- k-Clique Percolation Clustering of Co-reference Network

Introduction

This appendix explains the research procedures used to produce tables and figures in *Productions of Culture: Knowledge Survival in Art and Science*.

Compiling Database from Source Records

Thomson Reuters Web of Knowledge

Web of Knowledge (WOK) data that are available through regular subscriptions may be reported in a long, field-tagged, plain text format. The `wok2dbl.f` function recursively searches a directory for plain text batches of records and quickly imports them into R in a long format that may be easily queried. By default letter case standardization and deduplication by record ID is performed. The function returns a data.table object and optionally saves the same to the hard drive. Use of `data.table` provides accomodation for very large databases that perform poorly when treated as a `data.frame`.

By default the data.table is keyed by its WOK id (the "UT" field), then by the field. This makes querying easy. To see the authors of the first three articles, we might enter:

By default the value field is not keyed. While there are scenarios where this would be useful—e.g.~for calling every record by a particular author, or every record in a particular year—keying also sorts the data.table, and the original sort order is important for fields like "TI" (title) which may be broken across several observations.

The `wok2dbl` object remains in the long format of the original data source. We can see this by simply calling the `wok2dbl` object itself or by using `expand.grid` to query multiple keys at once. Here was ask for the source journal, publication year, number of references, and total citations for each of the first three records.

To sort results by record instead of field:

It is convenient to keep the original source data in a long format and to reshape it as necessary for use in different methods. This will be discussed in the *Formatting* section below.

JSTOR Data for Research

Where WOK data are superior for the study of citations, the JSTOR Data for Research (JSTOR) service provides much of the bibliographic information available in WOK and sometimes more accurately. This makes it useful as a cross reference when assessing the quality of a WOK sample, or for augmenting fields such as authors' names.

In addition to the usual variables, JSTOR data also provide ngram frequencies. These data are very valuable and allow limited full-text analysis using "bag of words" methods. The `jstor2dbw.f` function imports dfr.jstor.org records directly from the compressed files returned by queries to the service. Parallelization of the importing process is available and suitable for systems with fast disks. The function performs a standard set of text pre-processing procedures (e.g.~stemming and stop word, punctuation, digit and idiosyncratic word removal) on the ngram frequency tables contained in zip archives that include them. These ngram frequency tables are returned in the indexed format expected by the `stm` package, and all other bibliographic data available are returned as a `data.table`. A character vector attribute called `vocab` is attached to which the indexes in the `jstor2dbw$bow` refer.

Inspecting the `jstor2dbw` object without bags of words ("bow") or abstracts reveals the standard information, and in the conventional wide, flat file format. The only complex value here is author, and multiple authors are listed with names separated by commas.

While the `bow` variable contains the indexed ngram frequency table, which indexes the `vocab` attribute of the jstor2dbw object.

Identity Resolution

Also known as named entity recognition, identity resolution is a data quality problem preventing the researcher from identifying the same thing with a unique label. This happens whenever variations of a label exist. As a consequence the researcher may fail to connect two events to the same thing. When correcting for low identity resolution, the opposite error may be introduced, where two different entities are erroneously treated as the same thing.

The approach to identity resolution involves supervised machine learning. Because this method is not fully automatic it is difficult to implement as a straightforward routine. For now, the results of this analysis are exploited without a manual for conducting the resolution itself.

Formatting

Depending on the analysis or data manipulation to be performed, the `wok2dbl` and `jstor2dbw` objects may need to be converted to a different format, including network formats, which allow us to take advantage of records containing information on multiple units.

Flat File

The `reshape2` package makes it easy to return the wide or flat file format of a query of a `wok2dbl` object.

Many of the interesting fields in WOK records are complex, having multiple observations per record. Some are falsely complex, such as title (TI), which stores a single observation across several fields. Simple and falsely complex values are often trivial features of the document itself. Truely complex field usually store named entities to which the article is related. The most important complex fiels are author (AU and AF) and cited reference (CR). Source journal is an example of a named entity field that is always simple, because a document is only published in one source at a time, though it may have several authors and citations.

Network Formats

The simplest network data format to work with is an edgelist. An edgelist typically has two columns, the name of the node sending an edge in the first column and that of the node receiving the edge in second column.

Bipartite Edgelist

When considering the different relationships among things that could be treated as a network, the `wok2dbl` object is naturally in the format of a bipartite edgelist. For instance we may treat the sender as the paper (UT record id) and the receiver is the citation (CR) to create a citation network.

Or we could treat the author as the reciever to create a bipartite co-authorship network.

However, because of several problems of identity resolution of the CR field in particular, we recommend using the `dbl2bel.f` utility, which normalizes citation codes through case transformation, removal of digital object identifiers, and deduplication. It also optionally allows for data reduction of citations by flagging citations referenced only once (pendants). A report of the results of pendant treatment is printed.

The `dbl2bel` object is appropriate for import into methods designed for bipartite graphs. Because of the nature of record keeping, each complex unit is relateable to others only indirectly by virtue of common inclusion in an article-level record. With a few lines of code we could merge an article to author data.table to an article to citation data.table to yield an author to citation edgelist.

Monopartite Edgelist

A more common operation however is to reduce a bipartite graph to a monopartite graph. This is called a reprojection of the graph, and involves a trivial loss of data. Because many network methods assume monopartite data, we include the `bel2mel.f` utility. The function expects a two column matrix, so when choosing to drop pendants you must do so explicitly and leave off the pendant column.

Assuming that there is at least one 2-star (node of degree two or more) in the bipartite graph, `bel2mel.f` will by default return both monopartite projections. Each projection is the inverse of the other in the sense that what are nodes in the first projection are edges in the second, and vice versa.

Bag of Words

The `jstor2dbw` object contains a variable `bow` and associated attribute `vocab` which can be fed directly to the `stm` package for topic modeling. Usage will be described below.

Merging

Survival

Analytical Method

Clique Percolation

Topic Modelling

The `jstor2stm.f` function is a simple wrapper for the `stm` package for stuctural topic modeling.

Survival Analysis

Reporting