The objective of this task is to extract the information from the PDF files which contain press releases, company report, shareholder report and company statements. We have obtained these documents by combining a company name with a set of keywords.

These keywords are presented below:

“gender inequality” OR “gender issue” OR “gender justice” OR “gender pay gap” OR “gender proud” OR “gender gap” OR “gendered violence” OR

“sexual harassment” OR “Sexism” OR “sexual assault” OR “sexual crime” OR “sexual harassment” OR “Sexual misconduct” OR “Sexual violence” OR “sexual abuse” OR “sexual Accusation” OR

“domestic violence” OR “Intimate partner violence” OR “misogyny” OR “hashtag activism” OR “Feminism” OR “feminist” OR “Misogyny”

“Gender Equality” OR “Chief Diversity Officer” OR “Diversity and Inclusion“ OR “Diversity & Inclusion” OR “D&I” OR “diversity delegate” OR “sexism awareness” OR “Code of conduct on sexual harassment” OR “anti-harassment policy and complaint” OR “anti-discrimination and harassment policy” OR “Sexual Harassment Prevention Training” OR “bystander intervention” OR “Equal Employment Opportunity” OR “whistleblower complaint” OR “zero-tolerance diversity policy” OR “parental leave” OR “freedom from discrimination” OR “freedom from harassment” OR “freedom from violence” OR “reporting workplace harassment” OR “Commitment to women empowerment” OR “gender-sensitive language” OR “diversity” or “

You will find the following files:

* A zip file containing 131 PDF documents
* Excel file with the headlines of the documents and some info in columns A-D.

In the file name you find:

1. Company name
2. Headline of the file
3. Date of the file
4. Source (e.g., KRT)

Most of the files look alike and have a very similar structure.

**We have generated two manual examples (row 2,3 and 12) of which kind of information shall be reported.**

* **Date of the event:** note that this might or might not correspond with the actual date of the news piece (column A). The news can be reported the day after or refer to an event earlier on.
* **Who wrote the piece:** This information is found at the end of the document.
* **Subject:** This information is found at the end of the document.
* **Relevant text:** this is the paragraph that contains one or more of our keywords; If there is more than one paragraph, please report them separated by | symbol (see example row 3).
* **Keywords mentioned**: please report the keywords mentioned in the paragraph.
* **“Notes” column (K):** includes the observations on whether the keywords are not central to the news piece. For instance, the news piece 2, is not directly about our keywords (i.e. donation to a hospital) but the keywords are just mentioned in the news piece.

**Hints**:

1. If the file only includes “Reference Link” or just a link in the body of the text do not click on the link but rather focus on the content even if it’s only the title (heading)
2. Some articles do not have subjects-> Do not report the subject for these articles.
3. Some files will have the same title but will have different dates/time: these are NOT duplicates.

For report

1. Libaries I’ve tried that did not work: PyPDF2, pdfplumber
2. Library that do work: pdfminer3
3. Issues: PDFs have varying formats, making it harder to create a generic script
   1. Some PDFs list the source after “SOURCE”, “Source:”, “CONTACT:”, or not at all making it much harder to find the source
4. Will not be possible to automate relevant / not relevant notes unless a machine learning model is used which I am not familiar with.
5. Date of the event of the relevant keywords will be hard to extract unless it is in consistent format
   1. May be possible to write a parser that looks at several different date formats, but it would be a good amount of work and will require some trial and error
   2. Date of the

Wynn Automatic Text Extraction Report

I did some research and found several Python libraries that allow for text extraction from PDF documents. I tested PyPDF2 and pdfplumber before settling on pdfminer3 which worked the best. After testing, I found that I was able to automate parts of the text extraction process with the Wynn documents, but could not automate the entire process. I will list what is currently possible with automation and what is not possible below:

What can be automatically extracted:

1. **Who wrote the piece**
   1. I can currently search for the author of the piece by parsing through the text of the PDF document and searching for lines that contain “SOURCE”, “Source:”, “CONTACT”, etc. There are a few documents (such as example given in row 2) where I have noticed that the author of the document is not listed in such a manner. This makes searching for author more difficult and I have yet to find a way to entirely automate this process. Manually searching for the author here may be necessary in some cases. I am still thinking of different ways to approach finding the author.
2. **Subject**
   1. This is easy to extract from the PDFs as I can search for the line titled “Subjects:” and pull all the listed subject categories from the document (if they are present in the document).
3. **Relevant Text**
   1. This is also fairly easy to extract from the PDFs as the library I am using preserves the relative formatting and layout of the document as it is converted to text. This method allows me to pull the entire paragraph where a keyword is mentioned. The other two libraries that I used before did not have this capability.
4. **Keywords mentioned**
   1. This is very easy to extract as I can just keep track of all the keywords that were found in the document.

What needs to be manually extracted:

1. **Date of the event**
   1. The date of the event is quite difficult to extract from the relevant text because date formats may not be consistent throughout all of the documents. For example, under the title of the documents the time and date of the document is written in a standard format:
      1. **Day, Date, Year, Time, Time zone**
   2. However, the date listed for the occurrence of the event does not always follow this format. For example, the document used in row 2 has an event date of February 2018, which has the format:
      1. **Month Year**
   3. I am looking into ways to extract the date from the relevant text (if it exists) by using regular expressions or some other libraries.
2. **“Notes” column (K)**
   1. It is difficult to determine whether the relevant keywords that appear are central to the document through an automated process. One idea I had is to look at the title of document and see if any insights could be gained there. This may not work in all cases, however. To properly determine relevance, I believe it might be necessary to look at some more NLP and machine learning methods. Depending on the scale of the project, I believe it might be best to manually assign “relevant” or “nonrelevant” to the keywords found in the document. If there are too many documents for this to be feasible, I can continue looking into some alternative implementations.