CSIC 5011

Arc Diagrams for Co-Appearance Data in Novels, Visualization of Dream of Red Mansion characters

Xavier Plantaz, Department of Computer Science, The Hong Kong University of Science and Technology, xpfplantaz@connect.ust.hk

Abstract

Understanding characters relations in novels is an open problem and not well explored yet. In the report, I present an approach to investigate relations between main and secondary characters in Dream of Red Mansion characters, one of China's Four Great Classical Novels. Exploring different views to represent the dataset of 374 characters and their 26205 interactions, I define a qualitative method to find a 1D embedding resulting in a arc diagram visualization which is meaningful to understand the novel.

Motivation

Different modern visualization explore the idea of transforming complex relations data. For example, many papers focus on how to use paper citation graphs to find patterns in collaborations between universities. Books, the same way, give great opportunities to explore the complexity of the relations between their characters.

Problem Definition

How might we…

Create a systematized process

For artists and data visualization engineers

To find patterns in the relations between the characters

By displaying in an efficient way the relations graph of a given novel

So that we understand how they are linked and how they cluster

Source Code

Open source project in public domain
https://github.com/xpfio/CSIC5011
License CC0 1.0 Universal (CC0 1.0)

Data

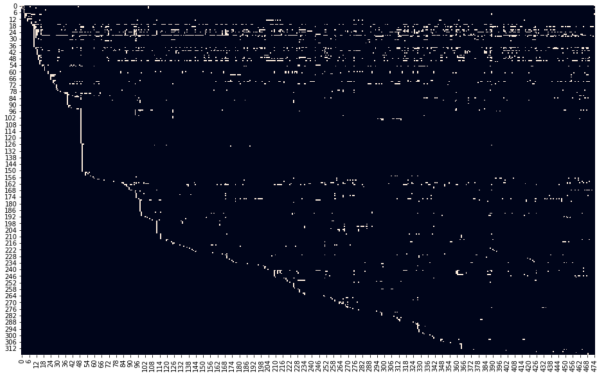
Dream of the Red Chamber, one of the Four Great Classical Novels of Chinese Literature was written by CAO, Xueqin in the middle of 18th century. The data frame provided contains 475 rows and 375 columns. It records the appearance of 374 characters in 475 scenes. The dataset was collected via crowdsourcing in the classes of Mathematical Introduction to Data Analysis and Statistical Learning, taught by Prof. Yuan YAO at Peking University. Data is available as R format, it is structured as a R DataFrame and can be transformed into a Pandas Data Frame using a Python library called rpy2.

	_name	appears_in_scene_001	appears_in_scene_002	appears_in_scene_003	appears_in_scene_004	appears_in_scene_00
0	chap80	1	1	1	1	1
1	贾演	0	0	0	0	0
2	贾源	0	0	0	0	0
3	贾代化	0	0	0	0	0
4	贾代善	0	0	0	0	0

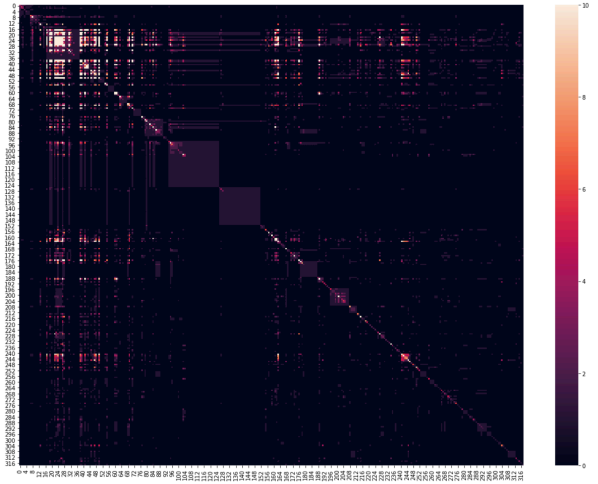
Data Frame: (374+1) characters x 475 events

Data Cleaning and Timeline View

A lot of Data Cleaning has to be done before starting working on the process, details are available in the notebooks, leading to this view, of 318 characters where we can see the apparition of characters over time (ex: for 贾宝玉, ■□■■□ … at line 18)



