LDA Mallet Model

1. ldamallet\_script.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](https://github.com/IshaShah27/dfg-humanrights/blob/master/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](https://github.com/IshaShah27/dfg-humanrights/blob/master/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](https://github.com/IshaShah27/dfg-humanrights/blob/master/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](https://drive.google.com/file/d/1q6iE1EozIos6P7g5h2_rn8aj1c3Qa2cb/view?usp=sharing) (stored in google drive)
   * Model output with each datapoint labeled with a topic
   * Used to conduct post model data analysis
6. ldamallet\_topics\_results
   * Word document of all 20 documents from initial LDA mallet model
   * Recommended topics to keep and remove are listed at top of document

[word\_embeddings.ipynb](https://github.com/IshaShah27/dfg-humanrights/blob/master/word_embeddings.ipynb): tested out-of-box word embeddings on dummy K-means model

[SVM\_model\_sasb.ipynb](https://github.com/IshaShah27/dfg-humanrights/blob/master/SVM_model_sasb.ipynb)

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

Next Steps:

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