10/11/2020
CS 410 Text Information System

Team APlus Project Proposal

Section 1:

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Full Name	NetID	Responsibility
Difan Gu	difangu2	Captain
Yanyue Wang	yanyuew2	Member
Long Wen	longw2	Member

2. Which paper have you chosen?

Generating Semantic Annotations for Frequent Patterns with Context Analysis 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

3. Which programming language do you plan to use?

Python

4. Can you obtain the datasets used in the paper for evaluation?

Yes. Download Link: https://dblp.org/xml/dblp.xml.gz

Section 2:

1. What is the function of the tool?

The tool will generate the semantic annotations for frequent patterns that are detected in the text documents.

2. Who will benefit from such a tool?

There are many people and entities who can utilize the tool. For example, librarians, companies, text mining researchers, and students can benefit from the tool. One good example can be low-quality review detection. Low-quality or fake reviews can be detected if there are significant amounts of similar patterns shared between them. Then the tool is able to mark those behaviors out and potentially eliminate them.

3. Does this kind of tools already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?

We chose to replicate the paper: Generating Semantic Annotations for Frequent Patterns with Context Analysis. So instead of comparing a new tool from existing ones, we are going to compare our results with results from the paper and explain possible discrepancies.

4. What existing resources can you use?

For the dataset, we have obtained an updated version of the same dataset used in the paper from https://dblp.uni-trier.de/xml/.

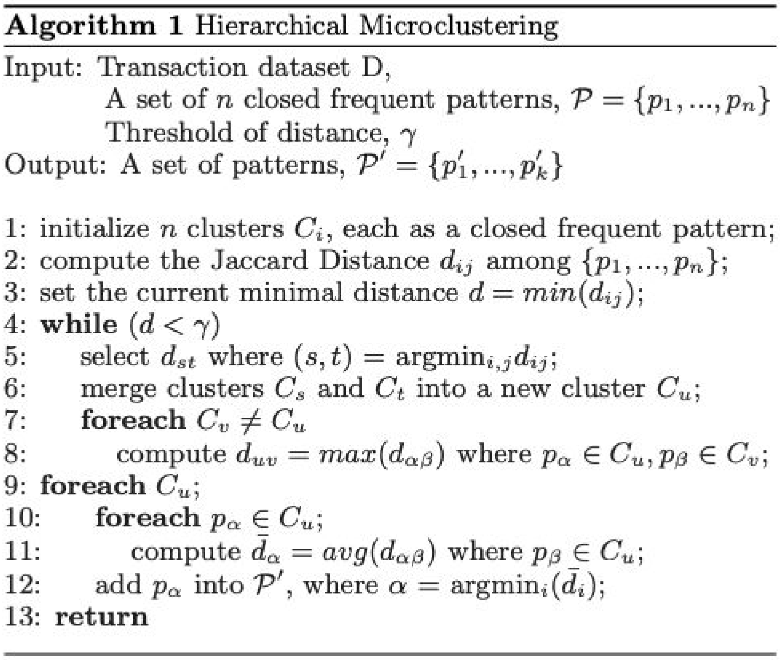
Natural Language Toolkit (NLTK)

