Document Term Matrix

Prof. Eric A. Suess

April 7, 2021

Document Term Matrix (DTM)

The Document Term Matrix is one of the most common data storage formats for text based data.

The DTM is based on the bag-of-words model. Each word is a feature in the data set. This leads to **sparse matricies**.

Document Term Matrix (DTM)

Some of the most popular R libraries and Python packages use DTM.

- R: tm, quanteda, CRAN Task View: NLP
- Python: NLTK, Gensim, Spacy

DTM

- each row represents one document (such as a book or article)
- each column represents one term
- each value (typically) contains the number of appearances of that term in that document

DTM

tidytext to DTM

```
> tidy() # to tidy format
> cast() # to DTM
```

Problem

5.3.1 Example: mining financial articles does not run!