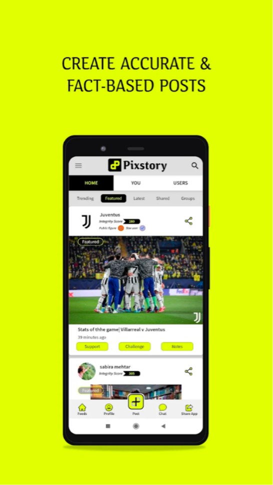
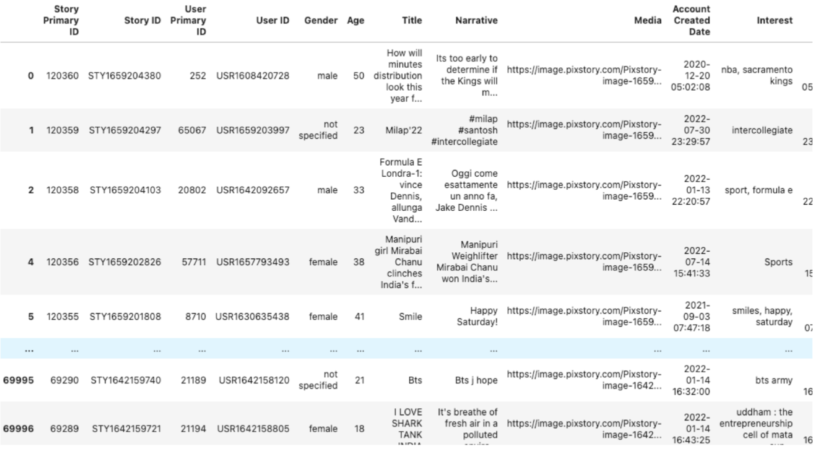
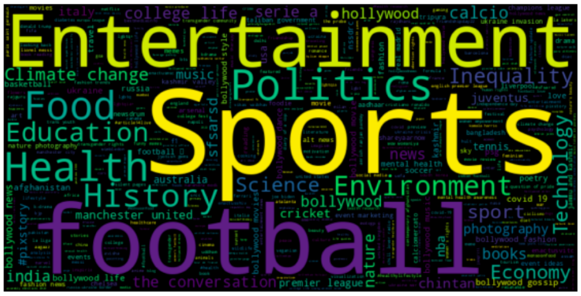
Homework: Analysis of the Newest Social Media Dataset Focused on "Clean" and Less Toxic Social Media: The Pixstory Dataset
Due: Friday, March 10, 2023 12pm PT

