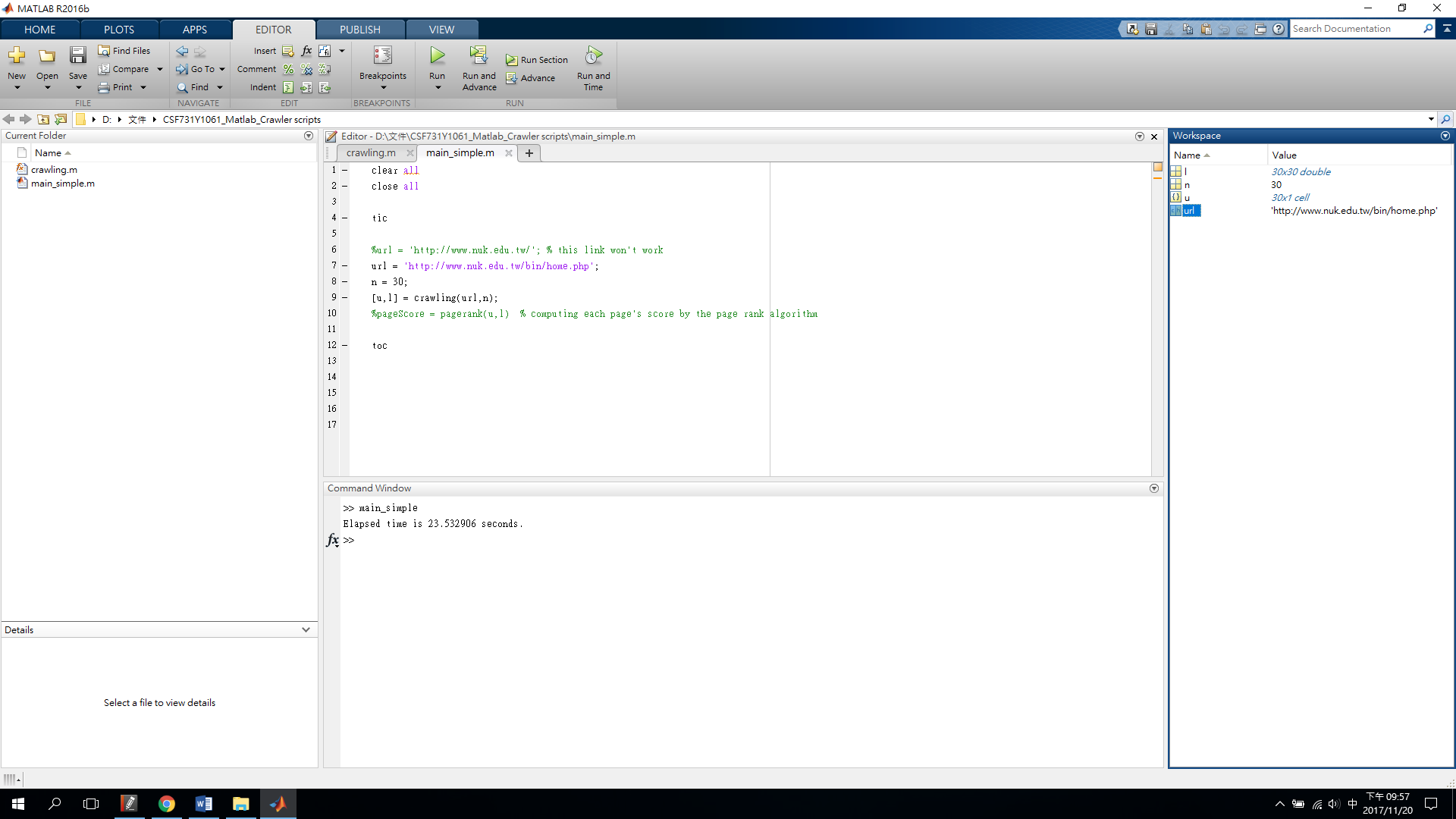
A1045516 - Homework1 爬蟲程式

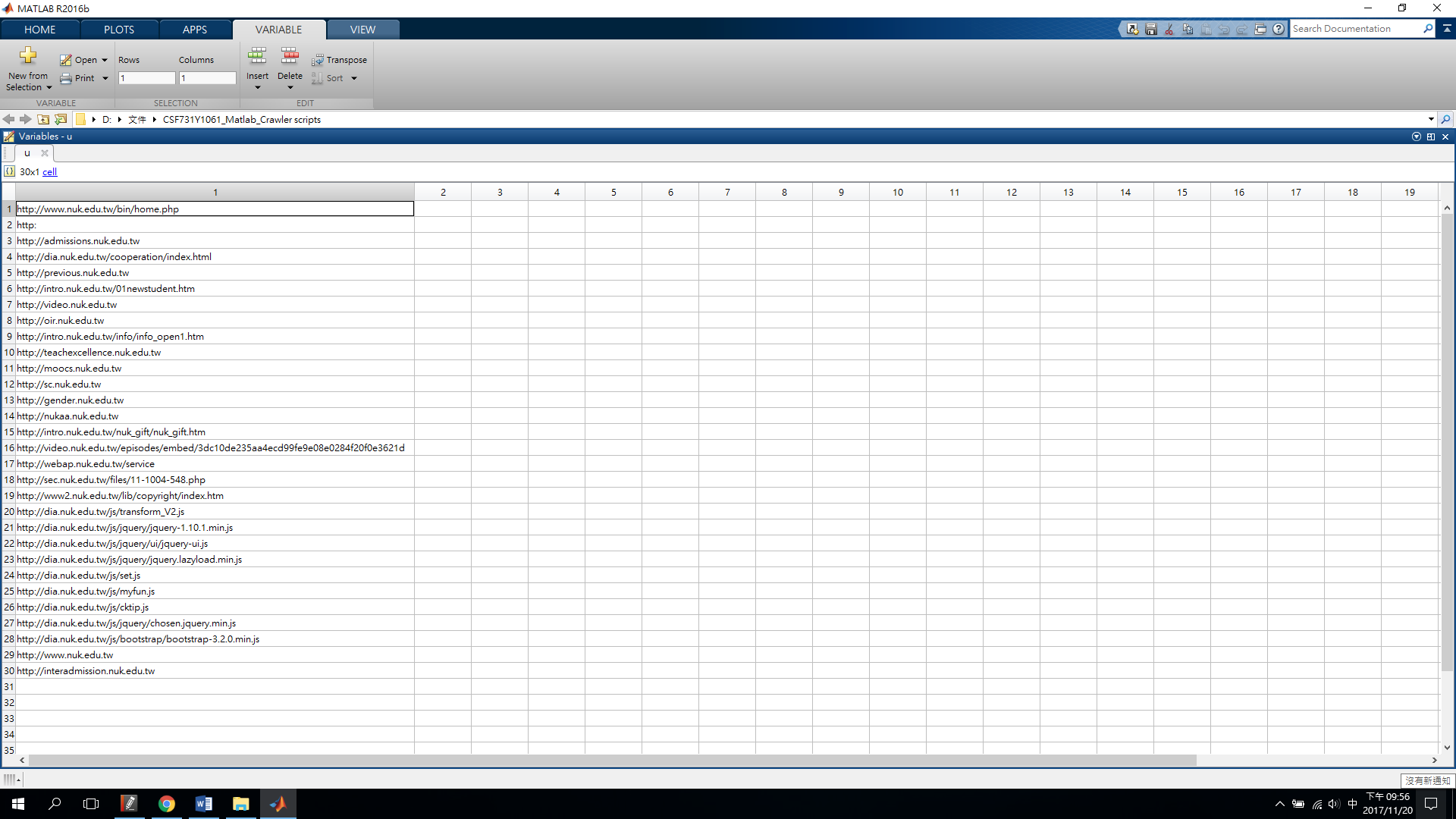
Problems:

