The Concept Lab: Viewer App

2018-12-06

Introduction

This is a guide to using the Shiny web 'Viewer' application created as part of The Concept Lab project. The purpose of the app is to visualize and explore the architecture of concepts inferred from large text corpora by means of statistical measures of the co-association of words in the text.

This document refers to the 'October' version of the app, completed in October 2018. A paper describing in full the natural language processing methods and some of the implementation details is available in the proceedings of IWCS 2017 [1].

The October version of the app is a Shiny R web application[2] hosted on Amazon Web Services and served through port 3838. On some public wireless networks, this port may be restricted — if the app fails to load try to access it from an internet connection that does not restrict this port.

The app pane is composed of a sidebar (on the left) and a main panel, with a tab menu along the top of the screen to switch between panels showing different aspects of the app.

When the app is first opened in a browser, the sidebar and Configuration Panel are displayed.

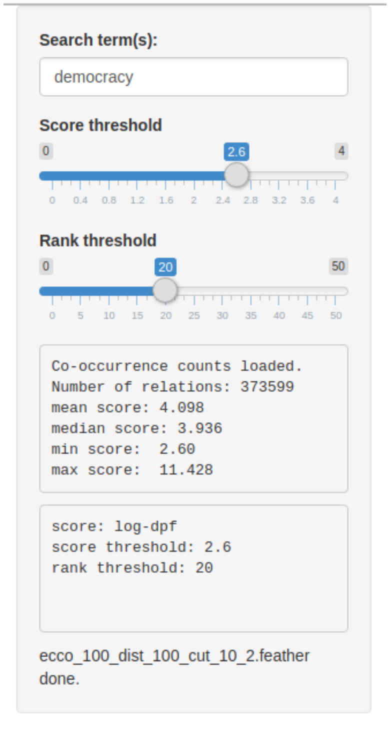
Sidebar

The contents of the sidebar may change depending on which panel is selected, but the following options appear on most panels:

- `Search Terms`: Multiple search terms may be entered here to specify which parts of the network will be displayed in the network panels or used to calculate measures in the other panels. Words should be entered in lower case, separated by spaces or commas. On the network visualization panes, displayed is that containing the neighborhood network of degree `n` of all of the search terms, where is `n` is chosen using the `Steps from search nodes (radius of ego network)` slider on the network panels sidebars.

- `Score Threshold`: The network is created by connecting nodes that have a score above this threshold according to the measure selected in the `measure` radio buttons on the configuration panel (default log-pmi).

- `Rank Threshold`: nodes are only connected if they are ranked



Search term(s):

democracy

Score threshold: 2.6

Rank threshold: 20

```
Co-occurrence counts loaded.
Number of relations: 373599
mean score: 4.098
median score: 3.936
min score:  2.60
max score:  11.428
```

```
score: log-dpf
score threshold: 2.6
rank threshold: 20
```

ecco_100_dist_100_cut_10_2.feather
done.

Figure 1: The sidebar

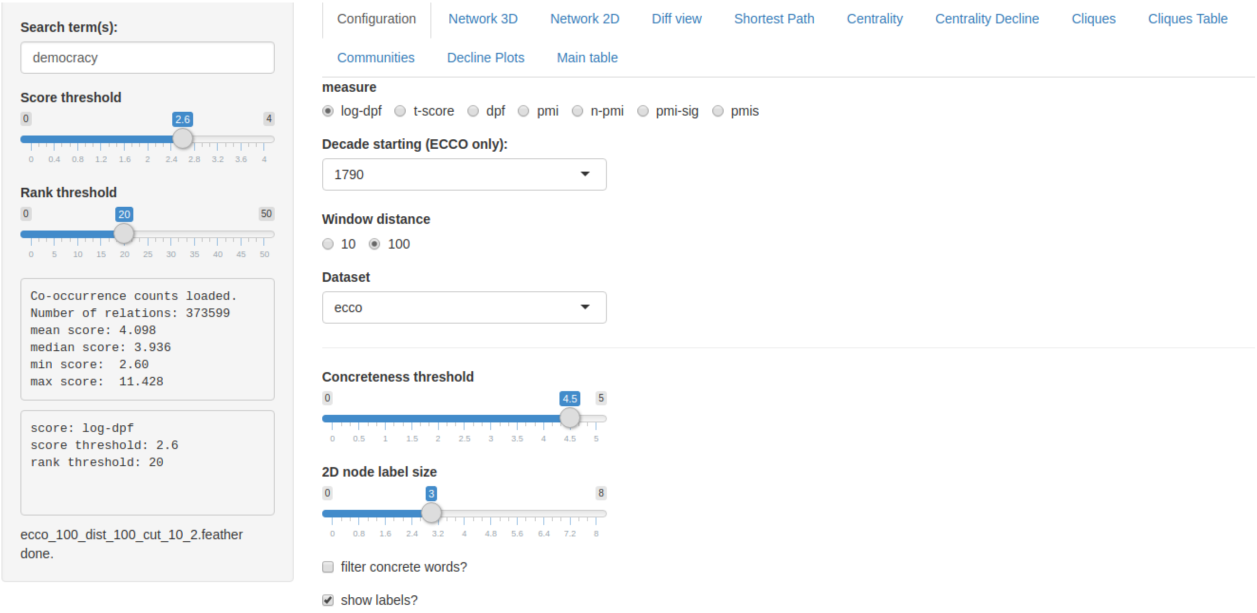
[1] Nulty, Paul. (2017). "Network Visualizations for Exploring Political Concepts".Proceedings of the 12th International Conference on Computational Semantics (IWCS).

