DATA 598: Replication Project

# Introduction

For examing the complexities in replication we chose the paper (Huffer and Graham 2017). This paper is a study of the online trade of human remains on Instagram. The paper focuses on understanding what is happening, where it is happening, and how these human remains are framed as collectible objects so that archaeologists, cultural heritage professionals, museums and so on are better equipped to engage with this desire and to channel it productively. The paper says that due to the accessibility of social media for the use of selling, displaying and trading human remains is so feasible, it has led to treating human remains as consumer products for the collector’s markets rather than objects of archaeological, ethnographic or anatomical value which is a cause of great concern. Many state or national-level jurisdictions have introduced specific restrictions related to the sale and transport of human remains, above and beyond the requirements laid upon signatories to UNESCO (1970) cultural heritage conventions, which applies to most source and demand countries. The study in the paper is to better help these organizations understand the dynamics of human remains trade so that they can take informed actions and make effective policies. It is important to note that the paper does not explore the legality of the trade of human skulls and bones as that is dependent on the jurisdiction of the seller. In that respect the authors of this paper have dropped the field pertaining to the location of the post.

The authors of the paper use several methods employing data mining techniques to query the “noise” surrounding the topics related to the trade of human remains on Instagram. The paper demonstrates that these methods show great promise as opposed to time and labor-intensive manual search methods. Some of the methods the paper explores is analyzing the images associated with the Instagram posts in the dataset for similarities in features such as hue, color, brightness, etc. The paper also analyzes the sentiment of the posts to uncover differences in posts with darker and lighter tones. Finally, the paper also performs network analysis on the accounts that follow the posters of the Instagram content in the dataset.

We have chosen to replicate one such method in this project which involves topic modeling. The purpose of this technique is to extract semantic structure from posts. These topics can be used to classify and group different posts relating to trade. This creates a foundation to perform much deeper research and analysis into understanding the different trends and patterns associated with different aspects of the trade such as buying and selling, sales prices, the material being traded and other mechanics of trade. The methods used in the research take advantage of the open data as the images, captions, location, user id, the selling price, etc. mentioned in the post to study the characteristics of the online trade. Although they have removed the usernames from the published data and figures but have used them in the analysis.

# Replication Source

From the paper (Huffer and Graham 2017), we are replicating Figure 1 (Figure 5 in the paper). This figure shows different topics modeled from the posts from the account of a single trader/user on Instagram. The key conclusions from this figure are that this particular trader is interested in “real” or authentic human bone artifacts. The original method used by the authors of the paper for the topic modeling is through the use of the ‘mallet’ package in R. We have used ‘gensim’ package in python for the same. We make this change of library so as to replicate the study, as this is like changing the tools/environment in which analysis is done but it should not change the results.

