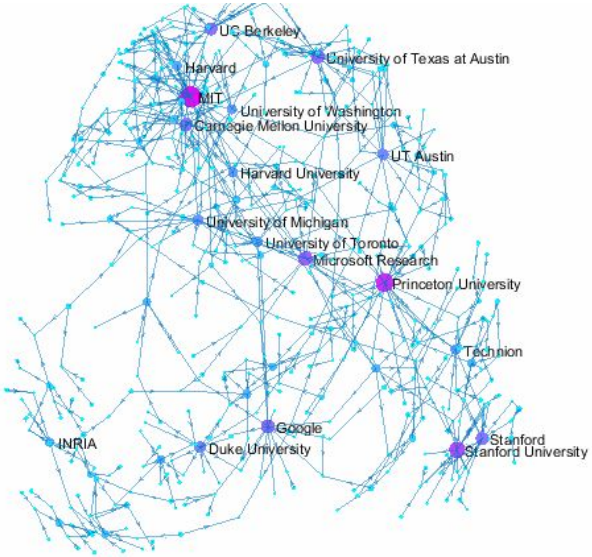
Semantic Annotations for Frequent Patterns with Context Analysis

Difan Gu
Wen Long
Yanyue Wang

ILLINOIS

Agenda

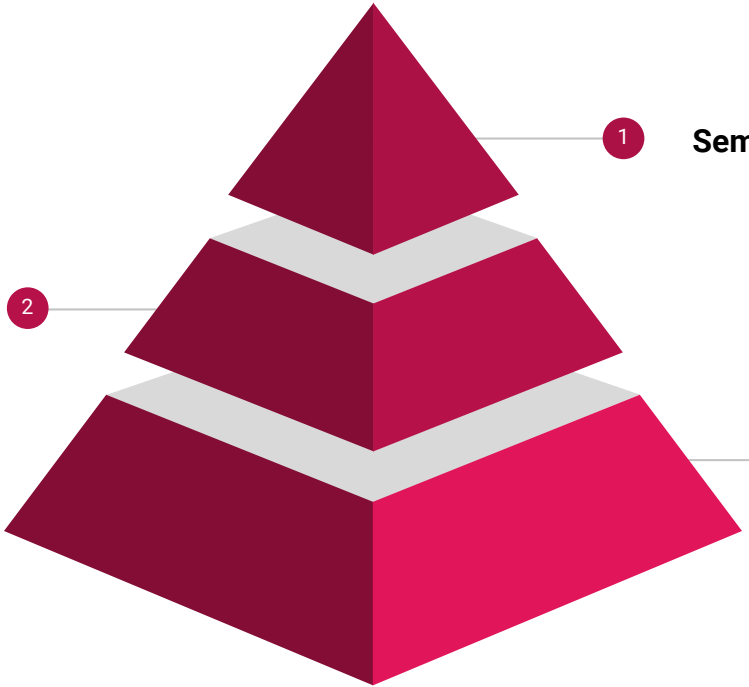
1. Project Objective
2. Methodology
3. Use Case: DBLP
4. Setup
5. Demo
6. Result



Loren on the Art of MATLAB

Project Objective

Goal: to generate semantic annotations by providing contextual information to help readers summarize ill-defined or vague information.



ILLINOIS

Methodology

Data Preprocessing

- Extract Data from XML
- Tokenize Text Data
- Remove Stop Words
- Extract Frequent Terms

Pattern Mining

- Pattern Preprocessing
- Extract Frequent Pattern
- Remove Pattern Redundancy

Ranking

- Rank by Mutual Information
- Word Conversion
- Generate Result

Use Case: DBLP

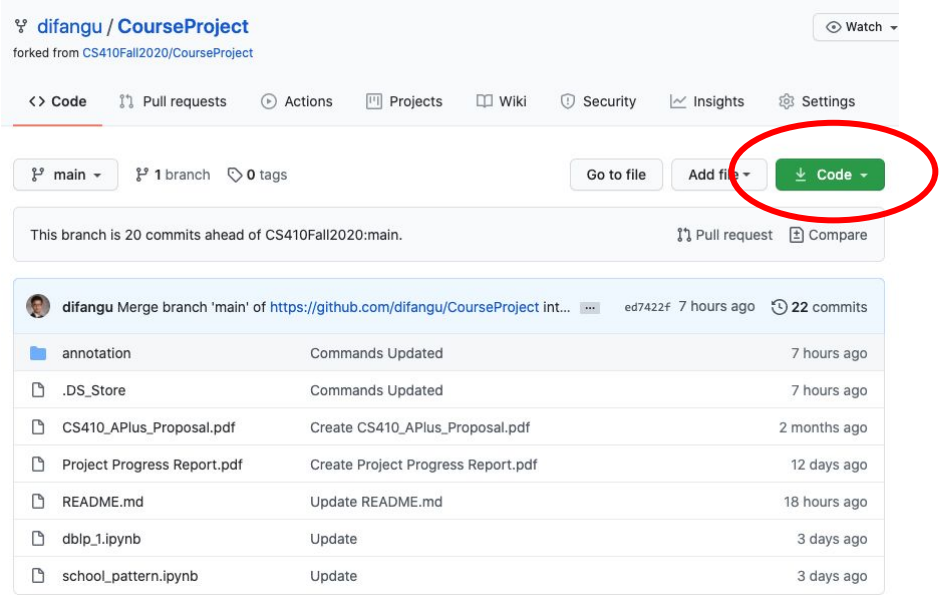
Given a large set of data, what are the most distinguished **academic focuses** of each college in major computer science conferences?

Desired Result:

- MIT: Artificial Intelligence, Robotics
- GT: Algorithm, Database
- UoM: Data Management, System Design

Setup

1. Go to https://github.com/difangu/CourseProject
2. Hit green “Code” button and download to any directory
3. Memorize the directory the “annotation” folder you saved



ILLINOIS

Demo

Demo

1. Open Terminal and cd to the location where holds the folder “annotation”

2. Run the following code in the oder:

    a. python school_pattern_creator.py
    b. java -jar spmf.jar run CloSpan ./school_output/school9.txt ./pattern_output/output9.txt 1%
    c. java -jar spmf.jar run CloSpan ./school_output/school2.txt ./pattern_output/output2.txt 0.5%
    d. java -jar spmf.jar run CloSpan ./school_output/school8.txt ./pattern_output/output8.txt 3%
    e. python pattern_decipher.py

3. Check out the result either in terminal printout or in the “outcome” file under “annotation”

Result

Top 5 semantic annotation for each school:

MIT	GT	UoM
System Control	Programming	Resource, Wireless
Power, Energy	Framework	Embedded, Systems
Control, Analysis	Management	Network, Resource
Power, Applications	Problems	Recognition, Object
Systems, Large	Approach	Social

Thank You

Reference

KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining August 2006 Pages 337–346 https://doi.org/10.1145/1150402.1150441

CloSpan: Mining Closed Sequential Patterns in Large Datasets, by X. Yan, J. Han, and R. Afshar. Proc. of 2003 SIAM Int. Conf. Data Mining (SDM'03), 2003