CS 345: Homework 2 (Due: Sep 30, 2013 before 10:01am)

Rules. This is to be completed no later than the start of class on the day it is due. Earlier submission is also possible by email (read Handout 1).

Problem 1. (30 points) **Estimating the corpus size of Google and Bing**
We are asking you to search in queries 1-6 certain uncorrelated terms, and in queries 7-9 to improvise and determine yourselves something better than 1-6. Choose two words for `w1` and `w2` that are not correlated to obtain as high and accurate estimate values for the corpus size of Bing and Google as possible. For a counterexample a poor choice for the pair (w1,w2) would be (tropical, fish) or (tropical,storm). The textbook suggests (tropical,Lincoln). Use something else (eg. your last name for w1 or w2) and NOT along the lines of those three examples. Because this is not a verbatim query for words `w1, w2` we use the $\lceil\rceil$ symbols in queries 7-9 instead of $\langle\rangle$.

No	Query Term(s)	GOOGLE No of hits	BING No of hits
1	$\langle$ *tiger* $\rangle$		
2	$\langle$ *algorithm* $\rangle$		
3	$\langle$ *tiger algorithm*$\rangle$		
E1	What is the corpus size based on 1-3?		
4	$\langle$ *cs*345 $\rangle$		
5	$\langle$ *delicious* $\rangle$		
6	$\langle$ *delicious cs*345$\rangle$		
E2	What is the corpus size based on 4-6?		
7	$\lceil w1\rceil$		
8	$\lceil w2\rceil$		
9	$\lceil w1\ w2\rceil$		
E3	What is the corpus size based on 7-9?		

Table 1: Table for Q3.7

Based on the hits reported for 1-3, 4-6 and 7-9 estimate through probabilistic reasoning as specified in class, textbook, and the notes what your estimate of the corpus size of Google and Bing would be by filling in lines `E1,E2,E3`. Are these estimates consistent with those reported elsewhere say in Homework 1? Explain. Can you improve upon them some other way? Explain.

Problem 2. (40 POINTS) **Crawling...**

(a) **Unix command line wget.** `Fetch a copy of the course web-page.`

(a.1) Login on an AFS account say on `afsconnect1.njit.edu`. You are going to use the `wget` command to get a copy of the course web page (a subset of all the files that are accessible to you). For help about the options of wget type `man wget` under AFS or `wget --help`. Then work as instructed below.

(a.2) Although wget by default can be used to copy a remote file locally, in order to grab information from a remote site that consists of more than one file (eg directories) you need to use some recursive options of wget. Read the manual for details on options of `wget`. Note also than on a UNIX system by typing `du -s dirname` you can get info about the file size of all files in the directory (and subdirectories) named `dirname`. The size may be in kilobytes or mutliples of 512B (read the manual). Certain details are missing because you are expected to read the manual pages and also to improvise. Handout 0 should help as well.

(a.3) After you get familiarized with `wget` and its options do the following.

a.3.1 Grab the course web-page located `... alexg/courses/cs345/index.html`. Don't ask why part of the URL is missing. If you do it properly, you will collect close to 3MiB of information (note that i typed MiB not MB). Use proper options of wget to achieve the objective. What is the command (all arguments to wget) that you used to collect this information? **(5 points)**

a.3.2 Wget generates a log file while transferring files. You can capture into a file by redirecting output to af file as partially shown next

```
wget ....   alexg/courses/cs345/index.html > & capturefilename
```

a.3.3 How many (remote) files did you capture locally? **(5 points)**

a.3.4 What was the total size (volume) of those files ? **(5 points)**

a.3.5 How many URLs were not found to contain information (and thus no file was transferred locally) ? What error did they generate? **(5 points)**

a.3.6 How many URLs were protected behind a password authorization? What error did they generate? **(5 points)**

(b) **webcrawler.** Download the zip file **webcrawler_2_1.zip**. It is available in the handouts section of the course web-page. Unzip it by typing `unzip webcrawler_2_1.zip` and it will unfold 7 Python files. You will run Python on the same machine you used in part (a). There are several versions of Python on AFS but the one that works correctly is version 2.6. Thus your invocation will involve something along the lines

```
/usr/bin/python2.6 modular_crawler.py
```

There are some additional options (2 to 3) required but definitely use the last option which is a numeric value that gives an upper bound for the number of files to be downloaded. The tricky part of this assignment is to formulate precisely the first and second options. The previous problem's issues (eg a.3.1 migh offer some insight). (You don't want risk exploring all of the web.)