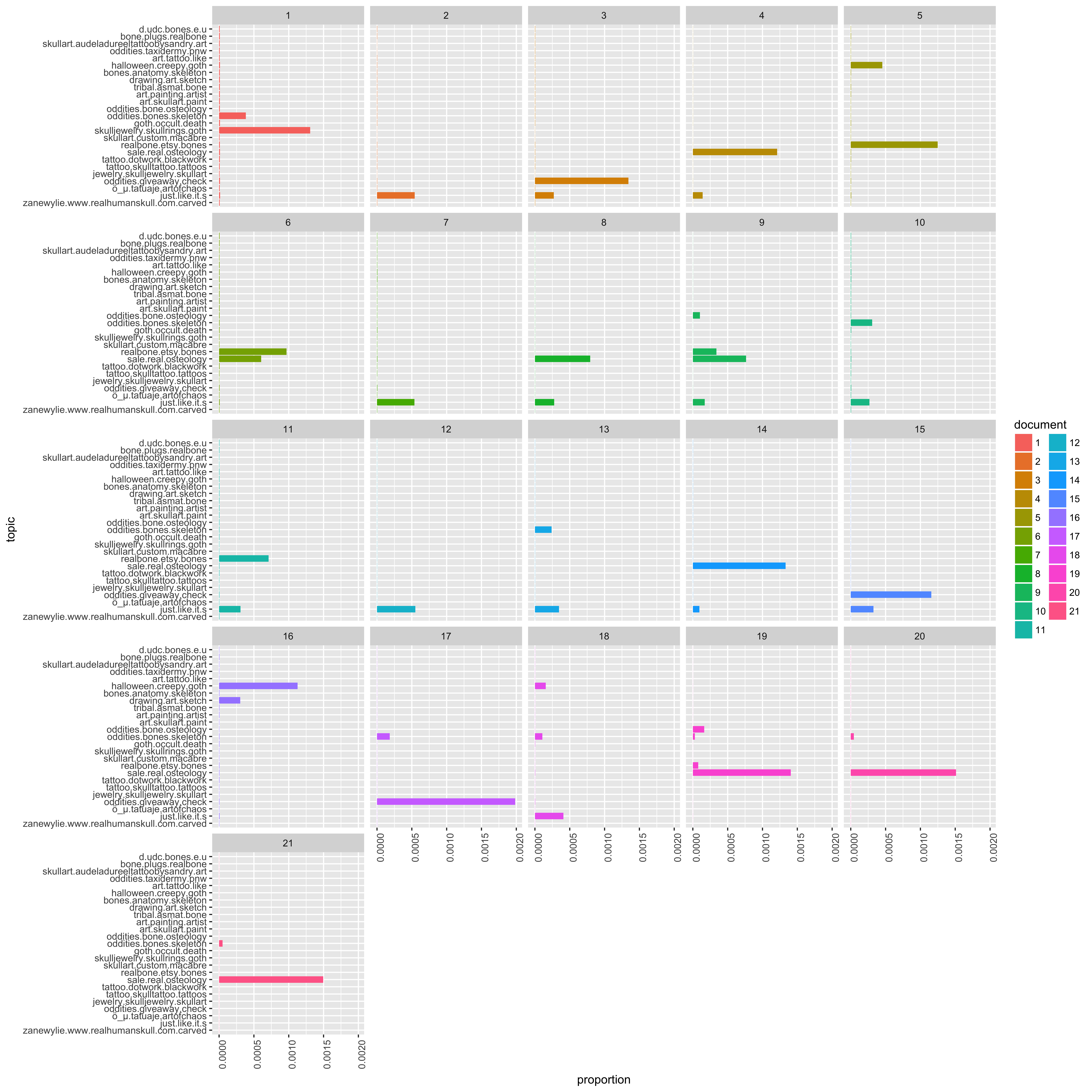


Figure 1: Original figure from (Huffer and Graham 2017)

# Replication Method

Figure 5: Topics within the posts of a single Instagram account, is related to topic modeling and was originally carried out using the ‘mallet’ package in R. We wanted to explore how topic-modeling was carried out in python. Python is a very powerful tool and performing a topic-model on it would be an important learning aspect. But the visualization powers of R is undeniable. So to club the strong aspects of both the languages, we process the data and bring it to a suitable form using python and then further using this result to create a visualization. We decided to use the ‘gensim’ package in Python to perform topic modeling and the ggplot library in R to draw the visualization. Steps: 1. We used the ‘reticulate’ package which helps us install python packages in R. The main package to be used is the python Gensim package. 2. We read in the csv file and removed stopwords as specified in the paper. 3. We converted texts from each of the Instagram posts into a vector of words and stored it in a variable called ‘data\_words’. We used gensim.utils.simple\_preprocess for this functionality. 4. We created a variable called ‘corpus’ that maps every word to id and calculates the term-document frequency. 5. We built an LDA model using the ‘gensim’ package and specified the parameter for the number of topics as 25. This model is named ‘lda’. The number of topics was selected to be 25 by the authors of the paper because they found that 25 was the ideal topic count where posts were easily being delineated between hobby collectors and professional sellers in the trade. 6. We used lda. lda.show\_topics() to generate the labels for the top 25 topics. 7. We used lda.get\_document\_topics() to generate a matrix of the documents and the corresponding probabilities of the 3-worded topics. 8. We wrote down our results in a preliminary CSV called ‘prob.csv’. Then we read the CSV with our R chunk, performing data type conversion and manipulation in the same way as in the original paper. 9. We select only a single username ‘234396855’ which is also the one used in the paper, for uniform comparison of 25 topics and top-3 topic words in each topic. 10. We used qqplot to approximate the visualization in R. 11. We also displayed all the words relating to a particular topic in R.

