*Elizabeth Fabio*
*BDA-594*

# Web Exercise 7: Topic Model and Video

1. What is the YouTube URL of your short demo video? Which software do you use to create this video? What are the advantages and disadvantages of the video software you used?

   **YouTube URL**: https://youtu.be/XFPyL1wFrSY

   The software I used to create this video is Zoom for the screen recording and iMovie to cut out a couple seconds near the end. I have, at first, used Camtasia to edit the video and added annotations such as a circle to emphasize a button that needed to be pressed and highlighting a section of the screen. However, the watermark applied to the video for free trial use was very distracting and I decided to edit the video through iMovie on my MacBook instead. However, the disadvantage of using iMovie is that there's no annotations and no option to screen record videos. Also, the added bonus of using Zoom is that you can record your meetings as well.

2. What are the key procedures for cleaning texts in Twitter messages?

   The key procedure for cleaning texts in Twitter messages begins with creating a corpus. A corpus is a large, structured set of machine-readable texts. When creating a corpus with the TWEET_TEXT column, you are able to perform the following cleaning procedures in R Studio with the tm package:

   - Converting capitalized letters to lower-case.
   - Remove punctuations.
   - Remove numbers.
   - Remove URLs.
   - Remove non-ASCII characters (e.g. emojis, IDN).

   Some of the text cleaning functions above may not be needed depending on your application. For instance, leaving numbers in your corpus may be needed to analyze a certain numerical pattern.

   After performing some preliminary text cleaning, you can filter out your corpus by removing stopwords. Stopwords is a list of common words to be removed from a corpus before processing the text data. In this case, the English stopwords from the SMART information retrieval system were removed from the Tweets_corpus. Afterwards, you can also remove specific words as well. In this case, the chosen words removed are: "london", "im", "ive", "dont", and "didnt".

   After removing stopwords and specific words from the corpus, excess whitespaces are removed from the corpus and word stemming is performed. The process of "stemming" reduces words to its root form and prevents words with the same meaning from being treated as different key words. For instance, "add" remains after removing "-ing" from "adding".

After all the data preprocessing is completed, the data structure from the corpus is converted back to a character list and, again, extra whitespaces are removed. In summary the key procedures for cleaning text in Twitter messages are:

1. Create a corpus.
2. Perform text cleaning functions.
3. Remove stopwords and selected words.
4. Remove excess whitespaces.
5. Perform the stemming process.
6. Convert corpus back to a character list.
7. Remove excess whitespaces again.

3. What is LDA? How to interpret topic-term distributions parameter (β) and document-topic distributions parameter (α)?

LDA, or Latent Dirichlet Allocation, is one of the most popular topic modeling methods. Topic modeling is a method of unsupervised modeling in finding some natural groups of items/topics even when we're not sure what we're looking for. So specifically LDA finds topics a document belongs to based on the words in it. In this lab exercise, these documents are the tweet text.

