Text Analysis Pipeline Implementation

Can AI Preserve our Science Legacy?

Shockwave Surfers - 2022 NASA Space Apps Challenge
Team credits: Anshuman Tekriwal (Ansh3101#9783), ayushgupta0010#2342, Sparsh Rastogi#6995, Tacoma-Local Lead - Joe Devlin (northdecoder#2610)

2-Oct 2022

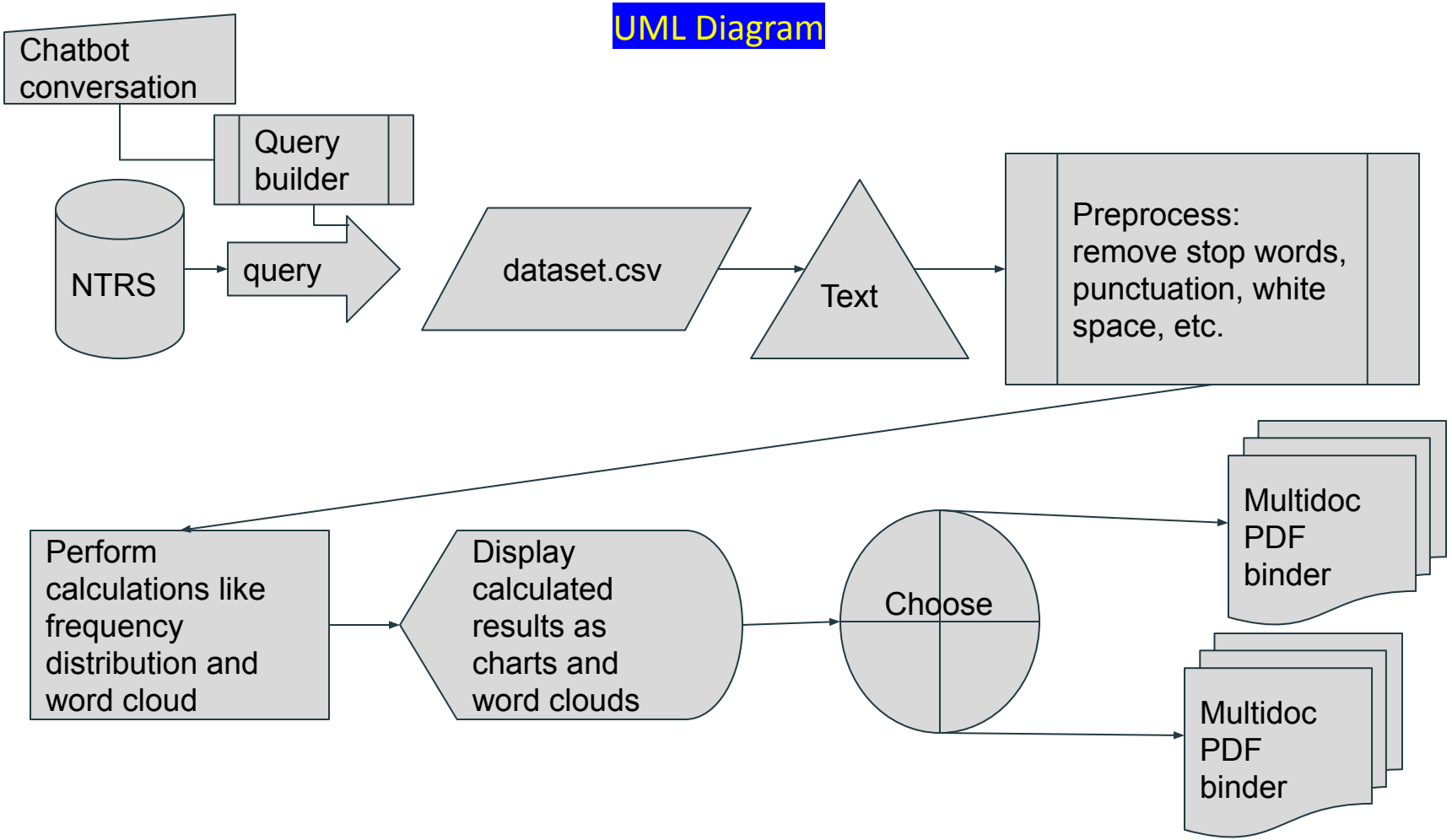
User story

As a researcher I would like to go to the library to request assistance from a subject matter expert librarian in curating appropriate research documents for studying. I expect the documents will be organized in the order of most importance first and will be bound, indexed and bookmarked with unique posties where appropriate.

