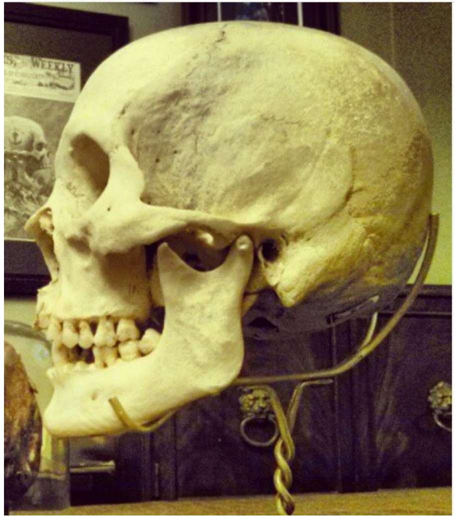
Reproducibility in Data Science: Replication Project

Ankit Tandon, Ankita Pal, Chavi Gupta, Kirti Kharb, Mabel Li

Replication Repository:
https://github.com/UW-MSDS-DATA-598-Reproducibility-WI20/gupta-kharb-li-pal-tandon-replication-project

Introducing “Insta-dead”

- We decided to replicate *“The Insta-Dead: The rhetoric of the human remains trade on Instagram”*
- *“Insta-Dead”* explores the trade of human skulls and bones on Instagram
- Dataset consists of 132,225 instagram posts pertaining to hashtags related
  - Each data point contains the link to the post, the image, the caption, and time of post
- Paper does not explore legality of trade, that varies based on jurisdiction
- They attempt topic modeling on the captions to impose structure and derive semantic meaning



Methods and Findings from “Insta-Dead”

- After removing stop words, LDA Mallet modeling is used to derive meaning from the posts using the ‘topicmodels’ package in R
- Authors found that 25 was the ideal number of topics to divide topics into people interested in skulls and bones for art and those who wish to possess and collect the pieces
  - Example of art:
    - **Another one of the skulls I made for our art show last week.** #skull #skullpainting #humanskull
    - **Drawing I did today. Quite pleased, it's not quite right but I don't draw very much. So I'm happy enough!** #skull #drawing #sketch #artist #pencildrawing #fairlyhappy #happyface #coukdbeworse #couldbebetter
  - Example of collecting pieces:
    - **First human bone piece for my oddities collection. Human vertebrae.** #bone #oddities #humanbone #odditiescollection #vertebrae #macabre

