Mozdeh User Manual

This is a practical guide for using Mozdeh to gather and analyse texts. Please email Mike Thelwall if you spot any mistakes or omissions.

Contents

[1. Getting started with Mozdeh: Overview 3](#_Toc162157814)

[1.1. Downloading Mozdeh and installing the supporting files 3](#_Toc162157815)

[1.2. Starting a new project 3](#_Toc162157816)

[1.3. Collecting comments on YouTube videos 4](#_Toc162157817)

[1.3.1. Sign up for YouTube API Key 4](#_Toc162157818)

[1.3.2. Downloading YouTube comments for a new project 8](#_Toc162157819)

[1.4. Collecting Reddit posts 9](#_Toc162157820)

[1.5. Importing text data 12](#_Toc162157821)

[1.5.1. Hints for A: 12](#_Toc162157822)

[1.5.2. Hints for B: 12](#_Toc162157823)

[2. Quick overview of text analysis methods for a Mozdeh project 13](#_Toc162157824)

[3. Analysing posts 15](#_Toc162157825)

[3.1. Searching the posts 15](#_Toc162157826)

[3.2. Creating a time series graph of the posts 16](#_Toc162157827)

[3.3. Identifying keyword spikes in the data 17](#_Toc162157828)

[3.4. Co-word analysis – comparing topics or genders 18](#_Toc162157829)

[3.5. Word Association Analyses 22](#_Toc162157830)

[3.5.1. Set A vs. the rest: Word associations individual queries and/or filters 23](#_Toc162157831)

[3.5.2. Set A vs. Set B: Comparing one query against another 24](#_Toc162157832)

[3.5.3. Project vs. reference set: Word associations for an entire topic in Mozdeh 25](#_Toc162157833)

[3.5.4. Statistical approach: Familywise statistical tests 25](#_Toc162157834)

[3.6. Word association thematic analysis 26](#_Toc162157835)

[3.6.1. Examples of publications using WATA 26](#_Toc162157836)

[3.6.2. WATA instructions 29](#_Toc162157837)

[4. Network analysis 30](#_Toc162157838)

[5. Web text thick description analysis 32](#_Toc162157839)

[6. Marking and Removing Spam 34](#_Toc162157840)

[6.1. Getting started 34](#_Toc162157841)

[6.2. Marking individual items as spam 35](#_Toc162157842)

[6.3. Marking search results as spam 35](#_Toc162157843)

[6.4. Marking search results as not spam 36](#_Toc162157844)

[6.5. Marking duplicates as spam 36](#_Toc162157845)

[6.6. Working with the spam-free texts 36](#_Toc162157846)

[7. Old Twitter/x instructions 37](#_Toc162157847)

[7.1. Generating a full-scale set of queries 37](#_Toc162157848)

[7.1.1. Brainstorm a set of potentially relevant queries 38](#_Toc162157849)

[7.1.2. Pilot the queries by gathering matching posts 39](#_Toc162157850)

[8. FAQ 42](#_Toc162157851)

[8.1. Can Mozdeh collect all tweets relevant to a topic that is in the future or less than a week old? 42](#_Toc162157852)

[8.2. What happens to the data if Mozdeh crashes or if there is a power cut? 42](#_Toc162157853)

[8.3. Can I merge two projects? 43](#_Toc162157854)

[8.4. Does Mozdeh continue collecting data if the computer/laptop goes into Sleep or Hibernate mode? 43](#_Toc162157855)

[8.5. Can Mozdeh find YouTube comments with timestamps? 43](#_Toc162157856)

[8.6. Can I create a subset of my existing texts based on date? 43](#_Toc162157857)

[8.7. Can I create networks of texts from a given date range? 44](#_Toc162157858)

[8.8. Does Mozdeh store the Tweet URLs? 44](#_Toc162157859)

[8.9. Does Mozdeh store the gender/sentiment/country information with the texts? 44](#_Toc162157860)

[8.10. Can I preview a Mozdeh project while it is collecting data? 44](#_Toc162157861)

[8.11. Can I create a copy of my project with different time slicing (day/hour/month)? 44](#_Toc162157862)

[8.12. Why is REPLY @user in the Mozdeh tweet but not in the original tweet? 45](#_Toc162157863)

[8.13. Why do I see #NAME in my data? 45](#_Toc162157864)

[8.14. Why does Mozdeh get truncated retweets, limited to 140 characters? 45](#_Toc162157865)

[8.15. Why does my YouTube key not work or why has it stopped working? 45](#_Toc162157866)

[8.16. Can I find which country users are from? 45](#_Toc162157867)

[8.17. Can I find nonbinary users? 46](#_Toc162157868)

[8.18. Can I see user self-descriptions? 46](#_Toc162157869)

[8.19. Can I count word frequencies per user instead of per text (e.g., for Twitter timeline data)? 46](#_Toc162157870)

[8.20. Why am I getting a proxy server error? 46](#_Toc162157871)

[8.21. Why do I always get a file busy error when running Mozdeh? 47](#_Toc162157872)

# Getting started with Mozdeh: Overview

## Downloading Mozdeh and installing the supporting files

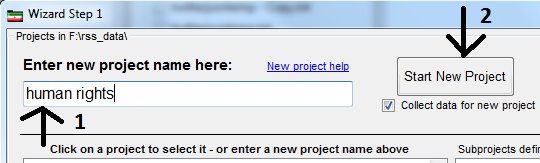
This section is a quick start and practice with the main features of Mozdeh. Section 2 onwards describes the steps necessary for a full-scale analysis with Mozdeh.

1. Download Mozdeh from <https://github.com/MikeThelwall/Mozdeh> or <http://mozdeh.wlv.ac.uk/> and save it to your Windows computer or USB stick. It does not need installation. It only runs on Windows and should be saved somewhere that data you collect can also be stored.
2. Double click Mozdeh and follow the instructions about selecting a folder in which to store your data. Mozdeh will ask to create a folder called Moz\_data on your computer. This folder will be initially empty but you will eventually populate this with new subfolders, one for each Twitter project.
3. Once you get to Wizard Step 1, download the zipfile of gender and sentiment files and default vocabulary, which is needed for some analyses (SentimentAndGender\_subfolders\_of\_moz\_data.zip). After starting Mozdeh for the first time, it will create a subfolder called moz\_data. Unzip the zipfile into this folder so that the three folders in the zipfile are unzipped inside of moz\_data. Make sure that they are directly inside moz\_data and not inside a subfolder of Moz\_data called SentimentAndGender\_subfolders\_of\_moz\_data.
4. See the next section on starting a new project.

## Starting a new project

This section describes how projects work and how to name them.

1. If this is the first time you have started Modzeh, go to the above section. Otherwise double-click Mozdeh on your computer to start it and enter a name for your project (e.g., *Human Rights*).



1. Information: Each project in Mozdeh must be given a name. This project name is also the name of the folder created by Mozdeh inside the main Moz\_data folder. The Wizard Step 1 lists all project names and you will be able to open the project at any future stage.
2. Information: New projects can be populated with data that is collected by Mozdeh or that you have already collected. If you want Mozdeh to collect data, click Start New Project. If you already have data, click Import Data.

