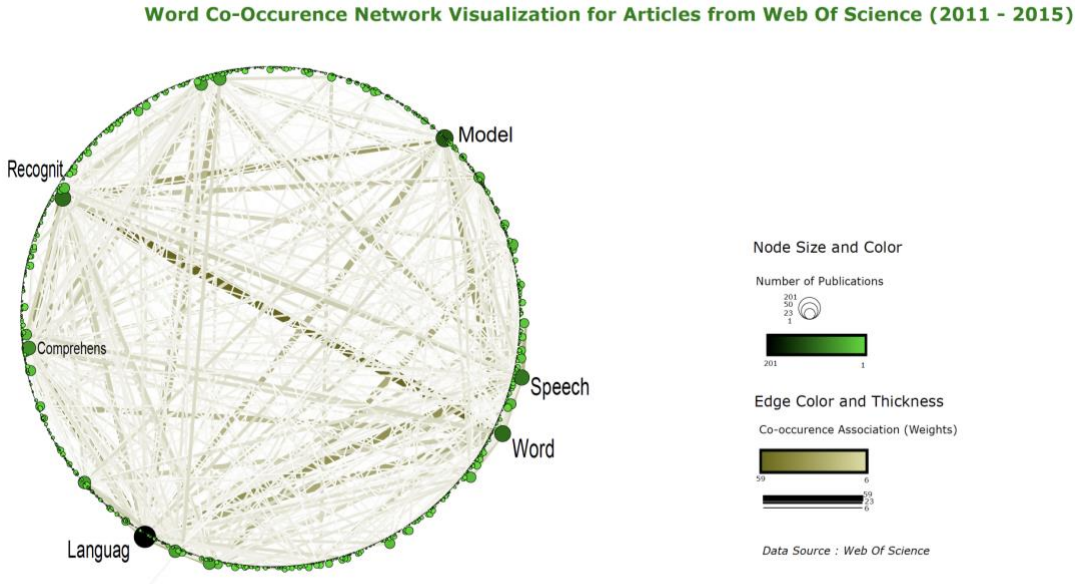
Assignment 4

"What": Topical Data

Problem Statement: Create a word co-occurrence network visualization for a collection of 743 journal articles from Web of Science that are related to computational linguistics for the period 2011-2015. The goal of this visualization will be to highlight the important keywords and topics for this field of research.

For the final visualization, be sure to highlight important terms and relationships in the topic network that will help someone viewing this network gain some insight. You may use any measure of importance or significance that you like, such as: bibliometric statistics, like the number of publications associated with a topic or the weight of a link between topics; or by network statistics, such as node degree or betweenness centrality measurements, or through community detection and clustering algorithms.

Data: The data consists of 13 columns and 743 rows of journal articles from Web of Science. A column named "New ISI Keyword" consists all the non-duplicate topic keywords which was used to build the network.



Process and Inferences:

The network consists of nodes that represent words; and edges that represent the association of any two words. Using the Circular Hierarchy Layout and DRL algorithm I got a beautiful visualization where nodes that frequently occur together were close to each other and the edge connection indicates which two words co-occur. For example, Word and Recognit(ion) have a thick edge. This indicates a strong co-occurrence. Since it is computational linguistic data, we would expect “Language” and “Model” words to be referred the most in all the research papers. And that is what we got to see from this visualization. Some words have no co-occurrences (no edges) because those words would be single words. I have labeled only first six topic keywords based on the number of publications.

Note on Color Preferences:

To emphasize on words with high co-occurrences, I used an extremely light color for word with low co-occurrences. The node color is easy to view with the range of black to light green.