Technischer Bericht
Tech4Germany

16. Oktober 2019

Zusammenfassung

This is the paper's abstract ...

1 Exploration

1.1 Ähnlichkeit von Berufen

Die Bundesagentur für Arbeit stellt in ihrem Portal BERUFENET Definitionen von Berufen zur Verfügung. Um eine Ähnlichkeit zwischen diesen zu ermitteln, muss die textuelle Form in einen Merkmals-Vektor überführt werden, der genauere Informationen zu den einzelnen Berufen liefert. Zwar könnte man die Levenshtein-Distanz[1] verwenden, allerdings verwendet diese keinerlei Informationen über die eigentlichen Berufsinhalte.

Anhand der Berufs-Bezeichnungen können durch vortrainierte Modelle wie bspw. Word2Vec[2] oder GloVe[3] bereits Merkmals-Vektoren erzeugt werden. Diese Modelle erstellen Merkmals-Vektoren anhand von Attributen, die anhand von verschiedenen Korpussen wie bspw. den Wikipedia-Daten gelernt wurden. Auch wir haben diese Verfahren verwendet um Merkmale für Berufe zu generieren, allerdings verfügen diese über keine kontextuellen Informationen zu einem Beruf, d.h. die Tätigkeiten sowie verschiedene Berufsbeschreibungen bleiben unberücksichtigt. Ein anderer Nachteil dieser Verfahrer ist, dass ein Merkmals-Vektor stets nur für ein einzelnes Wort generiert werden kann. Auch wenn die Vektoren über verschiedene Wörter gemittelt werden können, gehen sehr viele Informationen verloren.

Um zusätzliche Informationen zu generieren, haben wir deshalb die Tätigkeitsbeschreibungen analysiert. Dabei lässt sich mit einem Bag-of-Words Ansatz ein Korpus generieren, der alle relevanten Wörter durch Verwendung eines Stemmers in ihrer Rohform beinhaltet. Zu Beachten ist dabei, dass häufige Wörter rausgefiltert werden müssen, um die Relevanz der einzelnen Merkmale nicht zu gefährden. So müssen Bindewörter und Pronomen extrahiert werden, wofür sich bspw. die Python-Bibliothek NLTK[4] sehr gut eignet. Nach der Erstellung des Korpuses können nun die Merkmals-Vektoren der einzelnen Berufe berechnet werden. Dazu kann ein TfidfVectorizer bspw. von Scikit Learn[5] verwendet werden, der die relative Häufigkeit der gestemmten Wörter innerhalb einer Tätigkeitsbeschreibung betrachtet.

So können Merkmals-Vektoren für Berufts-Tätigkeiten bestimmt werden, die sehr hochdimensional sind. Da auch die meisten Tätigkeitsbeschreibungen nur einen Bruchteil des Korpuses abdecken, eignet sich die euklidische Distanz ($d_e(p,q) = \sqrt{\sum_{i=1}^{n}(p-q)^2}$) nicht, um Ähnlichkeiten (bzw. Distanzen) zwischen den Berufen zu berechnen. Stattdessen eignet sich die Kosinus-Distanz ($d_c(p,q) = 1 - cos(\theta) = \frac{p \cdot q}{||p|| ||q||}$), da dieser statt dem Pfad zwischen p und q den räumlichen Winkel zwischen diesen betrachtet. Um diese

[1] vgl. http://www.levenshtein.de
[2] vgl. https://code.google.com/archive/p/word2vec/
[3] vgl. https://nlp.stanford.edu/projects/glove/
[4] vgl. https://www.nltk.org
[5] vgl.

Merkmals-Repräsentation greifbar zu machen, eignen sich verschiedene Algorithmen, die die Dimensionalität des Merkmals-Raums reduzieren. In unserer beispielhaften Implementierung haben wir T-distributed Stochastic Neighbor Embedding[6] verwendet, da dieses Verfahren eine ansprechende Visualisierung als bspw. eine PCA oder eine LDA liefern.

Damit die Distanzen nicht zur Laufzeit bestimmt werden müssen, exportieren wir eine Distanzmatrix, die die paarweise Distanz zweier Embeddings speichert. So kann diese Matrix beim Starten der Anwendung geladen werden und so direkt auf 'dist_matrix[i][:]' zugegriffen werden, um die paarweisen Distanzen des Embeddings i zu den anderen Embeddings zu bestimmen.

[6] vgl. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html