Association Network of World War One Primary Sources

Motivation

Humanities is an area in which numerical data is used as historical evidence, but there is a reliance on statistics when examining events of the past. Analyzing historical texts is traditionally subject to the interpretation of the historians. Any one document can be subject to various interpretations. Another issue with relying on human interpretation is the limitations on the data size, a single person can only read and interpret so many sources. Being able to quantify texts will allow for a different perspective on vast amounts of written data.

Individually, primary documents from WWI are useful when investigating specific aspects of the war. However, what about the connections between these documents and the motivations behind what has been written. An understanding of where various texts from this period in relation to others will provide further insight to what we understand about the war.

Text Analysis

Various forms of text analysis exist so that text can be quantified and analyzed away from human bias. These include sentiment analysis, topic modeling and association networks. Sentiment analysis involves analyzing the 'feel' of various texts. This textual analysis technique involves looking at the text in the context of its individual words and thereby determining the sentiment through the frequency of the sentiment words within the text.[1] On the other hand topic modeling helps with the identification of patterns that are not initially prevalent when initially looking at the data. This technique, which is "similar to the clustering on numeric data", looks at finding natural groups for the data being analyzed.[2]

For this project however, association networks will be used to try and identify the connections between various texts from WWI.[3] The goal is to identify any significant relations between the different texts and what, if any, significant factors are the cause for any groupings.

Literature

After a search on the topic of association networks regarding historical texts, there does not seem to be any significant work done in this area.

Aims

1. Retrieve various texts from WW1. One source of data will be from the state library of NSW where diary entries will be retrieved. Various other data will also be retrieved but this is a starting point.
2. Learn how to use the tidy text package on R. This will be achieved through studying the textbook *Text mining with R*. (This can be accessed via https://www.tidytextmining.com/)
3. Clean the data that has been found and start trying to find word associations
4. Create networks to visualise connections between the texts
5. Conduct exploratory data analysis of networks, particularly the homology of the network, i.e. is there more connection between textual source from similar sources.

---

[1] Silge, J. and Robinson, D. (2017). *Text mining with R*. Sebastopol, CA: O'Reilly Media. https://www.tidytextmining.com/.

[2] Silge, J. and Robinson, D. (2017). *Text mining with R*. Sebastopol, CA: O'Reilly Media. https://www.tidytextmining.com/.

[3] Kolaczyk, E. and Csárdi, G. (2014). *Statistical analysis of network data with R*. New York, N.Y: Springer. 1-11.

Research proposal:
A computational approach to the conjugacy problem

Amelia Lee

1 Background

Very recent work by Boyle and Steinberg [1] has shown that certain equivalence problems for C^* algebras are decidable. This means that, in principle, an algorithm exists which can determine if these equivalences hold and that this algorithm will finish execution in a finite number of calculations.

This work relies on an algorithm developed by Grunewald to solve the "conjugacy problem" [2]. There is no implementation of Grunewald's algorithm. This project aims to describe such an algorithm in pseudo-code and then implement it. In particular, this project will focus on a special case of the conjagacy problem over the general linear group of invertible $n \times n$ matrices. Given $T, \hat{T} \in GL_n(\mathbb{Q})$, the algorithm will determine the existence of, and find $X \in GL_n(\mathbb{Z})$ such that

$$XTX^{-1} = \hat{T}. \tag{1}$$

2 Research questions

- Is it possible to transform Grunewald's constructive proof of (1) into a deterministic algorithm that, given T, and $\hat{T}$, will return a matrix X if it exists?
- What is the average complexity of the algorithm and how does it compare to the theoretical upper bounds to the number of computational steps it takes to solve this problem?
- How might the algorithm be optimized and is there an approximate upper bound on the size of n for which the algorithm is computationally feasible?

3 Methodology and Timeline

Week 1 - 2 Understand the scope of the problem and Grunewald's proposed algorithm. Develop a psuedo-code deterministic method for solving (1). Present this algorithm in the report.

Week 3 - 4 Research packages that handle arithmetic in $\mathbb{Q}, \mathbb{Z}$ and finite groups and develop the algorithm using a test driven, object oriented approach.

Week 5 - 6 Use manual and automated testing to confirm the validity of code and evaluate the algorithm's complexity. Present final findings in the report.

References

[1] Mike Boyle and Benjamin Steinberg. Decidability of flow equivalence and isomorphism problems for graph C*-algebras and quiver representations. *arXiv e-prints*, page arXiv:1812.04555, Dec 2018.

[2] Fritz J. Grunewald. Solution of the conjugacy problem in certain arithmetic groups. In S.I. Adian, W.W. Boone, and G. Higman, editors, *WORD PROBLEMS II*, volume 95 of *Studies in Logic and the Foundations of Mathematics*, pages 101 – 139. Elsevier, 1980.