4 Steps for Completion

4.1 Timeline

Deadlines



1. **Team Formation and Registration:** Deadline - **November 21**, via a single email to both: georgios.peikos@unimib.it and david.labarbera@unimib.it.
 - (a) By the deadline, the communication leader must send an email to the instructor including:

- ◊ A team name [Optional].
 - ◊ The selected project (for example, Financial Question Answering).
 - ◊ The name and student ID of each team member.
- (b) After registration, teams should start working on the Initial Proposal Submission (see Section 4.2).
- (c) Teams can also start working on Development Phase I (see Section 4.3).
2. **Initial Proposal Submission:** Deadline - **December 5**. Teams have to send via email the PDF file described in Section 4.2.
3. **Ideas Evaluation:** Deadline - **December 9**, each team will receive an email with further materials and comments on their proposed ideas.
- (a) After the email communication, teams can continue working on Development Phase I and start working on Phase II (see Sections 4.3 and 4.4).
4. **Final Project Submission:** Deadline - **January 12 by 23.59**; this will include a single **.zip** that contains your project, your (final/draft) presentation, and your source code. The **.zip** file must be sent via a single email at georgios.peikos@unimib.it and david.labarbera@unimib.it.
5. **Project Presentations:** Deadline - **January 15, from 14:30 - 17:30 in U6-01f**.

4.2 Initial Proposal Submission

This initial proposal is the first stage of the project. Your submission should be a concise yet comprehensive document, spanning 2-3 pages in the provided format (i.e. 12pt Font size). The PDF should be send by the team's communication leader via email to georgios.peikos@unimib.it and david.labarbera@unimib.it and encompass the following components:

Describe the task by outlining the situation in which it takes place and the specific information need it addresses. Explain who the intended users are, what goals they pursue, and what challenges they face when searching. Summarize the main characteristics of the document collection, such as its source, structure, and content type. This exercise will help you to understand the context in which your search engine will operate and guide you to the development process.

- ◊ **Task Description:** Provide a clear and concise explanation of the task. Describe who the intended users of your search engine are and explain why this task is valuable or beneficial for them.
- ◊ **Why the task is challenging?** Based on your opinion and your research, what makes your retrieval task challenging?
- ◊ **Why Solving This Task is Important:** Explain, based on your understanding and readings, why addressing this task has value. Describe the potential impact of your solution and identify who would benefit from it and how.

After understanding the task and its context, the next step is to start thinking of a solution, or even better, several potential ones. Let your creativity flow and explore different ways the problem could be tackled. The more ideas you come up with now, the better your chances of finding something original and effective later.

- ◊ **Report a List of Ideas:** Each team member should propose three ideas for a retrieval system that could address the chosen task. Include them in the PDF and discuss them together.
- ◊ **Select at Least One Idea to Develop:** As a team, agree on at least one idea to explore further and transform into a concrete project plan. Write in the PDF a short explanation why you selected this solution.
- ◊ **Define a Research Question:** Formulate one clear research question based on your selected idea, focusing on what you aim to investigate through your system [Can be the title of your project!]. **Never written one? Have a look <https://www.youtube.com/watch?v=42-d2HdbyS8>**

As the saying goes, a picture is worth a thousand words, often expressing ideas more clearly than lengthy explanations.

- ◊ **Method Outline (i.e. your idea/s):** Present an overview of your proposed methodology. You are encouraged to use visual aids such as diagrams or flowcharts to enhance clarity. For creating these, <https://www.draw.io/> or <https://excalidraw.com/> can be helpful tools. The latter allows online collaborative drawing.

Make sure you give proper credit to all sources, ideas, and materials that have contributed to your work.

- ◊ **Resource References:** Include citations or references to the resources that have helped you develop your ideas, your understanding of the problem, and inspired your ideas. This may include academic papers, online tutorials, or any other relevant material.

Ideas are exciting, but you should also assess their feasibility and ensure your goals can realistically be achieved within the available time!

- ◊ **Technical Aspects:** List and describe the technical elements and tools you intend to utilize in your project. This could include, GitHub libraries, frameworks, or algorithms, or tools you would like to learn.
- ◊ **Feedback:** You will receive feedback on the scope and feasibility of your ideas, such as whether your approach seems too ambitious, too simple, or well balanced for the project goals. We will also share additional resources to help you refine and implement your ideas more effectively.

4.3 Development and Implementation: Phase I

The initial phase of development focuses on implementing **three retrieval runs** that will serve as **baselines** for further progress. A baseline retrieval experiment represents a relatively simple retrieval pipeline designed to establish reference performance. For your projects, the baselines can also be the three simplest ideas you had in the initial proposal phase.

The first step in developing a search engine, after understanding your users and their context, is to become familiar with your data. In offline evaluation of Information Retrieval Systems, this includes the document collection you will index and the queries that represent how users express their information needs. Finally, you should examine the relevance assessments available in your collection to understand how documents are judged in relation to queries.

