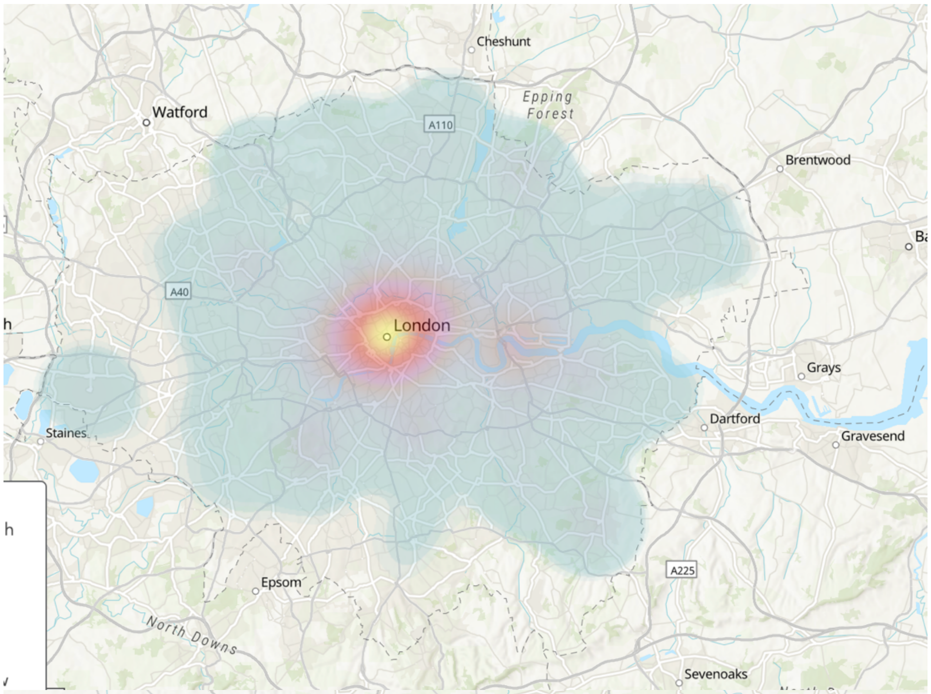
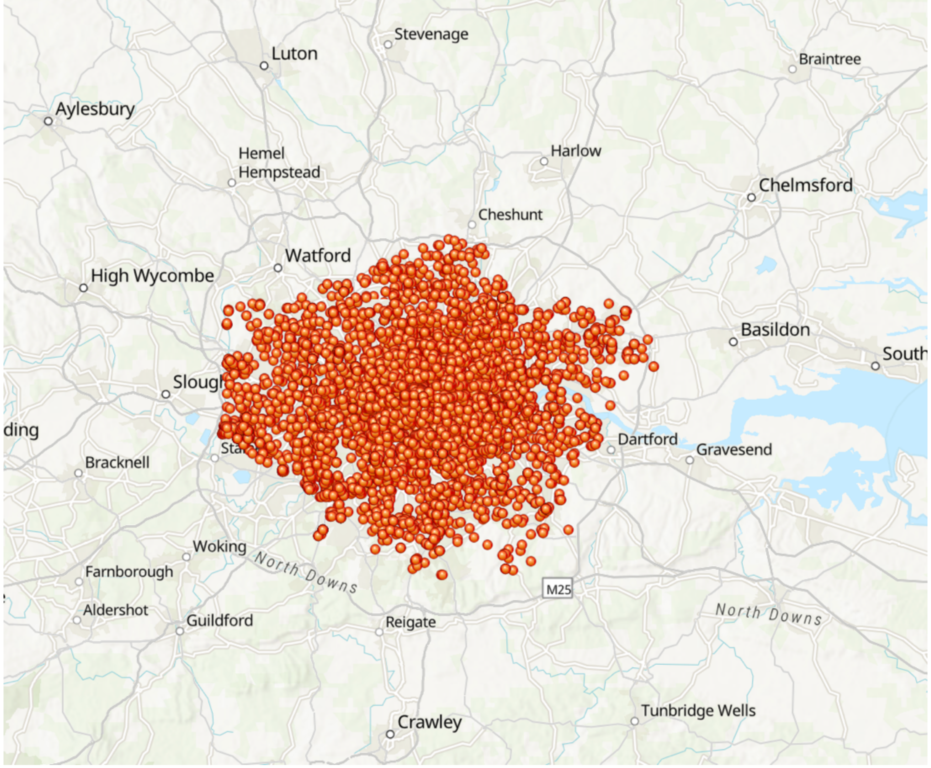
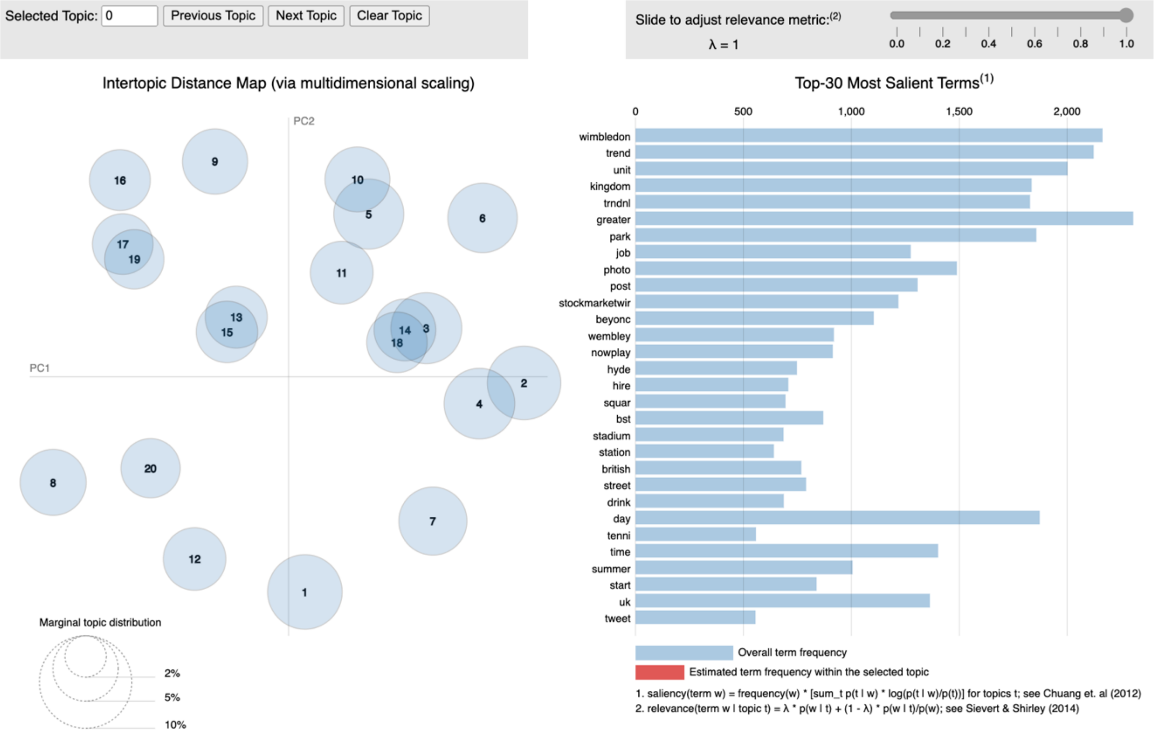
LAB 07