1. From the practice above, is there anything you learn? This is an open question, so use your imagination.
2. 程式碼 :

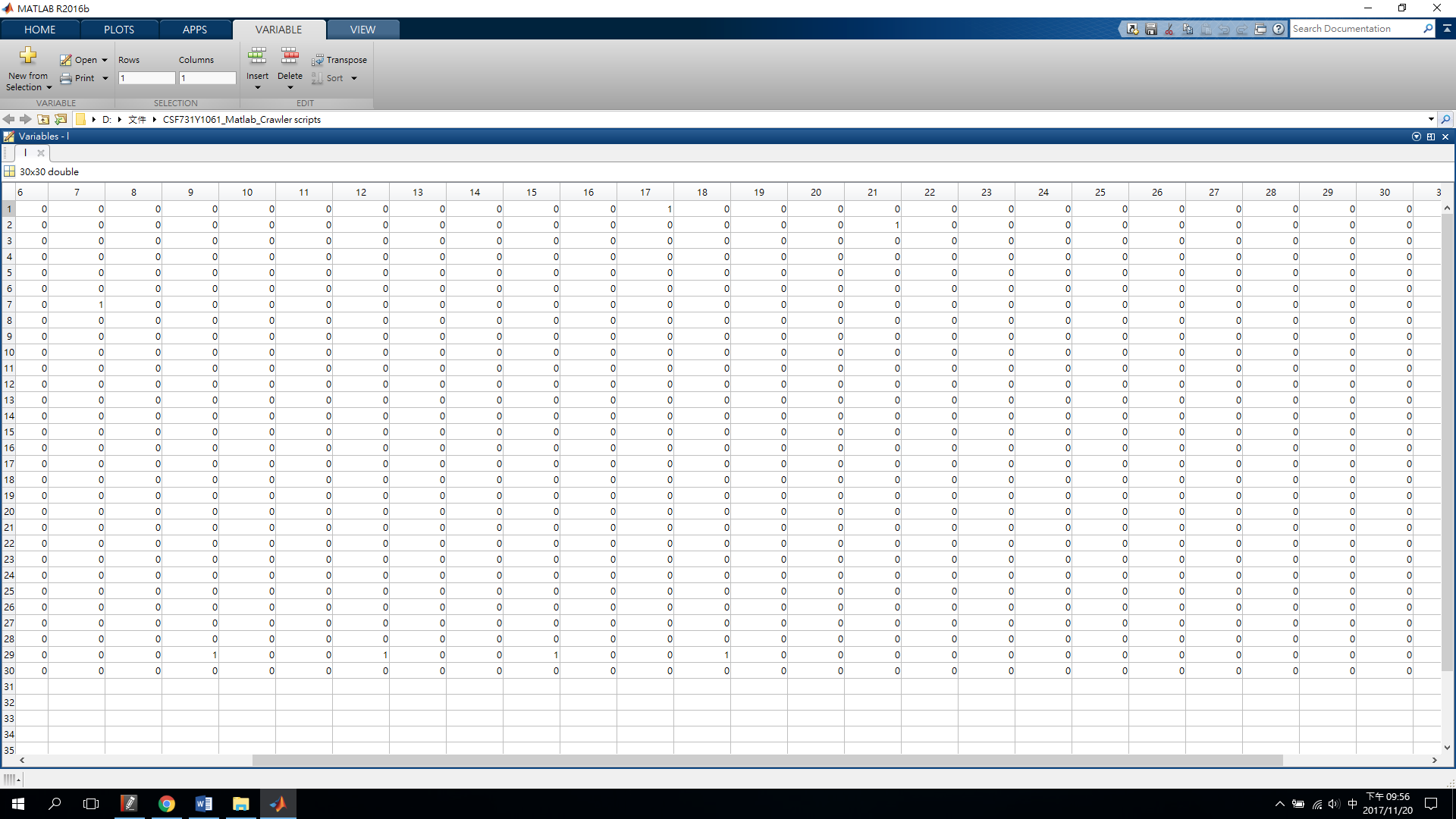