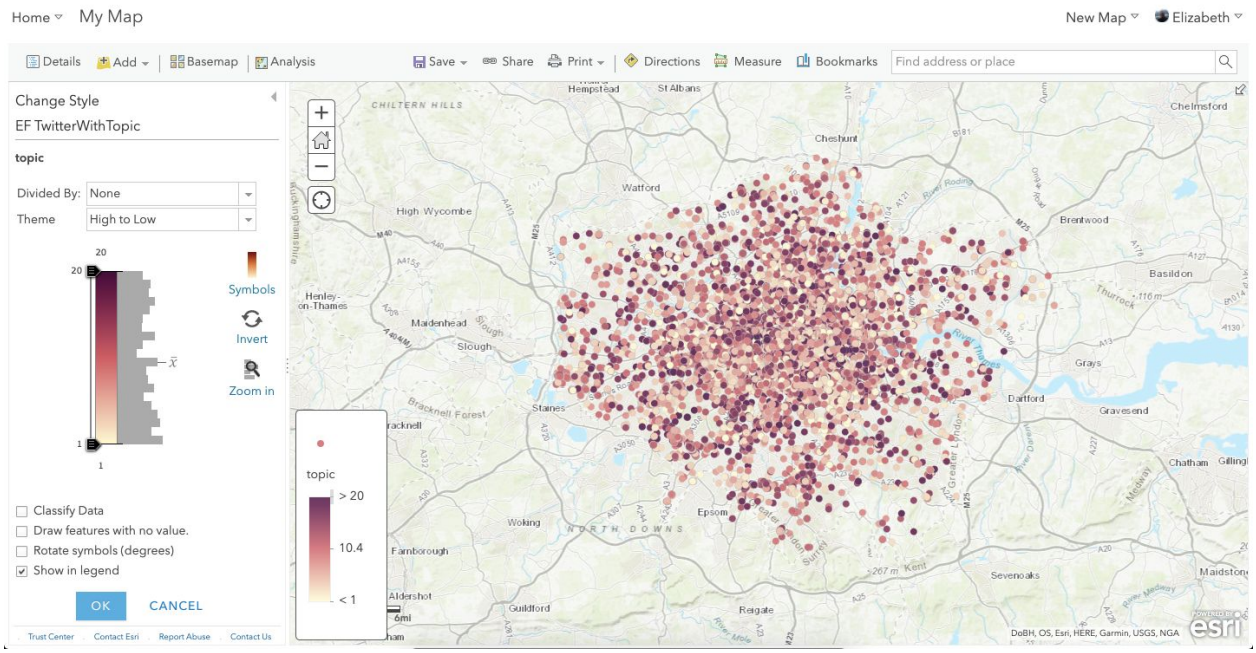
Topic-term distributions parameter (β) and document-topic distributions parameter (α) are called hyperparameters. Hyperparameter is a parameter value that controls the learning process. α determines the proportion of topics for a given document and β determines the distribution of words per topic. In tuning down α, documents will likely have less of a mixture of topics. This means documents will have one topic. In tuning α up (closer to a value of 1), documents will likely have more of a mixture of topics. This means documents will have a uniform distribution of topics. For β, tuning down will lead to less words for a topic and tuning up will lead to more words for a topic.

Looking at the Dirichlet distribution further, α adjusts p(topic| document) and β adjusts p(word | topic). p(topic | document) is the proportion of words in a document that are assigned to a topic and p(word | topic) is the proportion of assigned topics over all documents that come from this word. In the lab exercise, α and β values are set to 0.1. Thus, each tweet will have a uniform distribution of topics and there will be less assigned words for a topic.