1. What is the YouTube URL of your short demo video? Which software did you use to create this video? What are the advantages and disadvantages of the video software you used?
   The URL is: https://www.youtube.com/watch?v=Xcdp-UJjU0g
   I used Camtasia and it was very very simple! I found more advantages than disadvantages. Definitely loved that I was able to screen record and record my audio simultaneously. I found it very easy to edit and export. I loved the feature which allows you to post directly to youtube.
2. What are ALL the key procedures for cleaning texts in Twitter messages?
   In order to clean texts in twitter messages you must:
   1. Remove lower-case letters using the tolower function
   2. Remove punctuation, with the removePunctuation
   3. Remove numbers using removeNumbers
   4. Remove URLs, using gsub
   5. Remove non-ASCII characters using iconv
   6. Next, is removing stopwords such as “is”, “in”.
   7. Remove whitespace/Reformat using stripWhitespace
   8. Remove suffixes
   9. Convert the data structure from corpus back to a character list with extra white spaces removed.
3. What is LDA? (Please describe the basic concepts and possible applications at least over 100 words). How to interpret topic-term distributions parameter (β) and document-topic distributions parameter (α)?
   LDA is an unsupervised topic model called Latent Dirichlet Allocation, which is used for research in natural language processing, text mining and social media mining. This can be created in R, using the “lda” package to fit all data within the model. The outcomes of this model include assignments, topics, topic_sums, document_sums. LDA is a Bayesian model, where each topic is modeled over a set of topic probabilities. It is essential a “bag-of-words” model, and it doesn’t matter what order everything is in. Each topic is just characterized by various distributions of work. This model can help with preprocessing needed for machine learning. Interpreting the topic-term distributions parameter is based on the size. A bigger β means there is more similarities between the words in various topics. The document-topic distributions parameter can be interpreted by its size as well, meaning a bigger α points to more similarity between topics in various documents.

4. Include an ArcGIS Online Screenshot of the London Map (points without different colors) with the LDA and the results of “serVis” display (points with the assigned colors). (Please use different colors or markers to display different topics). Select one color (topic) to explain the possible represented topic and keywords.
   I chose a heatmap to represent the Latitude of the TwitterWithTopic.


Cheshunt
Epping Forest
Watford
A110
Brentwood
Ba
A40
h
London
Grays
Staines
Dartford
Gravesend
h
A225
Epsom
North Downs
Sevenoaks


Stevenage
Luton
Braintree
Aylesbury
Harlow
Hemel Hempstead
Chelmsford
Cheshunt
Watford
High Wycombe
Basildon
South
Slough
ding
Dartford
Gravesend
Bracknell
Woking
North Downs
Farnborough
M25
North Downs
Aldershot
Guildford
Reigate
Tunbridge Wells
Crawley



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)