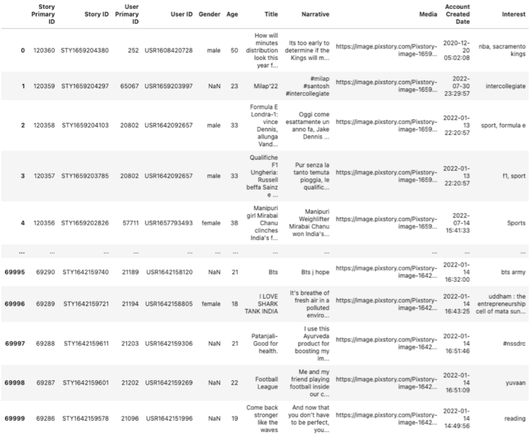
1. Overview





	Story Primary ID	Story ID	User Primary ID	User ID	Gender	Age	Title	Narrative	Media	Account Created Date	Interest	
0	120360	STY1659204380	252	USR1608420728	male	50	How will minutes distribution look this year f...	Its too early to determine if the Kings will m...	https://image.pixstory.com/Pixstory-image-1659...	2020-12-20 05:02:08	nba, sacramento kings	05
1	120359	STY1659204297	65067	USR1659203997	not specified	23	Milap'22	#milap #santosh #intercollegiate	https://image.pixstory.com/Pixstory-image-1659...	2022-07-30 23:29:57	intercollegiate	23
2	120358	STY1659204103	20802	USR1642092657	male	33	Formula E Londra-1: vince Dennis, allunga Vand...	Oggi come esattamente un anno fa, Jake Dennis ...	https://image.pixstory.com/Pixstory-image-1659...	2022-01-13 22:20:57	sport, formula e	22
4	120356	STY1659202826	57711	USR1657793493	female	38	Manipuri girl Mirabai Chanu clinches India's f...	Manipuri Weighlifter Mirabai Chanu won India's...	https://image.pixstory.com/Pixstory-image-1659...	2022-07-14 15:41:33	Sports	15
5	120355	STY1659201808	8710	USR1630635438	female	41	Smile	Happy Saturday!	https://image.pixstory.com/Pixstory-image-1659...	2021-09-03 07:47:18	smiles, happy, saturday	07
...	...	...	...	...	...	...	...	...	...	...	...	
69995	69290	STY1642159740	21189	USR1642158120	not specified	21	Bts	Bts j hope	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:32:00	bts army	16
69996	69289	STY1642159721	21194	USR1642158805	female	18	I LOVE SHARK TANK	It's breathe of fresh air in a polluted	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:43:25	uddham : the entrepreneurship cell of mata	16





	Story Primary ID	Story ID	User Primary ID	User ID	Gender	Age	Title	Narrative	Media	Account Created Date	Interest
0	120360	STY1659204380	252	USR1608420728	male	50	How will minutes distribution look this year f...	Its too early to determine if the Kings will m...	https://image.pixstory.com/Pixstory-image-1659...	2020-12-20 05:02:08	nba, sacramento kings
1	120359	STY1659204297	65067	USR1659203997	NaN	23	Milap'22	#milap #santosh #intercollegiate	https://image.pixstory.com/Pixstory-image-1659...	2022-07-30 23:29:57	intercollegiate
2	120358	STY1659204103	20802	USR1642092657	male	33	Formula E Londra-1: vince Dennis, allunga Vand...	Oggi come esattamente un anno fa, Jake Dennis ...	https://image.pixstory.com/Pixstory-image-1659...	2022-01-13 22:20:57	sport, formula e
3	120357	STY1659203785	20802	USR1642092657	male	33	Qualifiche F1 Ungheria: Russell beffa Sainz e ...	Pur senza la tanto temuta pioggia, le qualific...	https://image.pixstory.com/Pixstory-image-1659...	2022-01-13 22:20:57	f1, sport
4	120356	STY1659202826	57711	USR1657793493	female	38	Manipuri girl Mirabai Chanu clinches India's f...	Manipuri Weighlifter Mirabai Chanu won India's...	https://image.pixstory.com/Pixstory-image-1659...	2022-07-14 15:41:33	Sports
...	...	...	...	...	...	...	...	...	...	...	...
69995	69290	STY1642159740	21189	USR1642158120	NaN	21	Bts	Bts j hope	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:32:00	bts army
69996	69289	STY1642159721	21194	USR1642158805	female	18	I LOVE SHARK TANK INDIA	It's breathe of fresh air in a polluted enviro...	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:43:25	uddham : the entrepreneurship cell of mata sun...
69997	69288	STY1642159611	21203	USR1642159306	NaN	21	Patanjali-Good for health.	I use this Ayurveda product for boosting my im...	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:51:46	#nssdrc
69998	69287	STY1642159601	21202	USR1642159269	NaN	22	Football League	Me and my friend playing football inside our c...	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 16:51:09	yuvaan
69999	69286	STY1642159578	21096	USR1642151996	NaN	19	Come back stronger like the waves	And now that you don't have to be perfect, you...	https://image.pixstory.com/Pixstory-image-1642...	2022-01-14 14:49:56	reading

