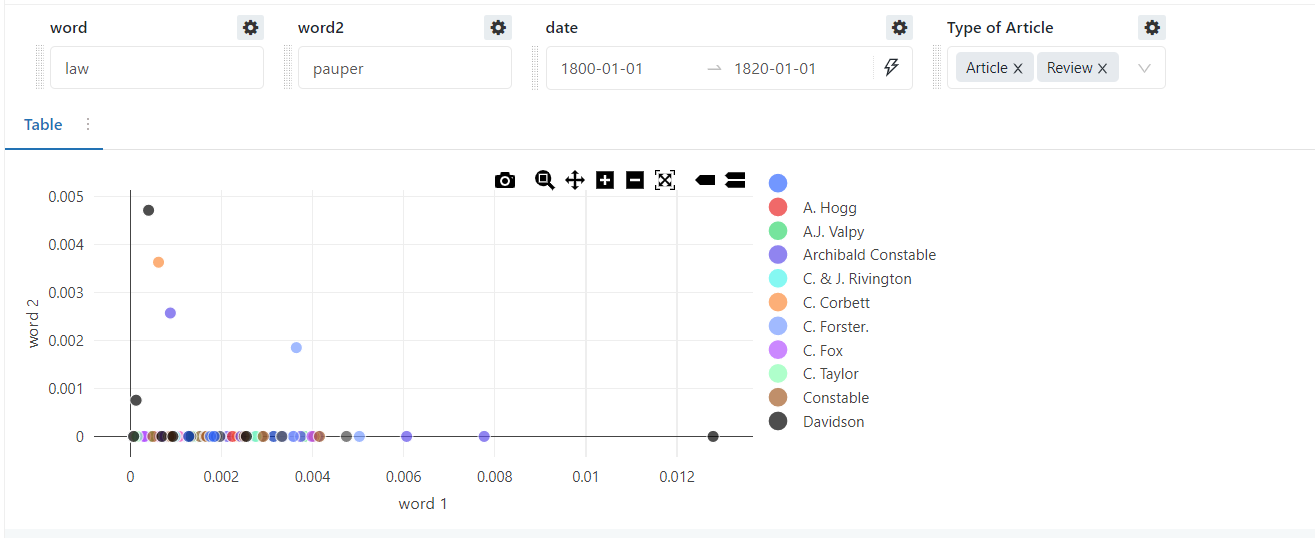
I want to create a visualization that meaningful represents the entire, or a subset of the British Periodicals collection (shows where and when a word is prominent in the archive)

* I can build a **tf-idf** based visualization

1. For each article, calculate the tf-idf
2. Create a bubble plot, x-axis = time, y-axis = journals, bubble size = tf-df
3. **Bubble plot showing tf-idf of articles in different journals**

**

Query:

SELECT

\*,

(SIZE(SPLIT(lower(text), lower({{word1}}))) - 1) \* (

log(

2,

1063728 / (

1 + (

SELECT

count(\*)

FROM

default.century1\_avro

WHERE

lower(text) rlike lower({{word1}})

)

)

)

) / wordCount as word\_1\_tf\_idf,

(SIZE(SPLIT(lower(text), lower({{word2}}))) - 1) \* (

log(

2,

1063728 / (

1 + (

SELECT

count(\*)

FROM

default.century1\_avro

WHERE

lower(text) rlike lower({{word2}})

)

)

)

) / wordCount as word\_2\_tf\_idf

FROM

default.century1\_avro

WHERE

(

(SIZE(SPLIT(lower(text), lower('law'))) - 1) \* (

log(

2,

1063728 / (

1 + (

SELECT

count(\*)

FROM

default.century1\_avro

WHERE

lower(text) rlike lower('law')

)

)

)

) / wordCount > 0

)

OR (

(SIZE(SPLIT(lower(text), lower('pauper'))) - 1) \* (

log(

2,

1063728 / (

1 + (

SELECT

count(\*)

FROM

default.century1\_avro

WHERE

lower(text) rlike lower('pauper')

)

)

)

) / wordCount > 0

)

AND (

to\_date(cast(NumericPubDate as string), 'yyyyMMdd') BETWEEN DATE '{{ date.start}}'

AND DATE '{{ date.end }}'

)

AND find\_in\_set(TypeOfArticle,"{{Type of Article}}") > 0

LIMIT

250