## Collecting comments on YouTube videos

### Sign up for YouTube API Key

You must do this to collect any YouTube data.

**VERY IMPORTANT**: Follow the instructions below to 1) register as a Google developer, 2) add the YouTube Data API to your Google developer profile and enable it, and 3) create a credentials key to use the YouTube API.

PLEASE DON'T SKIP 2 (add the YouTube Data API to your Google developer profile and enable it), or your key will not work. If your key does not work, delete the key (in credentials, click the bin icon), disable the YouTube Data API v3(in Dashboard), then add the YouTube Data API v3 (again: in Dashboard click ENABLE APIS AND SERVICES, then find and enable the YouTube Data API v3).) and create the key again (i.e., steps 2 and 3).

**Step 1**: Log on to Google with your Google email account (get one if you don't have one) and then control-click on this link [developer start page](https://developers.google.com/youtube/v3/getting-started) to open it in a new browser tab (so that you still have this instruction window open).

[A screenshot of a computer

Description automatically generated](https://developers.google.com/youtube/v3/getting-started)

**Step 2**: Click on the Google Developers Console link in the middle of the Google Developer start page that you have just opened. This gives the screen below.

A screenshot of a computer

Description automatically generated

**Step 3**: Click the project button at the top left and then the Create Project link, giving the project your own name (e.g., YouTubeTest). Now wait 10 seconds for YouTube to process your request.

A screenshot of a computer

Description automatically generated

**Step 4**: If not selected already, click on the Library button on the left hand side to see all the APIs. Find the **YouTube APIs** section and click the YouTube Data API link.

A screenshot of a computer

Description automatically generated

**Step 5**: Enable the YouTube Data API by clicking the **Enable** button so that it looks like below.

A screenshot of a computer

Description automatically generated

**Step 6**: Click the Credentials section at the top left, click the **Create credentials** button on the right.

A screenshot of a computer

Description automatically generated

**Step 7**: Select **Other UI** from the drop-down list, and **Public data** option and click **What credentials do I need?**.

A screenshot of a web page

Description automatically generated

**Step 8**: You should now be shown an API key. Copy it (Ctrl-C) and email it to yourself so that you will not need to repeat these steps to get it again.

A screenshot of a computer

Description automatically generated

**Step 9**: You are now ready and can close this window. This key allows you to access YouTube videos forever. Paste (Ctrl-V) it into the yellow YouTube API key box in Webometric Analyst or Mozdeh to run searches.

### Downloading YouTube comments for a new project

This section assumes that you have signed up for a [YouTube API key](http://lexiurl.wlv.ac.uk/searcher/YouTubeKeyRegister.html) for permission to submit automatic YouTube queries (see the section above). This may take half an hour. If the key does not work first time, [get a new key and try again](file:///C:\Users\Mike%20Thelwall\Dropbox\My%20CS3%20Web%20sites\Mozdeh\faq.html#keyfail).

1. Start Mozdeh, enter a project name and click New Project.
2. Select the YouTube tab in the Data Collection interface (see below). Enter your YouTube Data API v3 key in the yellow box and then in the large white box at the top, enter one of (a) one or more queries, (b) one or more YouTube video URLs, or (c) one or more video channel IDs. Also select the appropriate radio button underneath and then click Get Comments on YouTube videos and wait for it to finish (may take minutes, hours, or days, depending on how many comments need to be downloaded).

A screenshot of a computer

Description automatically generated

1. Once Mozdeh has finished downloading the comments, it will ask you a series of questions about processing the texts. Here are the recommended options: (a) **Yes** to filter out duplicate texts, (b) **1** for the Which language question if your comments are mainly in English, (c) OK to the Preparing project warning, (d) OK to the indexing options dialog box (accept the defaults), and (e) wait for five minutes to several days for the comments to be processed, depending on how many there are.
2. You will end up at a large search filter interface that can be used to explore the comments. See the section of this manual on analysing texts. The comments are also saved to a plain text file inside a folder called \_Temp inside the project folder.

## Collecting Reddit posts

This page describes how to download up to 1000 posts in each of one or more subreddits with Mozdeh. It assumes that you have already downloaded and started Mozdeh.

1. Start Mozdeh, enter a project name and click New Project.
2. A screenshot of a computer

   Description automatically generated
3. Select the Reddit tab in the Data Collection interface (see below). Enter a list of subreddit names, one per line, in the big box near the top. Tick the type of content you want to collect (recent, new links and/or top links). Click Collect Reddit Posts and wait for it to finish (may take minutes, hours, or days, depending on how many posts need to be downloaded).
4. A screenshot of a computer

   Description automatically generated . A screenshot of a computer

   Description automatically generated
5. Once Mozdeh has finished downloading the posts, it will ask you a series of questions about processing them. Here are the recommended options: (a) **Yes** to filter out duplicate texts, (b) **1** for the Which langauge question if your comments are mainly in English, (c) OK to the Preparing project warning, (d) OK to the indexing options dialog box (accept the defaults), and (e) wait for five minutes to several days for the comments to be processed, depending on how many there are.
6. You will end up at a large search filter interface that can be used to explore the comments. See the sections of this manual on searching and analysing text. The posts are also saved to a plain text file inside a folder called raw\_data inside the project folder.

## Importing text data

Mozdeh can import data as long from any source as long as it is in a standard plain text format or one of the other pre-defined formats.

To import data, start Mozdeh, enter a project name and then click Import Data (instead of New Project). You will then be asked a series of questions about your data. This page describes the format required for Mozdeh to recognise text.

For general texts, your data must be in one or more files as follows:

EITHER

A1) Plain text, tab delimited format AND  
A2) The first line must contain the names of the columns AND  
A3) One column must contain text data AND  
A4) One column must contain the date in standard format, such as: 2010-12-01 or Mon Apr 04 20:13:22 +0000 2016 AND  
A5) One column must contain a text label for the data, such as a topic. These can be the same for all or some of the texts.

OR

B1) Plain text, one text on each line and no tabs anywhere.

### Hints for A:

If your data is in a different format, it may be possible to load it into Excel and then choose Save As and then Text (tab delimited) as the format.

If your time is in a different format, it may be possible to load it into Excel and then make a new column with the first date format above using the formula below, hwere A2 is the Excel location of the original date.

=CONCAT(YEAR(A2),"-",MONTH(A2),"-",DAY(A2))

### Hints for B:

For B choose the plain text only import option in the Import Data Wizard (probably number 11).

All texts will be given today's date as their date.