Classy Librarian

The subject matter expert will be an intelligent chatbot that guides the discussion with the patron to help them query the NTRS database tagged CDMS, with resulting keywords, and n-grams curated by the librarian class software.

UML Diagram

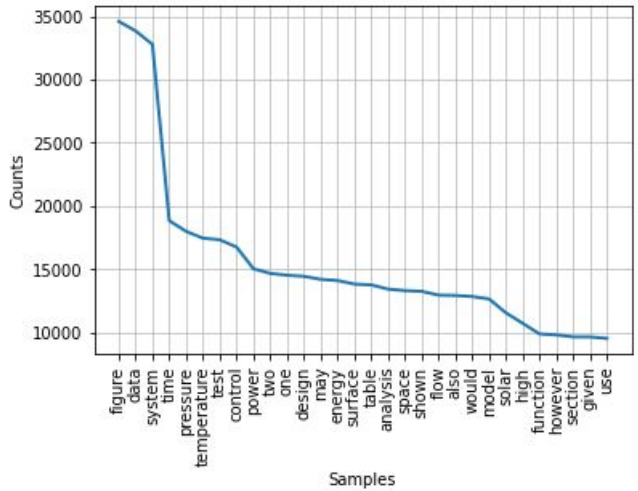

Chatbot conversation
Query builder
NTRS
query
dataset.csv
Text
Preprocess: remove stop words, punctuation, white space, etc.
Perform calculations like frequency distribution and word cloud
Display calculated results as charts and word clouds
Choose
Multidoc PDF binder
Multidoc PDF binder

Display Calculated results as a graphical representation

Query results ~= 3,187,533 (just over 3 million records !)

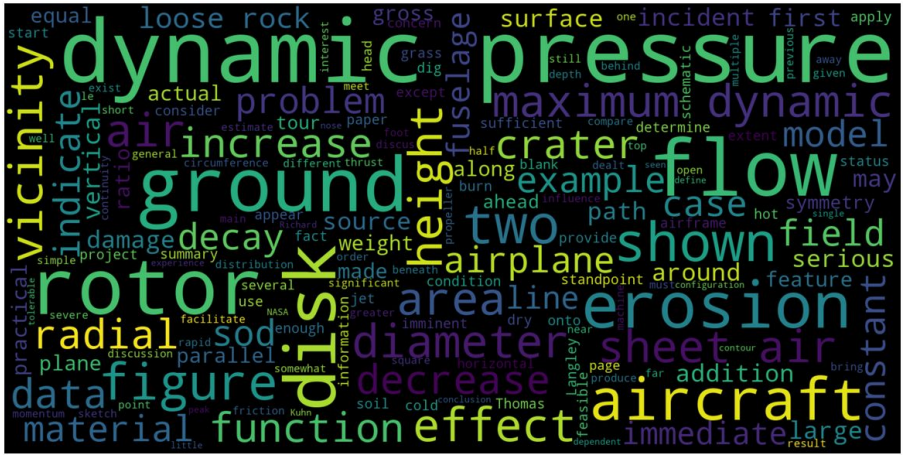
Word frequency distribution

```
fdist.plot(30)
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7f335069dd90>
```

Wordcloud example run on one example document



Key Take Aways

1. The library patron will receive an indexed and categorized pdf binder of the resulting queried documents.
2. There was so much data that a commercial search company refused to ingest it. It will probably be a dedicated customized niche solution.
3. Some queries return results that have sparse results, ie metadata but few documents.

References:

1. **Pypdf2** for converting from PDF to text files https://pypdf2.readthedocs.io/en/latest/index.html
2. **NLTK** for writing and analyzing the corpus https://www.nltk.org/
3. **Wordcloud** https://github.com/amueller/word_cloud
4. Inspiration to accept this years challenge, came from the 2015 Challenge Winners, Team NYSpacetag; https://github.com/jonroberts/nasaMining ; https://2015.spaceappschallenge.org/project/nyspacetag/
5. The term “will” in this document refers to a forward looking statement that may or may not actually occur.
6. **NTRS**: NASA Technical Reports Server; https://ntrs.nasa.gov/
7. **CDMS**: Cluster Data Management System; https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.5188&rep=rep1&type=pdf
8. **N-grams**: https://en.wikipedia.org/wiki/N-gram
9. Project page: https://github.com/space-apps-tacoma/cluttered_spark