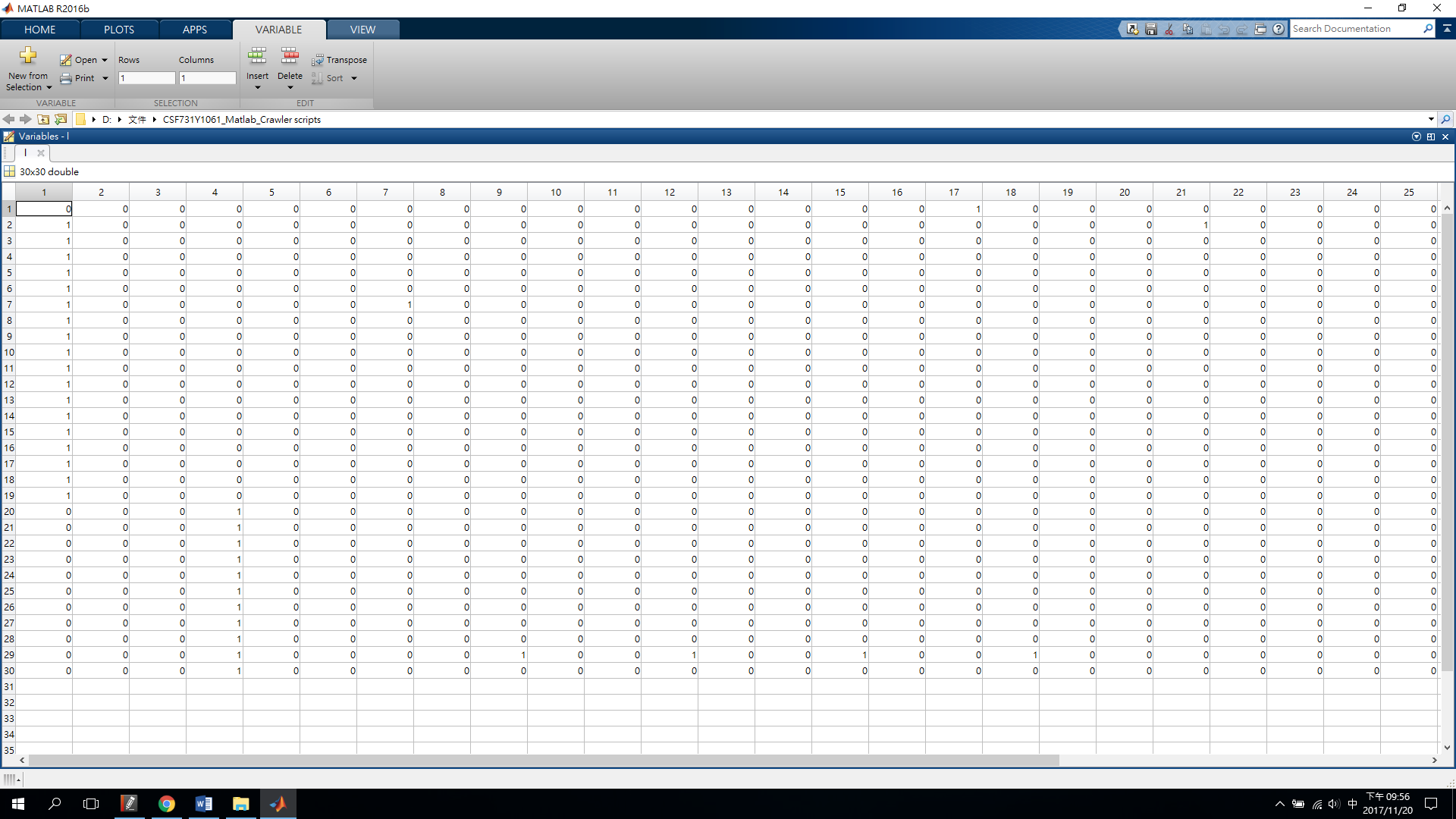
這份程式是透過老師給的main\_simple.m下去修改的，因為前面的練習要求說把n的數值改成30，所以老師原本寫的數值是10改成了30，其餘的部分都維持不變。