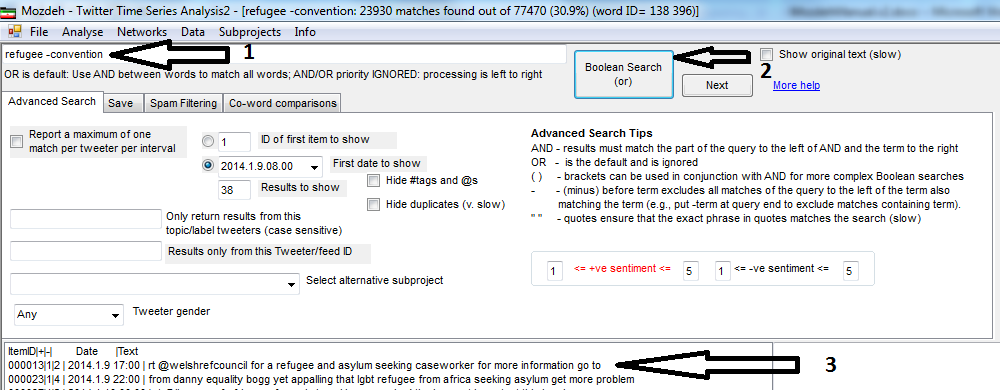
The label, author and URL of each text will be the filename. If you have different sets of texts that you want to compare, put them in different files in the same folder, give each file a meaningful name before importing into Mozdeh. The different files can be compared through Mozdeh's label function.

For this option, the author and date functions will not be useful.

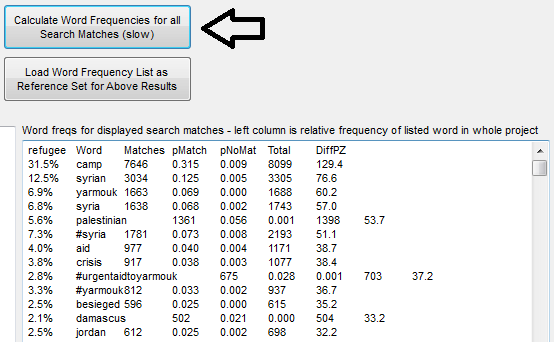
# Quick overview of text analysis methods for a Mozdeh project

This section gives a brief overview of the analysis methods available in Mozdeh. Other sections give more information about some of them.

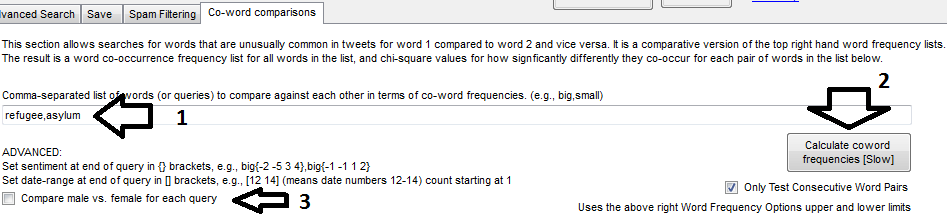
1. **Searching**. The initial screen for Mozdeh is its main search screen, which allows you to query the posts that you have downloaded. Test this by entering a query relevant to the posts you gathered (e.g., *refugee*) and then clicking Boolean Search to see posts that match that search (pointed to by the arrow 3 below – only the first 2 are shown but all matching posts can be browsed through). If there are matches that you are not interested in, try subtracting a term from the search. For example, putting *–convention* at the end of your query excludes all posts containing the term *convention*. (In the example below, this removed all posts about the UN convention on refugees.)

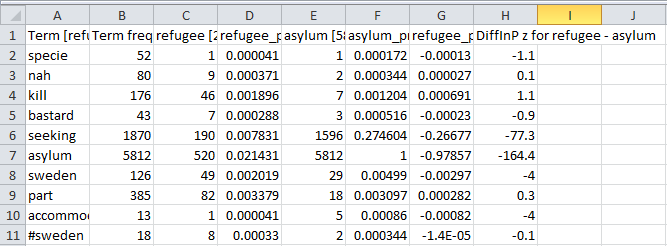


1. **Viewing the posts and posters**. The above screen shows the first page of results for the query *refugee* *–convention*. Note that the query results are a bit strange – they reflect the words indexed by Mozdeh and the form in which they are indexed, which is not always the same as the original text. For instance, plural words are converted to singular versions. The original versions can be read by checking the *Show original text (slow)* option and clicking a text in the search results list. This also shows the name of the poster.
2. **Words associated with your searches**. This list of words is a very powerful feature to suggest the main topics of interest relevant to your search within your posts. (warning: this may not be visible or may be partly obscured on your computer if it has a small screen). Entering a search in the main search box (top left-hand corner) and then clicking on the *Calculate Word Frequencies for all Search Matches (slow)* button gives a list of terms that occur more often for posts matching your search than for posts not matching your search. For example, the list below is for the search refugee. In other words, the listed terms occur more often in posts containing refugee rather than the other posts (e.g., containing asylum seeker, seeking asylum, migrant). The pMatch number shows the proportion of refugee posts containing the term (e.g., 0.315 or 31.5% of posts containing *refugee* also contain *camp*) and the pNoMatch number shows the proportion of non-refugee posts containing the term (e.g., 0.09 or 0.9% of posts not containing *refugee* also contain *camp*). Use this list of words to suggest topics of interest relevant to your search within your posts. Ignore the DiffPZ value at the end, it is a differences in proportions test z value used to rank the words in order of importance.

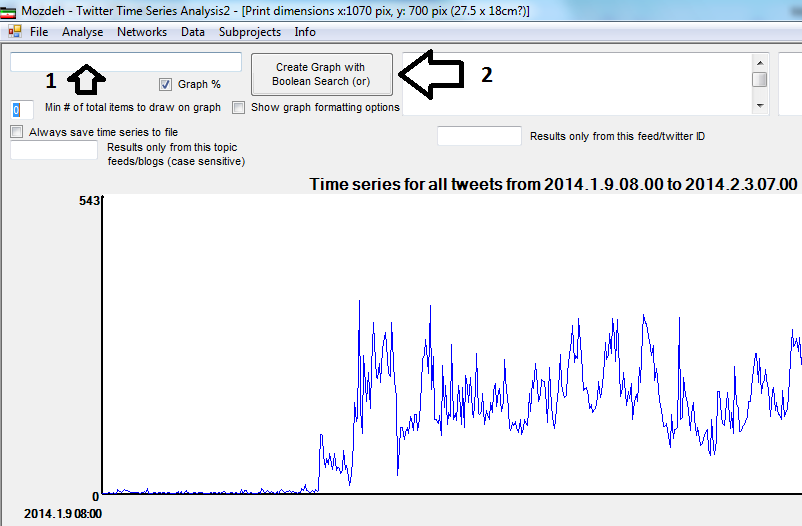


1. **Differences between searches, genders or sentiments**. This is another very powerful feature to suggest differences between main topics of interest, sentiment or gender within your posts. Mozdeh's co-word analysis feature is similar to the word frequency list above except that it lists words that are different between two or more searches, sentiments or genders. This is a bit trickier to use because you have to open up the results in a spreadsheet program. Click on the *Co-word comparisons* tab, enter refugee, asylum in the co-word box and click on the *Calculate coword frequencies* button. Try also entering one word and checking the Compare male vs. female for each query button, then clicking the *Calculate coword frequencies* button and accepting the filename it suggests. When it finishes it will open Windows explorer in a folder containing the results, which will have a filename something like refugee,asylum\_freqs - refugee AND asylum\_diffp.txt. Open this file and paste it into Excel or another spreadsheet to see word frequency information. This can be sorted by the *DiffInP z* column to get the words that are most important for one topic compared to another at the top (or the bottom) of the list. In the (unsorted) example below it is clear that seeker is much more strongly associated with asylum than with refugee.





