DATA.STAT.840 Statistical Methods for Text Data Analysis
Exercises for Lecture 2: Basic text processing

Student Name: Ahmad Sharif
Student ID: K436765
ahmad.sharif@tuni.fi

Exercise 2.1: Data acquisition - Building a better web crawler.

The web crawler code shown on Lecture 2 is poor in at least two respects:
1. It can crawl the same page multiple times, if a link on a later crawled page points to the already-crawled page.

2. It inserts all links from each page in order as pages to be crawled. If some page contains thousands of links, the crawling will crawl those first and may never get to the links from the next page, especially if the total number of pages are limited.

To fix this duplicate issue, at first before insert into the list, I check if the url already exist list data-structure.

```
import requests
import bs4

webpage_url = "https://www.sis.uta.fi/~tojape/"
webpage_html = requests.get(webpage_url)
webpage_parsed_html = bs4.BeautifulSoup(webpage_html.content, 'html.parser')


def getpageurls(webpage_parsed):
        pagelinkelements=webpage_parsed.find_all('a')
        pageurls = [];
        for pagelink in pagelinkelements:
                pageurl_isok=1
                try:
                        pageurl = pagelink['href']
                except:
                        pageurl_isok=0
                if pageurl_isok == 1:
                        if (pageurl.find('.pdf') !=-1)|(pageurl.find('.ps')!=-1):
                                pageurl_isok = 0
                        if (pageurl.find('http') ==-1 )|(pageurl.find('.fi')==-1):
                                pageurl_isok = 0
                if pageurl_isok == 1 and pageurl not in pageurls:  # Before Append we need to check
                        pageurls.append(pageurl)
        return(pageurls)
mywebpage_urls = getpageurls(webpage_parsed_html)
print(mywebpage_urls)
```