Progress Report: Generating Hierarchical Topic Taxonomies over Text

- Gathered dataset from my workplace. There are two levels in the hierarchy. The top level contains verticals such as “auto,” “technology,” “health,” and so on. There is a second level below. For example, “VR headset” would fall under “technology.” The texts under “VR headset” are a collection of short keywords such as “VR gear,” “virtual reality goggles,” “virtual reality games,” and so on.
- Tried to read hPLSA paper – A hierarchical model for clustering and categorizing documents, Gaussier et al.
    - Could not understand how the model worked after several hours, unfortunately. I feel like this idea might be beyond my current scope
- Wrote several text vectorizers, including one that can use pretrained word embeddings and fine tune them on new data.
    - I will use these in conjunction with clustering methods like kmeans and see if this approach can generate a sensible hierarchy over my dataset given only the keywords/documents
- Instead of hPLSA/hLDA, I will pivot to building hierarchical models using “flat” clustering models and previously mentioned text vectorizers
    - Perhaps even classification models could be used (if we frame every combination of topic/subtopic as a class)
    - My final hierarchical model implementations will form my submission