Hello everyone, my name is Qiang Tong. I am a graduate student from Computer Science Department. I am going to talk about how we collect the data.

We did not use Twitter API, because we hope to get the tweets when Apple products were released, but the Twitter API can only search tweets about 8 or 9 days ago. So we gave up Twitter API.

Then we thought maybe it's a good idea to use Twitter searching URL. Basically, you can use this url with the content to search on twitter, there are also two variables in this url, since and until. So we can get tweets which were tweeted like 3 years ago. But there is another problem, we can only get 20 tweets if we use urllib to analyze this url, because the other results would loaded automatically when we scroll down the page.

Then we found out another website to search tweet.
If you open this url in the browser, you can search tweets about iOS 4 in English between June 21, 2010 and June 22, 2010.
As we can see, iOS%204 means the key word is iOS 4.
AND language is English.
AND offset equals zero zero mean this is the first page of the results, you can set offset to 10, 20, 30 and so on and so forth to see load the result of next page.
AND minimum time and maximum time are a serials of numbers, which is called "timestamp", so you can set minimum time to another number to get result of another time. Technically, adding 3,600 to this number means an hour later.

However, there is another problem. When we tried to sea source HTML code of this URL, we cannot get the html of tweets, all we got is a bunch of Javascript code, which means all the results are generated by javascript.

Then we used another tool called "Selenium". Using selenium, we can open up a browser, and get the generated HTML of the tweets we need, and then parse it with BeautifulSoup and then save it into the database. I'd like to show you a demo how this works.