**Samtla-Char-NER Report**

**Implementation of Character-based Named Entity Recognition into the Samtla System**

*Student: Matthew Ralph,*

*MSc Computer Science project report, Department of Computer Science and Information, Birkbeck College University of London*

[*mralph02@dcs.bbk.ac.uk*](mailto:mralph02@dcs.bbk.ac.uk)

*Supervisor: Dr Dell Zhang*

*With thanks to Dr Martyn Harris*

*This report is substantially the result of my own work, expressed in my own words, except where explicitly indicated in the text. I give my permission for it to be submitted to the JISC Plagiarism Detection Service. The report may be freely copied and distributed provided the source is explicitly acknowledged.*

Table of Contents

[0](#_Toc520908609)

[2. Acknowledgements 2](#_Toc520908610)

[3. List of figures and tables 2](#_Toc520908611)

[4. Abstract 2](#_Toc520908612)

[5. Introduction (including background) 2](#_Toc520908613)

[6. Overall Results (trailer) 2](#_Toc520908614)

[7. Software Architecture 2](#_Toc520908615)

[8. Testing 2](#_Toc520908616)

[9. Summary and Conclusions 2](#_Toc520908617)

[10. What’s My Work 2](#_Toc520908618)

[10.1. Summary of exactly what is my contribution in the repo 2](#_Toc520908619)

[11. References 2](#_Toc520908620)

[12. User Manual 2](#_Toc520908621)

[13. Appendix: Code 2](#_Toc520908622)

# Acknowledgements

# List of figures and tables

# Abstract

# Introduction (including background)

# Overall Results (trailer)

# Software Architecture

## The Pipeline of tasks and Invoke

## Named Entity Downloading

## Raw Hansard downloading

## Hansard processing

## Hansard interpolation

## Formation of Tensors

## Overview of files in project and what they do

# Implementation issues

## Wikipedia data cleanliness

## TWFY API suspect return values

## NLTK span\_tokenize bugs

## Toy dataset model – tensor sparsity

# Testing

## Unit testing

## Manual evaluation

## Model cross-validation

## Overall evaluation

# Summary and Conclusions

## Pre-processing is hard

## Labelling is hard

## Sentence tokenization is hard

# References

# User Manual

# Appendix: Code

# What’s My Work

## Summary of exactly what is my contribution in the repo