1. **Time series graphs**. To create a time series graph of the posts, select *Graph Time Series* from the Mozdeh *Analyze* menu. Enter a blank search and click Create Graph with Boolean Search to generate a graph of the whole corpus. **WARNING: Your graph will be very simple or non-existent because for the pilot test you will only have data for one or a few hours**. You will get more interesting graphs if you collect posts over a few days or weeks with a full scale study. To create a graph of just the posts matching a particular search, enter the query and click the same button.



1. To switch back to the main search interface, select *Search* from the *Analyze* menu.

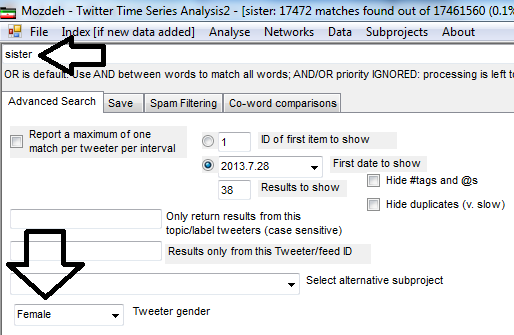
# ****Analysing posts****

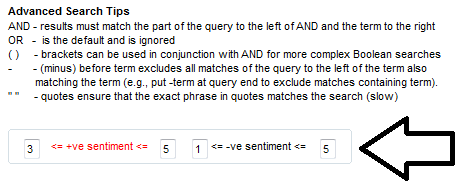
The posts can be explored by searching them, graphing them or automatically searching for keyword spikes. These are all described briefly below, although their application in research is not discussed here. Whichever is required, begin by starting Mozdeh, selecting the project, indexed as above and clicking the Open selected project button.

## ****Searching the posts****

Once Mozdeh is opened as above, the search screen will be shown. If it is not shown, select Search from the Analyse menu (see the arrow pointing to sister in the picture below). Simple queries can be entered into the text box near the bottom of the screen to search the indexed posts. The results are displayed in chronological order and there are various options for advanced searching, such as to jump to results from specific days. Clicking on a result will reveal the full text of the post. The default for searching is OR, so if you search for justin bieber then this will match posts containing either justin or bieber or both. To search for posts containing both justin and bieber, search for justin AND bieber instead.

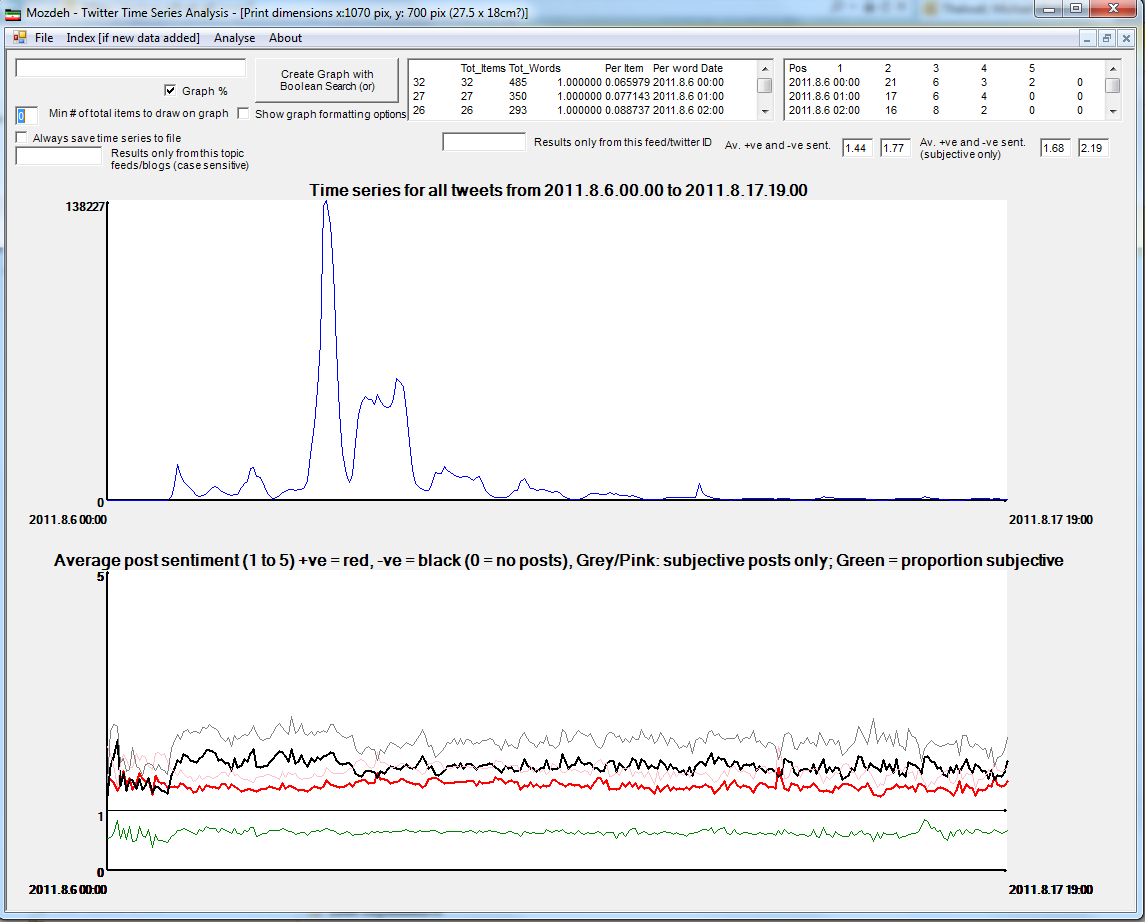
It is possible to search by poster gender and by the strength of sentiment of the posts (see below).





