Contribution ID: **30** Type: **"Hands-on" presentation**

Chaotic_Neural: Improving Literature Surveys in Astronomy with Machine Learning

Friday 28 May 2021 15:45 (20 minutes)

Literature surveys in astronomy are greatly facilitated by both open-access preprint servers (ArXiv) and online tools like the Astrophysics Data System (ADS). However, the astrophysics literature often uses specialised jargon, sometimes using multiple identifiers for the same phenomena. For example, the terms SFR-M_* correlation, Star Forming Sequence and Star Formation Main Sequence, all mean the same thing in the galaxy context, not to be confused with just Main Sequence which pertains to stellar evolution. This can often be challenging for young researchers to parse, and can cause even established astrophysicists to sometimes miss relevant references. Other issues include in-group bias towards referencing and citing literature in a paper, or papers sometimes getting overlooked due to the large volume of new literature.

To help circumvent these issues and provide agnostic, context-aware searches for relevant literature, we present chaotic_neural, a public python package that trains a Doc2Vec model on abstracts from the ArXiv to enable finding relevant literature. The model works by using a neural network to transform abstracts into a high-dimensional vector space. An input vector is generated using an abstract or a set of keywords. Relevant literature can then be searched for by looking for papers that lie in the vicinity of the input vector. Since the computation happens in a vector space, the search can be further refined with linear algebra using keywords. This introduces the possibility of adding and subtracting keywords and/or papers from other keywords and/or papers. The model also provides utility beyond literature surveys, creating a discovery space for future analysis and hypothesis testing. The currently available model (available at https://github.com/kartheikiyer/chaotic_neural) is trained on a large galaxies dataset (https://arxiv.org/list/astro-ph.GA), but can easily be adapted to other fields and datasets.

Author: IYER, Kartheik (Dunlap Institute for Astronomy and Astrophysics, University of Toronto)

Presenter: IYER, Kartheik (Dunlap Institute for Astronomy and Astrophysics, University of Toronto)

Session Classification: Afternoon 2

Track Classification: Literature