Archiving Small Twitter Datasets for Text Analysis: A Workshop Tutorial for Beginners

Dr Ernesto Priego, Centre for Human-Computer Interaction Design, City, University of London
Tuesday 26 June 2018, DH2018, Mexico City
ernesto.Priego.1@city.ac.uk
@ernestopriego

Fair warning: the steps here may have changed, may change or may have to be adapted on the go. Trial and error is OK: feel free to experiment and try out different combinations. I suggest trying out TAGS 6.0 and 6.1, though 6.1 is easier to set up.

Creating an Application to Get Data from the Twitter API

Before You Start1

Before you start this exercise you will need to:

1. Use Firefox or Chrome as your default browser. Please avoid Internet Explorer.
2. Be logged in your Twitter account.
3. Create a simple text file using Notepad (Windows) or TextEdit (Mac). They can be found in your list of programs in Start (Windows) or in Applications (Mac). Save this file in My Documents or similar folder you have set up for the conference or workshop. Save it as 'API Keys'. It's better if you save as .txt (simple text) instead of.rtf (Rich Text Format). You will use this file to paste your Consumer Key and Consumer Secret and keep it safe for future reference.

This exercise is composed by two tasks. The first task will show you how to create an application to tap into Twitter's API, in our case to get Twitter data. This means that creating this application is necessary to obtain access to Twitter's API. The second task will show you how to use a Google spreadsheet to capture streaming (live) data from Twitter in order to archive it.

Task 1. Creating an App Step by Step

All third-party (non-Twitter) services using Twitter data must do this in order to get permissions to Read only, Read and Write or Read, Write and Access direct messages. In our case we need permission to read only, as what we will eventually be doing (in session 5) is searching and archiving (i.e. 'reading') Tweets from Twitter's Search API.

Completing this task will provide you with two "keys": two strings or series of characters called

- Consumer Key (API Key)
- Consumer Secret (API Secret)

1. You need to have a Twitter account already. Make sure you are logged in Twitter. Then in a different browser tab, go to https://apps.twitter.com/app/new. You will see the page says "Application Management" as a title on the top right, and it should display your Twitter account profile picture on the top

left. If you have not updated your Twitter account profile picture, it will show the default picture of an egg (tsk, tsk).

2. In the same page you will see an online form with the following fields:

   - **Name**. This can be anything you like. Mine is “Ernesto’s TAG Archives”. Try not to use “Twitter” in the Name. 32 characters max.
   - **Description**. Anything you like as long as you use enough characters. Between 10 and 200 characters max. Mine is “App to collect and archive tweets”.
   - **Website**. If you have a personal blog your own blog’s URL. Your university URL works too. Be sure to begin the URL with http://
   - **Callback URL.** You must input https://spreadsheets.google.com/macros in this field. Make sure you insert this address correctly. We will be using a Google spreadsheet to collect Tweets, hence this callback URL, as it’s where the data will go to. (This may be different for TAGS 6.0).

3. Once you’re finished filling in the form, you will have to accept Twitter's “**Developer Rules of the Road**” These are Twitter’s terms and conditions. Very rarely will anyone read these!

4. We have no choice but accept them, so click on “**Yes, I agree**” box. Ideally we will have read them…

5. Then click on the “**Create Your Twitter Application**” button. If it shows you an error try again and ensure you have followed the steps above correctly.

6. You will be taken to a page that will show your newly-created application. You will see a navigation menu with the following tabs:

   - **Details (this will be the active tab)**
   - **Settings**
   - **Keys and Access Tokens**
   - **Permissions**

7. Click on the second tab, “**Settings**”. It will take you to a page where you can edit them.

8. Check that all your details are correct. Tick on the “**Allow this application to be used to Sign in with Twitter**” box. Do not click on the hyperlinked phrase at this stage. (You can do that later if you are curious).

9. On the same page if you scroll down a bit you will be able to replace the default application icon with an image of your choosing. This would be nice. (Default icons are an indication of carelessness). If you have images in your computer you can reuse and upload (for example the image file you used for your Twitter’s profile picture) click on “**Browse**” under “**Change Icon**” and follow the instructions. Uploading an image works the exact same way as attaching a file to an email or uploading media to your blog. **This is optional.**

10. You can also add your **organisation name** (for exampley your university’s name) and your **organisation web site** if you want. **This is optional.**

11. Scroll down and click/tap on the “**Update Settings**” button.

12. Now scroll up again to the top of the page and click on the third menu tab, “**Keys and Access Tokens**”.

13. You will see a section called "**Application Settings". It will provide you with the following information:**

    - Consumer Key (API Key) – [A string of characters]
    - Consumer Secret (API Secret) [A longer string of characters]
    - Access Level Read-only (modify app permissions) – [No need to do anything here!]
    - Owner [should be your Twitter username]
    - Owner ID [should be your Twitter username ID]

14. What you need is the Consumer Key and Consumer Secret. Copy and paste them onto your "API Keys" text file, making sure you identify which one is which and you don't miss any characters when copying and pasting.