- ◊ **Collection, Query, and Relevance Analysis:** Teams can present several statistics that can inform the development of their IR system.
- ◊ **Purpose:** This analysis not only guides the design of your retrieval system but also helps you understand why the system may not perform as expected and identify areas for improvement.
- ◊ **Usage of ASPIRE:** To help you in your analysis, you can use ASPIRE or you can write your own code.

In any case, here are some simple ideas your teams can try as baselines.

- ◊ **Indexing Strategies:** Experiment with various indexing approaches. For instance, one approach could involve indexing the entire content of documents, while another could focus on indexing only titles or specific sections (if applicable in your task).
- ◊ **Query Formulation and Expansion:** Experiment with different ways of expressing user queries, ranging from simple keyword-based formulations to more structured or natural language queries. Explore query expansion techniques such as adding synonyms, related terms for query expansion, or automatically retrieved relevant words (e.g. using the RM3 model).
- ◊ **Retrieval Models:** Start with classic lexical-based retrieval using TF-IDF and BM25.
- ◊ In total, you should implement **three** simple retrieval experiments.

Here are specific guidelines to help you structure and carry out this phase effectively.

- ◊ **Experiments:** You must index **the whole collection**, i.e. all documents.
- ◊ **Evaluation Measures:** To ensure consistent comparison of retrieval effectiveness across your retrieval systems and across teams, you must report the following metrics in your experiments: **P@1, P@5, P@10, R@5, R@10, nDCG@5, nDCG@10, and MAP**.
- ◊ **Starting Resource:** Each team can start their experiments based on the colab notebook we have prepared for you here: Download the collection [here](#).

This phase is important for setting a foundational understanding of the retrieval process, upon which more advanced techniques will be built in subsequent phases of the project. Please ensure that your final PDF report goes beyond merely stating the results by providing clear explanations and interpretations of the obtained outcomes. Use this opportunity to reflect on what your results reveal about the behavior of your retrieval system and the characteristics of your data. Discuss possible reasons behind performance differences across runs and suggest what could be improved or explored in the next phase.

4.4 Development and Implementation: Phase II

During this second phase, teams will move forward from the foundational work established in Phase I to the development and evaluation of their unique ideas.

Here are specific guidelines to help you structure and carry out this phase effectively.

- ◊ Conduct **at least two** additional experiments (E1, E2) that extend beyond the baselines. One of these (E1) can involve a neural approach, such as a bi-encoder retriever or a neural re-ranker.
- ◊ Alternatively, E1 can be an experiment using LLMs for tasks like query expansion or reformulation to enhance retrieval effectiveness.
- ◊ For the E2, you must develop and evaluate your proposed idea in depth, clearly describing the methods used, the rationale behind your choices, and the impact on retrieval performance.
- ◊ **Experiments:** You must index **the whole collection**, i.e. all available documents in the provided collection.
- ◊ **Evaluation Measures:** To ensure consistent comparison of retrieval effectiveness across your retrieval systems and across teams, you must report the following metrics in your experiments: **P@1, P@5, P@10, R@5, R@10, nDCG@5, nDCG@10, and MAP.**

5 Minimum Requirements Expected

Each project should demonstrate both understanding and experimentation. In summary, we expect you to:

- ◊ **Show understanding of the task and its context:** Explain how your search engine will operate, describe the benchmark collection, and identify the main challenges of your task.
- ◊ **Develop three baseline experiments:** Implement and evaluate **three** simple retrieval pipelines that serve as baselines for comparison.
- ◊ **Develop two advanced experiments:** Design and test **two** sophisticated retrieval approaches, such as neural ranking/re-ranking, and your own idea!

Your **final submission** must be a **.zip** file containing:

- ◊ The source code implementing your retrieval experiments (this may include scripts or Google Colab notebooks).
- ◊ The final project report, prepared using the provided template. You can download the template in word format from [here](#), or the latex version from [here](#). The report must be up to 5 pages of content.
- ◊ A draft or the final version of your presentation.

6 Assessment Criteria

Credits. The entire project carries a total of 3 credits. The initial part, which includes understanding the task and its challenges, proposing your ideas, and completing the development described in Phase I, will account for **1.5 credits**. The subsequent Phase II is allocated the remaining **1.5 credits**. In Phase I, the focus is on evaluating your understanding and application of the fundamental concepts covered in the labs. Phase II presents an opportunity for each team to focus on an area of their personal interests. It encourages **exploration** and **innovation**, as teams have the freedom to choose and develop their own ideas.

Bonus: The team that develops the most complete, innovative, and high-performing retrieval system within each project will receive an additional **1 credit**.

The projects will be evaluated based on their:

- ◊ **Functionality:** The system should be able to retrieve relevant documents; the absolute performance is not the main focus, as long as the retrieval process functions correctly.
- ◊ **Innovation:** Originality of the proposed solutions.
- ◊ **Technical Implementation:** Quality of coding, algorithm design, and data handling.
- ◊ **Team Collaboration:** Communication and teamwork.
- ◊ **Presentation:** Written and oral presentations.

7 Additional Information and Support

For any additional information or support, please send an email at georgios.peikos@unimib.it and david.labarbera@unimib.it.