b.1 Run webcrawler on the course web-page and explore all options for a URL as specified in Handout 0. For the more productive and complete of these options, and the last option held at 100 what is the URL of the 100th page crawled? (If you do it properly the log file would grow to a size of at least 150KiB, if you do it incompletely it would be around 50KiB.) **(5 points)**

b.2 How many times is there a URL that contains the digit sequence 71854? What does it relate to? **(5 points)**

b.3 You can redirect the log into a file the way you did it before in a.3.2. A command `sort filename` on a UNIX system will sort the lines of filename. A `sort -u filename` will also eliminate duplicates. What is the total number of unique URLs listed? (Among the lines of the log the URLs start from character have http as the first 4 characters of a line.) **(5 points)**

Problem 3. (20 POINTS)
Perform the following study for Google and Bing. We execute two queries in Google and Bing and the collected results can be found in the last 4 pages of this homework as well as in the Handouts section of the course web-page. One query(Q1) is `CS 485 NJIT` and the other(Q2) is `CS 485 NJIT Web-search`. The CS 485 course number is a generic special topics one and thus other NJIT CS 485 courses that ARE NOT RELEVANT TO the CS 485 NJIT Web-Search offered in Fall 2010 and Fall 2011 are output. Thus for Q1/Q2 relevant documents are only those of to a CS 485 course offered at NJIT with subtitle Web-search as can be recognized either by having the name or web-page of the instructor listed or the title of the course in the 3-5 lines of each result. (Each one of the 4 pages contains 10 results.)

The first query is not very well thought-out. The second is more refined. You are asked to evaluate those two queries and also Google's and Bing's capabilities in generating relevant documents. We compared both engines in HW1. We evaluate the engines' search capabilities now. **When you examine the output results tabulate by starting with Google Q1, then Q2, and then move to Bing and do not deviate from this order! Do not include ad links in considering relevant documents!**

(a) Tabulate for each engine and query and engine, the number of hits reported in the result page attached (easy always the same), the number of relevant documents (be cautious when you read the entries), and the precision for the single result page reported (do not get confused with the huge number reported by Google or Bing!).

(b) Fill the following table with data (columns 2 and 3). If Google is better give Google 1pt and Bing 0 in columns 4 and 5. If Bing is better, give Bing 1pt and Google 0. If it is a tie, give each one 1pt each. Add up the points of columns 4 and 5 and provide sum in the last row. Who wins?

	Values		Points	
	Google	Bing	Google	Bing
# Retrieved docs for Q1				
# Relevant docs for Q1				
Precision for Q1				
# Retrieved docs for Q2				
# Relevant docs for Q2				
Precision for Q2				
Number of point wins (sum)	-	-		

