**Reproducing Paper on Generating Semantic Annotations for Frequent Patterns with Context Analysis**

**Project Overview:**

In this project, we have tried to reproduce the model and results from the following published paper on Pattern Annotation.

**Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, and ChengXiang Zhai. 2006. Generating semantic annotations for frequent patterns with context analysis. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2006). ACM, New York, NY, USA, 337-346. DOI=10.1145/1150402.1150441**

The goal is to annotate a frequent pattern with in-depth, concise, and structured information that can better indicate the hidden meanings of the pattern. This model will automatically generate such annotation for a frequent pattern by constructing its context model, selecting informative context indicators, and extracting representative transactions and semantically similar patterns.

This general approach has potentially many applications such as generating a dictionary like description for a pattern, finding synonym patterns, discovering semantic relations, ranking patterns, categorizing and clustering patterns with semantics, and summarizing semantic classes of a set of frequent patterns.

**Experiment from the paper reproduced:**

Given a set of authors/co-authors, annotate each of them with their strongest context indicators, the most representative titles from their publications, and the co-authors and title patterns which are most semantically similar to them.

**Implementation Approach:**

Here, the general approach taken to automatically generate the frequent patterns and structured annotations for them by the following steps:

1. **Derive the frequent patterns Author/Co-Author pattern** from the Database using FP Growth algorithm.

2. **Define and model the context of the pattern**:

* Derive the context units for the Author/Co-Authors by mining the Closed Frequent Pattern to avoid any redundant pattern
* Derive the context units for the Titles by mining the Sequential Closed Frequent Pattern using Prefix Span Algorithm
* Select context units and design a strength weight for each unit to model the contexts of frequent pattern

3. **Derive the list of representative transaction** by finding the cosine similarity between the Frequent Pattern & the Frequent Transactions

4. **Derive the Semantically Similar Author Pattern** by finding the cosine similarity between the Context Units of the Frequent Pattern with the other context units and annotating it with contextually similar Frequent Author/Co-Author

5. **Derive the Semantically Similar Title Pattern** by finding the cosine similarity between the Context Units of the Frequent Pattern with the other context units and annotating it with contextually similar Title Words

**Installation and Usage:**

The following **packages** must be installed for the successful execution:

https://test.pypi.org/simple/ krovetz

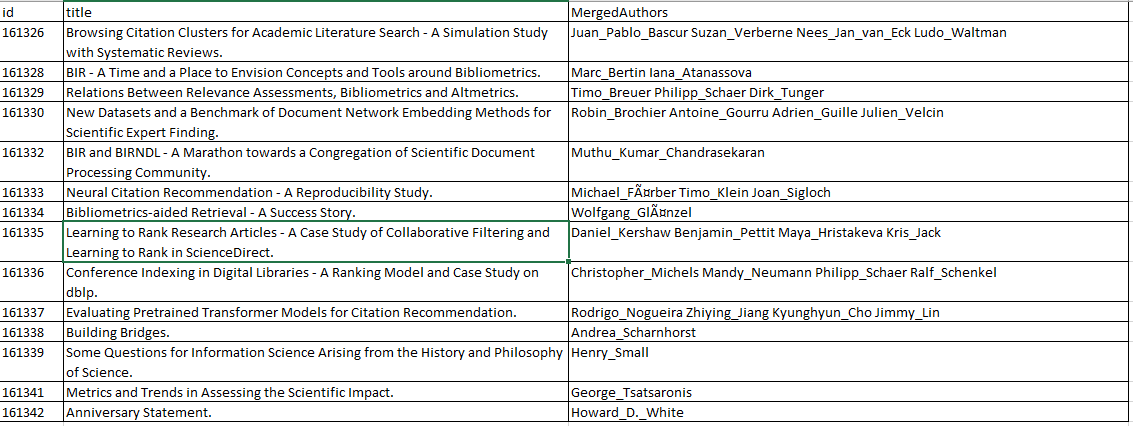
Prefixspan, re, csv, pandas, numpy, mlxtend, sklearn

