Dataset of Pathway2Text

This is the dataset of our NAACL 2022 paper:

Pathway2Text: Dataset and Method for Biomedical Pathway Description Generation
Junwei Yang, Zequn Liu, Ming Zhang* and Sheng Wang*

Please cite this paper when you use our dataset.

This dataset contains 2,367 pairs of biomedical pathways and textual descriptions. It can be used for automatic pathway description generation. In our paper, we showed it is also appropriate for Text2Graph and BioNER.

To construct this dataset, we collected biomedical pathways and their associated textual descriptions from three biomedical databases: Reactome [1], KEGG [2], and Pathbank [3]. We aligned nodes in graphs to entities in UniProt [4] and ChEBI [5] for retrieving missing node descriptions.

This dataset contains 2 JSON files.
The `mapping_database_to_pathway2text.json` file is organized as: `{name of database: list of graph identifiers}` , indicating the mapping between pathway graphs and data sources.
The `pathway2text.json` file is organized as:

```
{
    Graph identifier: {
        "Name": ,
        "Graph description": ,
        "Node_dict": {
            Node identifier: {
                "class": ,
                "label": ,
                "description": .
            },
            ...
        },
        "Arc_list": [
            {
               "arc_sourse": Node identifier,
               "arc_target": Node identifier,
               "arc_class": .
            },
            ...
        ]
    },
    ...
}.
```

The node classes include Submap, Macromolecule, Process, Complex, Multimer, Simple Chemical and Others. And the Others is the union of several classes occurring only in a single database (e.g., Unspecified Entity, Association in Reactome and Transport in Pathbank). Nodes in this class account for 7% over the whole dataset. The edge classes include Catalysis, Consumption, Stimulation, Inhibition, Production, Logic Arc and Belong To. If you want to use this dataset for SingleGraph2Text task, we suggest you to exclude Submap nodes as what we did in experiments.

- [1] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. 2022. The reactome pathway knowledgebase 2022. Nucleic acids research, 50(D1):D687–D692.
- [2] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 45(D1):D353–D361.
- [3] Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z, Patron J, Lipton D, Cao X, Oler E, Li K, Paccoud M, Hong C, Guo AC, Chan C, Wei W, and Ramirez-Gaona M. 2020. Pathbank: a comprehensive pathway database for model organisms. In Nucleic Acids Res.
- [4] The UniProt Consortium. 2020. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1):D480–D489.
- [5] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2015. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res, 44(D1):D1214–9.