LDA Mallet Model

1. modeling\_script\_LDA mallet. Ipynb
   * Main script for running LDA mallet model
   * Includes preprocessing, corpus/dictionary building, model running, and post model data analysis
2. dict\_07\_01\_20.dict
   * Dictionary for ldamallet\_model\_07\_01\_20
   * Use stored dictionary after 1st run on csv
3. corpus\_07\_01\_20.mm.index/ corpus\_07\_01\_20.mm
   * Corpus for ldamallet\_model\_07\_01\_20
   * Use stored dictionary after 1st run on csv
4. ldamallet\_model\_07\_01\_20
   * Stored model output from modeling\_script\_LDA mallet
   * Use stored model to rerun analysis/clusters
5. Model\_results\_07\_02 (stored in google drive)
   * Model output with each datapoint labeled with a topic
   * Used to conduct post model data analysis

Word Embeddings: tested out-of-box word embeddings on dummy K-means model

SVM\_model\_sasb

* SVM model created from labeled SASB data
* Includes preprocessing, model running, and post model data analysis

Next Steps:

1. Refine SVM model to improve precision/recall
2. Apply word embeddings to labeled SASB corpus to additionally improve precision/recall of SVM Model
3. Run LDA Model to remove unimportant paragraphs based on cluster topics for a reduced unlabeled proxy statement corpus
4. Append unlabeled proxy statement corpus to labeled SASB corpus and apply SVM model (or most successful classifier based on precision/recall)