IST736 Text Mining

Professor Bei Yu  
Spring 2025

HW\_1 Lexicon-based Text Mining  
With Whom Do People Spend Happy Moments?

**Data**

HappyDB is a corpus of 100,000+ crowd-sourced descriptions of happy moments. HappyDB is available on GitHub [1]. The cleaned data set is located at [2].

**Task (10 points in total; four subtasks, 25% each)**

1. HappyDB provided a “people dictionary” [3], which is a lexicon of common social relationships. Use this lexicon to find the top three social relationships mentioned in happy moments, e.g. spouse, parents, children, friends, or someone else.
2. Use your world knowledge to assess the strengths and weaknesses of the “people dictionary” in terms of answering the question “with whom do people spend happy moments?” You can define “people”, “strength” and “weakness” in your own way here.
3. Modify the “people dictionary” to fix the weaknesses that you have identified. Use the revised lexicon to redo task #1.
4. Extract five context words before and after the most mentioned people, sort by frequency, and explore patterns in the 100 most frequent context words, for example, any words indicating activities that people do in happy moments?

**Submission**

1. Submit a Python script that includes your code and your comments to explain your analyses and findings.
2. Add your name in both filename and comment on top of your code.
3. If you used co-pilot or other AI tools for this homework, add your prompts as comments in the submitted script. Prompts are not graded. They are collected to inform the instructor on students’ needs in programming help.

Grading criteria:

* Accuracy and reproducibility in analysis
* Clarity in explaining the analytical process and result

References:

[1] <https://github.com/megagonlabs/HappyDB>

[2] <https://raw.githubusercontent.com/megagonlabs/HappyDB/master/happydb/data/cleaned_hm.csv>

[3] <https://github.com/megagonlabs/HappyDB/blob/master/happydb/data/topic_dict/people-dict.csv>