*Note: To use the Krovetz Stemmer installed, Visual Studio C++ is needed to be installed. Team members had issues running the program without it.*

**Input data:**

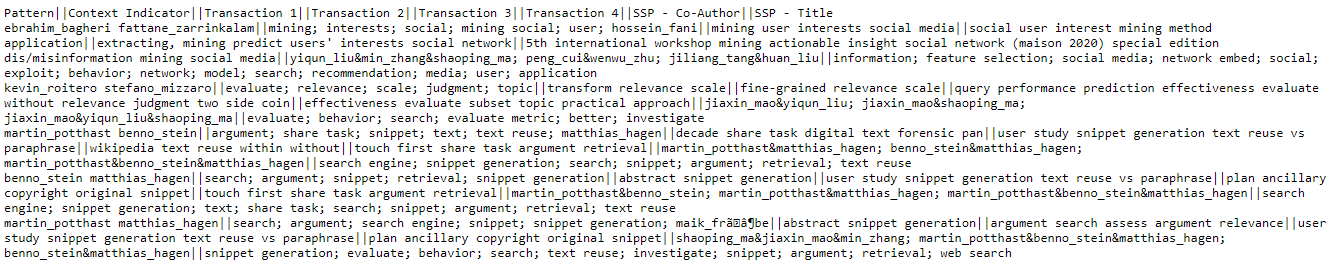
Here the input dataset (DBLP Dataset) is in a specific format. DBLP dataset (a subset of around 12k transactions/titles papers from the proceedings of 12 major conferences in Data Mining; around 1k latest transactions from each of such conference).

Each row represents a title presented in a conference. It has 3 fields – id (numeric), title (String) and MergedAuthors (string). The authors and co-authors associated with the title has been merged into a single column for easy analysis.

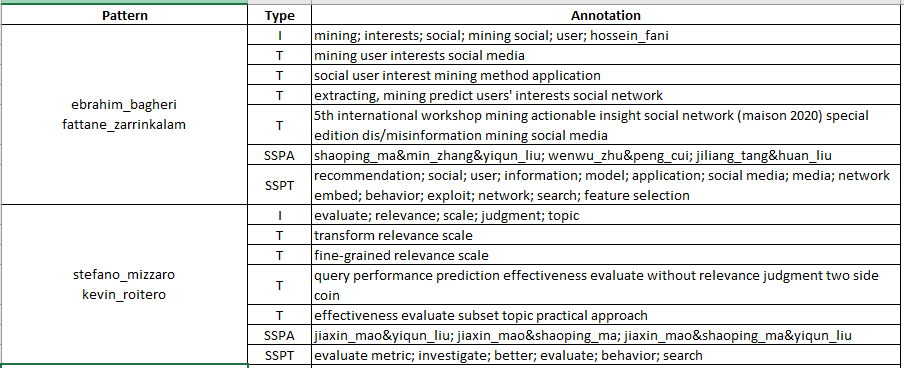


**Output:**

Once the program **patternAnnotation.py** is executed, it takes the **DBLP\_Dataset.csv** as the input and generates the output.txt file in the same path where source code exists. This output file contains all the closed frequent patterns and their most representative Context Indicators, most representative transactions (capped as 4), Semantically Similar Patterns (SSPs) as per co-author patterns and title term patterns. Each record in the output file represents one closed frequent pattern and their associated details.



Manually converted output to output.xlsx, with summary information of the annotated pattern is available as well:



I – Context Indicator; T – Representative Transactions; SSPA – Semantically Similar Author Pattern; SSPT – Semantically Similar Title Pattern

**Implementation Details:**

**Derive the frequent itemsets of Author/Co-Author:**

* **Algorithm used:** FP Growth
* **Output:** 64 Frequent Author/Co-Author itemsets (Removed any frequent author without co-Author)

**Define and model the context of the pattern:**

For each of the 64 Frequent Itemsets from the above step:

* **Derive the context units for the Author/Co-Authors:** Mined closed Author/Co-Author itemsets from already mining frequent itemsets by removing any itemset with redundant support
* **Derive the context units for the Title:** Algorithm used: Prefix Span; We got limited Title Patterns, hence we did not need micro-clustering to further reduce the titles.
* **Final Context Indicator:** Merged the context units of the above two steps
* **Context Indicator Weighting:** Here we slightly modified our approach to weighting as the mutual information value was not giving the appropriate weightage, based on our analysis & understanding. We use an approach similar to IDF, where when a context unit is present in more transaction, the weightage is reduced & any context unit that uniquely appears in the transaction will have the highest weightage.

