**Project Info**

* We will be investigating digitized periodicals that were published between 1770-1835 to investigate the culture of litigation.
* Periodicals are popular texts that not only contain *what* topics many British people were talking about, but also *how* they talked about it.
* We are specifically looking at the culture of litigation, the set and patterning of expressions used to describe British law.
* To create presentations of the collection of British Periodicals that could prove useful to academic researchers such as Mark Schoenfield.

**Resources**

* Proquest *-* Vanderbilt subscribed to Proquest to gain access the their British Periodicals Database
* Apache Spark *-* software built on Hoodoo, that uses parallelization technique to distribute data across multiple systems (making it much easier to work large datasets)
* Allows computation to be done in memory as opposed to on a disc drive
* Databricks - an application that allows us to write code Python as opposed to Scala (what Spark is written in) - PySpark

**Timeline**

Week 1 and 2:

Set up access to Databricks

Introduction to Standard Query Language

Started exploring British Periodical Archive

-------------------------------------------------------------------------------------------------------------------------------

Week 3:

Learn difference between starter and production endpoint

Examine new “century” tables taken from BP archive (1750 - 185)

Introduction to Regular Expressions (RegEx)

* centurymaintexts\_avro - No advertisements, front or back material
* centuryadverts - advertisements only
* Metadata- collection of famous books
  + Boccaccio’s Fiammetta
  + Dante’s Di
  + vine Comedy
  + Hyacinth and The Green Mouse
  + The Masque of Red Death

Challenge Problem: “Levels of Aggregation”

Write and SQL expression that lists:  
1) Every journal that published between 1840 and 1850

2) The number of articles that journal published

-------------------------------------------------------------------------------------------------------------------------------

Week 4:

Building visualizations that make sense

* Cliff showed **box-and-whisker** plot depicting the “lives” of Journals in terms of their publication numbers
* Pie chart of the distribution of journal types

Explore different visualizations

Parameterization - How to build interactive dashboards

Next Problem: How to build a dashboard out of multiple queries?

-------------------------------------------------------------------------------------------------------------------------------

Week 5:

* Cliff showed how to add an parameterized query to a dashboard
* Investigated Amy’s Interactive Query

Week 6:

* Presentation of queries
* Check out Advertisements
* Design and refine new query
* Check in with Ali

Week 7:

We presented each of our queries

* I presented a query “Average WordCount Per Article”

Suggestions:

* Check out advertisements (see if there are any interesting questions)
  + Check in with Ali

Learn how to create a dashboard (some struggles with parameterization)

Week 8:

Learned how to create multi-value query parameter

* example with Television News Archives

Learned how to create a fork of a query

Learned how to count occurrences of word

* Replace instances of a word with ‘ ‘

Week 9/10:

Create a group dashboard

ADD a select all box to query and set this as default!

**Quick Links**

*Archive*

Databricks sign-on [Login - Databricks](https://vanderbilt-vube.cloud.databricks.com/login.html?next_url=%2Fsql%2Feditor%2Fca358599-1598-4a8e-acbe-998e8506114d%3Fo%3D2267446239642753%26quickstart-table%3Ddefault.bp17701850)

Proquest [British Periodicals](https://www.proquest.com/britishperiodicals/history/fromDatabasesLayer?accountid=14816)

Fellowship - [Context](https://www.vanderbilt.edu/datascience/2022/01/18/analyzing-british-periodicals-to-understand-legal-discourse/)

*SQL*

SQL Basics Cheat Sheet - [Letter SQL Basics Cheat Sheet (learnsql.com)](https://learnsql.com/blog/sql-basics-cheat-sheet/sql-basics-cheat-sheet-letter.pdf)

Databricks SQL Release Notes: <https://docs.databricks.com/sql/release-notes/index.html>

Socratica SQL Youtube Series - [youtube.com/socratica](https://www.youtube.com/watch?v=nWyyDHhTxYU&list=PLi01XoE8jYojRqM4qGBF1U90Ee1Ecb5tt)

ReGex = <https://regexr.com/>

RLike - <https://www.sqlines.com/mysql/regexp_rlike>

[Adding Parameters to Queries](https://docs.databricks.com/sql/user/queries/query-parameters.html) - Databricks

Creating [SQL Functions](https://docs.databricks.com/spark/latest/spark-sql/language-manual/sql-ref-syntax-ddl-create-sql-function.html)

*Articles*

The Epicurean - <https://en.wikipedia.org/wiki/The_Epicurean>

Funny [Comic](https://www.proquest.com/britishperiodicals/docview/1827351003/CD538C8944D14DC9PQ/1?accountid=14816&imgSeq=1):

*Digital Humanities*

<https://carletonu.pressbooks.pub/digh5000/chapter/chapter-4-text-analysis/>