**CSI 5V93, Assignment 1, Parsing Big Data**

**Due Date: January 29, 2019**

|  |
| --- |
| **Background** |
| Ontology is a typical example of hierarchical data. Gene Ontology (GO) is one of the most widely used ontology databases in Bioinformatics, and provides a framework to elucidate biological roles of genes by semantic analysis. GO contains terms in a structured format (i.e., Directed Acyclic Graph) on three domains: biological processes (BP), molecular functions (MF), and cellular components (CC). |

|  |
| --- |
| **Description** |
| 1. Download the most recent version of ontology data (the OBO file). 2. Parse BP, MF and CC ontologies, separately, and build those 3 ontologies with "is-a" and "part-of" relationships. 3. Correct the errors that might exist in the data set. For example, remove the relationships between two different ontologies. |

|  |
| --- |
| **Submission** |
| * Source files * Answer the following questions:   1. How many terms does each ontology have?   2. What is the term ID of the root of each ontology?   3. How many leaf nodes does each ontology have?   4. What is the longest path length from the root to a leaf node?   5. What is the longest path length among the shortest paths from the root to leaf nodes? |