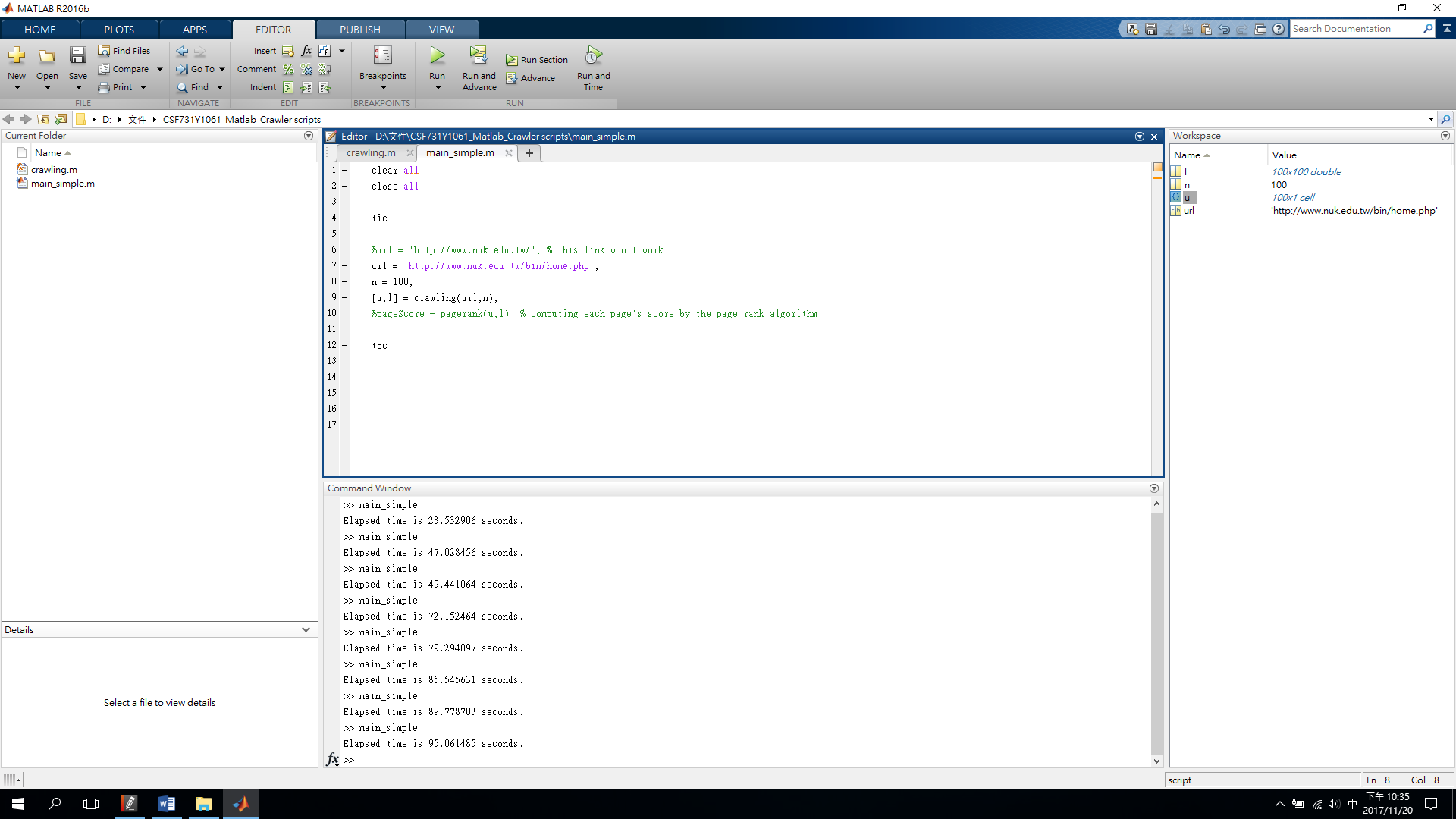
1. 抓取到的網頁 :



1. sparse binary adjacency matrix :