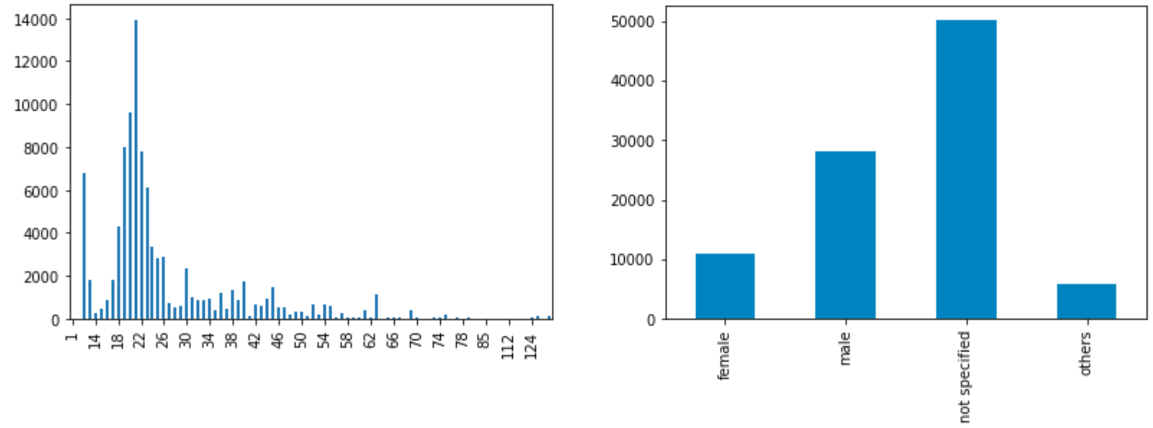
Figure 1: The PixStory application: creating accurate and fact-based posts and associated dataset of 95,000 rows x 11 columns and the most frequently occurring topics in the dataset show as a word cloud (bottom left).

In this assignment we will explore several of the topics discussed in the early portion of class – Big Data – MIME types and their taxonomy – Data Similarity – and so forth. To do this, we will leverage the dataset highlighted in Figure 1 – a set of 95,000 user posts on the new "clean" Social Media app, Pixstory, dating from January 2020 to December 2022. The posts contain several features which are highlighted below:

- Story ID (a numerical identifier for the user "story". Stories are a "feed" of posts for the user associated with a Story Primary ID. Each Story Primary ID can have many posts).

- Story Primary ID (a string readable story that can be used to aggregate Stories)
- User Primary ID (a string readable user ID that can be used to aggregate user posts)
- User ID (a numerical user identifier that uniquely identifies the user)
- Gender (Male, Female, Unspecified / null, Other)
- Age (Age of the person making the post)
- Title (the title of the posted Story, or post within the story)
- Narrative (the text of the post within the story, analogous to e.g., the text body of an Instagram or Facebook post)
- Media (a URL pointer to the associated media with this post)
- Account Created Date (the date that the user made the post from this particular account)
- Interest (a string separated list of interests associated with the post; users must identify what topic or interest their post is about)

The Pixstory dataset is a rich dataset with high variation in its features and properties. For example, as can be seen from Fig 1., most users in the dataset are between the ages of 14 and 26, and there are approximately three times as many males as females in the dataset's population with the remaining genders unspecified or other. Given Pixstory's stated goals of appealing to younger users, and with a strong focus on recruiting minority and other users to the platform, we can leverage this information to study the effectiveness of the platform.

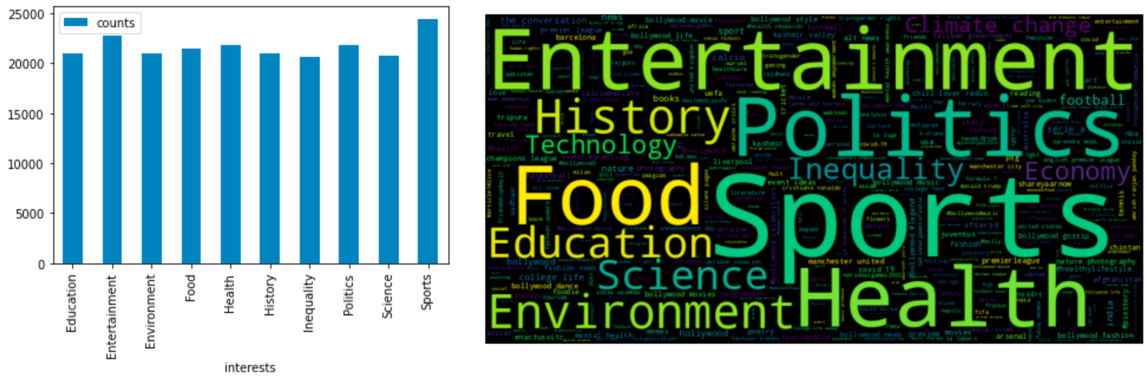


User Age Distribution in Pixstory Dataset

Gender Distribution in Pixstory Dataset

Figure 1. User age and gender distribution based on posts in the Pixstory dataset.