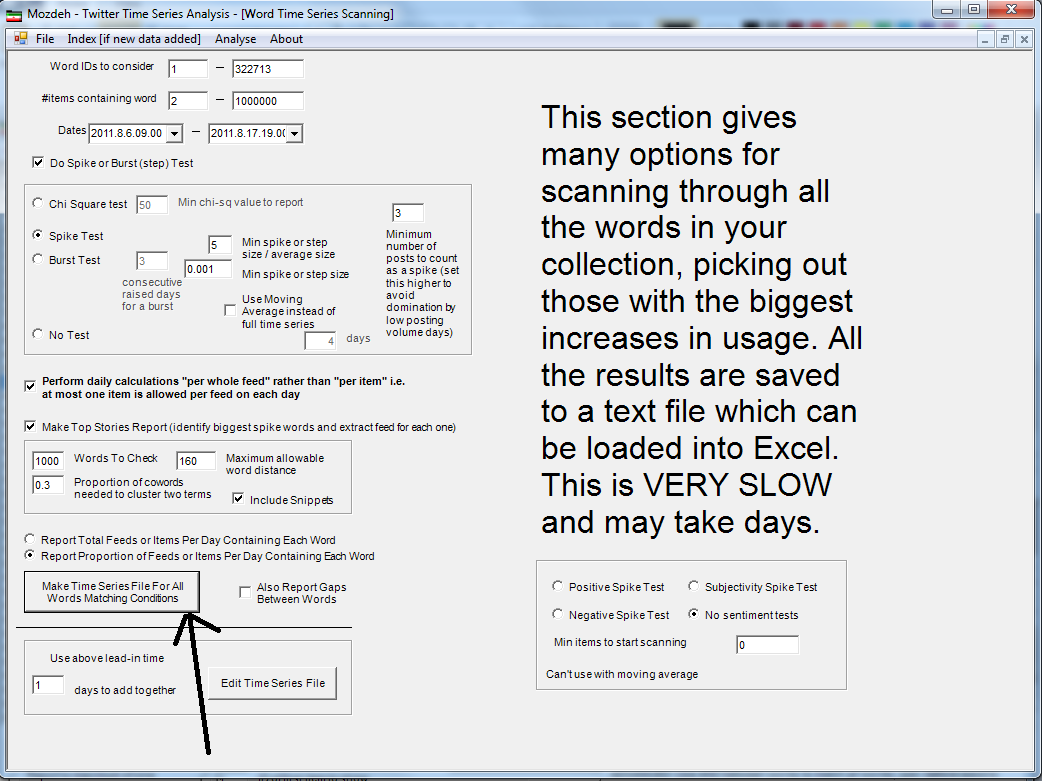
## ****Creating a time series graph**** ****of the posts****

To obtain a time series graph of the posts, select Graph Time Series from the Mozdeh Analyze menu. Enter a blank search and click Create Graph with Boolean Search to generate a graph of the whole corpus. To create a graph of just the posts matching a particular search, enter the query and click the same button. This is only worth doing if you have several weeks of posts so that there is a reasonable chance that the graph can reveal relevant patterns in the data.



## ****Identifying keyword spikes in the data****

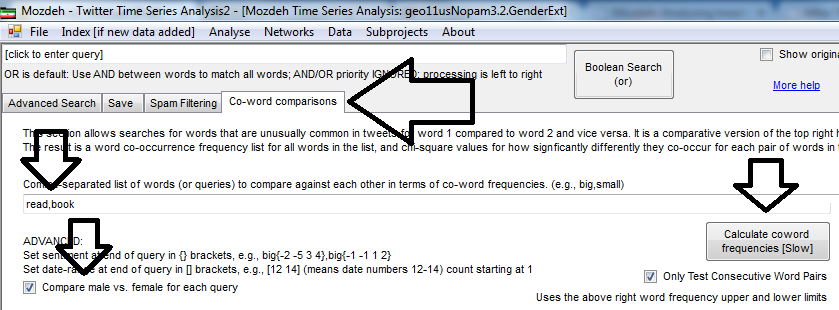
To obtain a list of the keywords that create the biggest spikes in the data, and are normally associated with significant events within the corpus, select Time Series Scanning from the Mozdeh Analyze menu. Click the Make Time Series File For All Words Matching Conditions button to run the search, which may take hours, and then load the results into a spreadsheet to interpret them.



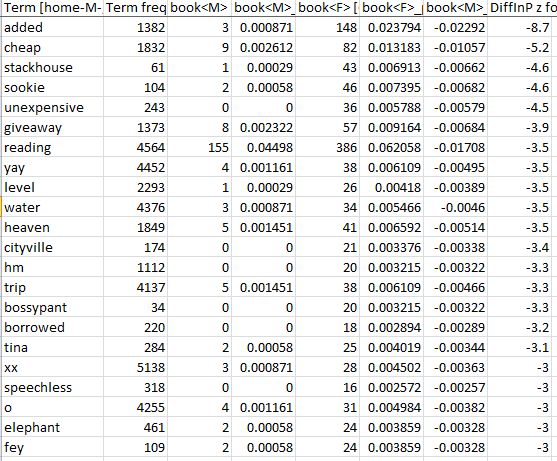
## Co-word analysis – comparing topics or genders

The co-word comparisons tab allows automatic comparisons of different topics within the posts or comparisons within topics (or within all the posts collected) by gender.

To compare two topics, construct queries for each one and enter them in the Comma-separated box in the Co-word comparisons tab, separating them by a comma. In the example below the two topics are "book" and "read". Clicking on the Calculate coword frequencies button will then produce a list of terms that tend to occur more often in posts containing "book" than in posts containing "read" and vice versa. These lists are plain text files that must be loaded into a spreadsheet program in order to be easily seen.

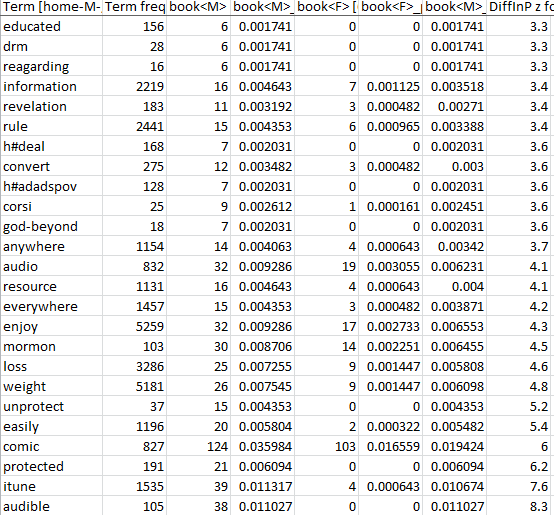


To compare male and female perspectives on a topic, enter one or more queries for the topic and also click the *Compare male vs. female* checkbox. This will produce lists of terms in posts matching the queries disproportionately often for male users compare to female users, or vice versa. The figure below is an extract from the results of this. The text file created by the Calculate coword frequencies button has been loaded into Excel and then sorted on the last column (DiffInP). The words in the list occurred in posts containing the word book disproportionately often for female users compared to male users. Some of them make sense (e.g., yay is more of a female word than a male word) but some do not (e.g., added, cheap). The top two suggest a possible influence for spam so this would need to be checked out and additional spam filtering done to remove the offending posts. The final column, DiffInP z is a statistical value that indicates the significance the difference between male and females is for the term, with larger values indicating that the difference is more significant (it is a differences in proportions test z value).

