Ben Ballard (bkq5nt)

Exploratory Text Analytics

Project Final

# Deliverables

To receive full credit for the assignment, you must produce ​a digital analytical edition​ of a corpus. This edition should include the following deliverables to be uploaded or linked to the Assignment for the Final Project in Collab.

* A ​collection of source files​ ​compressed​ in an archive (e.g., zip or tar.gz) and ​hosted​ on your UVA Box account.
* A ​manifest​ file describing those sources files, including their:
* Provenance​: Where did they come from? Describe the website or other source and provide relevant URLs.
  + Location​: Provide a link to the source files in UVA Box.
  + Description​: What is the general subject matter of the corpus?
* Format​: A description of both the ​file formats​ of the source files, e.g., plaintext, XML, CSV, etc., and the ​internal structure​ where applicable. For example, if XML then specify document type (e.g., TEI or XHTML).

**Data sources:**

1. AnnapolisPD.csv – F0 datasource
   * **Document its provenance, location, description, and format in your manifest.**
2. **Processed Data F1\_data** 
   * **Your extracted incident reports are your F1 Machine Learning Corpus Format.**
   * ****
3. **Standard Text Analytic Data Model (F2)**
4. **NLP Annotations (F3)**
5. **Vector Representations (F4) Use TFIDF to create vector representations of corpus.**
6. **Analytical Models (F5) Save term document matrix and add resulting TFIDF values to term.csv**

# Text Analytics Project Manifest

## Provenance: I built a web scrapper that pulls my local police reports from Annapolis, MD. The police provide a document nearly daily which provides information on crimes concerning the public as well as general service announcements.

## URL: <https://www.annapolis.gov/list.aspx?PRVMSG=253>.

## Description: Scrapper pulls over 2000 reports over the last 7 years. I had some trouble fully extracting the data from 2020-21, but since this wasn’t the focus of the course, I continued with the data I had.

Figure 1: Annapolis PD Report Page

## A screenshot of a report Description automatically generated

## F0 Structure

* document\_id: Unique identifier for each document.
* title: Title of the daily report from the link.
* link: URL linking back to the original report for the content.
* date\_sent: The posted date and time.
* content: The full content of the report.

## F1 Structure (Machine Learning Corpus Format)

## Description: Extracted incident reports from raw police data

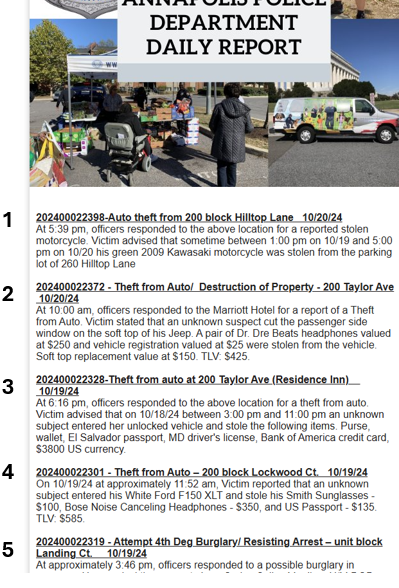
****Each report contains multiple blocks of information, including incidents that are assigned an incident ID, which seem to be related to specific crimes/cases. These often come with a kind of ID, and that’s what I searched for using regex, to pull these incidents out as a minimally recursive document.

Figure 2: Example of a Daily Report containing many Incidents

### F1 Structure

1. document\_id: Unique identifier linking to the original document in F0.
2. incident\_id: Identifier for specific incidents. Provides granularity beyond the document level.
3. report\_date: Date the report was filed.
4. incident\_date: Date the incident occurred.
5. text: Contains the report text, primary content for analysis.
6. link: URL to the full report. Maintains provenance.

## Extraction Method: regular expressions

### F2 Structure: STADM

**F2\_Doc\_Table: I’m not its improved from F1\_data. That might be enough or close. Watch the doc table explanation.**

F2\_Term\_Table:

A screenshot of a computer

Description automatically generated

F2\_Token\_Table:

A screenshot of a computer

Description automatically generated

## F3 Structure: NLP Annotated

* Parts of speech
* Named entities
* Sentiment scores
* [Need to specify which specific NLP features are included]

## F4 Structure (Vector Representations)

* TFIDF vectors for documents
* Term-document matrix
* [Need to specify storage format and structure]

## F5 (Analytical Models)

* Topic modeling results
* Principal components
* Word embeddings [Need to specify exact format and structure]

**Still to do:  
Key areas that need attention:**

1. **F2 Structure Details:** 
   * **You need to explicitly define the structure of all three required tables (DOC, TOKEN, TERM)**
   * **Each table should include its indexing scheme**
2. **F3-F5 Documentation:** 
   * **Need more specific details about what NLP annotations you're including**
   * **Need to document how vector representations are stored**
   * **Need to specify analytical model outputs and their storage format**
3. **Table Relationships:** 
   * **Document how tables are linked (shared indices)**
   * **Specify any OCHO (One Column Header Only) indices**

**Would you like me to help you detail the structure for any specific part of the pipeline? The F2 tables would be a good place to start since they form the foundation for F3-F5.**