Statusreport zur Bachelorarbeit 'Semantische Beziehungen'

Ruben Müller

Hinweis: Referenzangaben zum Projektplan sind kursiv und in runden Klammern im Format *(x,xx)*

KW13-

Am 25.03.2015 um 11 Uhr fand das Kick-Off Meeting*(1,01)* im Raum 321 der Hochschule der Medien mit Prof. Dr. Johannes Maucher, M.Sc. Andreas Stiegler (via Skype) und Ruben Müller statt. Hier wurde der Umfang und Inhalt der Bachelorarbeit besprochen.
Am 27.03.2015 habe ich den Projektplan(*2,01*), anhand einer Vorlage von `www.meinevorlagen.com`, angefangen und am 28.03.2015 fertig gestellt.
Am 30. und 31. März 2015 habe ich den Abstract*(2,02)* geschrieben.
Am 02. April 2015 habe ich mit der Einarbeitung in Word2Vec*(3,01)* und dem Wikipedia-Korpus*(3,02)* begonnen. Nachdem ich in PyCharm dann eine Out-of-Memory-Exception hatte, habe ich die neueste Version des Programms heruntergeladen (Version 4.0.5), die auch eine 64-Bit Version bereitstellt. Hier konnte ich dann die Größe des Speichers ausreichend vergrößern.
Mit dem Perl-Skript von `http://mattmahoney.net/dc/textdata.html`, unter Appendix A, kann der Wikipedia-Dump preprocessed werden. Allerdings ist dann der komplette Text in einer Zeile und enthält keine Satzzeichen mehr, die aber für Word2Vec gebraucht werden, da der Input hier Sätze sind. Hier ist das Perl-Skript anzupassen, was nicht viel Aufwand war. Da in der jetzt wegfallenden Regel auch sämtliche Wiki-Formatierungszeichen entfernt werden, müssen diese jetzt wieder in das Skript eingebaut werden.

KW1x-