Date Posted: 9/23/2013

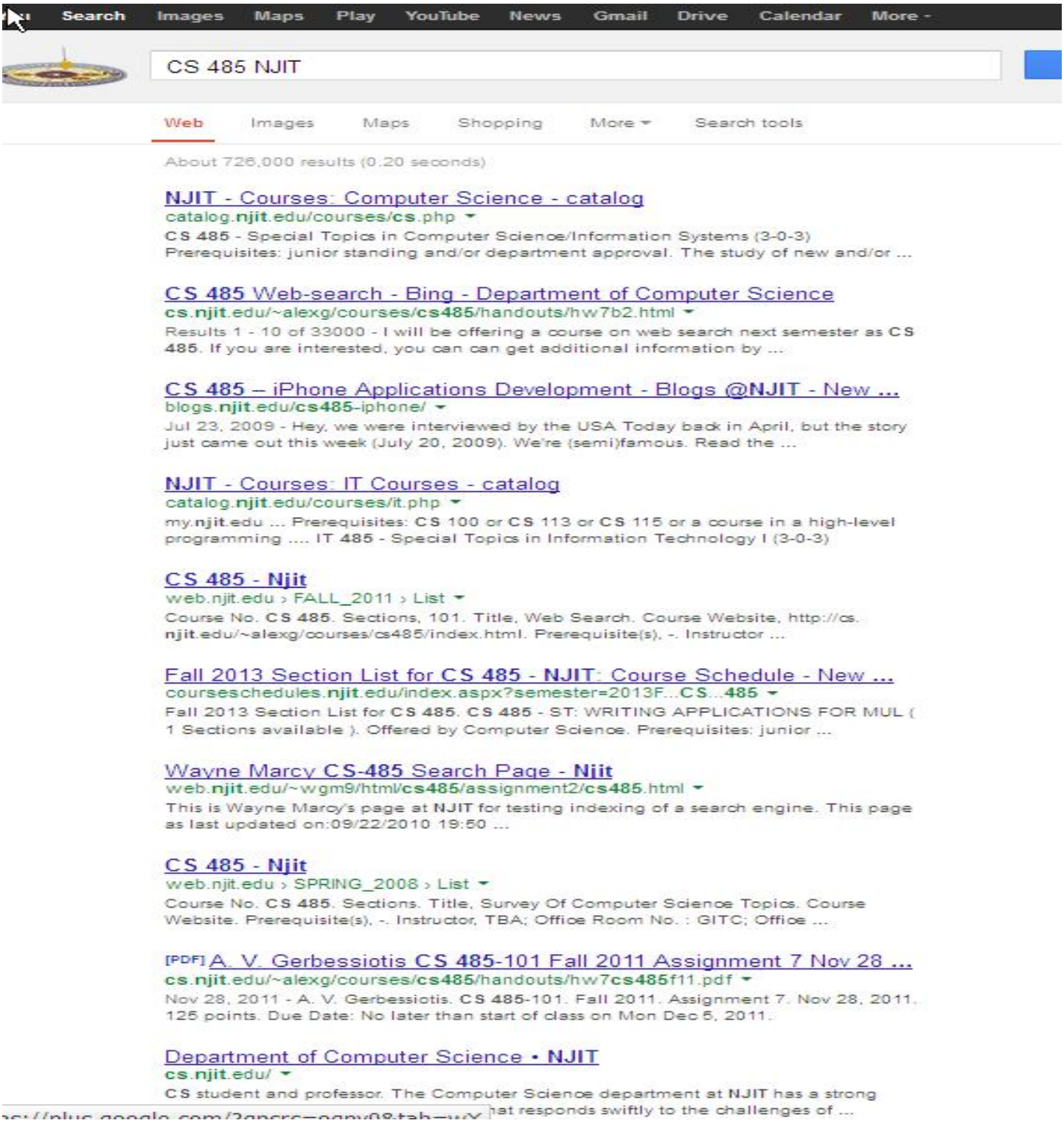

Figure 1: Google for `CS 485 NJIT`

