WordNet is a semantic lexicon for the English language that computational linguists and cognitive scientists use extensively. For example, WordNet was a key component in IBM’s Jeopardy-playing Watson computer system.

WordNet groups words into sets of synonyms called synsets.

For example, {AND circuit, AND gate} is a synset that represent a logical gate that fires only when all of its inputs fire.

WordNet also describes semantic relationships between synsets. One such relationship is the is-a relationship, which connects a hyponym (more specific synset) to a hypernym (more general synset).

For example, the synset { gate, logic gate } is a hypernym of { AND circuit, AND gate } because an AND gate is a kind of logic gate.

The WordNet digraph. Your first task is to build the WordNet digraph: each vertex v is an integer that represents a synset, and each directed edge v→w represents that w is a hypernym of v.

The WordNet digraph is a rooted DAG: it is acyclic and has one vertex—the root—that is an ancestor of every other vertex. However, it is not necessarily a tree because a synset can have more than one hypernym. A small subgraph of the WordNet digraph appears below.

The WordNet input file formats. We now describe the two data files that you will use to create the WordNet digraph. The files are in comma-separated values (CSV) format: each line contains a sequence of fields, separated by commas.

List of synsets. The file synsets.txt contains all noun synsets in WordNet, one per line. Line i of the file (counting from 0) contains the information for synset i. The first field is the synset id, which is always the integer i; the second field is the synonym set (or synset); and the third field is its dictionary definition (or gloss), which is not relevant to this assignment.

For example, line 36 means that the synset { AND\_circuit, AND\_gate } has an id number of 36 and its gloss is a circuit in a computer that fires only when all of its inputs fire. The individual nouns that constitute a synset are separated by spaces. If a noun contains more than one word, the underscore character connects the words (and not the space character).

List of hypernyms. The file hypernyms.txt contains the hypernym relationships. Line i of the file (counting from 0) contains the hypernyms of synset i. The first field is the synset id, which is always the integer i; subsequent fields are the id numbers of the synset’s hypernyms.

For example, line 36 means that synset 36 (AND\_circuit AND\_Gate) has 42338 (gate logic\_gate) as its only hypernym. Line 34 means that synset 34 (AIDS acquired\_immune\_deficiency\_syndrome) has two hypernyms: 47569 (immunodeficiency) and 48084 (infectious\_disease).