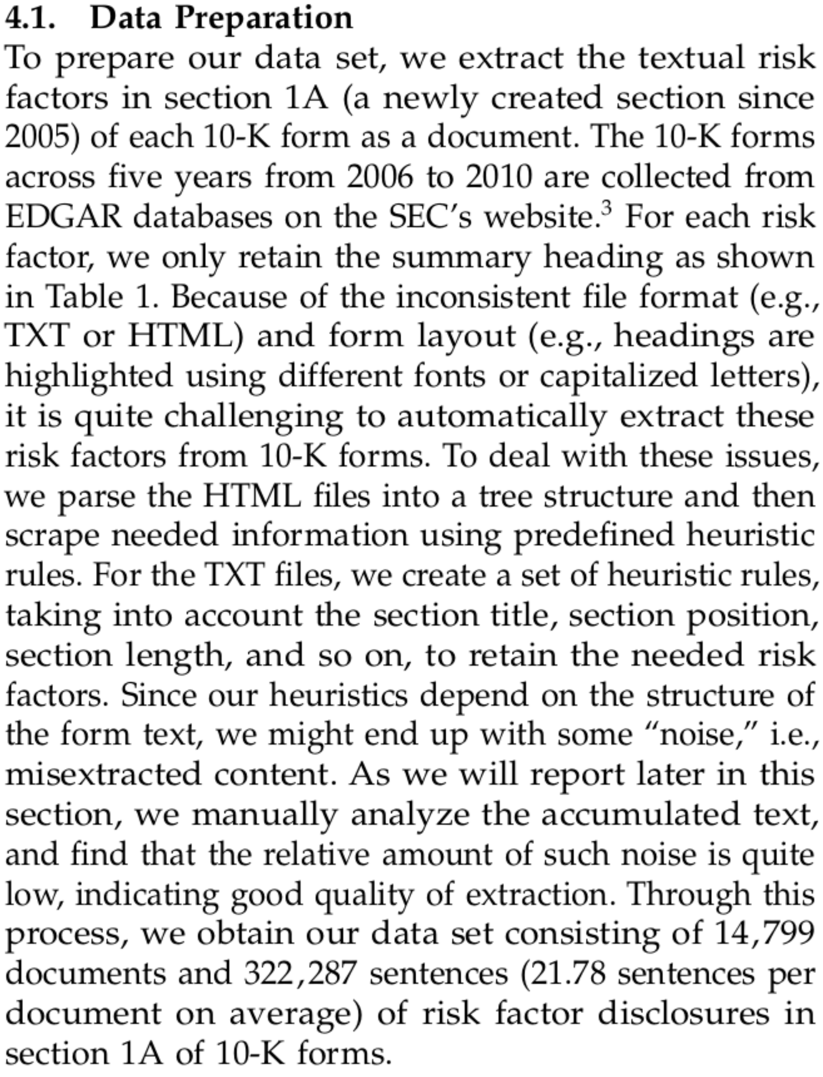
Business insights from Multimedia data: Text and Audio

The project objective is to gain business insights from firms' unstructured data: textual and audio, using big data technology and machine learning method.

Milestone 1: collect, parse and organize firm 10-K forms (text).

Given a list of company names, download all 10-K filings. Parse 10-K filing and only keep Item 1 (Business), item 1A (risk factors), item 7 (management's discussion and analysis of financial condition and results of operations. Text only). Save data into database.

(Bao and Datta 2014) describes a similar data preparation step in section 4.1.



4.1. Data Preparation

To prepare our data set, we extract the textual risk factors in section 1A (a newly created section since 2005) of each 10-K form as a document. The 10-K forms across five years from 2006 to 2010 are collected from EDGAR databases on the SEC's website.[3] For each risk factor, we only retain the summary heading as shown in Table 1. Because of the inconsistent file format (e.g., TXT or HTML) and form layout (e.g., headings are highlighted using different fonts or capitalized letters), it is quite challenging to automatically extract these risk factors from 10-K forms. To deal with these issues, we parse the HTML files into a tree structure and then scrape needed information using predefined heuristic rules. For the TXT files, we create a set of heuristic rules, taking into account the section title, section position, section length, and so on, to retain the needed risk factors. Since our heuristics depend on the structure of the form text, we might end up with some "noise," i.e., misextracted content. As we will report later in this section, we manually analyze the accumulated text, and find that the relative amount of such noise is quite low, indicating good quality of extraction. Through this process, we obtain our data set consisting of 14,799 documents and 322,287 sentences (21.78 sentences per document on average) of risk factor disclosures in section 1A of 10-K forms.

- A possible tool to download 10-K data (there are many others): https://github.com/joeyism/py-edgar
- A sample 10-K form: https://www.sec.gov/Archives/edgar/data/104169/000010416918000028/wmtform10-kx1312018.htm#s1b91fe1784b0499386019b733c9d8f7d
- U.S. SEC website: https://www.sec.gov/edgar/searchedgar/companysearch.html
- Wikipedia Form 10-K: https://en.wikipedia.org/wiki/Form_10-K (This link explains format of form 10-K)

- Bao Y, Datta A (2014) Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Manag. Sci.* 60(6):1371–1391.