Text Mining Toolkit for Digital Humanities

Olga Scrivner
olgas2@illinois.edu

Irina Trapido
trapido2@illinois.edu

Jay Lee
lee9@illinois.edu

ABSTRACT

Access to large digitized collections present a new opportunity to digital humanities researchers. At the same time, traditional methodology (e.g. close reading or keyword search) becomes very inefficient for literary data analysis. On the other hand, most available mining tools are built for different purposes and often require programming skills (even for software installation). Our project addresses these needs by optimizing a Shiny web application that allows researchers to explore visually and interactively their data collection. In addition, this application does not require any installation. While the core of this tool has been built, a lot of tuning and evaluation is needed, similar to META tool, in which the user is able to select parameters. Finally, the current tool uses various topic modeling algorithms, each with a different parameter setting, thus providing their evaluation and guidance to a non-technical user is necessary.

Keywords

text mining, visualization, digital humanties

1. INTRODUCTION

2. TEXT MINING TECHNIQUES FOR DIGITAL HUMANITIES

3. TOPIC MODELING TUNING

4. CLUSTER ANALYSIS

5. GUI OPTIMIZATION

6. DISCUSSION AND CONCLUSION

Keeping Tables and figures as illustration how to use them in this format

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Figure 1: A sample (.jpg format).

THEOREM 1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

Citation [1] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

7. REFERENCES

[1] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.

Table 2: Some Typical Commands

Command	A Number	Comments
`\alignauthor`	100	Author alignment
`\numberofauthors`	200	Author enumeration
`\table`	300	For tables
`\table*`	400	For wider tables

Figure 2: A sample figure that needs to span two columns of text.

Figure 3: A sample black and white graphic (.ps format) that has been resized with the `psfig` command.