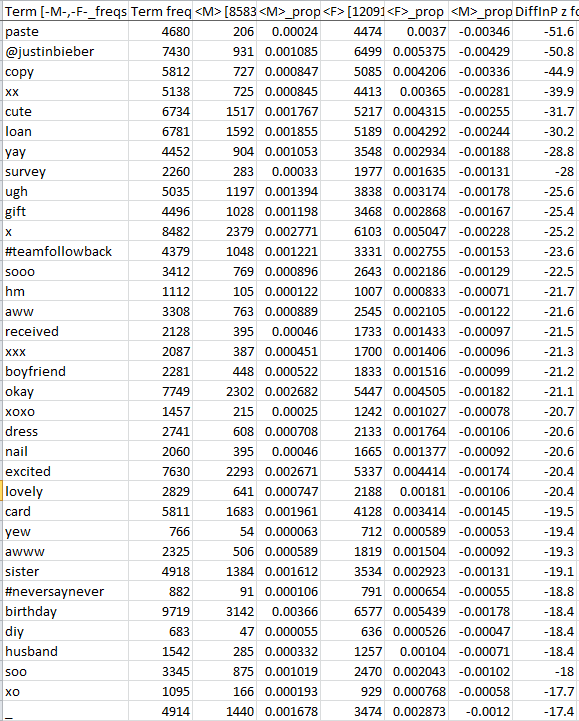
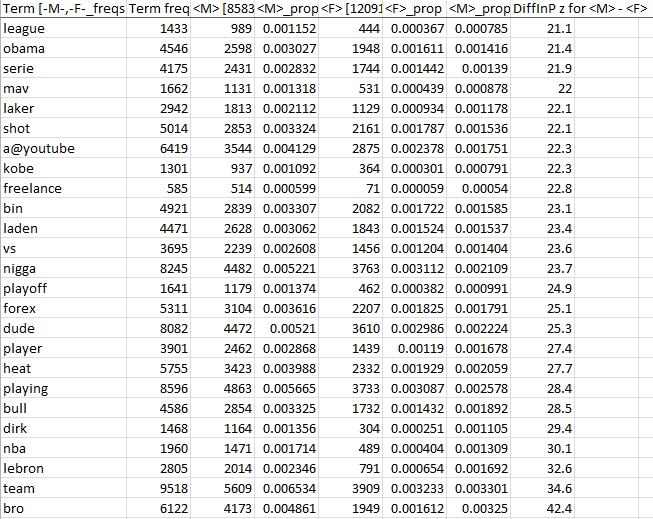


The non-spam results are words that suggest gender differences. These are useful to quickly suggest important differences, although the co-word method is only capable of identifying substantial differences that are reflected using common words.

The table below is the bottom half of the above table and shows words that occur disproportionately often in male posts. The results suggest that males are particularly interested in comic books, for example, because of the word comic in this list.



The tables below are the result of running a co-word test for gender with no words in the comma-separated text box. This points to gender differences for the whole collection of posts. Can you tell which is the male list and which is the female list?



## Word Association Analyses

There are three types of word association analyses that can be applied to texts gathered by Mozdeh. These can detect the following.

* **Set A vs. the rest**: Words that are more common in texts matching a keyword (or text query) and/or filter (gender, time, popularity, sentiment, users) than in the remaining texts.

A diagram of a computer

Description automatically generated

* **Set A vs. set B**: Words that are more common in texts matching one keyword (or text query) and/or filter compared to texts matching another keyword (or text query) and/or filter.

A diagram of a computer

Description automatically generated

* **Project vs. reference set**: Words that are more common in texts in the whole project than in a reference set of texts.

These tests all rely on the same statistical approach.

The tests are described in detail below for queries and on other pages for [gender](file:///C:\Users\Mike%20Thelwall\Dropbox\My%20CS3%20Web%20sites\Mozdeh\WordFrequencyAnalysesGender.html), country [Twitter only], [time](file:///C:\Users\Mike%20Thelwall\Dropbox\My%20CS3%20Web%20sites\Mozdeh\WordFrequencyAnalysesTime.html), retweets/likes/citations, sentiment, and users/source queries. They can also be applied to combinations (e.g., positive tweets from UK females vs. positive tweets from US females).

### Set A vs. the rest: Word associations individual queries and/or filters

It can be useful to identify the words that are unusually common in posts that match a particular query/filter because they may point to important aspects of the topic discussed. This splits the data into two parts: A and the rest. It finds words that occur more often in A than in the rest of the texts. This is achieved as follows.

1. Enter the query in the normal top-left search box and/or any gender, sentiment or other features, just as if you were going to run a search.
2. Click Mine associations for search and/or filters (slow).
3. Examine the bottom right text box (see below) to see a list of words that associate with your query and/or filters. These words are unusually common in texts matching the query compared to the remaining texts in the project listed in descending order.
4. [optional] To copy the results to a spreadsheet or word processor for analysis, right click in the box of results, and click Select All from the menu that appears, then click Copy from the menu and then paste the results Ctrl-V) into your spreadsheet or word processor document.

In the example below the query is spring and no filters were selected. At the top of the list, ignoring spring, the term roll associates with spring because 23.4% of the texts containing "spring" also contain "roll". In contrast, 0.7% of the remaining posts contain "roll". The chi-square is a statistical test of this association.

A screenshot of a computer screen

Description automatically generated  
The chi square value indicates how statistically significant the difference is between the frequency of the terms in the topic-specific collection of texts compared to all texts collected, with higher values indicating more significant results.

### Set A vs. Set B: Comparing one query against another

The above methods find words that associate with one term and/or filters compared to the rest of the project. You might want to compare a set of queries against each other. This splits the data into two parts: A and B. It finds words that occur more often in A than B and words that occur more often in B than in A. To do this, select the Association mining comparisons tab and enter your queries in the new text box, separated by commas. Then click Compare words matching the above queries (slow). The results will be saved into a set of files (not displayed on screen) that need to be loaded into a spreadsheet to be analysed. The list below will find words that associate with bus in comparison to train, with bus in comparison to car, and train in comparison to car.

A screenshot of a computer

Description automatically generated

After loading into a spreadsheet and formatting, you might see results like below. This is for the query like [1 62] against the query like [63 67]. It is a comparison of terms associating with the term like earlier posts (months 1 to 62) against words associating with it in later posts (months 63 to 67). Square boxes can be used to delimit dates with queries in this way. Benjamini-Hochberg statistics are given for these for the whole set of tests at once (i.e., not separately for positive an negative).

A screenshot of a spreadsheet

Description automatically generated

### ****Project vs. reference set****: Word associations for an entire topic in Mozdeh

To identify words that associate with your entire project, a generic list of word frequencies is needed that the topic words can be compared against. This should be a plain text file with a list of words and their frequencies from a common, generic collection of texts. A file of [word frequencies for a large collection of UK and Ireland tweets](file:///C:\Users\Mike%20Thelwall\Dropbox\My%20CS3%20Web%20sites\Mozdeh\resources\vocabularyFromUKirelandTweets.zip) will be used unless you have your own. Click the Load Word Freq List Reference Set button on the bottom right of the screen, follow the instructions, clear the search box, and click Mine associations for search and/or filters (slow). The results will appear in the text box below it.