1. 心得 :



n = 1000

n = 30

n = 400

n = 500

n = 600

n = 700

n = 800

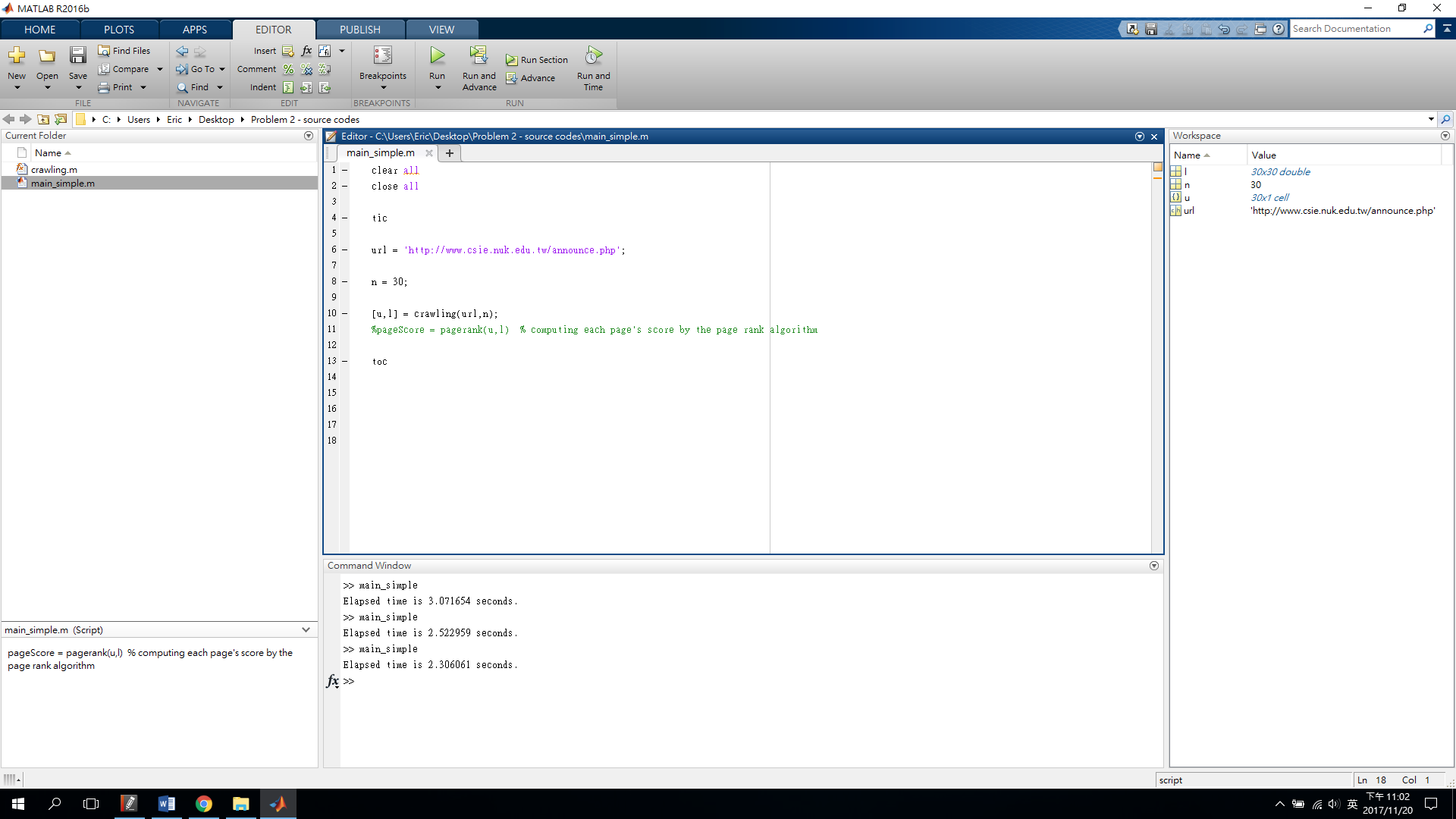
n = 900

在這份爬蟲的練習當中，我嘗試了去改變n值的部份來去看看是否有任何的變化產生，從上面這張圖所測試出來的結果可以得知n值與執行上的時間關係是呈現正比的關係，當其n值逐漸的變大的時候，其所需的執行時間是逐漸的增長，而且還有一個很微妙的變化，每一次的n值都是不斷的加10上去，然而其執行時間上的變化確實有不斷增長的趨勢，可是卻有著不同的增長幅度，在某幾個區間其執行時間是呈現倍數的成長，而在某些區間則是呈現出只是單純多幾秒的增長，這可以說是一個十分特別的現象，理論上而言，當n值不斷的增大，代表說爬蟲程式爬的深度勢必會越來越深，並且廣度也會越來越廣，那如果照這樣的趨勢的話，理當每次的增長幅度都應該是越來越大，可是從上面測試出來的結果卻不是如此，而我在猜測會造成這樣的狀況的主要原因可能有三種，第一種是爬蟲程式在爬的過程當中剛好有爬到幾個網頁是特別的複雜，導致整體在時間上的花費就瞬間增長了起來。第二種是從上圖可以發覺到一點，在執行時間上有倍數的情況發生感覺都是發生在最一開始所爬的網頁部份，雖然到後n是不斷的增大，可是呈現出來的執行時間和前一次的執行時間是沒有相差很多的，這也意味著說除了前半部分的網頁相對比較複雜之外，其餘後面的網頁就相對是比較的簡單，才會導致後半部的執行時間的增長幅度並不像前半部一樣如此的巨大。而第三種就是因為過於的頻繁去訪問同一個網頁，導致其網站的伺服器懷疑有惡意攻擊的情況發生，從而啟動了防衛的機制，讓其爬蟲的程式無法爬到所有範圍內的東西，而這也就導致所花費的時間就相對地減少，所以就正如同老師上課所講的在用爬蟲來去爬其他網站的時候要小心為慎，否則就有可能會產生這樣的情況發生。

