データセットの候補

研究トレンド予測を行っている論文から使えそうなデータの候補を得る。Predicting research trends with semantic and neural networks with an application in quantum physicsの論文で行われている方法は、二つのコンセプトが同じ論文に共起されていたら、一つの論文であっても、関係なくリンクを貼っている。しかし、重みについても考慮したい。そのためのデータセットを取得したい。

基本的に、**新たな研究トピックに利用できるベンチマークデータセットが公開されていない**ことが難しい点である。

概要

新たな研究トピックに利用できるベンチマークデータセットが公開されていないので、基本的に新しい問題設定で、グラフを作成し直すには、新しくデータを作り直す必要がある。データの提供元として、次のサイトが候補になる。

- arXiv(https://arxiv.org/)
- Web of Science(https://clarivate.com/ja/solutions/web-of-science/)

この二つから、作成する方法を調べることが必要かもしれない。

詳細

Embedding technique and network analysis of scientific innovations emergence in an arXiv-based concept network

データの発行元：掲載・データの作り方：あり・データの配布：なし
arXivで作ったデータを使って、解析している。内容はよくわからないが、データの作成をしなければならない時、参考にできる

Forecasting Emerging Trends from Scientific Literature

https://aclanthology.org/L16-1066/
LREC議事録をデータとしているが、機械学習の文脈では、不適切。

Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval

https://www.sciencedirect.com/science/article/pii/S0306457318304199?via%3Dihub
データの発行元：掲載・データの作り方：なし・データの配布：なし
WoS(Web of Science)の使っている。参考にはできないかも。

S2ORC: The Semantic Scholar Open Research Corpus

https://arxiv.org/abs/1911.02782
コーパス。

S2ORC、1を紹介します。これは、多くの学問分野にまたがる81.1Mの英語の学術論文の大規模なコーパスです。 コーパスは、豊富なメタデータ、論文の要約、解決された書誌参照、および810万のオープンアクセス論文の構造化された全文で構成されています。 全文には、引用、図、表の自動的に検出されたインラインの言及が注釈として付けられ、それぞれが対応する紙のオブジェクトにリンクされています。 S2ORCでは、何百もの学術出版社からの論文とデジタルアーカイブを統合されたソースに集約し、これまでに機械可読な学術テキストの最大の公的に利用可能なコレクションを作成します。 このリソースが、学術テキストを介したテキストマイニングのためのツールとタスクの研究開発を促進することを願っています。

Topic diffusion analysis of a weighted citation network in biomedical literature

https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23960
データの発行元：掲載・データの作り方：なし・データの配布：なし
PubMed and PubMed Centralからデータを取得。化合物にのみ適応可能？ バイアスチックなことをやっている。
発行元によって引用数は変化してしまうことから、人気度の新しい考え方を提案。

Predicting the Impact of Scientific Concepts UsingFull-Text Features

https://deepblue.lib.umich.edu/bitstream/handle/2027.42/134425/asi23612.pdf?sequence=1
データの発行元：掲載・データの作り方：なし・データの配布：なし
論文の重要度を測る際に、論文のメタデータと論文の中身を両方用いると、片方を用いる場合よりも、うまくいくことを示した論文。

Emerging research topics detection with multiple machine learning models

https://www.sciencedirect.com/science/article/abs/pii/S0048733314000298?via%3Dihub
データの発行元：掲載・データの作り方：なし・データの配布：なし
データの作り方の部分は、少し乗っていたので、参考になるかも。

A bibliometric model for identifying emerging research topics

https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23930
データの発行元：掲載・データの作り方：なし・データの配布：なし

- A new methodology for constructing a publication-level classification system of science
  http://www.ludowaltman.nl/classification_system/
- The inconsistency of the h-index
  https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21678

よくわからない

Combining deep neural network and bibliometric indicator for emerging research topic prediction

sciencedirect.com/science/article/abs/pii/S0306457321001072#bib0049 データの発行元：掲載・データの作り方：論文引用(https://www.nature.com/articles/s41587-019-0275-z には記載なし）・データの配布：なし
次に来る論文のトピックを、将来のトピックの論文数などから判断し、予測する。

Identifying emerging topics in science and technology

https://www.sciencedirect.com/science/article/pii/S0048733314000298?via%3Dihub#bib0030
データの発行元：掲載・データの作り方：なし・データの配布：なし