Text Mining Seminar

GloVe: Global Vectors for Word Representation

Munich, 29.11.2017

Department of Statistics
Ludwig Maximilian University of Munich

Degree course: M.Sc. Statistics

Student:	**Advisor:**	**Subadvisor:**
Daniel Schalk (11470019)	Prof. Dr. Bernd Bischl	Janek Thomas

Contents

1 Introduction

A important thing when it comes to text mining is to create word embeddings. Those embeddings are used for further analyses or more general as features in statistical models. GloVe is a technique to create word embeddings out of a given corpus. A very interesting thing is that we want to learn from text without having labels. Hence, we have an unsupervised task.
- word vectors (image ...) - first terminology
In the following we want to discuss some important topics related to GloVe. After deriving the model with a more theoretical point of view in section 2. In section 3 we also want to take a look at how to evaluating given word vectors which is quite interesting since we handle an unsupervised task. After that we take a short look at the data and common sources for text corpora in section 4.

Then we know how to evaluate word vectors and have an idea about the data. With that knowledge we may ask how different methods to create word embeddings or different data sources influences the quality of the word vectors. This is discussed shortly in chapter 5. After that we take a short look how we can use GloVe embeddings for further text classification techniques followed from a small conclusion in section 6.

2 GloVe Model

Some mathematical stuff:

- How to derive the model

3 Evaluating Word Vectors

- Distances
- Curse of Dimensionality

4 About the Data

5 Real Word Vectors

TEXT HERE

6 Outlook and Conclusion

6.1 Outlook

6.2 Conclusion

List of Figures

List of Tables

Bibliography

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. R version 3.4.0.