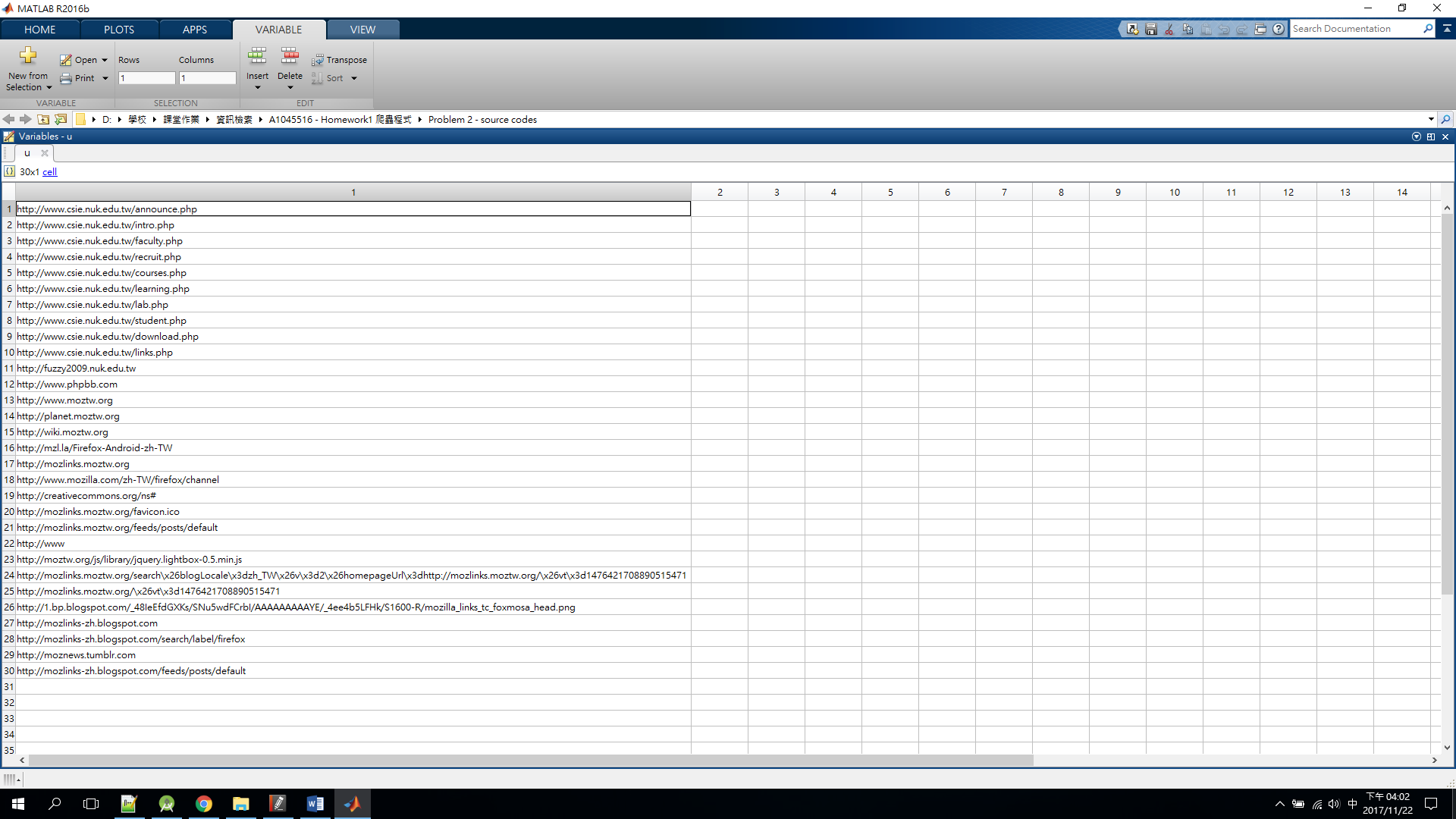
還有雖然其n值從10改成了30，但可以發覺到說其所獲得的sparse binary adjacency matrix是更加的稀疏，從這可以發覺到說其實網路上所存在的網頁其實是十分的分散的，通常要所有網頁彼此都有關聯性的話其實是有一定程度上的難度，但這種情況雖然可以促使我們獲得更多的資訊，可是伴隨而來的就是許多無意義的資訊，往往當我們輸入了關鍵字下去之後，我們都可以找到一堆的網頁，但在這些網頁當中實質上對我們是有幫助的其實是不多的，這也就導致我們資料的查找從原本的少數且集中變成了大量且分散的情況發生，從而導致我們要查找相關的資料所需花費的時間是越來越多，而這讓我更加深刻地感受到資訊檢索這門課存在的意義，因為他不單單只是要把這些分散且毫無相關的資料形成一個結構化的資料，更多意義上是要去大大的提升查找到對自己是有幫助且也是自己所期待的資料，所以資訊檢索這個領域可以說是佔領著一個十分重要的地位，並且這個領域對於我們的影響勢必會越來越巨大。

