# Legal Verbiage and Unsupervised Classification

Chris Han – Attorney

Anthony Chi – Data Scientist

**Last Updated: August 26st, 2018**

**Topic Modeling**

We are using the gensim library to perform LDA (Latent Dirichlet Allocation). LDA assumes that documents are probability distribution over latent topics. Topics are probability distribution over words. LDA looks at a number of documents and assumes that the words in each document are related. It then tries to figure out the 'recipe' for how each document could have been created. We just need to tell the model how many topics to construct and it uses that 'recipe' to generate topic and word distributions over a corpus. Based on that output, we can identify similar documents within the corpus.

### In order to understand the LDA process, we have to know how LDA assumes documents are generated:

1. Determine the number of words in the document.
2. Choose a topic mixture for the document over a fixed set of topics (ie. topic A 20%, topic B 50%, etc).
3. Generate words in the document by:
   * Pick topics based on the document's multinomial distribution
   * Pick words based on the topic's multinomial distribution

Suppose we have a corpus of documents, and we want LDA to learn the topic representation of K topics in each document and the word distribution of each topic. LDA would backtrack from the document level to identify topics that are likely to have generated the corpus.

### LDA algorithm in a nutshell:

* First randomly assign each word in each document to one of the K topics.
* For each document, LDA
  + assume that all topic assignments except for the current one are correct
  + calculate two proportions:
    - proportion of words in document d that are currently assigned to topic t → P(topic t | document d)
    - proportion of assignments to topic t over all documents that come from this word w → P(word w | topic t)
  + multiply two proportions and assign w a new topic based on that probability. P(topic t | document d) \* P(word w | topic t)
* eventually we'll reach a steady state where assignments make sense

**Topic Modeling Visualization**

The numbers might not make sense. We are fortunately enough to have 2-D visualization that represent differences between topics by distances. We tried different number of topics using the LDA algo and visualizations to try to identify the optimal number of topics as the LDA input.

## III. Prelim Results

### We iterated through different NLP, keywords identification, and topic modeling methodologies. Topic modeling successfully identified unique topics. However, domain experts determined that the topics were not unique enough to determine whether language needed to be actioned or not. The conclusion was that the domain experts would label features as we would proceed with a supervised classification exercise instead of an unsupervised one. With the new approach and small amount of data, we can quickly determine if the new approach can effectively classify docket texts into two classes: action and no action.

back to the top