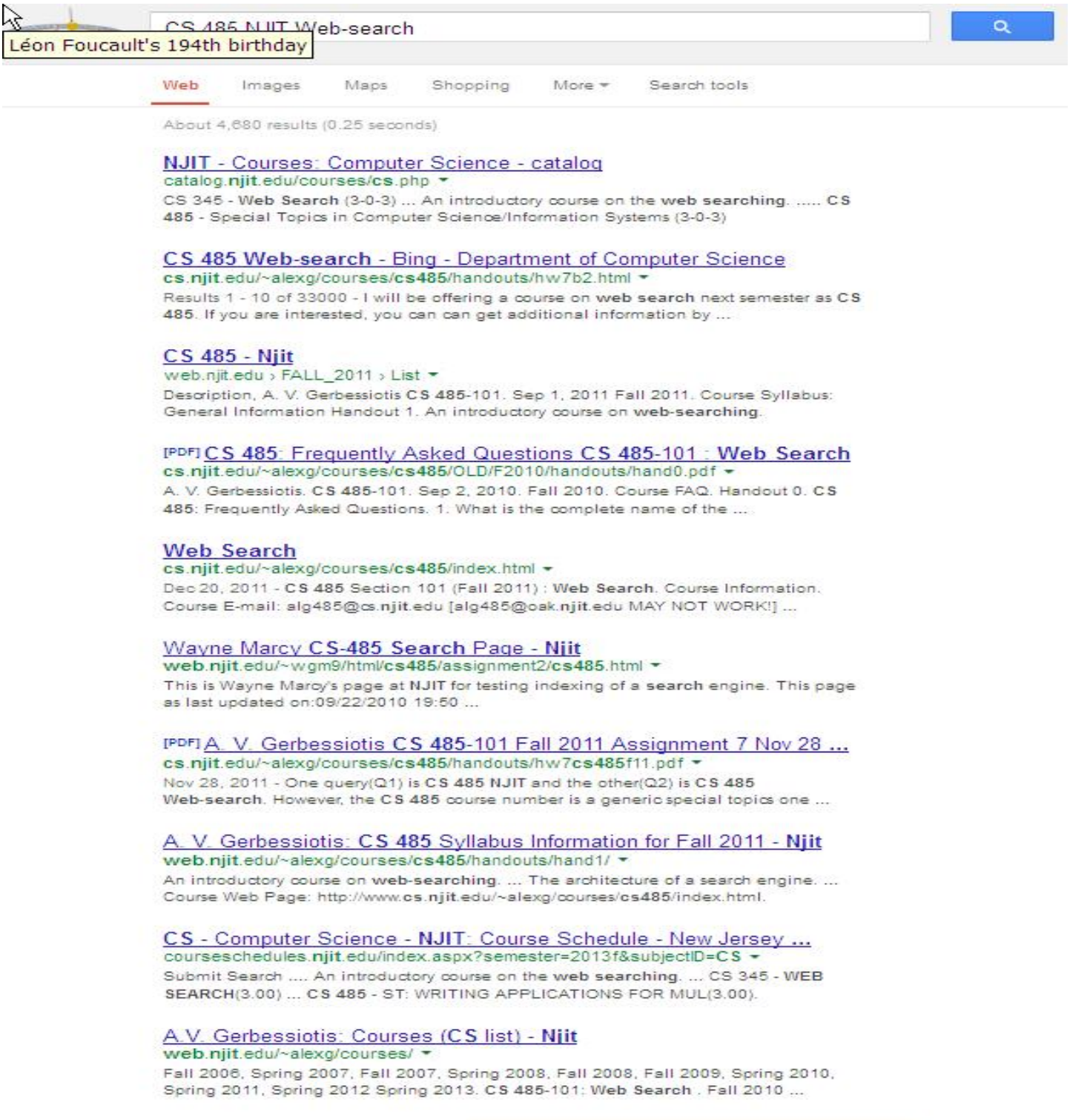

Figure 2: Google for `CS 485 NJIT Web-search`

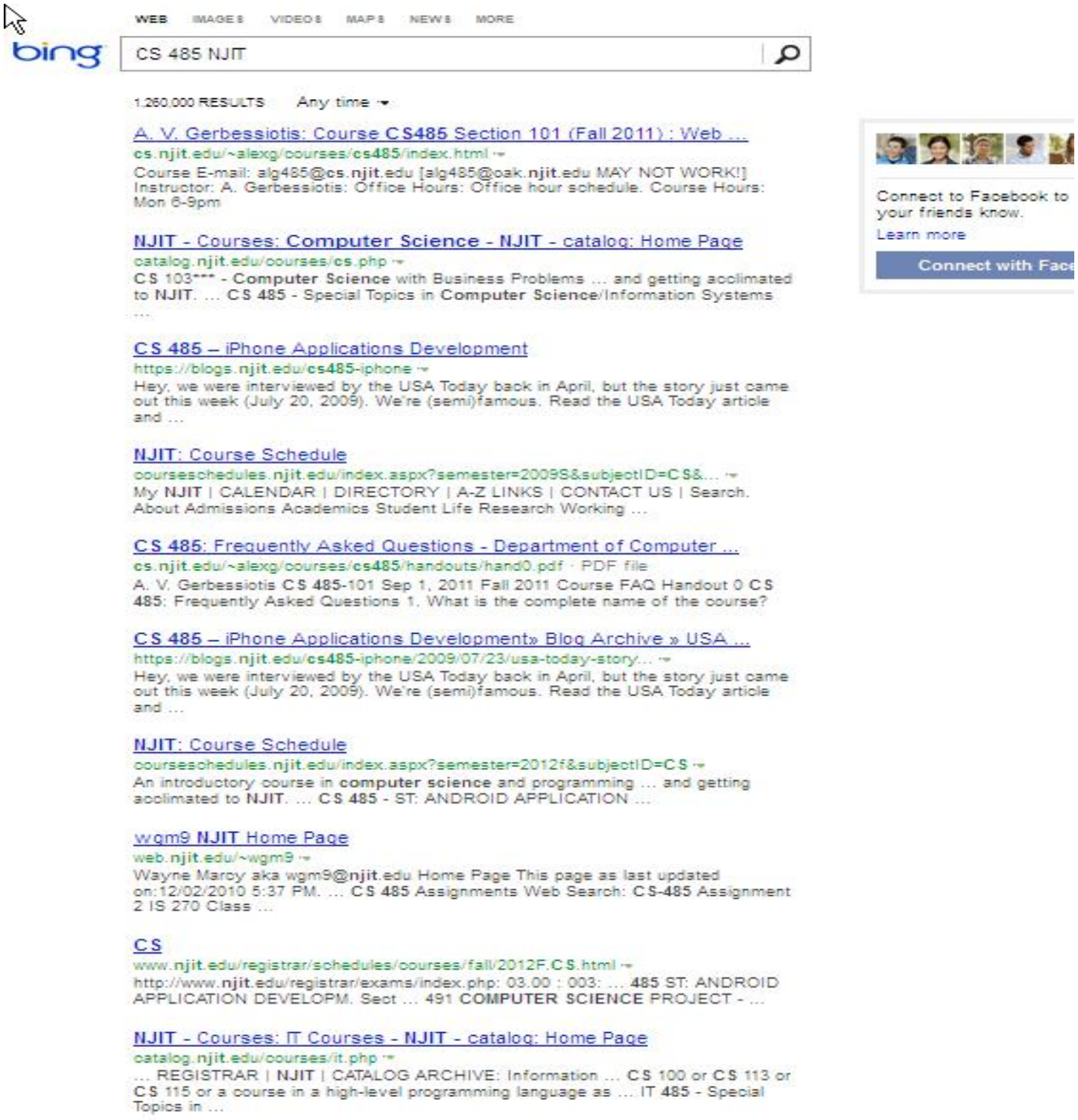

