LAIRHUB UOFT

24th June, 2024 - Scatter Gather

Minutes of MeetingDate: June 24, 2024

Attendees: Aravind Narayanan, Prof. Javed Mostafa

Subject: Web of Science API Expansion, Data Analysis, and Integration into PATTIE

Agenda:

1. Expansion of Web of Science API and Semantic Scholar data collection.
2. Data analysis and storage.
3. Integration of data into PostgreSQL and modification of PATTIE.
4. Topic modeling and LLM integration.
5. Workflow design for data acquisition and interaction with PATTIE.
6. Preparation for upcoming conference.

Minutes:

1. **Expansion of Web of Science API:**
    - Successfully expanded Web of Science API along with the starter API.
2. **Data Collection from Semantic Scholar:**
    - Obtained approximately 1,500 papers from Semantic Scholar, with around 800 including abstracts.
    - Manually skimmed through the papers and performed simple data analysis.

- Identified popular venues such as the Digital Humanities Conference.
- Used the field of study to filter out irrelevant information due to similar names of professors and authors across different fields.
- Stored all data in JSON format.
- Achieved an average of 1-2 requests per second using the Semantic Scholar API.

3. **Data Storage and Integration:**
   - Immediate next step: Port data into PostgreSQL.
   - Modify PATTIE code to reference PostgreSQL instead of PubMed to test initial functionality. This is the primary focus.
4. **Topic Modeling and LLM Integration:**
   - Brainstorming on topic modeling.
   - Consider using LLMs to improve cluster labels.
   - Design a workflow for the entire process from data acquisition to seamless integration into PostgreSQL, avoiding duplicates.
   - Ensure the pipeline works in tandem with the PostgreSQL database and formulate queries accordingly.
   - Reorganize data post-acquisition and create a flow for this as well.
5. **Workflow for PATTIE Interaction:**
   - Identify parts of PATTIE that specifically reference PubMed and adapt them to interact with PostgreSQL.
   - Before frontend implementation, use LLMs to expand cluster labels for better readability and meaningful descriptions. Ensure API token limits are set accordingly.
   - Use prompt engineering to improve the process; older GPT models might be sufficient and cost-effective.
   - Limit API calls as there will be a maximum of 10 clusters.
6. **Preparation for Upcoming Conference:**
   - Upcoming conference: Joint Conference on Digital Libraries in July/August.

- Deadlines: July 26th (regular) and August 10th (late-breaking).
- Aim to present a well-designed workflow that links LLMs for this application.
- Prepare flowcharts and flow diagrams to showcase the innovative system architecture.
- Highlight the novelty of using LLMs for better navigation.
- Implement a toggle switch for topic navigation, enabling switching between papers/authors, directly mapping to relevant papers within a topic.

Action Items:

□ Port data into PostgreSQL.

□ Modify PATTIE to reference PostgreSQL instead of PubMed.

□ Design a seamless workflow for data acquisition and integration.

□ Use LLMs to improve cluster labels and expand descriptions.

□ Prepare for the Joint Conference on Digital Libraries with relevant documentation and presentations.