Youth Recidivism Final Project Report

Wei Luo, Mengqiao Cai

Goal and Motivation:

The YES rehabilitation program has a lot of youth who formerly finished this program and reenter into this program again and we want to find out why this program attracts people who finishes this.

Data Aggregation

This step was done by previous group (see Codebook\_Case Management Notes.xlsx)

Topic Modeling Analysis

Step 1: Clean text

I started cleaning text by making all letter lower case, tokenize, lemmatize and remove all stop words that I previously defined, which returns a list of words in the original sentence.

Step 2: Make Bigram and Trigram of data

Bigram is the technique that looks at the combination of neighbor two words in a text, e.g. in a sentence: “I am happy”, bigram will look at “I am” and “am happy” in this case, and trigram by its name will look at the combination of neighbor three words.

Step 3: Create Corpus and build model

By using lemmatization, I create a corpus that contains only nouns, adjectives, and adverbs in the dataset. After finish building corpus, I build a LDA model using current corpus and group the caseworkers’ notes by 5 groups, which I will treat as ‘label’ in later analysis. I save this model and corpus to ‘lda.model’ and ‘corpus.txt’. Later I use PyLDAvis to generate a topic modeling visualization result(see lda.html). Below is a screenshot of visualization.

Chart, bubble chart

Description automatically generated