UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

MSc Seminar

Friday January 31, 2020 at 3:00PM in Reynolds, Room 2224

Document Embeddings Applied to Name Matching

Emanuel Felipe Principe de Carvalho

Advisor: Dr. Luiza Antonie
Advisory Committee: Dr. Andrew Hamilton-Wright
Advisory Committee: Dr. Dan Tulpan

ABSTRACT:

First described 60 years ago, Record Linkage has been studied by different fields such as linguistics, statistics, epidemiology, historians, and computer scientists. It is the task of matching multiple representations of the same entities along different data sources. Challenges arise from the fact that no single unique identifier is present most of the times, due to typographical errors or absence of such information. The matching is typically done for each attribute of the entry-pair evaluated resulting in a score associated to that pair.

Name Matching is particularly challenging as there is no simple way of generating mathematical representation of words. Most of the classical techniques rely on calculating a similarity metric based on number of edits necessary to pair both words (Levenshtein, Hamming, etc), Token based (Jaccard index, bag index, etc), or Linguistics observation of real data (Jaro-Winkler).

As an alternative, we aim to develop an alternative methodology to generate vectorial representation of names titled Name2Vec. Our project uses Word and Document Embeddings provided by FastText on a 250.000 names database. These embedding models act as dictionaries, translating each name into its vector, trivializing the matching process to the calculation of the cosine similarity of those vectors.