Issues

Indexing – Whilst indexing the documents, memory was an issue in both extracting the information, indexing the documents to Elasticsearch, and querying through kabana/other. When querying all documents, the elasticsearch server would shut down due to a JVM out of memory error

Tokenization – Data is of an inconsistent format. CSV uses ‘”’ characters to wrap around a cell that contains a contiguous block of text, but that text can contain ‘”’ characters in itself which is difficult to parse. For the most parts, only the description contains ‘”’ characters, but occasionally some other fields also contain these characters. When parsing data like this, it is usually the case that there is a single entity per line, so when parsing the text you can process an entity on a line by line basis. However, this is not the case. Some fields contain new line characters making processing an entity on a line by line basis impossible. Some entities did not have data stored in the correct column, for instance line 82, the Charlie Chaplin move, had ‘Comedy’ where the cast field should be.