One of the other important elements of Pixstory is that it is focused on clean social media, and not on toxic energy or topics that plague the other platforms. The goal eventually for Pixstory is to minimize hate speech on the platform and instead focus on positive energy and discussions. This can be seen when examining the key topics present in the dataset, the top 10 of which are shown below in Figure 2 and in text form.



Top 10 Interests shown as histogram and Word Cloud in Pixstory Dataset

Figure 2. The most frequently posted about topics in the Pixstory dataset (also shown in Word Cloud form on the right)

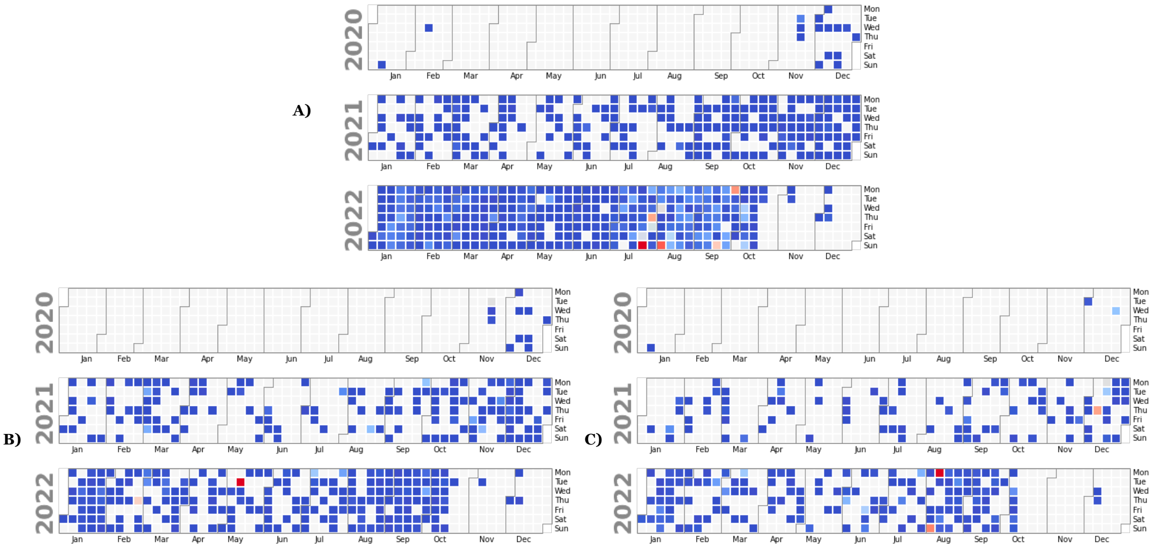
The largest discussed topics include sports, entertainment, politics, health, food, environment, history, education, science, and inequality. Given the age breakdown in terms of the data, and genders, these topics can help us decipher clues and questions about the users posting on the platform and their ancillary and other interests. For example, we can use this information to explore: do younger users prefer topics related to sports, and is there a breakdown between age, and gender related to this? Are older users focused on health topics, such as the pandemic, or COVID? Who is concerned with food, and the environment?

Additional insight can be gleaned from exploring the temporal properties of the user's posts on the platform. Are there particular days of the week when users' posts? What happens if you slice it down by gender, or age for example? Examine Figure 3 closely to determine if you can see any patterns.

2. Objective

Exploring the Pixstory dataset, some things may jump out at you. First off, Mondays before Summer 2022 never tend to be a good time to post. What might be the reason for that? Diving down into the topics, one of the things that may jump out is that the user's on this platform are very interested in sports, and in particular, Soccer (or Futbol). Looking at world events there were a number of matches related to the Arsenal soccer team that took place during that time. Could that have been the reason for the increase in posts? Additionally, given the age distribution and gender distribution of posts, what other things can you glean, even from the posts themselves? Are there particular topics and trends that make sense given the user population on the platform? Is the platform living up to its stated goals of being "clean social media"? And promoting discussion surrounding climate change, social justice and equality and other topics?