### Statistical approach: Familywise statistical tests

Mozdeh uses a 2x2 chi-squared statistical test to assess the statistical significance of the difference between the proportions of texts containing each word in the two sets compared. When many statistical tests are run at the same time then the chances of drawing a false conclusion from at least one of them is high. This problem occurs with Mozdeh when it calculates many chi-squared values at the same time (controlling the familywise error). To reduce the risk of falsely believing that a term is significant, Benjamini-Hochberg procedure is used by Mozdeh. This tests all the words at once and reports the significant terms using a single test, controlling the risk of false positives due to multiple tests. To use this procedure, look at the stars in the right hand column (below). One star \* is significant at the (familywise) 5% level, two stars \*\* is significant at the (familywise) 1% level and three stars \*\*\* is significant at the (familywise) 0.1% level.

Technical note: The total number of tests used in the Benjamini-Hochberg method is the number of words that have a high enough frequency to be capable of generating a statistically significant result. This is normally the number of words that occur in at least 2 or 3 different posts.

The example below illustrates the star system. At the 5% level, 11 terms are statistically significant (the top 11 terms, from 'leave' to '#copelandbyelection'). At the 1% level, 9 terms are statistically significant (the top 9 terms). At the 0.1% level, 5 terms are statistically significant (the top 5 terms).

A screenshot of a computer screen

Description automatically generated

## Word association thematic analysis

These instructions cover techniques to apply the word association thematic analysis (WATA) method. WATA identifies themes that occur more often in one subset of texts than another. For example, given a set of Covid-19 tweets it might find that male-oriented themes include sports and news, female-oriented themes include personal safety and mental health, and nonbinary-oriented themes include politics and identity.

How does it work? For any set of texts, you set filters in Mozdeh to split the texts into two non-overlapping sets, based on gender, country, label, sentiment, retweet count, or a query. Mozdeh then finds words occurring more often in one set than another. You then read lots of texts containing each word to identify its typical context and then use thematic analysis methods to group the words into themes. Full details are given in the following book.

Thelwall, M. (2021). Word association thematic analysis: A social media text exploration strategy. San Rafael, CA: Morgan & Claypool.

### Examples of publications using WATA

The table below illustrates some WATA studies. Starred papers do not use the name WATA.

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic** | **Data** | **Comparison** | **Example findings** |
| Gender differences in reactions to Covid-19 | Tweets mentioning Covid-19 | Female v. male | Females tweet more about safety, males more about politics (Thelwall & Thelwall, 2020). |
| Personal experiences of ADHD | Tweets about “my ADHD” | ADHD v. other disorders | The brain is discussed as if it is a separate entity (Thelwall, et al., 2021a). |
| Evolution of #BlackLivesMatter during Covid-19 | Covid-19 tweets about racism | Tweets in four different periods. | The George Floyd killing led to tweets about systematic racism (Thelwall & Thelwall, 2021). |
| Self-presentation on Twitter | UK Twitter profiles | Female v. male v. nonbinary | Nonbinary profiles more likely to mention games and sexuality (Thelwall et al., 2021b). |
| Autism on Twitter | US autism tweets during Covid-19. | Autism v. others | Autistic tweeters do not have distinctive reactions to Covid-19 (Thelwall & Thelwall, 2022). |
| Gender differences in museum interests | Comments on YouTube museum videos | Female v. male | Females are more explicitly positive about content (Thelwall, 2018c). |
| Discussions of bullying in YouTube | Comments on YouTube influencer videos | Bullying v. Others | Strategies used to address bullying include generalisation (Thelwall & Cash, 2021). |
| Interests on Reddit | Reddit posts | Female v. male | Females more likely to mention doctors in the science subreddit (Thelwall & Stuart, 2019). |
| Factors associated with success in SteemIt | Steemit (like Reddit) posts | Successful v. unsuccessful posts | Financial news is less likely to be rewarded (Thelwall, 2018b). |
| Nursing research | Nursing journal articles\* | USA v. other countries | Nursing administration and management is not studied in some countries (Thelwall & Mas-Bleda, 2020). |
| US research subjects | US journal articles\* | Female v. male | Lists of gendered research topics and styles (Thelwall, et al., 2019b). |
| UK research subjects | UK journal articles\* | Female v. male | Lists of gendered research topics and styles (Thelwall et al., 2020). |
| Indian research subjects | Indian journal articles\* | Female v. male | Lists of gendered research topics and styles (Thelwall, et al., 2019a). |
| Research quality | UK journal articles | High v. medium v. low quality | Lists of research topics and methods assocated with higher or lower scores in the UK Research Excellence Framework 2021 evaluations (Thelwall et al., 2023) |

References to papers using Word Association Thematic Analysis (not necessarily using that term).

Thelwall, M., Abdoli, M., Lebiedziewicz, A. & Bailey, C. (2020). [Gender disparities in UK research publishing: Differences between fields, methods and topics](https://doi.org/10.3145/epi.2020.jul.15). El Profesional de la Información, 29(4), e290415. <https://doi.org/10.3145/epi.2020.jul.15>  
Thelwall, M., Bailey, C., Makita, M., Sud, P. & Madalli, D. (2019b). [Gender and Research Publishing in India: Uniformly high inequality](https://arxiv.org/abs/1812.03744)? Journal of Informetrics, 13(1), 118–131.  
Thelwall, M., Bailey, C., Tobin, C. & Bradshaw, N. (2019a). [Gender differences in research areas, methods and topics: Can people and thing orientations explain the results](https://wlv.openrepository.com/handle/2436/621959)? Journal of Informetrics, 13(1), 149-169.  
Thelwall, M. & Cash, S. (2021). Bullying discussions in UK female influencers’ YouTube comments. British Journal of Guidance and Counselling, 49(3), 480-493. https://doi.org/10.1080/03069885.2021.1901263  
Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P. & Levitt, J. (in press). [Terms in journal articles associating with high quality: Can qualitative research be world-leading](https://arxiv.org/pdf/2212.05420)? Journal of Documentation. https://doi.org/10.1108/JD-12-2022-0261  
Thelwall, M., Makita, M., Mas-Bleda, A. & Stuart, E. (2021a). [“My ADHD hellbrain”: A Twitter data science perspective on a behavioural disorder](https://doi.org/10.2478/jdis-2021-0007). Journal of Data and Information Science, 6(1). https://doi.org/10.2478/jdis-2021-0007  
Thelwall, M. & Mas-Bleda, A. (2018).[YouTube science channel video presenters and comments: Female friendly or vestiges of sexism](https://wlv.openrepository.com/handle/2436/620985)? Aslib Journal of Information Management, 70(1), 28-46.  
Thelwall, M. & Mas-Bleda, A. (2020). [How does nursing research differ internationally? A bibliometric analysis of six countries](https://onlinelibrary.wiley.com/doi/abs/10.1111/ijn.12851). International Journal of Nursing Practice, 26(6), e12851. https://doi.org/10.1111/ijn.12851  
Thelwall, M. & Stuart, E. (2019). [She’s Reddit: A source of statistically significant gendered interest information](https://arxiv.org/abs/1810.08091)? Information Processing & Management, 56(4), 1543-1558.  
Thelwall, M. & Thelwall, S. (2020). [Covid-19 tweeting in English: Gender differences](http://www.elprofesionaldelainformacion.com/contenidos/2020/may/thelwall-saheeda.pdf). El Profesional de la Información, 29(3), e290301.  
Thelwall, M., & Thelwall, S. (2021). [Racism discussions on Twitter after George Floyd during Covid-19: A space to address systematic and institutionalized racism](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3764867)? Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3764867  
Thelwall, M., Thelwall, S. & Fairclough, R. (2021b). [Male, female and nonbinary differences in UK Twitter user self-descriptions: A fine-grained systematic exploration](http://manu47.magtech.com.cn/Jwk3_jdis/EN/10.2478/jdis-2021-0018). Journal of Data and Information Science, 6(2), 1-27.  
Thelwall, M., & Thelwall, S. (2022). [Autism Spectrum Disorder on Twitter during Covid-19: Account types, self-descriptions and tweeting themes](https://europepmc.org/article/ppr/ppr341891). Data Science and Informetrics, 2(2), 1-12.  
Thelwall, M. (2018a). [Social media analytics for YouTube comments: Potential and limitations](https://wlv.openrepository.com/bitstream/handle/2436/620685/Social%20Media%20Analytics%20for%20YouTube%20Comments%20TF%20r2c_accepted%20preprint.pdf?sequence=1&isAllowed=n). International Journal of Social Research Methodology, 21(3), 303-316.  
Thelwall, M. (2018b). [Can social news websites pay for content and curation? The SteemIt cryptocurrency model](https://wlv.openrepository.com/bitstream/handle/2436/620844/SteemPaper1r%20anon%20mainR1bR2%20-%20Preprint.pdf?sequence=1). Journal of Information Science, 44(6), 736–751.  
Thelwall, M. (2018c). [Can museums find male or female audiences online with YouTube](https://wlv.openrepository.com/bitstream/handle/2436/621751/Museums%20and%20gender%20on%20YouTube2w2R1bPreprint.pdf?sequence=1&isAllowed=n)? Aslib Journal of Information Management, 70(5), 481-497.