15. Click/tap on the "**Create Access Token**" button. This will allow Twitter to list your application in your Twitter account's Applications section.

16. Click/tap on the "**Test OAuth**" button on the top right corner of the page. It should show you your Consumer Key and Consumer Secret again. You can calso copy them from there. You can then close the window. Congratulations, you have now created an application to work with Twitter's API!

Now that you have successfully created your app, make sure you keep your "API Keys" file safe. Your Consumer Key and Consumer Secret are like passwords, so they should be kept safe and secret. They are purposefully hard if not impossible for humans to remember. You will have to refer to your file in our next lab session.

Let's check if your app is listed in your account.

1. Go to https://twitter.com/. Because you are already logged in, it should remember your details.
2. On the top navigation bar, locate your small Twitter profile picture to the right of the "**Search"** field. Click/tap on it. A drop-down menu will pop up.
3. Click on "**Settings**".
4. On the left side of the screen you will see an "**Account**" menu with different options under your username. Locate "**Apps**". Clic/tap on it.
5. You will see a list titled "**Applications**". You should see your DITA application at the top. You will see how the information you added in the form should be displayed.
6. Do you have any other Apps? Which are they? What kind of permissions do they have? What do they do? To find out more about third-party applications, see https://support.twitter.com/articles/76052 . Compare with your classmates. Ask if anyone has any other apps listed in their account.

Task 2. Searching and Archiving Twitter with a Google Spreadsheet

For this exercise we will be using an application developed by Martin Hawksey. The application is a mashup using the Twitter Search API and Google's API. It uses an online Google Spreadsheet to collect Tweets and their related metadata. This allows you to have the data in a text format which makes it easier to analyse and visualise.

You will need to be logged in Google and Twitter before starting this exercise. You will need your Consumer Key and Consumer Secret from Twitter, which you should have obtained in the first section of this exercise. As usual, please **do not** use Internet Explorer for this exercise.

My suggestion is to read this document before following the steps on the computer. This will familiarise you with the task and the steps you will need to follow. The process is intuitive and it will become almost second nature with some practice but if this is the first time it is worthwhile to be patient and careful. Once you have read this document you can start again, this time following the steps carefully.

Sometimes the script is slow to react so you might have to tap/click twice before it works, but please be patient; doing the steps carefully (it often means slowly) is likely to produce the results we want faster.

1. Let's start by going to https://twitter.com/search-home and searching for #DH2018. [If you are using this tutorial later on, please use a current hashtag]. Look at what you get. Explore the results for say five minutes.
2. Now go to http://tags.hawksey.info/. Do not click on anything yet. Take some time to look carefully at what you have on the screen in front of you. Scroll up and down if necessary.
3. Click on **Get TAGS** (blue box on the centre of the page). It will take you to http://tags.hawksey.info/get-tags/
4. As in number 2 above, do not click on anything yet (I know you want to!). Take some time to look carefully at what you have on the screen in front of you. Breath deeply twice or thrice. Scroll up and down if necessary.
5. Click on **TAGS (New Sheets).**
6. A Google Drive (previously Google docs) page will open in a new browser tab.
7. Click on "Make a copy" (blue box).
8. **The Google Spreadsheet** "Copy of TAGS v6.0os" will open on your browser. **Enter a new document name on the file name** field on the top left of the screen. Rename the file something like "Julia's TAGS exercise". If your name is not Julie, then replace that with your own name ;-) Click **OK** (blue button).
9. You can close the previous browser tab now (TAGS v6…). Be careful not to close your own document!
10. You are now on your document. This is a Google spreadsheet living online. Take some time to look carefully at what you have on the screen in front of you. Breath deeply twice or thrice. Scroll up and down if necessary. You are looking at the **Readme/Settings** sheet. There will be a tab with that label at the bottom of the document. Do not click on anything yet (I know you want to!).
11. At the top of your document you should see its name ("Julie's TAGS Exercise"). Next to it there will be a 'favouriting' star (you can click on this if you want) and the icon of a folder.
12. Click on the folder twice, slowly. A pop-up message, **Move to**, will appear. Unless you already have a dedicated folder on your Google Drive, click on **Create new folder**. A **New folder** pop-up window will appear. In the field type your folder's name, something like " Twitter Archives". Click **Create.** The folder will appear with a tick next to it. Now click **Move** (blue button).
13. We are back at your document. You will see a **Enable custom menu** button in row 7. Click on it once.
14. A pop-up will appear: **Authorisation Required. This app needs authorization to run.** Click **OK**.
15. A **Request for Permission** pop-up from Google's API will appear. Note it will show your Gmail account on the top right. Read what it says, scrolling down if necessary.

