A Survey on Domain-specific Embedding Models

CS5352 Course Project, Spring 2021

[This project accepts two students.]

**Contact**

Chenxu Niu (chenxu.niu@ttu.edu)

**Description**

Embedded vectors learn the distributional semantics of words and are used in different applications such as Named Entity Recognition (NER), question answering, document classification, information retrieval, and other semantic search applications. Typical examples can be seen from GloVe[1] and FastText[2] and Word2vec[3]. But the majority of work has focused on generic terms and general domain word embeddings. Domain-specific terms are still very challenging since there are few statistical clues in the underlying corpora for these items.

The goal of this research is providing a survey of different domain-specific embedding models and implement one model to show its effectiveness.

1. Literature study of the related papers about domain-specific embedding models.
2. Implement a demo to illustrate one model.

**Requirements**

* Proficiency with Python programming language
* Knowledge of Linux

**Students will have a chance to learn about:**

* Machine learning
* Natural Language Processing
* Graph algorithms

**References**

1. [GloVe: global vectors for word representation](https://www.aclweb.org/anthology/D14-1162.pdf).
2. [Enriching word vectors with subword information.](https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00051?source=post_page---------------------------&)
3. [Efficient estimation of word representations in vector space.](https://arxiv.org/pdf/1301.3781.pdf)
4. AGROVEC: <http://www.fao.org/agrovoc/publications/agrovoc-linked-dataset>