What ways could you explore this, by looking beyond this rich dataset and by applying lessons learned from class thus far where we have been studying the 5 V's, MIME types of associated datasets, and large datasets and their characteristics? Could you join for example, a dataset related to major international sporting events to this data? There are many such datasets such as: https://www.topendsports.com/events/games/list.htm for all events, past and present, or for example https://www.topendsports.com/events/calendar-2022.htm for 2022? Furthermore, could you also look at major entertainment events such as film festivals and their timelines, such as this dataset https://www.film-fest-report.com/home/film-festivals-2022? What trends emerge? Do entertainment posts increase when these festivals happen, but perhaps, sporting events take precedence because they are not coincident with the same events? Can you deduce age, or gender as a factor influence number of posts, posting times or content?



Temporal breakdown / heat map of posts for the Pixstory dataset. All 95k posts shown in part A) (top portion) broken down from January 2020 - Dec 2022. On the bottom left, part B) shows the temporal breakdown by gender **male** and the right portion of the diagram C) shows the temporal post breakdown by gender **female**

Figure 3. Temporal breakdown of Pixstory data - overall A), and by gender male B), and gender female C)

What can you learn about the platform's commitment to posting positive messages and keeping online dialogue posting and staying away from hate speech. Could you do a simple scan of the user posts and try to flag whether there is hate speech or not by examining whether any of the words from GLAAD's hate speech list appear https://www.glaad.org/hate-speech-listing or perhaps you could also look at this dataset from the ADL https://www.adl.org/resources/hate-symbols/search There are many such lists of these terms that you could flag the dataset with regarding the posts. Finally, you may want to take a look at the Sarcasm corpus (https://nlds.soe.ucsc.edu/sarcasm1) and try to mitigate when sarcasm is being used compared to actual hate speech! This is a difficult problem of course to solve generally but we are interested in simply collecting these features, not deciding on the actual speech itself. What other features can you think of?

You will choose at least three additional publicly accessible datasets along these lines to join the Pixstory to, and you must add at least three new features per dataset that you join. The datasets you select may not all belong to the same MIME top level type – that is –

you must pick a different MIME top level type for each of the three datasets you are joining to this Pixstory dataset.

Once the data is joined properly, you will explore the combined dataset using Apache Tika and an associated Python library called Tika-Similarity. Using Tika Similarity, you can evaluate data *similarity* (as discussed during the Deduplication lecture in class; and also during data forensics discussions). Tika similarity will allow you to explore and test different distance metrics (Edit-Distance; Jaccard similarity; Cosine similarity, etc.). And it will give you an idea of how to cluster data, and finally it will let you visualize the differences between different clusters in your new combined dataset. So, you can figure out how similar posts and users are, given their ages, genders, posting times, and other features, and explore your new augmented Pixstory dataset. For example, you may ask, how many posts about the topic of Arsenal came when the team was playing and if the posts contained any hate speech, or simply sarcastic or positive statements, and did gender or age have any influence on the posts or outcomes?

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

3. Tasks

1. Download and install Apache Tika
    a. Chapter 2 in your book covers some of the basics of building the code, and additionally, see https://tika.apache.org/2.7.0/gettingstarted.html
    b. Install Tika-Python, you can pip install tika to get started.
        i. Read up on Tika Python here: http://github.com/chrismattmann/tika-python
2. Download and install D3.js
    a. Visit http://d3js.org/
    b. Review Mike Bostock's Visual Gallery Wiki
        i. https://github.com/mbostock/d3/wiki/Tutorials
3. Download the Pixstory dataset
    a. We will provide you a Dropbox link in Slack for each validated team
    b. Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
4. Create a combined TSV file for your Pixstory dataset
    a. First, take each of the independent CSVs and combine them (on UNIX you can simply use the cat command, e.g., cat *.csv > combined.csv or on Windows you can use copy see: https://www.tomnash.eu/how-to-combine-multiple-csv-files-into-one-using-cmd/)
    b. Convert the CSV to TSV (here's a simple example of how to do this with Python https://unix.stackexchange.com/questions/359832/converting-csv-to-tsv)
5. Add and expand the dataset with the following features
    a. Associated sporting events – add any overlapping sporting events corresponding to the date which the account made the post. See https://www.topendsports.com/events/games/list.htm for all events, past

and present, or for example https://www.topendsports.com/events/calendar-2022.htm

   b. Associated entertainment film festivals, see https://www.film-fest-report.com/home/film-festivals-2022 and try to find similar lists for 2020 and 2021.
   c. Add a flag for hate speech detected from either the https://www.glaad.org/hate-speech-listing and the ADL https://www.adl.org/resources/hate-symbols/search so two features in total. Compare the text of the post with the filter lists and flag the post as containing that speech or not.
   d. Add a flag for sarcasm by comparing the post text to https://nlds.soe.ucsc.edu/sarcasm1. Compare the text of the post with the sarcasm corpus and add a flag for detected or not.