Target Figure

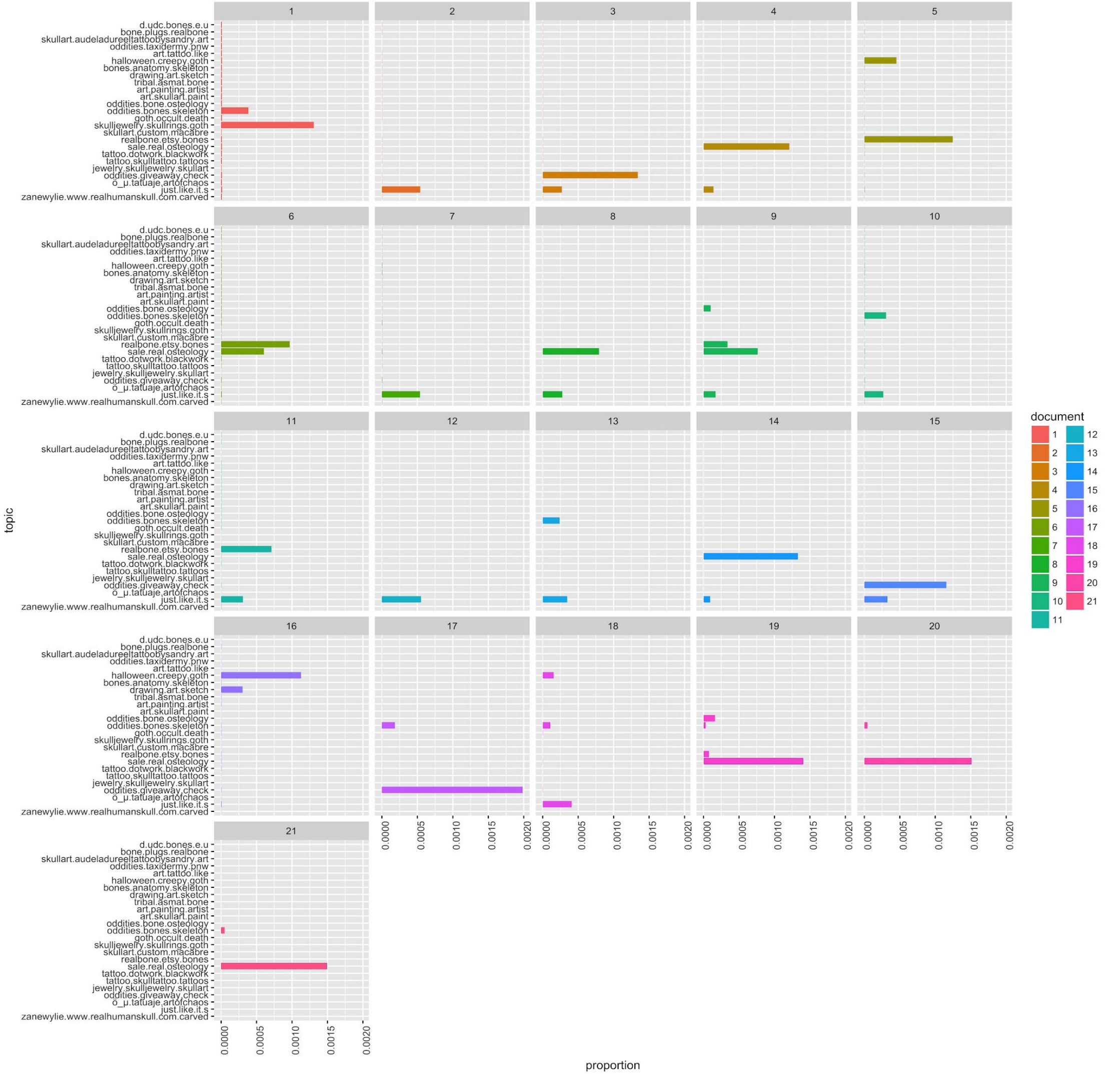
Figure represents posts from a “collector” user

21 boxes each representing a document

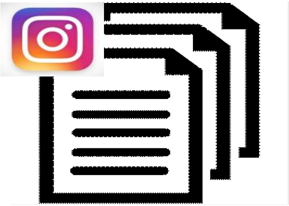
Topics are combinations of words, represented on the y-axis