16. Click **Accept** once. Be patient. The pop-up will then close.
17. A good tip: click once on the **Enable custom menu** button in row 7 again. A yellow caption box will run the script and say “finished script” then disappear. This ensures us our script will run OK. It might take several tries, depending on Google’s mood. This does not always work, but let’s try it out.
18. Locate your document’s top menu again. Locate **TAGS** (it’s next to Help). Click on it once. A drop-down menu will appear. The first option is **Setup Twitter Access.** Click on it.
19. Unless you have already used TAGS 6.0 before, a **Twitter Authorisation** pop-up message will ask you for your **Consumer Key and Consumer Secret**. Carefully copy and paste them from the text file you were asked to create on our previous lab session. Be careful not to leave any characters out when you copy. Click **Next**.
20. Click **Next** again. We are connecting to Twitter now.
21. A page on a new browser window will open. Click on **Sign in with Twitter.**
22. A page on a new browser window will open. Take some time to look at the URL of this page. If your Consumer Key and Consumer Secret were correct, it will ask you to authorize your application. Click on **Authorize app**.
23. You should see a “Success! Twitter has authenticated successfully. You can close this window” message. Act accordingly.
24. If the **Twitter Authorisation** browser window is still open, you can close this one too now.
25. Locate your document’s tab on your browser and go back to it. Be careful not to close it accidentally!
26. We will now enter the term we will be searching for and hopefully archiving. As this is our first exercise we will do #DH2018. Enter #DH2018 in row 9 (point number 3) in the search field.
27. We are almost there. On row number 15 you will find **Follower count filter**. I usually enter 2 in this field, so that only tweets from users that have at least 2 followers are included. Users with less than 2 followers tend to be automated accounts (bots, often spam bots).
28. Locate your document’s top navigation menu again. Click **File.** Drop-down menu appears. **Spreadsheet Settings.** In **Locale** and **Time Zone**, choose Mexico and Central Time (Mexico City).
29. Click **TAGS**. Drop-down menu appears. Click on **Run now!** Let the script run. Be patient.
30. Check that an **Archive** sheet tab has appeared at the bottom of your document, next to **Readme/Settings**. Click on **Archive.** What can you see? Take some time to explore the labels on row number 1 and the rest of your **Archive** sheet. Take your time.
31. Go back to **TAGS** on the navigation menu. Click on **Add Summary Sheet.** Allow the script to run. Check it out. Take your time.
32. Go back to **TAGS** on the navigation menu. Click on **Add Dashboard Sheet.** Allow the script to run. Check it out. Take your time.
33. You have now created an archive of your search collecting as many tweets as possible from the last 7 days. Well done! How does it compare to using Twitter Search?
34. Now go back to **TAGS** on the navigation menu. Click on **Update archive every hour**. This will ensure your document is archiving new tweets even when you are not there (until the Google spreadsheet cell limits is reached.

When cell limits are reached the script becomes unresponsive and the spreadsheet "breaks", which is why it is important to keep an eye on your live archives).

35. Go to **File** in the navigation menu. Locate **Publish to the web** and click on it. A pop-up appears. Click **Start Publishing**. Then click **Close.**
36. Go to the **Readme/Settings** sheet of your document. Click on **TAGS Explorer** and **TAGS Archive.** See what happens. This requires quite a lot of bandwidth so it might not always work. Be patient. If you like what you see, share on Twitter. Go meta: tag it with #DH2018!

Please experiment doing different searches. You will need to create a new TAGS document for every different search. I recommend that you already know what you want to search for before creating your document so you can give it that title (for example #DH2018 TAGS).

Text Analysis Exercise

You can download your Archive sheet from File- Download As. I recommend .cvs.