[2] `https://shiny.rstudio.com/`

above this value on either of each others' lists. Although score is symmetrical, ranking is not, so node `A` may be at position 15 on the list for node `B`, while node `B` is at position 25 on the list for node `A`. If either of the nodes' rankings are above this threshold, they will be connected by an edge.

Configuration Panel

The Configuration Panel is the main screen from which the dataset and several universal preferences are selected. When the app is opened, the ECCO dataset will be loaded by default. This takes a few seconds, and when it is complete the sidebar text display will show 'Co-occurrence counts loaded' as well as several properties of the loaded data (see Figure 1). The bottom of the sidebar shows the data file name from which the current co-occurrence counts are loaded, in this case 'ECCO_100_dist_100_cut_10_2'.



Search term(s): democracy

Score threshold: 0 – 2.6 – 4

Rank threshold: 0 – 20 – 50

```
Co-occurrence counts loaded.
Number of relations: 373599
mean score: 4.098
median score: 3.936
min score:  2.60
max score:  11.428
```

```
score: log-dpf
score threshold: 2.6
rank threshold: 20
```

ecco_100_dist_100_cut_10_2.feather
done.

measure: ● log-dpf ○ t-score ○ dpf ○ pmi ○ n-pmi ○ pmi-sig ○ pmis

Decade starting (ECCO only): 1790

Window distance: ○ 10 ● 100

Dataset: ecco

Concreteness threshold: 0 – 4.5 – 5

2D node label size: 0 – 3 – 8

☐ filter concrete words?

☑ show labels?

Figure 2: Configuration Panel

From the top down, the options available in the configuration panel are as follows:

- `measure`: The measure by which the association between words should be measured. Each of these measures are based on the number of times words co-occur in a certain context, adjusted for the number of times each word occurs independently.

$$DPF(A,B) = \frac{Co\text{-}occurrences(A,B)}{Freq(A) * Freq(B)}$$

DPF (Distributional Probability Factor) is a measure similar to pointwise mutual information, with an extra parameter to downweight the score of very infrequent words. By default the log-dpf option is selected, as the association scores calclated by DPF tend to have a power-law distribution

- Concreteness threshold / filter concrete words: If the `filter concrete words` checkbox is ticked, then only words which appear in a contemporary vocabulary rated for concreteness by human annotators[3] will appear. The words are rated on a scale from 0 (most abstract) to 5 (most concrete), and only words below the concreteness threshold set by the slider will be included. That is, if the slider is set to 4.5, words rated 4.5 and greater (the most concrete) will be filtered out.
- 2D node label size: This slider controls the size of the node labels for network panels that use 2D output. This is useful for setting an appropriate size for capturing readable screenshots for figures.

Network Panels

- Prune nodes of degree < : Only nodes with this number of connections or less will be displayed. The default setting is two, meaning that nodes that are only connected to one other node in the network are not shown.
- Centrality sample size: The centrality panel shows a table of nodes ranked by their centrality score in the co-occurrence network specified by the dataset, thresholds, options, and search terms specified in the sidebar and configuration pane. The betweenness centrality is calculated by finding the length of shortest paths between random nodes in the network. [4] As this algorithm is computationally intensive, only paths of length less than the value specified here are counted in the estimation. The higher the sample size, the more accurate the betweenness estimation but the longer the time taken to run.

Centrality

This tab shows a table of nodes ranked by their centrlaity score in the co-occurrence ntworek specified by the dataset, thresholds, options, and search terms specified in the sidebar and configuration pane.

[3] Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. ‘Concreteness ratings for 40 thousand generally known English word lemmas.’ Behavior research methods 46.3 (2014): 904-911..Note that this annotated word list excludes many archaic terms and proper nouns.

[4] Ulrik Brandes, A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology 25(2):163-177, 2001.