I, **Sandeep Ravi**, declare that the submitted work is original and adheres to all University policies and acknowledge the consequences that may result from a violation of those rules

Most difficult task :

My experience with the development went smooth. This is my first time performing web scraping. So considerably difficult part in assignment was understanding the concept of scraping. Starting from learning technology involved, the legal aspects of scraping, reading about the legal disputes involving web scraping was fun. Then for selecting a museum, it was a random choice and at last it turned out to be a good one. For selecting a tool, I wanted a tool which is based on a programming language I am comfortable. So I had to go through the architectures of many tools before selecting one and ended up with Scrapy.
Selecting only paintings created after 1700 was tricky. The date format used in the website was not fixed. I used three regular expressions for selecting the Year. There are few paintings which had very unique date formats, since writing a separate regular for each one of them was not feasible, I had to discard these paintings.

Tool Used : Scrapy

I wanted to use a tool which was based on a programming language, which I am comfortable with. Python seemed to be good language for this software and Scrapy was my best bet with python. Scrapy has good documentations and tutorials on the web, which made my job of learning web scraping much easier. Scrapy has a good community built around it, which helps while fixing issues. With Scrapy selecting an element at particular XPath is very easy. The only problem with Scrapy is when the website has Ajax content and thankfully the selected museum site did not have any Ajax data, which made Scrapy perfect choice. You give Scrapy a root URL to start crawling, then you can specify constraints on how many number of URLs you want to crawl and fetch,etc., and then Scrapy will take care of complete data scraping.

Sample Record :

```
<Art>
  <Date value="1966"/>
  <Title value="Untitled"/>
  <Artist value="Eva Hesse"/>
  <Medium value="Enamel paint and string over papier-m&#226;ch&#233; with elastic cord"/>
  <Dimension value="Overall approximately 33 1/2 x 26 x 2 1/2&quot; (85 x 65.9 x 6.4 cm)"/>
  <Image value="http://www.moma.org/collection_images/resized/919/w500h420/CRI_210919.jpg"/>
  <Status value="On view"/>
  <Credit value="Ruth Vollmer Bequest"/>
  <Category value="Painting and Sculpture"/>
</Art>
```

Fields Not extracted :
The structure of the selected website was very organised. The task of extracting data in this website was pretty straight forward (Except for the Date field). So I could extract all the fields, which I wanted.

Website : ***http://www.moma.org/***
Screen Shots