SUGGESTING ARXIV PAPERS BY SEMANTIC RELATION

CS224U PROJECT - LITERATURE REVIEW

Junshen Kevin Chen
Stanford University
jkc1@stanford.edu

Overview

For this project, I propose to design a system that evaluates the level of relation between two academic papers, and suggest possible related work given a text query.

Motivation Three main factors contribute to motivating this project. First, the current state of academic research tools such as Google Scholar functioning similar to a traditional search engine, which primarily leverages keyword matching and ranking by popularity (citation counts), while having being limitedly influenced by content semantics. Second, large data dumps of academic papers such as arXiv is available to extract relationships between academic work in the form of a citation graph. Finally, the research in transfomer models such as BERT [todo] provides powerful ways to encode semantics, and trained models are readily available to be fine-tuned for this specific task.

Problem Tentatively, I formulate this project into several following problems, with each latter depending on the former:

1. Given two articles, produce a probability of citation, or "how likely does one cite another"
2. Given two articles, predict a "citation distance", or how many edges it needs to traverse from one to another
3. Given a query text, possibly the abstract of the working paper, recommend related articles

In this document, I select related works that provide context to the proposed project, in hopes that the outcome of this project will be able to perform exactly this selection task.

1 arXiv Dataset and Citation Graph

References