MetaPy

BeautifulSoup

5. What techniques/algorithms will you use to develop the tool? (It's fine if you just mention some vague idea.)

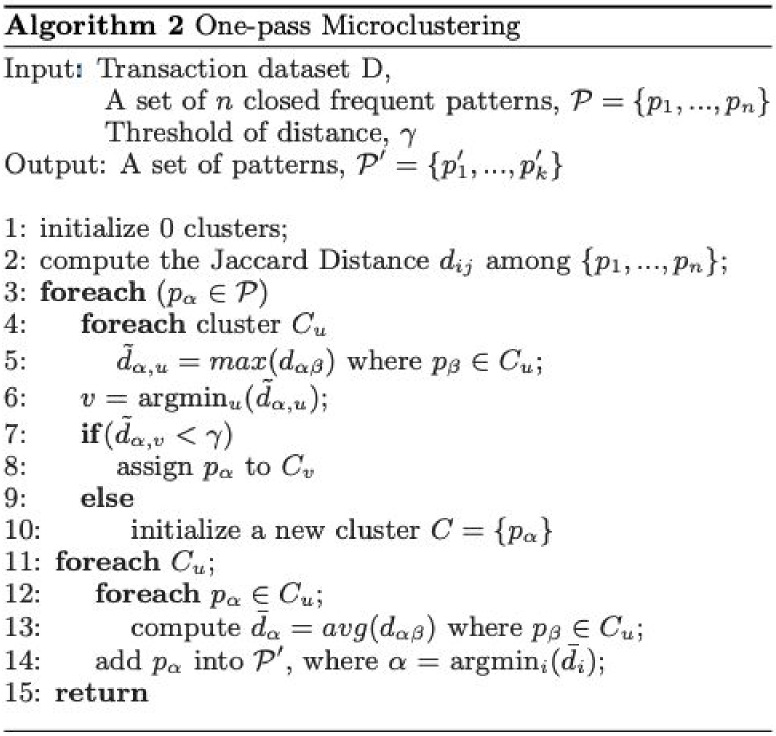
We will use the python to re-produce the algorithms mentioned in the paper.



Algorithm 1 Hierarchical Microclustering

```
Input: Transaction dataset D,
       A set of $n$ closed frequent patterns, $\mathcal{P} = \{p_1, ..., p_n\}$
       Threshold of distance, $\gamma$
Output: A set of patterns, $\mathcal{P}' = \{p'_1, ..., p'_k\}$

initialize $n$ clusters $C_i$, each as a closed frequent pattern;
compute the Jaccard Distance $d_{ij}$ among $\{p_1, ..., p_n\}$;
set the current minimal distance $d = min(d_{ij})$;
while $(d < \gamma)$
    select $d_{st}$ where $(s, t) = \text{argmin}_{i,j} d_{ij}$;
    merge clusters $C_s$ and $C_t$ into a new cluster $C_u$;
    foreach $C_v \neq C_u$
        compute $d_{uv} = max(d_{\alpha\beta})$ where $p_\alpha \in C_u, p_\beta \in C_v$;
foreach $C_u$;
    foreach $p_\alpha \in C_u$;
        compute $\bar{d}_\alpha = avg(d_{\alpha\beta})$ where $p_\beta \in C_u$;
    add $p_\alpha$ into $\mathcal{P}'$, where $\alpha = \text{argmin}_i(\bar{d}_i)$;
return
```



Algorithm 2 One-pass Microclustering

```
Input: Transaction dataset D,
       A set of $n$ closed frequent patterns, $\mathcal{P} = \{p_1, ..., p_n\}$
       Threshold of distance, $\gamma$
Output: A set of patterns, $\mathcal{P}' = \{p'_1, ..., p'_k\}$

initialize 0 clusters;
compute the Jaccard Distance $d_{ij}$ among $\{p_1, ..., p_n\}$;
foreach $(p_\alpha \in \mathcal{P})$
    foreach cluster $C_u$
        $\tilde{d}_{\alpha,u} = max(d_{\alpha\beta})$ where $p_\beta \in C_u$;
    $v = \text{argmin}_u(\tilde{d}_{\alpha,u})$;
    if$(\tilde{d}_{\alpha,v} < \gamma)$
        assign $p_\alpha$ to $C_v$
    else
        initialize a new cluster $C = \{p_\alpha\}$
foreach $C_u$;
    foreach $p_\alpha \in C_u$;
        compute $\bar{d}_\alpha = avg(d_{\alpha\beta})$ where $p_\beta \in C_u$;
    add $p_\alpha$ into $\mathcal{P}'$, where $\alpha = \text{argmin}_i(\bar{d}_i)$;
return
```

6. How will you demonstrate the usefulness of your tool?

We will apply the tool on the dataset mentioned in the paper, i.e. DBLP Dataset, and eventually deliver a presentation based on our findings to demonstrate what benefits and functionalities the tool can bring to the table.

7. A very rough timeline to show when you expect to finish what. (The timeline doesn't have to be accurate.)

Oct 24: Submit proposal and start researching
Oct 30: Complete researching and start coding
Nov 28: Progress Report Submission
Dec 7: Complete coding and start video recording
Dec 12: Final Submission of Code and Video Presentation