Hum ETM Capstone Update – 02/13

Griffin McCauley, Eric Tria, Theo Thormann, Jake Weinberg

A Process of Iteration

- Completed data processing and cleaning of entire 2022 dataset
- Successfully built, trained, and evaluated a baseline sequence classifier model for the most recent period of user engagement
- The model produced an accuracy of ~67% on the testing set, but the model's room for improvement seems capped by the quality and content of the available data
- Given the nature of collection and the way events are triggered, patterns in the sequence proved to be more difficult more the model to discern than expected
- Having discovered this, we were prompted to reevaluate the feature selection process and consider how the model could be adapted

Enhanced Feature EDA

Additional Attributes

- Since events and their time stamps do not seem robust enough, we began exploring additional features to incorporate
- Want to extract more information on content type, journal origin, article topics, and user identification

Relevant features uncovered

- Content Type (in Meta)
- URL (for articles)
- Content ID (for journal)
- Set User (for identification)
- Referer (for source)

Count Feasibility EDA

- The notebooks for this EDA can be found on our GitHub at the link below
- [EDA in Github]
- The following slides will also highlight the major findings as well

Enhanced Feature EDA

Event Meta Table: Content Type

```
META_NAME
content_type    1068970
day             2011605
description     1886980
image            279918
referer         2011605
tags             642979
title           1886909
utm_campaign      10174
utm_content       10174
utm_medium        10174
utm_source        10174
utm_term          10174
Name: ID, dtype: int64
```

```
META_VALUE
journal_article     1068966
microsite_home            4
Name: ID, dtype: int64
```

- Total Events: 10,863,469
- Journal Articles: 1,068,966
- Microsite Home: 4

Content Table: Content Type

	TYPE	EVENTS
0	issue	125641
1	journal_article	8012589
2	account_management	179789
3	search	197223
4	None	1623851
5	in-brief	39
6	self-serve	129483
7	cross-ref-citation	12
8	microsite_home	594842

Enhanced Feature EDA

Event Table: Referer

	REFERER_GROUP	EVENTS
0	OTHER	4166569
1	GOOGLE	2933301
2	PUBMED	1428813
3	RUPRESS	1355904
4	GOOGLE SCHOLAR	978882

Event Table: URL

	URL_TYPE	EVENTS
0	ARTICLE	9516546
1	OTHER	1346923

Enhanced Feature EDA

Event Table: Tags

	TAG	EVENTS
0	"mice"	2216857
1	"t-lymphocytes"	996746
2	"tissue membrane"	698284
3	"signal transduction"	619142
4	"antibodies"	585891
5	"neoplasms"	581571
6	"infections"	499700
7	"genes"	425608
8	"hum_immunopathogenesis"	414480
9	"mitochondria"	365549
10	"actins"	359561

Content Keyword Table: Keywords

	KEYWORD	EVENTS
0	mice	2500110
1	None	2125695
2	t-lymphocytes	1133698
3	tissue membrane	768383
4	signal transduction	703489
...	...	...
23662	cochlear implants	1
23663	phosphorylases	1
23664	confidence interval	1
23665	supraoptic nucleus	1
23666	hum_eleetrophoresis	1

Reconceived Model

Motivation

- Had hoped to identify dropout risk based off of most recent engagement cycle
- Since it turns out the data is not rich enough to support this approach, we seek to classify users in a more objective and feature-driven manner
- Can delineate users into two types: subscribed and anonymous based on Set User feature
- By using the previously processed sequence data in conjunction with newly found features, we aspire to predict currently unidentified users who are likely to subscribe and engage more intently

Engagement Through Subscription

- Based on a user's first X number of events, can we predict whether they are more similar to full-time subscribers or cyclical binge users
- Adapt current subsequence data to be fed in as just one input to a larger model
- Bring in features such as counts for each of the different journals engaged with, how many cycles they have performed, how many unique articles have been visited, what the primary referers are (e.g. Google, PubMed, etc.), and how similar engagement sequence is to subscribers
- These derived features can then be used as inputs to an MLP classification model
- Offers robust support for adaptation and tuning

Next Steps

- Define the exact derived features we need to pull from the data warehouse
- Clean and organize the newly identified additional data (can now be done more easily in parallel)
- Build a basic MLP architecture to perform classification given our redefined feature space
- Train and tune the model
- Determine if the sponsor would like any additional criteria included to make the transition easier if they hope to adapt the model for Reviewer Recommendation