DeepImpact: Impact Factor Predictor from Abstract Text

By: Xander Krohannon, Gayathri Panangipalli, and Sharmilla Selvaraj

**Introduction:**

Contributing to the library of human knowledge is a proud and noble tradition among the world’s scientists; as well as an important milestone for all aspiring researchers. In the modern era, this aspiration maintains all the classic obstacles: sufficient rigor and novelty of the work, mentor sponsorship and approval, acceptance of a journal to publish the work, and incorporating the feedback and criticism from peer evaluations, with the additional wrinkle of selecting the appropriate journal to submit to. Per the 2018 STM Report [1] there are over 33,100 peer-reviewed journals in print, each with their own specific research area and specifications. One heavily weighted metric used by journals is the impact factor, measured by the number of times an article is published within a set duration of its publish date, typically 6 months. As a result, a tool to predict a proposed article’s impact factor is desired, as a means to aid in researchers in selecting journals to submit their article to. Here we present, DeepImpact a deep learning model that takes an article’s abstract, as well as the primary author’s institutional association as features to predict the likely impact factor of that particular work.

**Problem Statement:**

Publishing scholarly works is a rite of passage for graduate students, and often a requirement for their advising faculty. In the modern world, there are a plethora of peer-reviewed journals selecting which 3 million articles will be published this year. Selecting which journal to submit a body of work to, in order to maximize the likelihood of it being published, is a daunting task with few guidelines beyond the intuition of experts, that only comes with experience. DeepImpact aims to serve as a tool to predict an article’s impact factor, thus aiding the aspiring author in determining which journals they should submit their work to.

**Data Source:**

PubMed is a repository for over 30 million biomedical articles from a wide range of journals[2]. To this end, the meta data of a random 5 percent of the articles contained in PubMed were downloaded in the XML format. This data was divided into training data and testing data, X percent and 100 – X percent respectively. An additional 1 percent of the articles were downloaded to serve as an independent validation set.

**Methodology:**

Data preprocessing was performed using an ad hoc Python script (version 3.7.6) utilizing the datetime, networkx, and xmltodict libraries. The XML files downloaded from PubMed were parsed into Python dictionaries using the parse command from the xmltodict library, resulting in each tag becoming a dictionary entry, with enclosed tags as keys, thus retaining the hierarchical structure of the data. The article’s PubMed ID, abstract, and the primary author’s school and department were extracted in the raw text format. The article’s publish date was extracted and stored as a datetime date object. Additionally, a list was composed of the PubMed ID’s of each article referenced by a particular article. This was combined with the previously extracted features and stored as a tuple in a list of all article information.

In order to get the true impact factor of each article the directed network module from networkx was used. Each article was created as a unique node within the network. Then directed edges were created originating at the citing article and pointed towards the article being cited, if the citing article was published within 6 months (180 days) of the article being cited. After all edges had been added, the number of edges facing towards a particular node, the number of other articles that cited that article within 6 months of its publication, was treated as the true impact factor of that article.

**References:**

[1] Johnson et al.(Oct 2018). The STM Report Fifth Ed.STM-Association. https://www.stm-assoc.org/2018\_10\_04\_STM\_Report\_2018.pdf

[2] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;46(D1):D8‐D13. doi:10.1093/nar/gkx1095