Week 3

Established: The means of taking a sentence and dividing it into weights and terms to be fit into a dictionary.

We will set up the process and proceed in each of the three directions, as usual. As this is the last week of Phase 1, we anticipate finishing up the IS and CAP research and moving towards experimentation, although some should be done this week.

RA

Storage and Visualization

The unstructured referenced data, just as reference

1:1 Customer, ReviewLookUp, TopicDM

Graphical user interface, diagram

Description automatically generated

We’ve done extensive research in word clouds, and alternatives in NoSQL require more work, ala independent study. Two good alternatives that we discussed but didn’t really build upon last semester are the document coloring and some graphs and word cloud variants

A picture containing timeline

Description automatically generated

And some tested code:

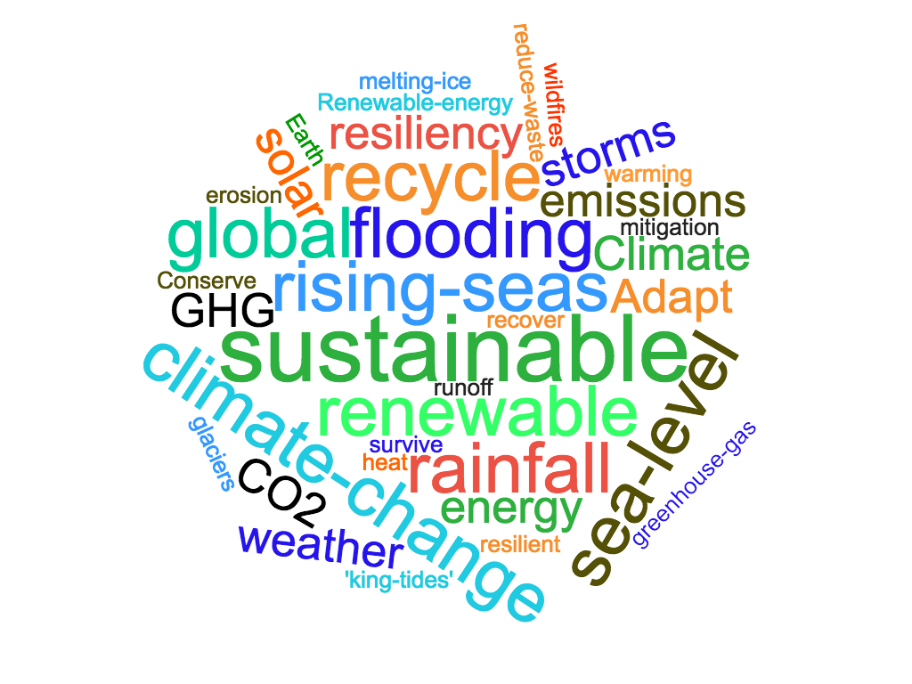
Text

Description automatically generated

Not very much variation in how limited word clouds can show information at a glance. While that’s what we want, could be TOO simple:



OR more standard



A bar chart, easier but less information conveyed at a glance than the word cloud:

Chart, bar chart

Description automatically generated

While alternatives can show up, we can compare the performance of each and determine statistically the optimal way to show. It’s kind of hard to represent each “correctly” this way since the difference is more qualitative than quantitative, but this can allow us to get a better handle on things we can already measure. For graphs, a good quote I saw [1] was “are the relationships more important than the data? If they are, that’s a good indicator that a graph may be the right tool.” Since this is not necessarily the case for our topics (as while they have association and weight, the records themselves to not rely entirely on the base documents).

There is one more way we can approach and look at the ways we can visually analyze the results, particularly if we treat it as a text mining and information retrieval problem, where we can use plots like a U-matrix or a hitmap to show areas that are related and have relation or weight. An example of such can be seen as follows:

Scatter chart

Description automatically generated

These matrix/hitmap signify connections and # of hits, but can be adapted to a select statement, MAYBE. Most likely though they can map out the trend over time aspect like we discussed previously with the word cloud. But it’s again not a very great visualization itself, just a great tool.

[1] https://blogs.perficient.com/2021/02/10/to-sql-or-to-nosql-that-is-the-question/