1. Determine how many pages are there in CSIE of NUK? (You may modify surfer.m to do this)
   1. 程式碼 :

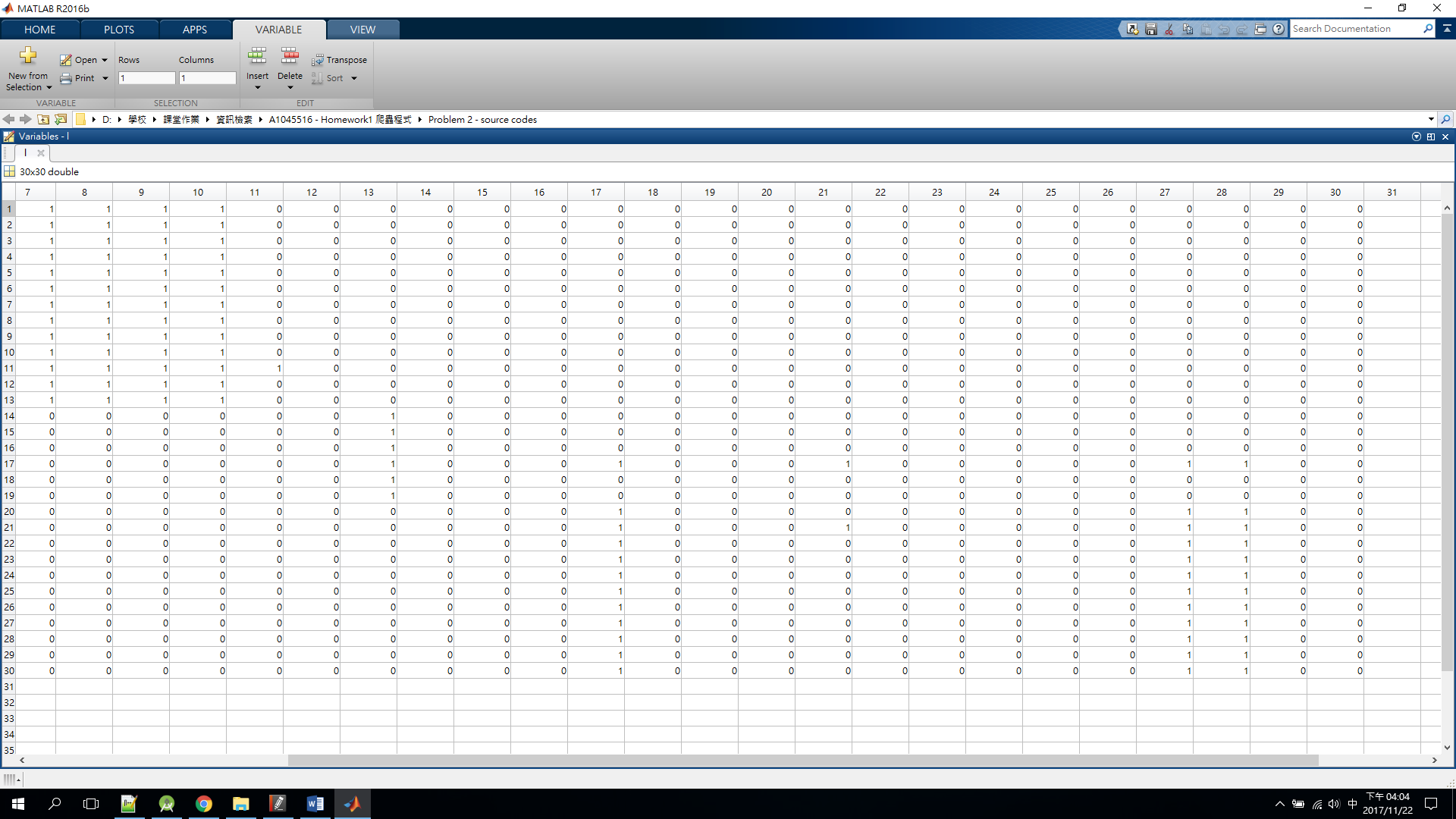
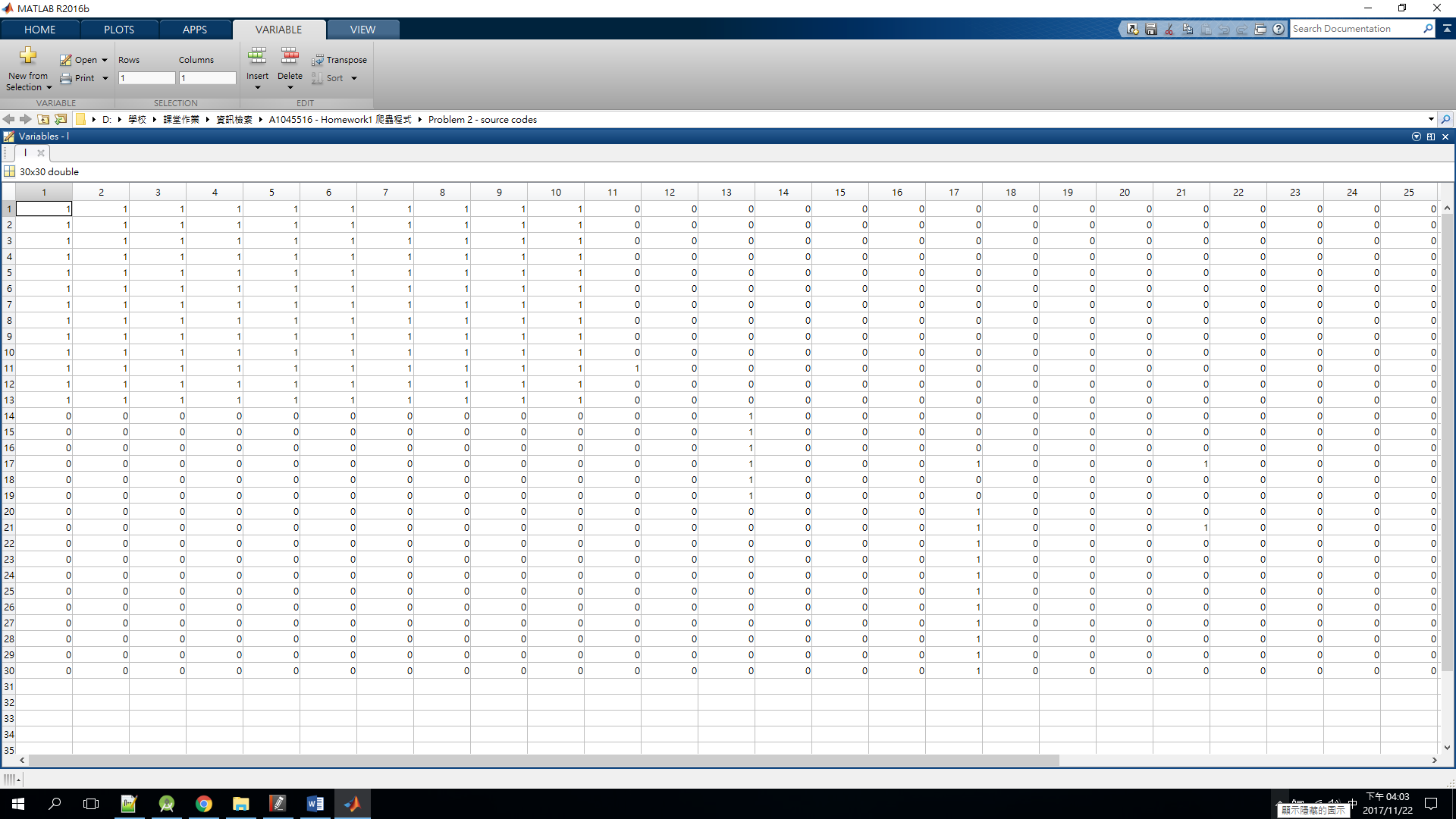
這個程式碼是取自老師所提供的程式碼，而做修改的部分就是將n從原本的10改成了30，其餘的維持不變。