We kept the same criteria for stopwords and the number of topics as the original study. We followed the same analytical procedure. The ‘gensim’ package in Python is for topic modeling and has similar functions to the ‘mallet’ package used by the original study. Since we use the same dataset and methods and only change the topic modeling library i.e. the tools and the environment of the experiment, the result should be the same.

Below is the code we have used for our replication.

### Load stop words from en.text to process and clean the data  
  
stop\_words = pd.read\_csv('..\data\en.txt',header=None)  
stop\_words=stop\_words[0].to\_list()  
df = pd.read\_csv('..\data\posts-formatted-for-topicmodelling.csv', encoding='latin-1')  
df.head()

## Unnamed: 0 ... text  
## 0 1 ... I brought Trevor to work!!! Haha #humanbone #t...  
## 1 2 ... My first #humanbone ! Thanx @USERNAME\_HERE #go...  
## 2 3 ... My new torso came in today! I love it!! #human...  
## 3 4 ... Just got 2 femurs and a hand!!!! Yaaaay! #huma...  
## 4 5 ... Dead hooker mistle toe...  
##   
## [5 rows x 4 columns]

def sent\_to\_words(sentences):  
 for sentence in sentences:  
 yield(gensim.utils.simple\_preprocess(str(sentence),deacc=True)) #deacc = True removes punctuation  
  
data\_words = list(sent\_to\_words(df.text))  
#print(data\_words)  
#print([word for word in simple\_preprocess(str(data\_words[0])) if word not in stop\_words])  
#print( type(simple\_preprocess(str(data\_words[0]))))  
def remove\_stopwords(texts):  
 return[[word for word in simple\_preprocess(str(doc)) if word not in stop\_words] for doc in texts]  
data\_words\_nostops = remove\_stopwords(data\_words)  
#print(data\_words\_nostops)

## Create Corpus  
id2word = corpora.Dictionary(data\_words\_nostops)  
texts = data\_words\_nostops  
corpus = [id2word.doc2bow(text) for text in texts]

## Build model with 25 topics  
lda = gensim.models.ldamodel.LdaModel(corpus, num\_topics=25,id2word=id2word,minimum\_probability =0, random\_state=100)

The following are the topics that were generated from the model:

## ['skulljewelry.skullart.silver.skull.hand.carved.humanskull.gear.witch.ring', 'inktober.skull.humanskull.alternative.handmade.la.ilustracion.unique.bnw.occult', 'skull.humanskull.bones.oddities.human.skeleton.skulls.medical.anatomy.osteology', 'skull.human.humanskull.username\_here.ring.u\_.real.www.sale.ca', 'humanskull.skull.art.drawing.sketch.illustration.artist.blackandwhite.painting.artwork', 'humanskull.fire.skull.skulls.bones.witch.portland.humanskulls.pdx.naturalhistory', 'ideas.regrann.candles.people.studio.meet.big.sagittal.lost.remember', 'skull.humanskull.skulltattoo.tattoo.halloween.today.realism.tattoodesign.blackandgrey.black', 'humanskull.skull.texture.terror.contemporaryart.night.flowers.fb.halloween.woman', 'username\_here.humanskull.skull.amazing.humanhead.photography.skeleton.skulls.skullart.corpse', 'humanskull.skull.de.tattoo.skulltattoo.sugarskull.portrait.progress.human.tattooer', 'skull.humanskull.skulls.realhumanskull.human.username\_here.sale.real.skullforsale.collection', 'message.humanskull.skull.war.dayak.mask.indonesia.ossuary.graveyard.happyhalloween', 'pencil.skull.humanskull.sketchbook.skulldrawing.drawing.cosplay.realistic.study.art', 'fire.selfie.usd.humanskull.skull.oil.pirate.end.prints.blxckink', 'darkart.skull.crafty.harley.slaughterskulls.humanskull.aprilslaughter.scarylady.biker.skullart', 'skullduggery.cross.deerskull.soul.spirit.humanskull.haven.color.ocean.beauchene', 'humanskull.brain.green.graffiti.wax.page.kutnahora.skull.update.shows', 'dark.skulls.horror.oddities.macabre.humanskull.art.oddity.bones.skull', 'goth.gothic.skullrings.jewelry.skull.vampire.gold.humanskull.skulls.silver', 'reference.medicalhistory.loveskull.pictures.god.makes.thcentury.university.bear.feels', 'logo.axe.followme.bad.french.stunning.haha.harley.giving.stuff', 'halloween.humanskull.skull.spooky.code.follow.history.makeup.photos.scary', 'username\_here.tattoo.tattoos.humanskull.skull.skullring.skulltattoo.rings.ink.inked', 'art.skull.johnny.humanskull.paint.skullart.skulls.picoftheday.instagood.ne']