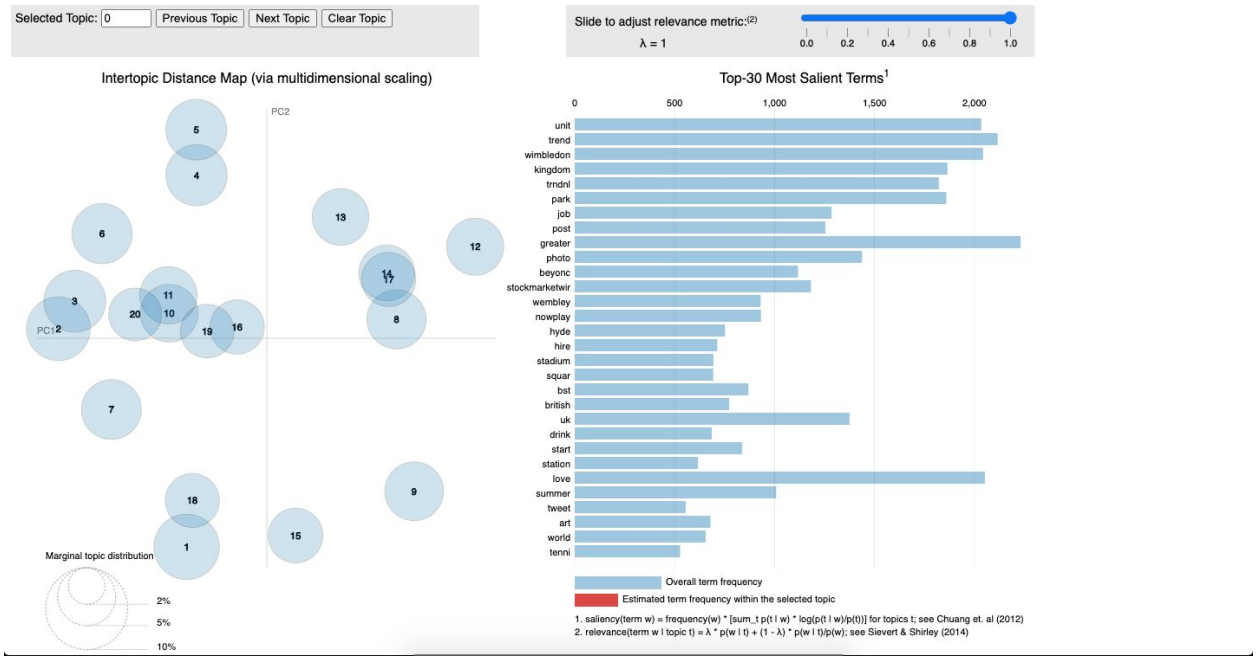
*Elizabeth Fabio*
*BDA-594*

4. Include an ArcGIS Online Screenshot of the London Map with the LDA and the results of "serVis" display. (use different colors or markers to display different topics). Select one color (topic) to explain the possible represented topic.