* 1. 抓取到的網頁 :



* 1. sparse binary adjacency matrix :



* 1. 答案 :

題目問說在NUK的CSIE裡有多少個網頁，由上面所顯示出來的結果可以推得有10個網頁是在NUK的CSIE裡面。

Ans : 10個

* 1. 心得 :

由上面得出來的結果可以得知到兩點，第一個，從抓取出來的網頁的結果顯示出前10個網頁是NUK的CSIE的內部網頁，而其餘第11到第30的網頁都是一些其他的網頁，所以由此就能夠去推得說NUK的CSIE裡面的網頁總共有10個網頁。第二個，從sparse binary adjacency matrix我們可以看到前10個與NUK的CSIE有關的網頁彼此都是互連的，而這確實和NUK的CSIE這個網頁所呈現出來的網頁內容是一模一樣的，由此可得知我們所得出的結果是正確無誤的，而且從Problem 1和Problem 2可以更加的確立一件事情，那就是網頁的散佈其實是很分散的，除了像這種學校系網彼此的網頁都有著極高的關聯性之外，絕大部分的網頁都是互不關聯的，所以這也才會促使非結構化的資料比結構化的資料還要來的多上許多，而這也是呼應到課堂上所呈現出來的圖表，兩者所要描述的事情是一致的。

做完了上面這兩個問題之後，讓我發覺到說爬蟲看似是一個簡單的概念與想法，可是在實作上卻是困難了許多，並且其中還淺藏著許多深奧的學問，比方說如何讓自己爬出來的結果更加的精確並且又是自己所需要的，還有就是如何讓自己避免被網站的伺服器以為我們是在做惡意的攻擊等等……，我覺得這都是一種經驗上的累積，在一開始實作爬蟲的時候勢必沒有像專業人士一樣能夠輕鬆爬到自己想要爬到的東西，這點我就有很明顯的感受，因為我在一開始實作爬蟲的時候，所爬到的內容與自己所期望的結果是有所出入的，而且又因為是第一次碰到爬蟲並且又是第一次碰到像這樣的問題，所以當下的我是十分的頭痛的，但在經由老師上課的教導與助教詳盡的解說之後，才漸漸地讓我對於爬蟲有了更進一步的認識，不過就目前我所會的還只不過是皮毛上的東西而已，真正更難的部分其實才是整個爬蟲實作最為精華的地方，而這些也是我必須再加強訓練自己的地方。