CAB431 Practical (Week 2)
Basic I/O operations for Processing Text Data

Objectives:

- Getting familiar with or selecting a programming language (e.g., Java, Python or C#) that you have studied for doing practical questions and assignment 2.
- Basic data processing for text documents.
- Please note that we do not teach the basic skills for using a programming language. Please contact the unit coordinator if you have any difficulty for using a programming language.

Data (selected 10 files from RCV1V2 document collection)

- We'll be working with a sample dataset which is a small subset of just 10 documents from the RCV1v2 document collection. More specifically, we'll be working with a subset of the pre-tokenized version of this (for convenience, and for copyright reasons). Our dataset can be downloaded from Blackboard learning resources practical page.
- PS: Prof Yuefeng Li has got the permission to access these dataset for his group. Please do not release the dataset to other people.

Task 1 for Week 2 practical: Write a program that loads (read) RCV1v2 collection, and prints out the **itemid** and the number of words in <**text**> of each file.

APPENDIX A - Sample Format of Document (a file)

<newsitem xml:lang="en" date="1997-07-20" id="root" **itemid="741299"**>
<title>BELGIUM: MOTOR RACING-LEHTO AND SOPER HOLD ON FOR GT VICTORY.</title>
<headline>MOTOR RACING-LEHTO AND SOPER HOLD ON FOR GT VICTORY.</headline>
<dateline>SPA FRANCORCHAMPS, Belgium</dateline>
<text>
<p>J.J. Lehto of Finland and Steve Soper of Britain drove their ailing McLaren to victory in the fifth round of the world GT championship on Sunday, beating the Mercedes of German Bernd Schneider and Austrian Alexander Wurz by 15 seconds.</p>
<p>Their victory enabled them to open up a 16-point lead in the overall standings over Schneider, who mounted a strong challenge on the struggling leaders in the final minutes of the four-hour race.</p>
<p>But Soper, struggling with the car's handling caused by a broken undertray, just managed to hold on for the win.</p>
<p>Lehto had opened up a lead of over 90 seconds during a mid-race downpour in the Ardennes mountains.</p>
<p>"I thought that everyone else was driving on dry-weather tyres," he joked afterwards.</p>
<p>"We swapped to rain tyres at exactly the right time and I was able to push hard and open up a big lead."</p>
<p>Third to finish was the Porsche of France's Bob Wollek and Yannick Dalmas and Belgian Thierry Boutsen.</p>
<p>The Belgian, a former Formula One driver, switched from the car he normally shares with German Hans Stuck following a power-steering failure on his own car.</p>
</text>
<copyright>(c) Reuters Limited 1997</copyright>
</newsitem>

APPENDIX B - NetBeans Java IDE

NetBeans IDE lets you quickly and easily develop Java desktop, mobile, and web applications, as well as HTML5 applications with HTML, JavaScript, and CSS. The IDE also provides a great set of tools for PHP and C/C++ developers. Please refer https://netbeans.org/kb/index.html to get a quick start.

How to install NetBeans on your system?

1. Go to https://netbeans.org/downloads/.
2. In the upper right area of the page, select the language and platform from the drop-down list. You can also choose to download and use the platform-independent zip file.
3. Click the Download button for the download option that you want to install.
4. Save the installer file to your system.
5. After the download completes, run the installer. Please refer to *NetBeans IDE 8.1 Installation Instructions* for details.

Example: Steps for Task 1:

1. Create a project named “cab431tut” under H: drive. Project location should be H:\Documents\NetBeansProjects, a main class could be created automatically, please assign a name for this main class.

2. Create a java class (or more classes if you like) in cab431tut project, which should contains some methods to read data from text files and print out the document IDs and the number of terms in each document.

The easiest way to read input is to create a new Scanner object from a passed File or FileInputStream object. The easiest way to write is to create a new FileWriter object from a passed File.

3. Edit your main class to call methods in above java class and run it to see your program is working.

Please do not panic, ‘google’ it!

We are lucky, there are better than enough resources from internet which can help you out.

e.g. http://www.tutorialspoint.com/java/java_files_io.htm

e.g. https://www.cs.swarthmore.edu/~newhall/unixhelp/Java_files.html