**EF_TwitterWithTopic.png**:

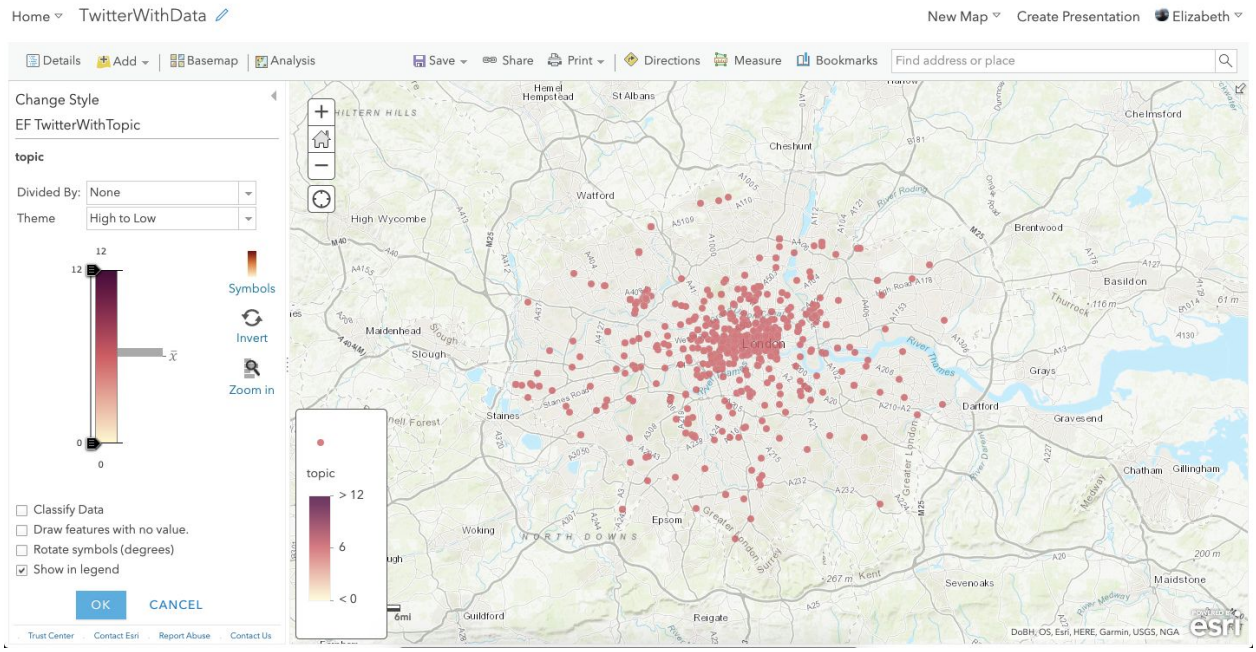



**EF_serViz.png**:

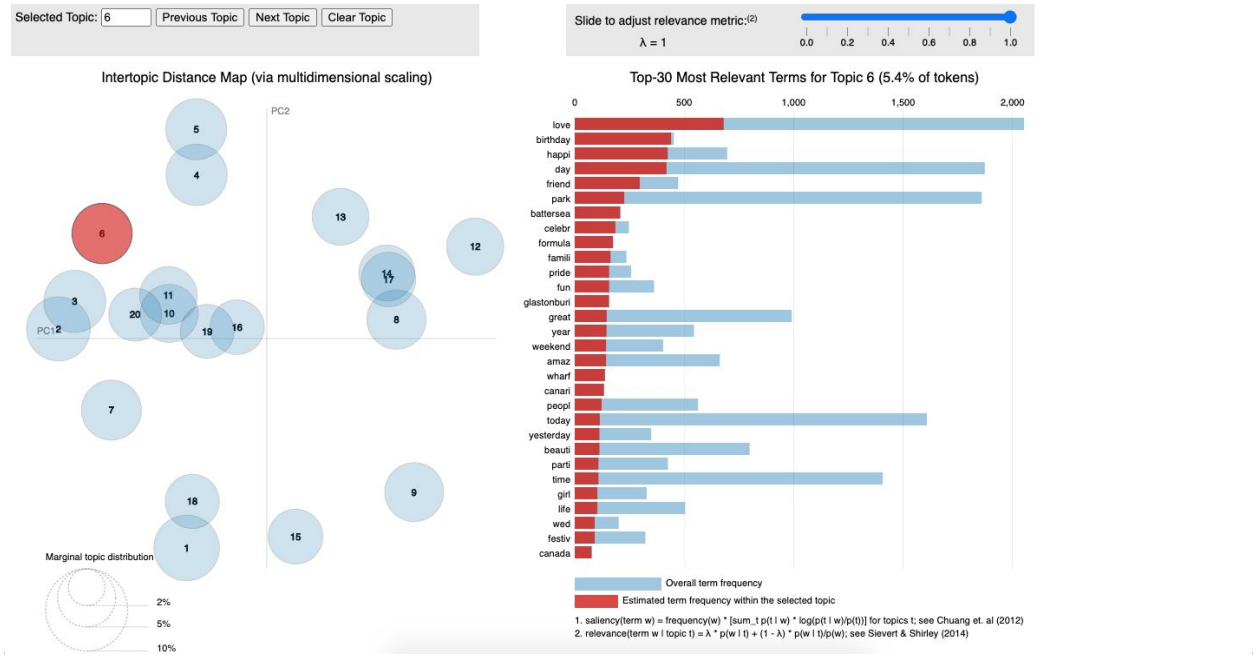


*Elizabeth Fabio*
*BDA-594*

## **EF_TwitterWithTopic_topic6.png**:




## **EF_serViz_topic6.png**:



*Elizabeth Fabio*
*BDA-594*

From observing topic 6, the words listed in the “serViz” appear to be celebratory words such as “love”, “birthday”, and “happi” as the top three words when λ = 1 (**EF_serViz_topic6.png**). Values of λ close to 1 have high relevance rankings to frequent terms within a given topic, whereas values of λ close to zero have high relevance rankings to exclusive terms within a topic. Looking at the TwitterWithTopic web map on ArcGIS Online, all tweets with a specific topic have a particular color to distinguish from other tweets with different topics (**EF_TwitterWithTopic.png**). After filtering points with only topic 6, the points on the map now should be the same purple-pink color (**EF_TwitterWithTopic_topic6.png**). These points are more concentrated in central London and are dispersed on the outskirts of the city. A cluster of tweets with topic 6 tend to be around areas where big events occur like football games in Wembley Stadium and areas where fun activities occur like the London Aquatics Centre and the Copper Box Arena. Also, topic 6 contains the word “park” and a good amount of points are set in different parks like Hyde Park and Regent’s Park.