Homework: Large Scale Social Media Extraction and Analysis for Pixstory
Due: Friday, April 7, 2023 12pm PT

---

1. Overview

Figure 1. Several posts and associated images from Pixstory

In the second assignment, you will build upon the work that you did in assignment 1 in identifying additional explanations and features that helped you understand the Pixstory social data, and whether posters were helping to uphold the values of the platform. You spent a lot of work in flagging whether the posts had hate speech information as identified by GLAAD, the ADL, and you looked at whether or not you could find hints of sarcasm in the posts. You also added features based on movie events and sports events overlapping with the dataset's time 2020 – 2023, and you tried to use features of the dataset including age, gender, interest, and posting date/time to slice the data and discover patterns and trends in the information. Further, beyond these 5 new columns, you were asked to select 3 datasets of 3 different top level MIME types and for each dataset, add 3 new features (total 9 new features) to your data. At the end of the new you added 14 new features to the data exploring the unintended consequences of Big Data, and the five 5 V's.

In this assignment you will focus on large scale content extraction and data science by exploring parts of the Pixstory dataset left unexplored in assignment 1. You will explore the unique properties of language in the dataset by deriving post language. As you noted in assignment 1, not all posts are the same language! For example, look at this post in Fig. 2.

Original Post (detected it): Pazzesco quanto accaduto al Vitality Stadium di Bournemouth. I padroni di casa sono stati battuti per 1-0 dal Boreham Wood, società semi professionista che milita nella 5ª divisione inglese, la National League. A siglare il gol decisivo è stato il trentasettenne Mark Ricketts che ha così estromesso dalla competizione la squadra di Championship. Tra meno di un mese, il Boreham Wood proverà a replicare l'impresa… questa volta in casa dell'Everton. Crediti foto: profilo twitter Bournemouth

Translated (to en): What happened at the Vitality Stadium in Bournemouth is crazy.\nThe hosts were beaten 1-0 by Boreham Wood, a semi-professional club in the 5th Division, the National League.\nThe decisive goal was scored by 37-year-old Mark Ricketts, who knocked the Championship team out of the competition.\nIn less than a month, Boreham Wood will try to replicate the enterprise... this time at Everton's home.\nPhoto credits: Twitter profile Bournemouth

Figure 2. Post text in Italian (top), and translated into English (bottom)

The post (top) originally is in Italian (ISO-639-2 language code: **it**). As we learned, many of the existing extraction methodologies assume English (ISO-639-2 language code: **en**), however if we had a methodology to automatically translate the post after detecting its language, we could turn all the posts into English and then further analyze them. Since the Pixstory dataset originally didn't provide a language column, we will add it (detected ISO-639-2 language, 2-character code as discussed in class) using Two methods: Tika's Language Detector along with the Google Langdetect Python Package. Additionally, since it would be useful to have the translated version of the post in English, we will leverage some recently developed software via IRDS and our recent USC graduate Dr. Thamme Gowda and his Ph.D. research in machine learning to automatically perform this translation. The software, called RTG (Reader Translator Generator) will be described in more detail later.

However, why stop there? There are plenty of **locations** mentioned in the translated (and even original) post text. As we have discussed in class during the advanced extraction lectures, as well as the metadata lectures, and clustering lectures, and as we will talk about during the Named Entity Recognition (NER) lecture, it is possible to use machine learning and natural language processing (NLP) to scan text and extract locations. For example, in this post (and its translation provided by RTG) we see the location, **India** mentioned in the translated text on the right of Fig. 3.

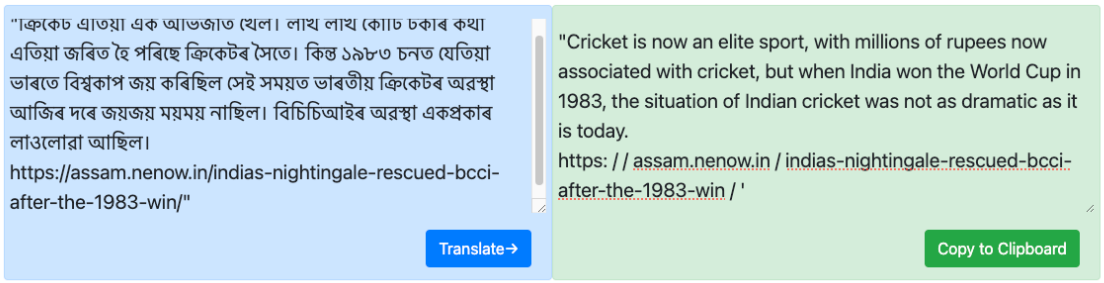

Figure 3. Post text in Bengali (left), and translated into English (right) by RTG. Notice the mention of location "India" in the text on the right.