In "The Hermeneutics of Screwing Around," Stephen Ramsay (2014) contrasts the controlled, order nature of conventional search, like Google, with a new research methodology characterized by "surfing and stumbling," or, in his terms, "screwing around". Ramsay argues that the result of this research methodology would be revelatory.

He asks: "Could we imagine a world in which 'Here is an ordered list of the books you should read,' gives way to, 'Here is what I found. What did you find?'"

Inspired by Ramsay's provocation, your mission (because you have decided to accept it) will be to substitute "books" for "data" and to do just that- to use the text analysis tools Wordle, Voyant and Many Eyes to apply the methodology of "screwing around" to any of the text fields in one of your TAGS Archive spreadsheets (datasets).

1. Getting your data ready

Before you try out the tools you should prepare your data. Locate your workshop folder in your computer and locate the archive you created.

Identify what is the file format of your datasets. You may have downloaded them as .xls, .xlsx, .csv or even .txt.

As we saw previously, while .xls and .xlsx formats can accommodate different sheets in a single file, .csv means it is just one sheet per file. Because .csv stands for comma separated values, you can also have data as comma separated values in a simple text file if you have the data that way, for example

Id_str, from_user, text, created_a, time, geo_coordinates, user_lang, etc

Data presented in this way can be opened in Excel or any text editor (there might be differences in how the data is presented, so bear this in mind).

TAGS Datasets

If you don't have your tweet archives in .csv already, open your spreadsheet with Excel and select the Archive sheet. You can then **Save As** and choose .csv as the file format.

You are supposed to have created at least two TAGS archives. Choose the Archive .csv file you are interested in. Select all the text of the the **text** column of your dataset (this should be column C) by clicking on the letter **C** tab. Then do Command C to copy, or use the **Edit** menu then **Copy**. This might be quite a lot of text or not too much depending on how large your TAGS archive was.

Example of selected column C in a TAGS dataset

Now **create a new Notepad simple text file** (TextEdit on Macs) and **save it** with an appropriate name for your dataset in the right DITA folder. This should be .txt

Now **paste the text** you had copied from your .csv onto your .txt file. This should be only one column as you have only copied column C. Have all the characters copied correctly? If not, can you guess why not?

You will use this .txt file to work with the text analysis in a moment.

2. Surfing and Stumbling

Now you should have two .txt files with text data from your datasets, one from TAGS and another one from Altmetric.

The tools chosen for this exercise are mostly intuitive; allow yourself to experiment with them and to click on options without fear. Be bold and explore.

We will be using the following tool to "surf and stumble". Take the time to write down notes about your experience using it.

- **Voyant Tools** (try out its different tools, it might work better if you upload the file to be analysed instead of pasting the text, but feel free to experiment). http://voyant-tools.org/ Of the tools above this one is the most interesting and promising; developed by two LIS/DH scholars.

Some of the tools above allow you to apply **"stop words"** and to even edit them before you apply them (for example, you might want to exclude "RT" or "t.co" from your TAGS text dataset). Which tools allow you to do this? Could you figure out how?

Do take screen shots and figure out how to export what you create. You will need this later for your blog post. If you don't know yet how to take a screen shot or screen capture, search for it using a search engine such as Google.

3. Identifying themes

With Voyant, you will be able to get a list of words and their frequencies with the Words in the Entire Corpus tool. Word clouds, when properly done and analysed and the corpus is interesting and diverse, can also give you an indication of patterns and trends. Behind the cloud there is a data table that has quantified the frequency of each word (therefore visualising the word in different size according to its frequency). The word list can provide a first clue about the nature of the text.

Questions you can ask of the word list may include:

1. What are the basic preoccupations or trends expressed in this corpus?
2. What is unusual in the corpus?
3. Are there any patterns in the tenses of words used?
4. Given any expectations, are there words missing from the word list?

By identifying themes from word lists mined from a corpus, we are performing a qualitative analysis from quantitative data, via what Franco Moretti would call "distant reading".

4. Blogging

After "surfing and stumbling" (i.e. analysing the text) in both your datasets, you should craft a blog post that includes:

1) A screen capture— or the actual embedded pane or pasted text export in the case of Voyant— that shows the best (or most interesting) instance(s) of what it was that you found; and

2) A written account of 350-450 words that explains which tools you found the most interesting, what they helped you found and how, and why you think what you found was interesting. **Try focusing on something more than word clouds— for example, with Voyant, try the other tools apart from the Cirrus tool (Summary, Words in the Entire Corpus, Corpus).**

Download this from https://dx.doi.org/10.6084/m9.figshare.6686798