Natural Language Processing		**Name:**	____________
Prof. Dr.-Ing. Gerard de Melo		**ID:**	____________

Assignment 1: Dataset and Model Analysis

1 Data Acquisition

Based on your proposed course project ideas, try to obtain at least one corresponding dataset. This can be a pre-existing dataset found online, or one that you collect yourself, e.g., using a Web crawler.

Briefly describe this dataset in around 1–2 paragraphs, with additional tables or plots as appropriate. In particular:

a) What is the source of your dataset?

b) What does it contain? What language(s)? What is the structure? What kinds of labels or annotations are provided, if any?

c) Provide basic statistics, e.g., number of words/sentences/documents, average sentence/document length, vocabulary size with/without simple normalization, etc.

2 Simple Linguistic Model

Train any kind of linguistic model on your data (or on a subset of your data), e.g., a simple word embedding model, an n-gram language model, a character-based neural language model, or a more powerful deep neural network (trained for language modeling or any other NLP, as long as you obtain a trained model that provides word embeddings, an encoder capable of being analysed, or a model that makes predictions of any kind).

a) Provide at least 5 encouraging examples of what this model predicts or assumes, e.g., nearest neighbours of word embeddings (even if is a model trained for some other task), language model completions, sentiment polarity predictions.

b) Provide at least 5 negative or concerning examples of what this model predicts: For example: Does this model have any biases? Does it confuse different meanings of words? Does it fail to properly deal with certain kinds of inputs?

3 Describe Next Steps for Project

Describe in 1–2 paragraphs your project goals and particular next steps you plan to take towards attaining the goals for your project. Optionally also describe any steps you may have already taken towards your course project goal.