Kurzfassung

Das *World Wide Web* beinhaltet eine nahezu unbegrenzte Menge an Informationen, jedoch liegt nur ein Bruchteil dieser Informationen in strukturierter Form vor. Der Großteil des Wissens ist in unstrukturierten Texten, wie bspw. Nachrichten- oder Wikipedia-Artikeln verborgen. Während ein Teil der strukturierten Informationen in Wissensdatenbanken, sogenannten *Knowledge Bases*, wie DBpedia oder in Open-Source-Projekten, wie Web Data Commons, aufbereitet und gesammelt werden, sind die meisten unstrukturierten Datenquellen noch weitgehend unerschlossen. Im Rahmen dieser Arbeit werden öffentlich zugängliche unstrukturierte Daten in Form von Nachrichtenartikeln aus dem Common-Crawl-Projekt unter Hinzunahme von Informationen aus DBpedia zum Generieren neuen Wissens genutzt. Dabei werden Methoden und Modelle aus dem *Information Retrieval* und *Natural Language Processing* dazu verwendet, die Möglichkeit und das Potential der Erweiterung von Knowlege Bases durch die enorme Menge frei zugänglicher Daten zu evaluieren. Zu diesen gehören das Extrahieren von Informationen per *Named Entity Recognition* und *Entity Linking*, die Reduzierung von Features mit dem Topic-Modell *Latent Dirichlet Allocation* und die Darstellung von Texten als numerische Vektoren, auf denen anschließend *Machine-Lerning*-Modelle zur Klassifikation trainiert werden. Der Kern der Arbeit ist die Klassifikation von Texten aus dem Common Crawl-Datensatz, die Firmen beinhalten, und daraus resultierend in einem Multi-Label Klassifikationsproblem eine oder mehrere Branchen zugeordnet bekommen. Dazu wird eine Datenaufbereitungs- und Modellbildungs-Pipeline aufgebaut, die zeigt, dass eine solche Klassifikation möglich ist. Als bestes der verwendeten Modelle schneidet ein lineares SVM-Modell ab, das auf Bag-of-Word-Features trainiert wird. Es erreicht eine *Exact Match Ratio* von 91,79%. Für die Multi-Label-, Micro- und Macro-Metriken erreicht es jeweils F1-Werte, die größer sind als 0,9. Die Pipeline ist so entworfen, dass sie per Cloud-Computing für große Datenmengen skaliert werden kann. Um das Potential der Pipeline für den Common-Crawl-Datensatz zu bewerten, wird dieser auf Texte mit enthaltenen Entitäten hin untersucht, indem ein Teil der Pipeline auf Apache Spark übertragen wird.

Abstract

The *World Wide Web* contains an almost unlimited amount of information, but only a fraction of this information is available in a structured form. Most of the knowledge is hidden in unstructured texts, such as news or Wikipedia articles. While a part of the structured information is prepared and collected in *knowledge bases*, like DBpedia or in open source projects, such as Web Data Commons, most unstructured data sources are still widely unexploited. In the context of this work, publicly accessible unstructured data in the form of news articles from the Common Crawl project are used with the addition of information from DBpedia to generate new knowledge. Methods and models from the *Information Retrieval* and *Natural Language Processing* are used to evaluate the possibility and potential of extending knowledge bases through the enormous amount of freely accessible data. These include extracting information via *Named Entity Recognition* and *Entity Linking*, the reduction of features with the topic model *Latent Dirichlet Allocation* and the representation of texts as numerical vectors on which *Machine-Lerning* models for classification are trained. The core of the work is the classification of texts from the Common Crawl dataset, which contain companies, and, as a result in a multi-label classification problem, one or more industries are assigned to. For this purpose, a data preparation and modelling pipeline is set up, which shows that such a classification is possible. The best of the models used is a linear SVM model that is trained on bag-of-word features. It achieves an *Exact Match Ratio* of 91,79%. For the multi-label, micro, and macro metrics, it reaches F1-Scores greater than 0,9. The pipeline is designed to be scaled for large amounts of data using cloud computing. To evaluate the potential of the pipeline for the Common Crawl data set, it is examined for texts with contained entities by transferring part of the pipeline to Apache Spark.