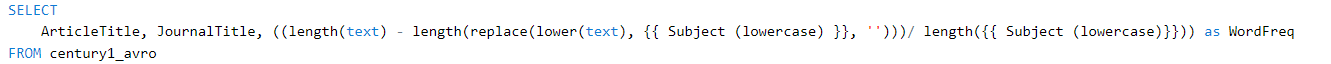
1. Bar chart showing number of cumulative number of times a substring (subject) appears in the text of a Journal
2. What if we controlled for number of articles?... TO be continued





2) Line graph displaying percentage of all words in article text that are a given subject

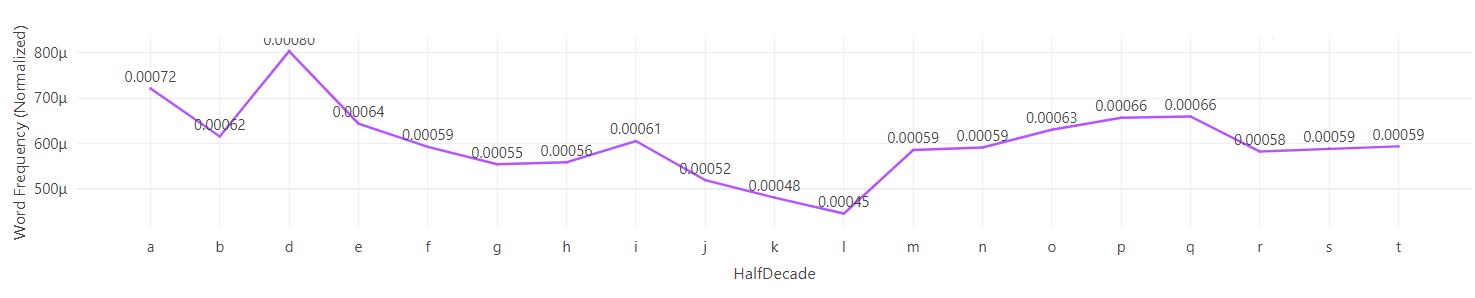
SELECT HalfDecade,

SUM((length(text) - length(replace(lower(text), {{ Subject (lowercase) }}, '')))/ length({{ Subject (lowercase)}})) / SUM(wordCount)

FROM century1\_avro

GROUP BY HalfDecade

\*Should read percentage



*Variation* - Comparing two subjects

SELECT HalfDecade,

SUM((length(text) - length(replace(lower(text), {{ Subject 1 (lowercase) }}, '')))/ length({{ Subject 1 (lowercase)}})) / SUM(wordCount),

SUM((length(text) - length(replace(lower(text), {{ Subject 2 (lowercase) }}, '')))/ length({{ Subject 2 (lowercase)}})) / SUM(wordCount),

FROM century1\_avro

GROUP BY HalfDecade

3) Stacked bar chart displaying the distribution of word frequency over time

1. Perhaps account for changes in the representation of each category

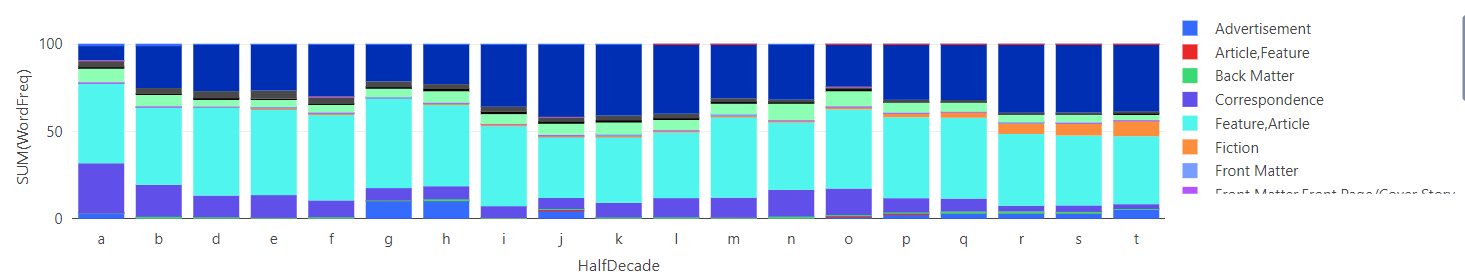
SELECT TypeOfArticle, HalfDecade,

-- This column contains the number of times the word appeared in the artcile

(length(text) - length(replace(lower(text), {{ Subject (lowercase) }}, '')))/ length({{ Subject (lowercase)}}) as WordFreq

FROM century1\_avro

ORDER BY TypeOfArticle



**Idea: Bubble Plot Showing presence of word across different journals**

Proxy for word presence: [tf-id](https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/)f

SELECT

-- Count of "subject" term in article

(length(text) - length(replace(lower(text), lower({{ Subject }}), '')))/ length({{ Subject }})

-- divided by number of words in artcile

/ wordCount)

--multiplied by the log base 2 of idf

\* log (2,

-- of total numbers of articles

count(\*))

FROM

century1\_avro