AI4M Assignment 1

To access this dataset, follow this link: AI4M_Dataset (please access from a UAL account)

The dataset I have collected for this assignment is images of illuminated manuscripts, a historical art form that combines text and iconography, generally associated with Western art and religion.
I collected these images using Pin down on Pinterest, a repo that allows for scraping images off specified Tumblr blogs and another that allows you to scrape images from the Internet Archive.

To get the pins from Pinterest, I typed in several key terms so that I could catch any images that would show up with just one, those search terms were 'illuminated manuscript', 'medieval manuscript', and 'marginalia'.

I found several blogs on Tumblr which were dedicated to archiving these manuscripts. Using the github repo (https://github.com/dixudx/tumblr-crawler) I was able to scrape the images from, *loveforillumiantedmanuscripts*, *beatufiulmedievalmanscripts,* and *Penn Libraries Manuscripts*, the latter being an archive of the University of Pennsylvania's manuscript collection.

Code/Software:
$ git clone https://github.com/dixudx/tumblr-crawler.git
$ cd tumblr-crawler
$ pip install -r requirements.txt
$ python tumblr-photo-video-ripper.py

Using the GitHub repo (https://github.com/terrybroad/internet-archive-downloader) I scrapped images from the Cleveland Museum of Arts collection.

Code/Software:
(https://archive.org/developers/internetarchive/installation.html)
$ sudo pip install internetarchive
!python search.py –collection=clevelandart –subject=medieval+art

The process from there was removing any potential duplicates that might exist in the various datasets, removing the work of any modern manuscript makers as I wanted this dataset to be made of purely historical datapoints. I also removed any non-western work because I didn't want to lump in artworks that might be considered aesthetically similar but that are used for different cultural purposes.
This dataset contains 7,662 images that are unlabelled.

Pictorial Examples: