April 20-2019

LDA in sklearn running over words only (unigrams). Cutoff of 3 2 topics pick number of top words it spits out.

Compare to XGBoost and SVM baselines. E.g. How often does the output from the SVM correspond to a particular topic from LDA? How often does the output from the XGBoost coresond to a particular topic from LDA?

Check out cases where topic modeler is not confident e.g. probabilities are close to 50. Maybe ones that are missed are on the borderline in the lDA model.

Could also run svm and get the coefficients for different word features.

Ken is running sampling experiments again. We're both running with 70,000 ig features 30,000 features and then 10,000 features

Ken is having some issues with cluster centroids. It has been running for 2 days and still hasn't finished downsampling

It appears to be using a sparse to dense conversion because it's using up to 50 gb of ram

---

Revision #1

Created Thu, Feb 6, 2020 4:39 PM by kenneth

Updated Thu, Feb 6, 2020 4:39 PM by kenneth