6. Identify at least three other datasets, each of different top level MIME type (can't all be e.g., text/*)
   a. Check out places including: https://catalog.data.gov/dataset (Data.gov)
   b. For each dataset, develop a Python program to join the data to your new Pixstory dataset
      i. For each non text/* dataset, be prepared to describe how you featurized the dataset
   c. Each dataset that you join must contribute at least three features (in addition to the features you are adding described in part 5)
   d. For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature
7. Download and install Tika-Similarity
   a. Read the documentation
   b. You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)
   c. Convert the TSV dataset into JSON using Tika Similarity's tsv2json tool
   d. Compare Jaccard similarity, edit-distance, and cosine similarity
      i. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?
   e. How to the resultant clusters generated highlight the features you extracted? Be prepared to identify this in your report
8. Package your data up by combining all of your new JSONs with additional features into a single TSV (tab separated values) file where the columns represent the features and the rows are the data samples.
9. (**EXTRA CREDIT**) Add some new D3.js visualizations to Tika Similarity
   a. Currently Tika Similarity only supports Dendrogram, Circle Packing, and combinations of those to view clusters, and relative similarities between datasets
   b. Consider adding
      i. Feature related visualizations, e.g., time series, bar charts, plots
      ii. Add functionality in a generic way that is not specific to your dataset
      iii. See gallery here: https://github.com/d3/d3/wiki/Gallery

iv. Contributions will be reviewed as Pull Requests in a first come, first serve basis (check existing PRs and make sure you aren't duplicating what some other group has done)

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. Please fill out the group details in the Google Form provided after class on Thursday, February 9th. Only one form submission per team. If you have any questions, contact the TA via his email address with the subject:
DSCI 550: Team Details.

4.2 Pixstory dataset

Access to the data is provided by a Dropbox link. The dataset itself is approximately 37.6Mb zipped and 100 Mb unzipped. You may want to distribute the data between your team-mates since the data is fairly small (for now).

4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the `tika-app.jar` from: http://tika.apache.org/download.html. You should obtain a jar file called `tika-app-2.7.0.jar`. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/2.7.0/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://cwiki.apache.org/confluence/display/TIKA/TikaServer. The Tika Python library is a robust REST client to the Java-side REST API.

You can also get more information about Tika by checking out the book written by Professor Mattmann called "Tika in Action", available from: http://manning.com/mattmann/.

5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the Pixstory data and its users, posts, topics, and features previously unanswered? What clusters were revealed? What similarity metrics produced more (in your opinion) accurate measurements? Why? What did the additional datasets suggest about "unintended consequences" related to users posting on the platform and the stated goals of the platform? You should also clearly explain which datasets you used to join the Pixstory data and how you extracted the new features from each dataset.

Thinking more broadly, do you have enough information to answer the following:

1. Are there clusters of users with similar features that tend to post about the same topics?
2. Does the time of day of the post matter?
3. Are specific ages or genders of the users more likely to post about specific topics?
4. Are specific ages or genders of the users more likely to post any hate speech? Did you detect any?
5. Is there a set of frequently co-occurring features that define a particular user or class of user?
6. What insights do the "indirect" features you extracted tell us about the data?
7. What clusters of users and/or posts made the most sense? Why?

Also include your thoughts about Apache Tika – 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: BIGDATA Homework: Team XX. So if your team was team 15, you would submit an email to dsci550.sp2023@gmail.com with the subject "DSCI 550: Mattmann: Spring 2023: BIGDATA 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 that you used to join three other datasets, and what you used to extract additional features.
- 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 and Tika Similarity, you should include those jar 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_BIGDATA.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_BIGDATA.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.sp23@gmail.com.

Important Note:

- Make sure that you have attached the file 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