What seed URL(s) you used for each task?

I created a list of urls using a for loop for the pages that I want to crawl. For movies, I used the page comedy genre of the all the movies. For cast, I grabbed pages for cast based on popularity.

How did you manage to only collect movie or cast pages?

I went through the list of comedy movies on IMDB and grabbed the urls for the movie listed. I did not follow any link in the movie site and only followed links on the list site. I did the same for casts.

Did you need to discard irrelevant pages? If so, how?

Yes, there was a page that gave me a 404 error when I was scraping casts. I just skipped that cast and moved onto the next person.

Did you collect the required number of pages?

Yes. IMDB – 5000 scraped. Casts – 5007 scraped