Dataset

NASA astronomical dataset in XML form

- Source XML Data Repository on the Computer Science & Engineering of the University of Washington
- 2435 reports with titles, authors, paragraphs, definitions, units, etc. marked

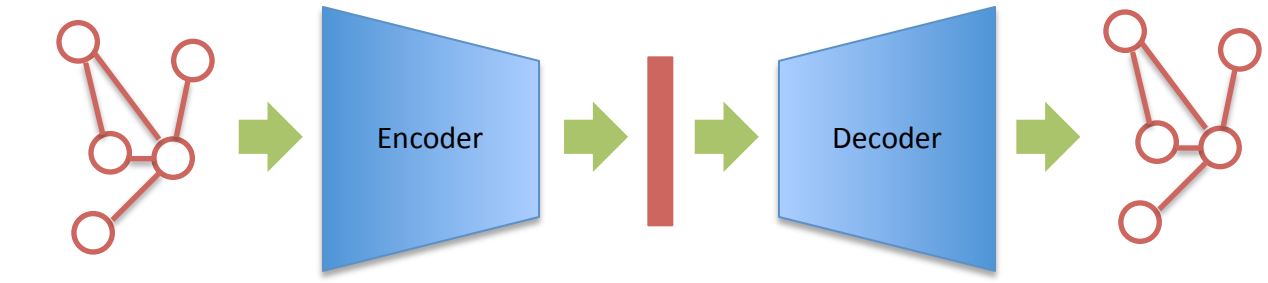
File documentation

Folder 'preprocess'

- 'split_xml.py' — Creates separate XML files. Outputs a total of 2435 XML reports from 'xml/nasa.xml' into the folder xml/nasa'.
- 'edge_parser.py' — Parses all the XML files in the folder xml/nasa'. Outputs edge integer vectors for each file in the folder 'output/edges'.
- 'vector_parser.py' — Parses all the XML files in the folder xml/nasa'. Outputs text vectors representative of node features for each file in the folder 'output/txt_vectors'.
- 'vectorize.py' — Parses all the text vectors in the folder 'output/txt_vectors'. Outputs text embedding vectors for each vector in the folder 'output/vectors'. Uses spacy to produce the embeddings.

Model architecture

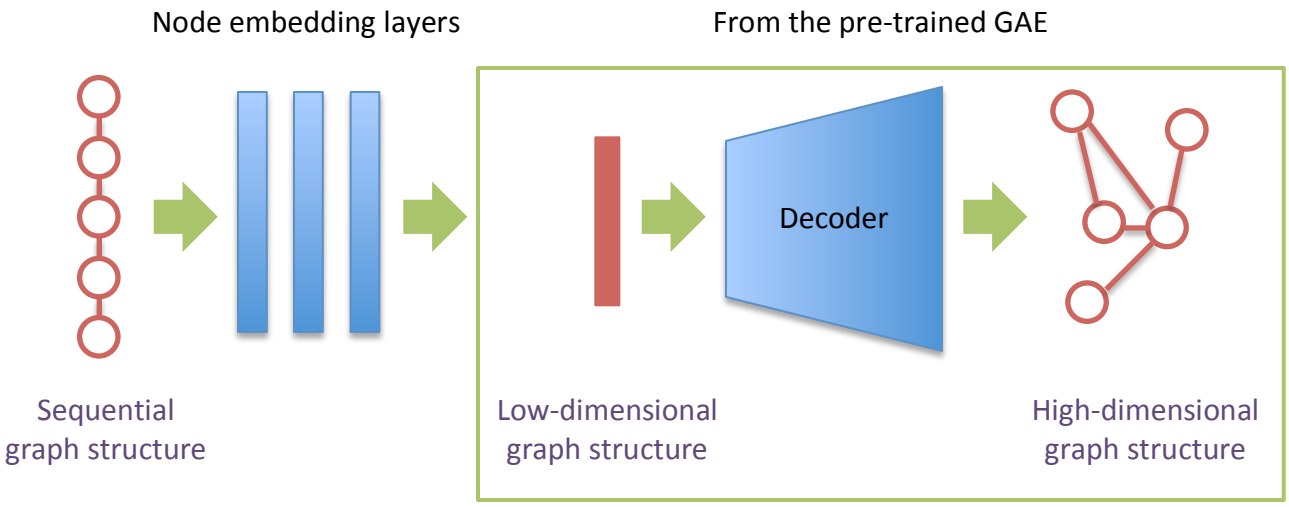
Graph Auto-Encoder (GAE)



Graph auto-encoder

- Encoder: high dimensional graph structure → low dimensional graph structure
- Decoder: low dimensional graph structure → high dimensional graph structure
- Output: reconstructed adjacency matrix
- Model: GAE, ARGA

Graph Generator System



Graph generator system

- Sequence → low dimensional graph structure
- Output: node embeddings
- Model: BigGraph, graph attention network (GAT, Attention Walks), bidirectional LST, CNN + CRF