Location names like "India", or "Los Angeles", and so on, can be geocoded, and their corresponding latitude and longitude can be extracted using a tool developed by the USC Data Science Group called GeoTopicParser. The tool can take some text, and then perform an analysis using the Geonames.org database, and custom NLP and NER parsing code, using Tika, generating output that looks like the following if the text contained in the post mentioned, e.g., "China":


```
[
   {
      "Content-Type":"application/geotopic",
      "Geographic_LATITUDE":"39.76",
      "Geographic_LONGITUDE":"-98.5",
      "Geographic_NAME":"United States",
      "Optional_LATITUDE1":"27.33931",
      "Optional_LONGITUDE1":"-108.60288",
      "Optional_NAME1":"China",
      "X-Parsed-By":[
         "org.apache.tika.parser.DefaultParser",
         "org.apache.tika.parser.geo.topic.GeoParser"
      ],
      "X-TIKA:parse_time_millis":"1634",
      "resourceName":"polar.geot"
   }
]
```


One last thing that we can do with the text, illustrating large scale content analysis is to run machine learning approaches on the Pixstory posting text to determine whether or not the post references so-called "toxic" content. We can do this in assignment 2 by leveraging machine learning, and the Detoxify library. Detoxify is a machine learning library based on Pytorch that will examine text and identify the level of toxicity, severe toxicity, obscenity, identity attacks, insults, threats, and the presence of sexually explicit content in the text. You can use Detoxify to compare against the flags you generated in assignment 1 from the ADL and GLADD and sarcasm screening words to see whether they correlate at all (their presence) to what Detoxify things of the posting text.

However, we're not stopping at the text, and in this assignment, you will take a look at the associated post image as well. For example, every post has a URL in it pointing at an image associated with the post. We can use a machine learning based Image Captioning algorithm called Show & Tell originating from Google to automatically caption and generate text features about the Pixstory posts and use these captions as additional features to describe the post. We will leverage two easy to use Tika Docker files to identify objects present in an image and to generate a textual (human readable) caption for the image. Both of these Docker Files are available in Apache Tika and they leverage Machine Learning and Deep Learning extraction techniques in particular Google's Tensorflow technology and custom Deep Learning models built in the USC IRDS group. You can see some examples of the Image Captioning and Image Object identification in action below in Figure 4a-c showing 3 automatically generated labels (with only generic training). We will integrate this Tika capability and generate labels and text captions for your Pixstory posts.

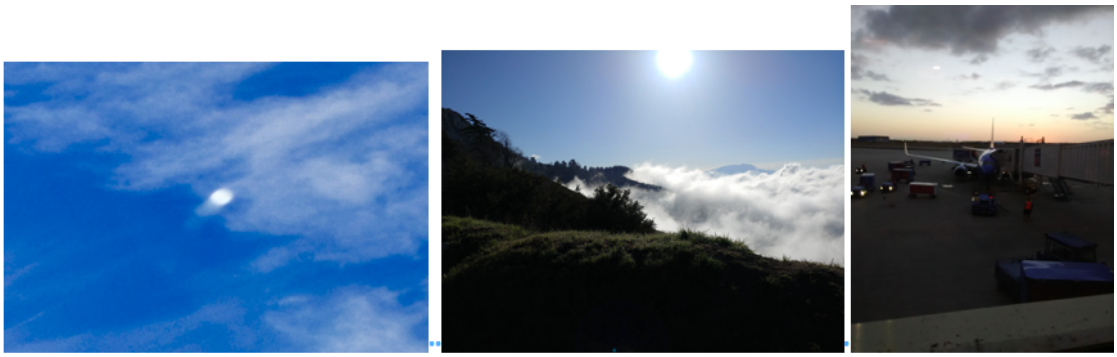

Figure 4: a) a light/orb shown in the daylight; b) an orb present against a mountain background; and c) an orb in a cloudy sky.

a plane flying in the sky over a field	*a view of a mountain range with a mountain in the background*	*an airplane is parked on the tarmac at an airport*

Machine Generated Labels for a). b). c)

The combination of these techniques will allow you to apply knowledge gained from the Parsing/Extraction Lecture, the lectures on advanced extraction (including Deep Learning and Metadata), and also topics discussed including Large Scale Content Extraction. In particular, please consider techniques discussed in class to embark on this assignment.

2. Objective

The objective of this assignment is two-fold. First, you will expose the richness of the posts and their associated languages by performing language analysis and generating the ISO-639-2 standard two-character language identification codes as we discussed in class. Exposing the language will help you in performing automated Machine Translation using the RTG neural machine translator built by the USC IRDS group. With the translated posts, you can identify location names using the GeoTopicParser and geocode the associated locations in the posts and then finally you can run Detoxify to automatically classify the toxicity of the posts. Ideally we would expect that users of the Pixstory platform are adhering to the stated goals of being a "clean" social media, and you will investigate that as part of the assignments. The location names will also aid you in comparing e.g., the posts, topics, and events, together from your prior assignment 1 work. In addition, using Detoxify will provide you with ground truth from which to compare the ADL, GLAAD and sarcasm screening flags you generated in assignment 1.

In addition, the other objective of this portion of the assignment is to leverage the richness in the underlying images associated with the post, along with large scale machine learning and data science, to generate text from the images (a caption), along with the list of objects present in the imagery. Both will provide additional features with which you can examine the posts, and compare: are these images capturing the essence of the tagged topic(s)? Are the images, captions, and objects in the images upholding the stated goals of the platform? Are there any trends in the image analysis with respect to age, gender, or posting dates and times? You will explore these questions in assignment 2.