Bars encode how much of each topic is present in each document

Documents are color coded

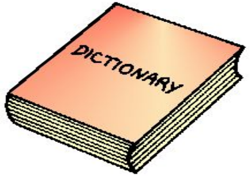


Our replication attempt

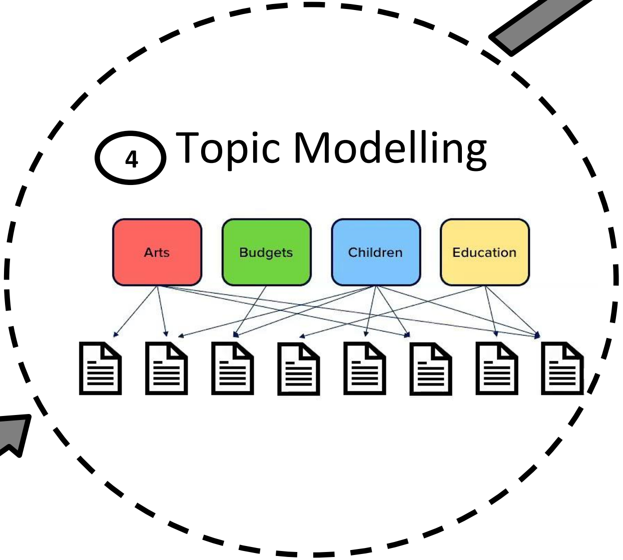


1 Insta-posts

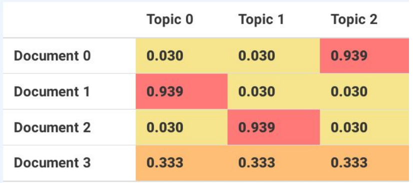
2 Text Preprocessing



3 Corpus

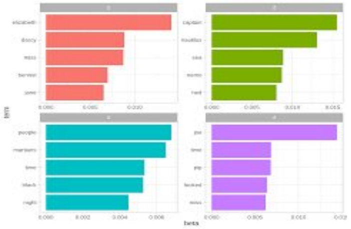


4 Topic Modelling



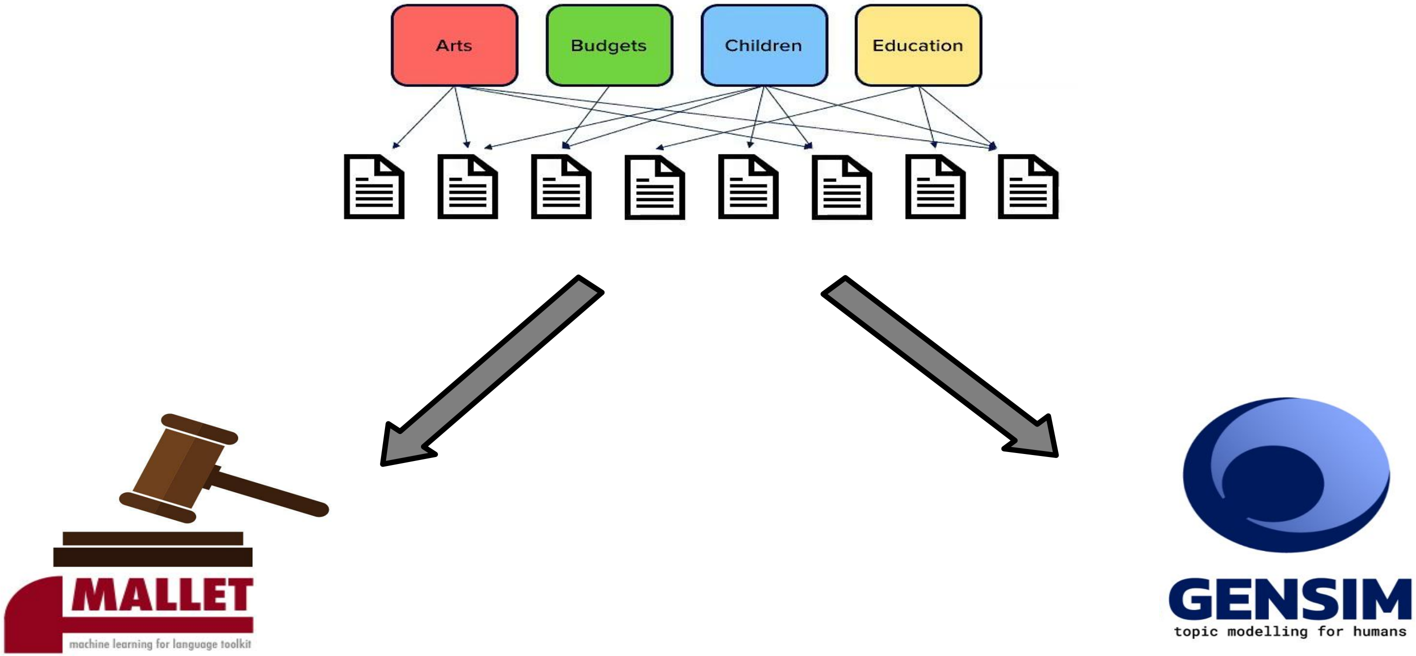
	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.939	0.030	0.030
Document 2	0.030	0.939	0.030
Document 3	0.333	0.333	0.333

5 Topic-Document Probability



6 Visualize Results!

4 Topic Modelling


Arts
Budgets
Children
Education
MALLET
machine learning for language toolkit
GENSIM
topic modelling for humans