**Derive the list of representative transaction:**

* Based on the context indicators identified for each of the Frequent Itemset, we identified top 4 transactions with the highest Cosine Similarity

**Derive the list of Semantically Similar Author/Co-Author pattern:**

* Based on the context indicators identified for each of the Frequent Itemset, we identified which of the other 63 context indicators are similar using Cosine Similarity. We took the top 3 similar Context Indicators & took their Author/Co-Author as the SSP

**Derive the list of Semantically Similar Author/Co-Author pattern:**

* Based on the context indicators identified for each of the Frequent Itemset, we identified which of the other 63 context indicators are similar using Cosine Similarity. We took the top 3 similar Context Indicators & took their frequent title pattern as the SSP

**Team Contribution:**

|  |  |  |
| --- | --- | --- |
| Name | ID | Contribution |
| Bipin Chandra Karnati | karnati3 | Analysis, Implementation of Extracting Title Patterns of Frequent Authors, Context Indicator Identification, Code Integration |
| Savitha Govindarajan | savitha3 | Analysis, Implementation of Extracting Itemsets of Frequent Authors, SSP Author Pattern, Co-ordination, Code Integration |
| Utpal Mondal | umondal2 | Analysis, Data Extraction & Clean Up, Implementation of Cosine Similarity, Detailed analysis on MI, Documentation , Code Integration |

**References:**

Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, and ChengXiang Zhai. 2006. Generating semantic annotations for frequent patterns with context analysis. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2006). ACM, New York, NY, USA, 337-346. DOI=10.1145/1150402.1150441

**FP Growth Algorithm implementation:** https://towardsdatascience.com/fp-growth-frequent-pattern-generation-in-data-mining-with-python-implementation-244e561ab1c3

**Prefix Span:** https://pypi.org/project/prefixspan/

**Cosine Similarity:** <https://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/>

**DBLP Dataset:** 12 Major Conferences on Data Mining. 1000 latest titles from each conference

**ACL** - Annual Meeting of the Association for Computational Linguistics (<https://dblp.uni-trier.de/db/conf/acl/>)

**ADBIS** - Symposium on Advances in Databases and Information Systems (<https://dblp.uni-trier.de/db/conf/adbis/>)

**CIKM** - International Conference on Information and Knowledge Management (<https://dblp.uni-trier.de/db/conf/cikm/>)

**ECIR** - European Conference on Information Retrieval (<https://dblp.uni-trier.de/db/conf/ecir/>)

**ICDE** - IEEE International Conference on Data Engineering (<https://dblp.uni-trier.de/db/conf/icde/>)

**ICDM** - IEEE International Conference on Data Mining (<https://dblp.uni-trier.de/db/conf/icdm/>)

**KDD** - Knowledge Discovery and Data Mining (<https://dblp.uni-trier.de/db/conf/kdd/>)

**PAKDD** - Pacific-Asia Conference on Knowledge Discovery and Data Mining (<https://dblp.uni-trier.de/db/conf/pakdd/>)

**SDM** - SIAM International Conference on Data Mining (<https://dblp.uni-trier.de/db/conf/sdm/>)

**SIGIR** - Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (<https://dblp.uni-trier.de/db/conf/sigir/>)

**WSDM** - Web Search and Data Mining (<https://dblp.uni-trier.de/db/conf/wsdm/>)

**WWW** - The Web Conference (<https://dblp.uni-trier.de/db/conf/www/>)