**DupliatesPipeline**: Possibly removing duplicates.

**HTMLPipeline**: Gets the HTML pre and post rendering. Splits them up into components. Stores images in one table and other content in another.

**ScreenshotPipeline**: Takes and saves the screenshots.

**TextAnalysisPipeline**: Collects the words and sentences into lists ready to be put into the DB. Finds the longest word and longest sentence. Uses these figures to determine the number of columns in the chart and the labels for each column. Compiles the url\_stats object to store the frequency of different word and sentence lengths in terms of the columns they will be in on the chart. Then the averages are calculated and added. The counts and averages are stored in the text\_data table. Each url in url\_stats is saved as its own JSON file. Another JSON file called extra\_data is saved – it contains the relationships between urls and their names, and lists of the different labels in order.