Figure 3: Bing for `CS 485 NJIT`

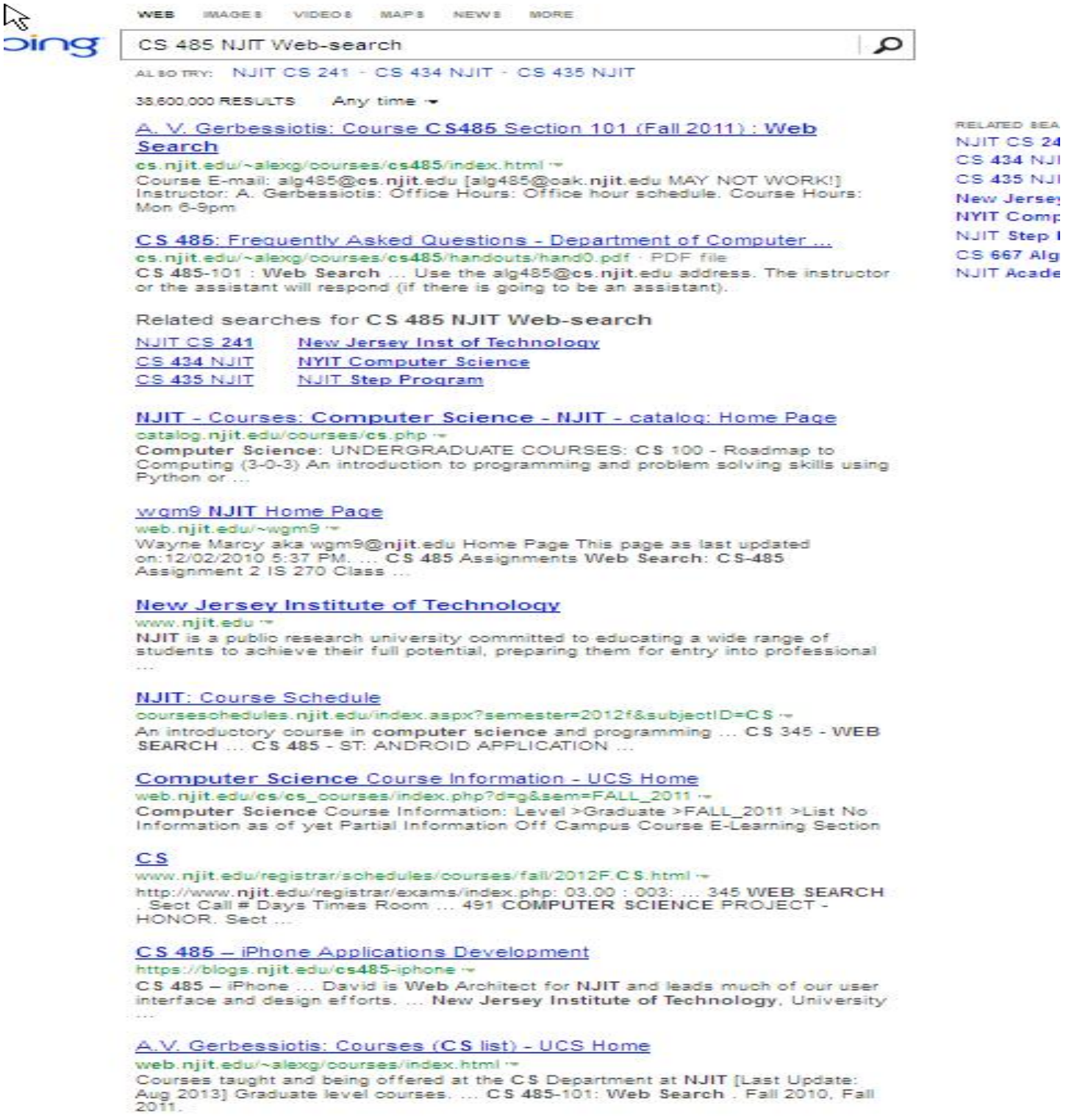

Figure 4: Bing for `CS 485 NJIT Web-search`