DATA 598: Replication Project

# Introduction

For examing the complexities in replication we chose the paper (Huffer and Graham 2017). This paper is a study the culture of trading human remains on Instagram. It uses several methods to automate data mining techniques to query the “noise” surrounding topics related to trade of human remains. We have chosen to replicate one such method in this project which involves topic modelling. 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 more deeper research and analysis into understanding the different trends and patterns associated with different aspects of trade such as buying and selling, sales prices, material being traded and other mechanics of trade.

# Replication Source

From the paper (Huffer and Graham 2017), we are replicating Figure 1 (Figure 5 in the paper). This figure shows different topics modelled 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 artificats.

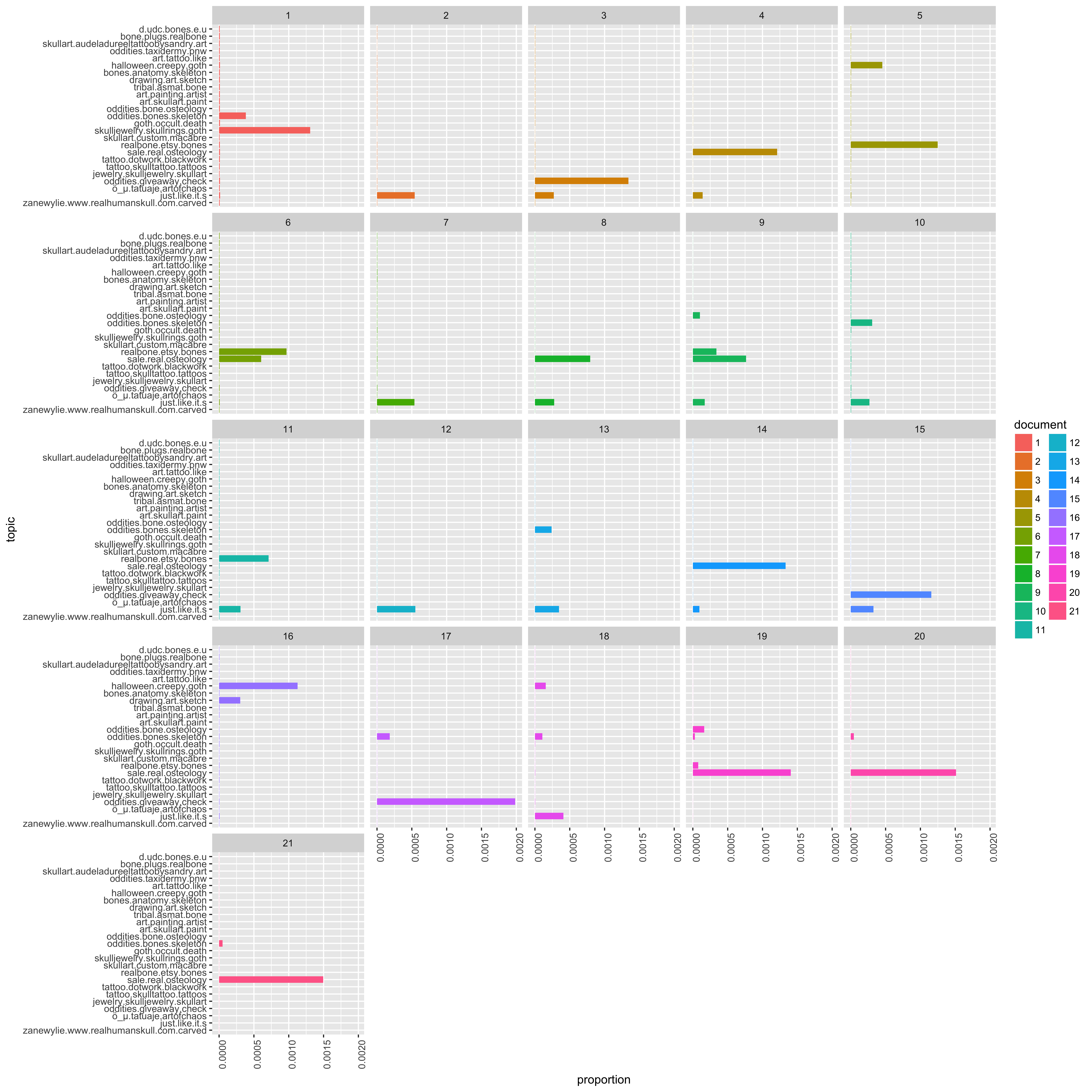


Figure 1: Original figure from (Huffer and Graham 2017)

# 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 decided to use the ‘gensim’ package in Python to perform topic modeling and replicate the figure. Firstly, we read in the csv file and removed stopwords as specified in the paper. Secondly, we converted texts from each of the instagram posts into a vector of words and stored it in a variable called ‘data\_words’. Thirdly, we created a variable called ‘corpus’ that maps every word to an id and calculates the term document frequency. Then 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’. Finally, we visualized the LDA model using pyLDAvis. Our replication uses different software and package, but it should not affect the overall result. 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 operating system, the result should be the same.

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)

# Replicated Figures



# Conclusion

# 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>.