Answers file – ex1 – Big Data Management:

Question 1:

Xpath:

//table//a[contains(text(),' ')][starts-with(@href, '/wiki')][preceding-sibling::span[@class='flagicon']]/@href

Example:

given https://en.wikipedia.org/wiki/Andy_Ram the XPath returns /wiki/Jonathan_Erlich

Question 3:

Most visited url first 100,000: https://en.wikipedia.org/wiki/Anders_J%C3%A4rryd

Most visited url last 100,000: https://en.wikipedia.org/wiki/Bob_Bryan

The reason we got two different results for the url with most visits in the first and last 100,000 sessions is that the process of jumping from one page to another is made with great randomness, when we start from a random page and in each jump we make a random jump to a page from the possible links (or to all possible nodes if no links) so it is only reasonable that we will get different results in the two sessions because the process heavily depends on randomness.