

**Problem 1** (Taxonomy Induction). In this exercise, we are working on taxonomy induction.

- **Input:** A set of terms, which can be both entities or classes.
- **Output:** A taxonomy built around the input. The taxonomy has to be a **tree** (a connected directed acyclic graph with one root, every node except the root has a single parent), with depth (maximum distance from leaf nodes to the root) **at least 4**. It must contain all terms from the input, and a maximum of 50 terms in total.

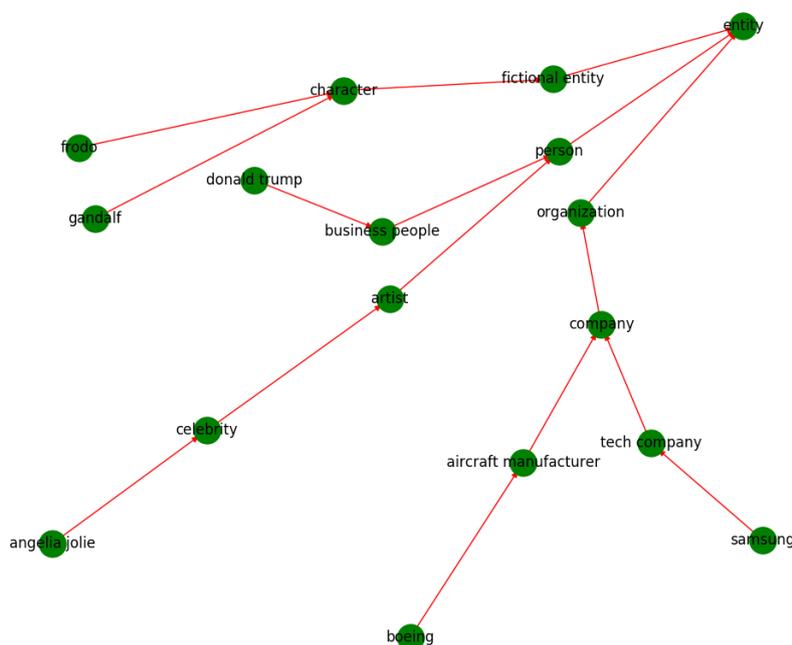
We provide a hypernymy dataset<sup>1</sup> (`webisalod-pairs.txt`) from WebIsALOD, which contains 4.5M hyponym/hypernym pairs along with confidence scores. The file has the format `hyponym ; hypernym [tab] score`. Use of other datasets is also permitted, as is use of cleaning steps on top of WebIsALOD.

**Program:** Your main program, called `run.py`, takes a file that stores the given terms (one per line) as input, and plots the resulting taxonomy tree. *You can use any library for plotting, such as `networkx`, `igraph`, `pydotplus`, etc.*

A valid sample solution would be:

**Input:** *donald trump, business people, angelia jolie, person, frodo, character, gandalf, samsung, tech company, company, organization, boeing, aircraft manufacturer.*

**Output:** See below.



**Evaluation:** We run your algorithm on three inputs: `input-1.txt` and `input-2.txt` from the link below, and one unseen input. We only run the command: `python run.py input-file` to evaluate your program. Please make sure that your program is able to plot the resulting taxonomy.

We evaluate by (1) correctness of the resulting taxonomy, (2) structural “beauty” (e.g., not all branches in bottom layer only), (3) resilience in working with the unseen input.

Your submitted files must include all necessary code and files, especially the main program file `run.py`. If you used any external libraries, please indicate them in a README file.

<sup>1</sup><https://www.dropbox.com/s/r816ll86rsioe9o/lab05-material.zip?dl=0>

Please submit all necessary files, which are compressed into a zip file named:

**Lab05\_MatriculationNumber\_Name.zip**

to the email address: [cxchu@mpi-inf.mpg.de](mailto:cxchu@mpi-inf.mpg.de) with title of the email: **[IE]Lab05\_MatriculationNumber\_Name**

**Deadline: 23:59 23.11.2019 (Saturday)**