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Generate a copy of your TSV v1 dataset. Call it "v2" or something similar. You will add your new columns for
    a. language identification by Tika
    b. language identification by Google LangDetect
    c. translated text from RTG
    d. GeoTopic name, along with associated lat/lng
    e. Image Caption generated by Tika's Show & Tell caption generator
    f. Detected objects in the image using Tika's Inception Rest service
2. Install Google's LangDetect using PIP and the instructions here https://pypi.org/project/langdetect/
3. Install RTG (Reader Translator Generator) using the instructions here https://gowda.ai/posts/2021/04/mtdata-nlcodec-rtg-many-english/ The result will be a REST service running on port 6000
4. Download and install Tika Python using PIP and the instructions at http://github.com/chrismattmann/tika-python
    a. With the RTG running, running tika's translate module will automatically work fine (since it will pick up the RTG server)
    b. The Tika language module (it's language detector) should also work fine
5. Install GeoTopicParser using the instructions here https://cwiki.apache.org/confluence/display/tika/GeoTopicParser
    a. The result of this should be the Lucene GeoGazetter REST server running as specified here: https://github.com/chrismattmann/lucene-geo-gazetteer
    b. You can connect the GeoGazetteer to Tika-Python using the instructions here: https://github.com/chrismattmann/tika-python#changing-the-tika-classpath
6. Install Detoxify using PIP and the instructions here: https://pypi.org/project/detoxify/
    a. Note that if you are using Mac and Python, using pyenv, and you run into issues installing Detoxify and torch with PIP, see this for an easy workaround https://github.com/pytorch/pytorch/issues/53601#issuecomment-967307449
7. Install Tika Image Dockers and generate captions for your Pixstory images posts
    a. To access the images, use the URL from the post and give it the URL prefix "/optimized", such as: https://image.pixstory.com/optimized/Pixstory-image-164416629024955.jpeg
    b. Download all 95k images associated with the posts
        i. Write a simple python script to do this
    c. Install Tika Dockers package for Image Captioning and Object Recognition
        i. git clone https://github.com/USCDataScience/tika-dockers.git and https://hub.docker.com/r/uscdatascience/im2txt-rest-tika
        ii. Read and test out: https://cwiki.apache.org/confluence/display/TIKA/TikaAndVisionDL4J
        iii. Read and test out: https://github.com/apache/tika/pull/189
    d. Iterate through all the Pixstory posts and add the generated image caption and the detect object(s) column to your dataset
8. Iterate through all the Pixstory posts and detect their ISO-639-2 two character language code using both Tika's language identifier and Google LangDetect/Python
    a. Write a Python program to do this
    b. Add the two new columns to your dataset
9. Iterate through all the Pixstory posts and use RTG to automatically translate all posts to English

    a. Write a Python program to do this
    b. Add the new column for the translated post text to your dataset
10. Iterate through all the Pixstory posts and then run Tika GeoTopicParser and extract out Location name, including Lat/Lng
    a. Write a Python program to do this
    b. Add the new column(s) to your dataset
11. Run Detoxify on all the Pixstory posts and generate scores for
    a. Toxicity
    b. Severe Toxicity
    c. Obscenity
    d. Identity Attack
    e. Insult
    f. Threat
    g. Sexual Explicit
    h. Add these columns and scores to each of your posts in your new dataset

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized at **minimum 2, and maximum 6**. You may reuse your existing groups from discussion in class. If you have any questions, contact the TA via his email address with the subject:
DSCI 550: Team Details.

5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. For example, the following questions are of interest.

1. Are there any age or gender or topic based correlations by location in the posts?
2. What is the most prevalent language in the posts, and least prevalent?
3. Is there a correlation between post language and identified mentioned locations?
4. Are there correlations between the sporting events, or the entertainment events with locations?
5. Do the Detoxify scores and associated GLAAD and ADL or sarcasm flags line up? Is there any relationship between the flags and the identified Detoxify scores?
6. Do the image captions accurately represent the image?
7. Are the identified objects present in the image described in the original post and/or the generated caption?
8. Are there any age, or gender specific trends you see in the text captions or identified objects in the image media?

Also include your thoughts about the ML and Deep Learning software like RTG, GeoTopicParser, Detoxify, LangDetect, Tika Image Captioning, etc. – what was easy about using it? What wasn't?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail dsci550.sp2023@gmail.com. Use the subject line: DSCI 550: Mattmann: Spring 2023: EXTRACT Homework: Team XX. So if your team was team 15, you would submit an email to dsci550.sp2023@gmail.comwith the subject "DSCI 550: Mattmann: Spring 2023: EXTRACT Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts and other notes and a readme file.
- Include your updated dataset TSV. We will provide a Dropbox or Google Drive location for you to upload to.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries other than Tika Python, you should include those files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_XX_EXTRACT.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  TEAM_XX_DSCI550_HW_EXTRACT.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds the Gmail's 25MB limit, upload the zip file to Google drive and share it with dsci550.sp2023@gmail.com.

Important Note:

- Make sure that you have attached the file the when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof