Author: André Fernandes dos Santos, andrefs@andrefs.com, up201700584

Project: Computing semantic relatedness from knowledge graph embeddings

Datasets

The Wikipedia Oriented Relatedness Dataset (WORD) is a concept relatedness dataset. It contains almost 20k pairs of concepts from Wikipedia scored by humans for their level of relatedness. It was compiled by the IBM Debater® project.

DBpedia is a knowledge graph containing structured content extracted from Wikipedia. It is available in RDF format, which means that all information in DBpedia is represented as triples Subject-Predicate-Object. DBpedia contains currently 3 billion triples and is maintained by research groups from the University of Mannheim and Leipzig University.

Goal

Develop machine learning algorithms to learn how to calculate relatedness values between DBpedia nodes.

Methods

One approach for automatically calculating semantic relatedness between concepts is based on knowledge graphs. This is typically done either by calculating metrics of the paths between two entities or by representing the entities (nodes) as vectors of numerical features, and comparing those. The vectorial representation can be obtained using algorithms such as *node2vec, rdf2vec* and similar tools. These are usually computationally expensive, but It is also possible to download pre-calculated node embeddings for DBpedia or other large semantic graphs.

In this project we will experiment with different algorithms to create a regression model capable of calculating relatedness values for pairs of DBpedia concepts. This will be achieved through the following steps:

1. Calculate or download pre-calculated node embeddings for DBpedia.
2. Find correspondences between DBpedia nodes and entries in the WORD dataset.
3. Train different models using pairs of concepts from DBpedia (represented as numeric vectors) and the corresponding relatedness reported by WORD.
4. Evaluate the results for pairs of concepts belonging to the Test portion of the WORD dataset.

Research questions

1. What performance can be achieved using machine learning for computing semantic relatedness for DBpedia concepts?
2. Which machine learning algorithm achieves the best results for this task?
3. Do the models obtained perform as well for named entities as they do for concepts?