Correlation View

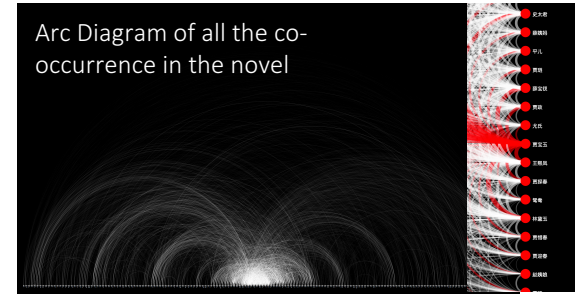


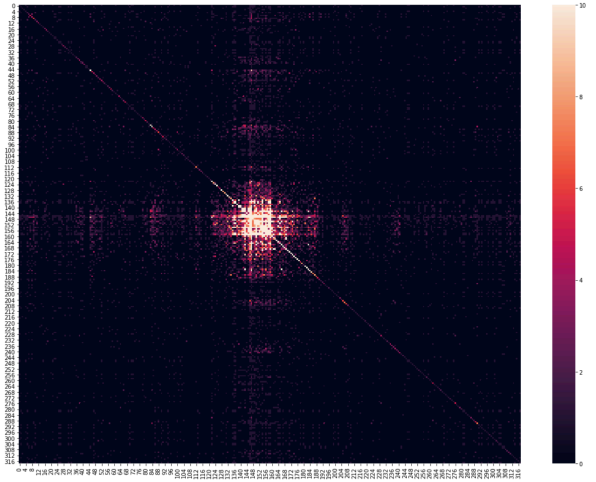
Correlation Matrix (Cij = number of scenes where i and j appear together)

This matrix show how complicated the relations between the characters are, as represented in the attached force layout, that doesn't help to understand who are the key characters and how they are connected.

1D-Embedded Arc Diagram

The main part of this project is the embedding of this correlation matrix in 1 dimension to create the final view (on the right).
To do so, I explored many possibilities including random projection, PCA projection, Isomap, LLE, LTSA, MDS, Totally Random Trees, Spectral embedding and t-SNE (see attached). Best results were obtained with MDS. I also provide in the notebooks the exploration of different datasets, such as the correlation from Les Miserables, exposing different possibilities for 1D embedding. Best results were obtained defining a custom distance, to create similarities between characters, defined as 1 divided by the number of scene were two characters appeared together.





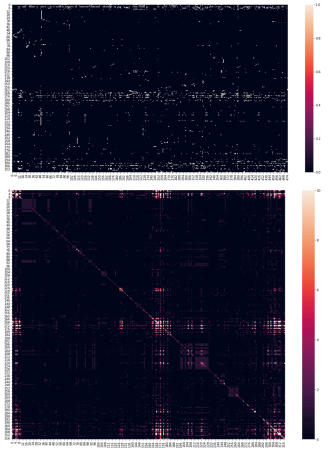
Conclusion

The selected 1D embedding transformation creates a very interesting arc diagram, flat enough and very relevant to pursue deeper character analysis. The highlighting function allow the user to select a given character, for example贾宝玉in the above picture and understand it's connection with the other characters in the novel. However, we loose the time dimension, and clusters inside chapters might be interesting to explore as future work. This work can also easily be extended to other relations graph such as social networks or the citations graph mentioned earlier.

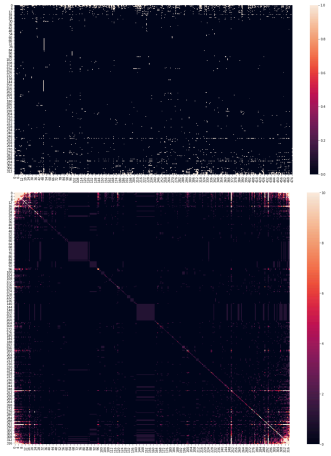
CSIC 5011

Arc Diagrams for Co-Appearance Data in Novels, Visualization of Dream of Red Mansion characters

