Machine Learning für einen textbasierten Event-Klassifikator

Studiengang: BSc in Informatik | Vertiefung: Computer Perception and Virtual Reality
Betreuer: Prof. Dr. Jürgen Eckerle
Experte: Dr. Frederico Flückiger (Eidgenössisches Finanzdepartement EFD)

Maschinelle Textverarbeitung (Natural Language Processing) machte in den vergangenen Jahren enorme Fortschritte und ist aus dem Alltag kaum mehr wegzudenken. Das automatische klassifizieren von Dokumenten findet zum Beispiel Anwendung bei Spam-Filtern, dem Erkennen von Sprachen oder Sentimentanalyse. Im Rahmen der Bachelorarbeit wurde diese Technik nun auch auf Facebook-Events angewandt.

Aufgabenstellung

In Projekt 2 von Kevin Suter wurde eine Website (Happens-Now) entwickelt, welche dem Benutzer basierend auf diversen Kriterien Veranstaltungen von unterschiedlichen Quellen vorschlägt. Diese Events besitzen vor dem Verarbeiten noch keine Kategorie – was jedoch für eine korrekte Empfehlung notwendig wäre.
Ziel dieser Bachelorarbeit war es, mit verschiedenen Mitteln aus dem Machine- und Deep-Learning Bereich ein Verfahren zu entwickeln, welches neue Events entsprechend verarbeitet und in die vordefinierten Kategorien einteilt.

Umsetzung

Da ein Supvervised-Learning Ansatz für die Klassifizierung gewählt wurde, mussten zu Beginn alle Testdaten entsprechend von Hand kategorisiert werden. Dies, damit der Algorithmus später anhand korrekt klassifizierter Daten lernen kann, welche Muster am wahrscheinlichsten der entsprechenden Kategorie zugeordnet werden können. Sind die Testdaten kategorisiert, müssen sie entsprechend aufbereitet und in eine für den Computer verständliche Sprache übersetzt werden.
Bei Textklassifikation werden die Dokumente jeweils vektorisiert um sie anschliessend miteinander zu vergleichen. Je näher die Vektoren beieinander sind, desto ähnlicher sind sich die Dokumente.

Für die anschliessende Klassifizierung wurden diverse Algorithmen aus dem Bereich des Deep- und Machine Learnings eingesetzt und deren Resultate untereinander verglichen. Konkret verwendet wurden Neuronale Netze, Naive Bayes und Decision Trees.

Ergebnisse

Über alle angewendeten Verfahren hinweg wurden 65% der Events korrekt klassifiziert. Dies wäre für einen produktiven Einsatz zu wenig und somit ungeeignet. Leider war das aber aufgrund der Menge und der Qualität der Trainingsdaten zu erwarten.
Bis zum Ende der Arbeit standen nur ein paar hundert

```
Results
======

Correctly Classified Instances         447               78.0105 %
Incorrectly Classified Instances       126               21.9895 %
Kappa statistic                          0.6523
Mean absolute error                      0.0737
Root mean squared error                  0.2672
Relative absolute error                 34.2743 %
Root relative squared error             81.6479 %
Total Number of Instances              573
```

Abbildung 1: Resultat mit Naive Bayes

Trainingsdaten über alle sechs Kategorien zur Verfügung. Für eine exakte Dokumentenklassifikation werden normalerweise pro Kategorie tausende von Testdaten verwendet.
Da es sich um Facebook-Veranstaltungen handelt ist auch die Qualität der Daten entsprechend schlecht. Von kurzen, nichts aussagenden Beschreibungen bis hin zu multilingualen Texten ist alles dabei.

Thomas Buchegger