Themenvorschlag für eine Masterarbeit in der Web Science Group
am Lehrstuhl Informationssysteme (Prof Naumann)

Sehen in 10k Dimensionen

Dimensionsreduktion 10 Jahre nach tSNE

Dieser Themenvorschlag entstand in einem attraktiven Forschungsumfeld am HPI (http://hpi.de/naumann/) und im Rahmen eines Drittmittelprojekts zur Exploration großer Datensammlungen.

Problem

- Wir leben in 3D
- Bildschirme sind 2D
- Daten sind 10000D
- Dimensionsreduktion von Daten zur Visualisierung ist nicht trivial
- Kontext soll auch in 2D erhalten bleiben

Lösungsansatz

- tSNE aufgreifen und in Fehlerfunktion in auto-encoder übertragen
- Vergleich verschiedener Algorithmen zur Vorreduktion für tSNE
- Was mit Deep Learning
- Ohne IoT
- Keine Blockchain

Hintergrund:

Unfortunately, our world has only three dimensions, computer screens only two even. To visualise high-dimensional data, we have to reduce it to see it. This process comes with the cost of losing information, as all the data is compressed into a much smaller space. In the case of embeddings (i.e. word2vec), this is very helpful. Mikolov et al. use simple neural networks to reduce a large vocabulary of words down to 50-300 dimensions and found that this lower dimensional space has useful properties. For example, the Euclidean distance between data points can be interpreted as their semantic similarity. To visualise the vocabulary, data scientists and researchers often use tSNE (Marten et al, 2009). While other dimensionality reduction approaches (e.g. PCA) may not reflect relatedness properties well, tSNE uses a cost function, which keeps points that are clustered in the high dimensional space together even in the low dimensional space.

In this master's thesis, we want to compare the influence algorithms for the first stage in tSNE. Furthermore, we would like to transfer the tSNE cost function to some kind of auto-encoder, so that we are able to reduce all data in one go with a single (simple) model. This thesis can be very low-level and math heavy (proper understanding of mathematical model behind neural nets & develop cost function) or engineering heavy (comparing lots of existing methods to find good constellations).

#DataVisualisation #DeepLearning

Kontakt:

Ich freue mich auch über Anregungen und komplett eigene Ideen, die mit unstrukturierten Daten aus Text und Web zu tun haben.
Buzzwords: Ingestion, Extraction, Clustering, Visualisation.

Klingt interessant? Dann schreib' doch einfach eine E-Mail: tim.repke@hpi.de
Bei Fragen stehen wir vorab gern zur Verfügung.