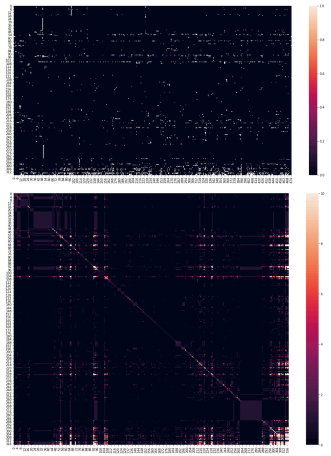
Xavier Plantaz, Department of Computer Science, The Hong Kong University of Science and Technology, xpfplantaz@connect.ust.hk



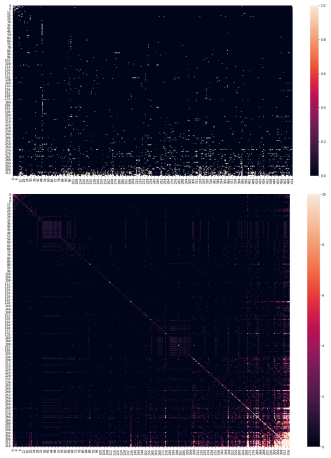
X_projected



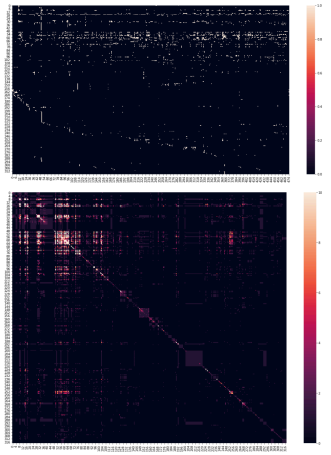
X_iso



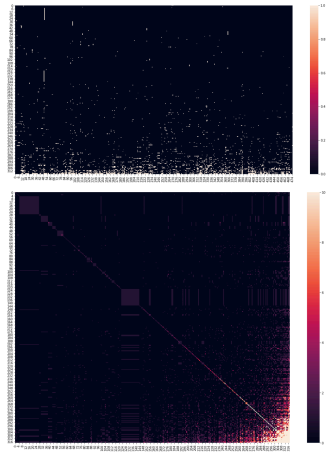
X_lle



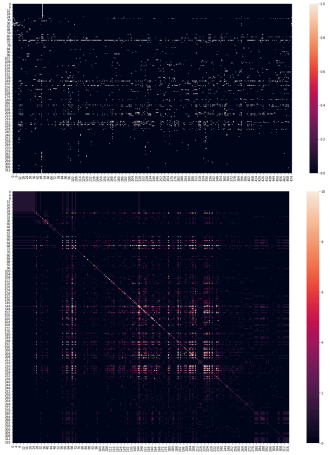
X_mlle



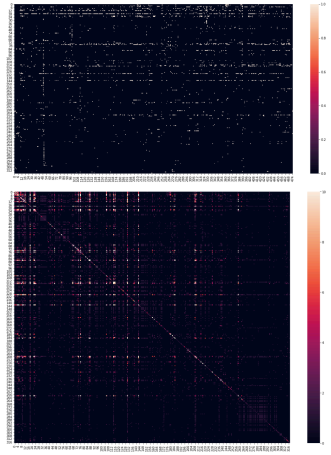
X_ltsa



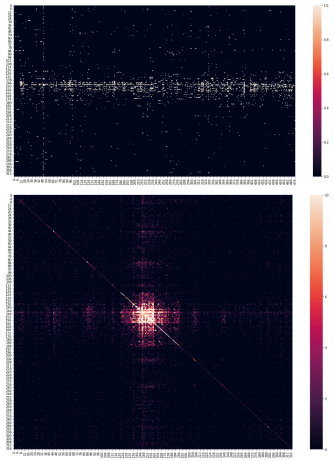
X_reduced



X_se



X_tsne



X_mds

1D Embeddings

This slide shows all the experiments made to find the best representation for the 1 dimension embedding.

The selected one is the highlighted one with a big centered cluster.

CSIC 5011

Arc Diagrams for Co-Appearance Data in Novels, Visualization of Dream of Red Mansion characters

Xavier Plantaz, Department of Computer Science, The Hong Kong University of Science and Technology, xpfplantaz@connect.ust.hk