My work at the Making and Knowing Project

Dana CHAILLARD, Ecole Polytechnique

I would like to thank everyone at the Project that I had the chance to work with, Pamela H. Smith, Naomi Rosenkranz, Terry Catapano, Roni Kaufman, Matthew Kumar, Gregory Schare and Tianna Helena Uchacz. It was a wonderful experience working with you all.

All my code can be found at :
https://github.com/danachaillard/manuscript-object
The readme should allow you to get everything running.

My work consisted in three tasks. The first one was to create word clouds from the categories in the thesaurus. This was done in the file word_clouds.py using a specific Python package. For each semantic tag, there is a word cloud in the word cloud file.

The second task was to create semantic trees from the vocabulary in the thesaurus. Once again category by category. These were created in the vocabulary_abstraction.ipiynb jupyter notebook. The trees are only to be found in the notebook. They were created using wordnet, which is a corpus that can give us the hypernyms of a word. For the the trees, I used only the tagged expressions in the thesaurus that are single words. I also had to clean the data to remove words not in the corpus or where the correct meaning we are looking for is not in the corpus.

The third task was to create sentence embeddings using a machine learning approach. For that, I used a pretrained ELMo model found on tensorflow_hub. ELMo generates a 1024 dimensions vector for each sentence. I then used PCA and t-SNE to reduce these dimensions to 2 and plotted that in an html file. All these files can be found in the sentence_encode directory. I also created a graph where two categories are plotted in different colors. It is called sentence_encode_dual.