## Generate topic labels from top 3 words of the topic  
x=lda.show\_topics(num\_topics=25,num\_words=3,formatted=False)  
topics\_words = [(tp[0], [wd[0] for wd in tp[1]]) for tp in x]  
l=[]  
for topic,words in topics\_words:  
 l.append(".".join(words))

**R function to generate the replication image:**

prob\_matrix <- read.csv("../data/prob.csv", header=TRUE)  
require(reshape2)  
  
plot\_figure <- function() {  
 topic\_docs <- prob\_matrix  
 captionstext <- read.csv('../data/posts-formatted-for-topicmodelling.csv', stringsAsFactors = FALSE)  
## kludge for when username comesthrough as numeric rather than character  
 captionstext$username <-as.character(captionstext$username)  
  
 documents <- data.frame(text = captionstext$text,  
 id = make.unique(captionstext$username),  
 class = captionstext$year,   
 stringsAsFactors=FALSE)  
 topic\_docs<- t(topic\_docs[,-1])  
 names(topic\_docs) <- documents$id  
# find top n topics for a certain author  
 df1 <- t(topic\_docs[,grep("234396855", names(topic\_docs))])  
 #8963295 is a person who has 'for sale' in her post  
 #255766488 natural\_selections - skullshop.ca  
 #361451583 ryan matthew cohn  
 #234396855 pandora's box, York  
 topic.proportions.df <- melt(cbind(data.frame(df1),  
 document=factor(1:nrow(df1))),  
 variable.name="topic",  
 id.vars = "document")   
   
 # plot for each doc by that author  
 require(ggplot2)  
 dpi=600 #pixels per square inch  
 png("../analysis/figures/output\_figure.png", width=14\*dpi, height=14\*dpi, res=dpi)  
  
 print({  
 p <- ggplot(topic.proportions.df, aes(topic, value, fill=document)) +  
 geom\_bar(stat="identity") +  
 ylab("proportion") +  
 theme(axis.text.x = element\_text(angle=90, hjust=1)) +   
 coord\_flip() +  
 facet\_wrap(~ document, ncol=5)  
 })  
 dev.off()  
}

# Replicated Figures



Figure 2: Final results of the top 3 words of the 25 most popular topics

# Conclusion

Using the python package ‘gensim’ for performing the topic modeling and R for compiling the data and plotting the figure, we are able to reproduce the experimental data and reproduce the Figure 1 (Figure 5 in the paper) to a large extent. Both Figure 1 and Figure 5 in the paper show three keywords for each topic and the frequency of each topic for different documents. However, the selected keywords for the 25 topics in our reproduced figure are slightly different than those in the original one, although they share many overlaps, such as “art” “skull” “tattoo”. This can be attributed to the differences in the internal workings of the packages in R and python. In the original publication, the authors created a topic trainer object, loaded the documents into the object, trained the model, and normalized the results. In our replication project, we broke the texts in each document into words, mapped these words to Ids, and trained the LDA model. The differences probably lay in how the LDA model in Python and MalletLDA model in R were run, since the body of the texts is the same. We have displayed all the words related to the 25 topics to aid the comparison of our results to the original results. Although we used the same analysis procedure, a different library, and a different running environment, we do not get the exact same figure, but one which is in close resemblance to the figure in the original publication. The diversions are likely due to detailed executions of the models rather than differences in principles.

# 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>.