Original Method

Replication Method

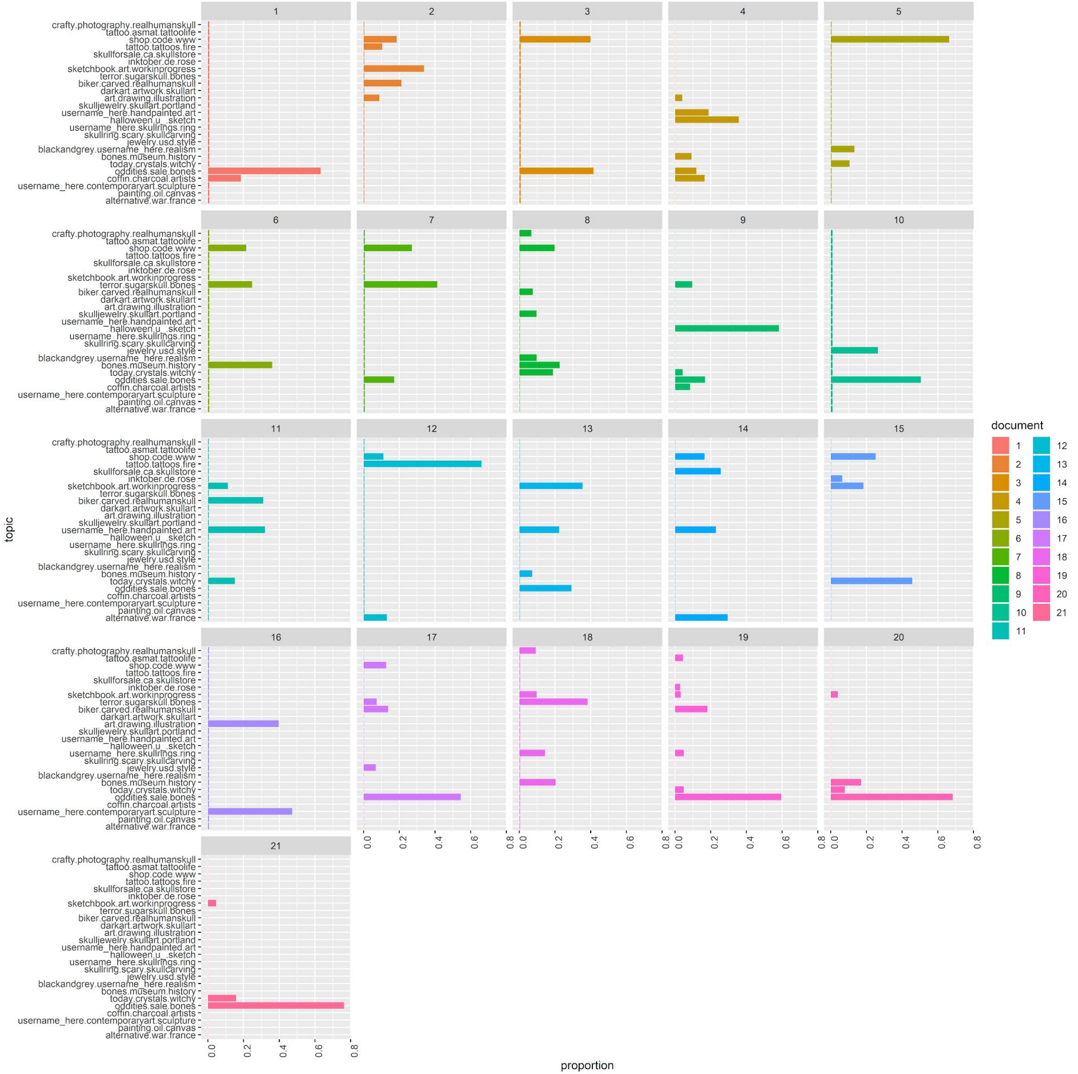
Replication Results

21 boxes each representing a document:

- Topics extracted are slightly different from the original source
- Topics identified for each document is slightly different
- Similar inferences:

Replication: 'Bones.museum.history.paris.catacombs. Skeleton.travel.photo.face.humanremains'

Original: 'd\udc bones e\u b\u death f\u a\u d\u paris catacombs'



Reflection on the replication process

- Had to change paper because the original paper had replication difficulties
- Finding an Alternative Topic Modelling Technique

  (finding similar functionality)
- Producing comparable results while changing technical details
- Installing Python Libraries for RMarkdown! (Extremely frustrating!)
- Inconsistent data files from original source! (Improper documentation?)

References

- Huffer, Damien, and Shawn Graham. 2017. “The Insta-Dead: The Rhetoric of the Human Remains Trade on Instagram.” *Internet Archaeology*, no. 45. https://doi.org/10.11141/ia.45.5.

- Rafferty, G. 2019. “LDA on the Texts of Harry Potter.” https://towardsdatascience.com/basic-nlp-on-the-texts-of-harry-potter-topic-modeling-with-latent-dirichlet-allocation-f3c00f77b0f5