See also: (word association analysis, but not full WATA):

Thelwall, M. (2021). [World Food Day on Twitter 2009-2020: Driven by UNFAO and aligned campaigns](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3807861). SSRN

### WATA instructions

Step 1: Collect your data (tweets, YouTube comments, other) with Mozdeh in the same way as for any other Mozdeh project. Go to the main Mozdeh search screen when you have finished.

Step 2: Decide what type of comparison you are making. If you are comparing the texts matching a filter against the rest (e.g., female-authored tweets against all other tweets) then follow the version of step 3 for a one-vs-remainder word assocation test (3a). If you are comparing one set of texts against another, but not the rest (e.g., nonbinary-authored tweets against female-authored tweets) then follow the version of step 3 for a A-vs-B word association test (3b).

Step 3a: Enter filters in the search screen to match your set (e.g, gender, country...), check the **Always save mine associations results..**. option in the **Advanced** menu and click the **Mine Associations for Seach and Filters (slow)** button. This should produce a file containing a list of words occuring more often in the set matching your filters than in the remaining texts. When a row is starred at the end, this means that the difference is statistically significant. The results are in a file in the folder that will appear when the procedure is finished. This list is sorted in descending order of statistical significance.

Step 3b. Select the **Association mining comparisons** tab. Enter two queries in the text box, following the instructions below. The two queries specify the two subsets to be compared (or enter one query and select the Male vs. Female option). Here are some examples.

\* The queries **nonbinary,transgender** are an instruction to compare texts containing "nonbinary" with texts containing "transgender".

\* The queries **<N>,<F>** are an instruction to compare texts authored by nonbinary people with texts authored by females.

\* The queries **<M>{{UK}},<F>{{UK}}** are an instruction to compare texts from the UK authored by males with texts from the UK authored by females.

\* The queries **our{{UK}},our{{USA}}** are an instruction to compare texts from the UK containing the word "our" with texts from the USA containing the word "our".

Now click the **Compare words matching the above queries (slow)**button. This should produce a file containing a list of words occuring more often in texts matching the first query than texts matching the second (and vice versa). When a row is starred at the end, this means that the difference is statistically significant. The results are in a file in the folder that will appear when the procedure is finished. This list is sorted in descending order of statistical significance.

Step 4: Configure the filters on the search screen for the first of the two queries compared, so that all texts to classify match the original queries. In the 3a case, keep the filters and/or queries used in step 3a. In the 3b examples, the following would be set.

\* Enter **nonbinary** in the search box.

\* Select **nonbinary** in the gender selection box.

\* Select **male** in the gender selection box and **UK** in the country selection box.

\* Enter **our** in the search box and **UK** in the country selection box.

Step 5: Click on the Save tab. Click the WATA button and select the file created by stage 3a or 3b (the version ending in diffp or diffp.txt). In reply to the questions, select column 1 and answer Yes to the question about search screen filters (unless you don't need to use them). This will produce a file that can be loaded into a spreadsheet that contains 100 randomly selected texts containing each of the first 100 words found by the word association tests (unless you changed the recommended answers).

Step 6a: Skip step 6.

Step 6b: Repeat steps 4 and 5, altering the filters in Step 4 to match those in the second query. For example, in the 3b cases this would be:

\* Enter **transgender**in the search box.

\* Select **nonbinary** in the gender selection box.

\* Select **female** in the gender selection box and **UK** in the country selection box.

\* Enter **our** in the search box and **USA** in the country selection box.

After Step 5, you don't need Mozdeh any more. Use the Step 4 file for the thematic analysis stage to find themes in the words found by the word association analysis. For example, if one of the top 100 word assocation words is "racist" then the file will contain 100 texts including the word "racist" and the context of the use of this word can be deduced by reading them. Repeating this for the other words and clustering the word contexts as part of a thematic analysis might put "racist" with an "Abuse" theme or a "Politics" theme, for example. **Important**: If you followed 3b/6b then you will have *two* files, one for each of the two queries. Only classify texts for a word from the file using the word more often. For example, if the term *stupid* is used more often by UK females than UK males for the third pair of queries above then text containing stupid should be classified from the first file (UK females) and the texts in the second file (UK males) should be ignored.

# Network analysis

The network menu in Mozdeh can be used to create networks of posters based upon their posts to other users with @username in their post. Try the options in the Network menu to see if this reveals interesting patterns in terms of users that post to others a lot or who are posted to a lot.

Two types of networks can be created by Mozdeh from the downloaded tweets.  
1) **Tweet network**. *This network illustrates which users tweet each other and how often.* If @user1 tweets mentioning @user2 then this creates an arrow from @user1 to @user2. If there are multiple tweets from @user1 mentioning @user2 then the width of the arrow is proportional to the number of tweets.

[March 2018 update] If a tweet is a reply to @user1 but @user1 is not mentioned in the tweet then Mozdeh adds REPLY @user1 to the start of the tweet at data collection time so that these replies can be included in the network.  
For example if @user1 - @user4 are in data set and there are two tweets from @user1:

* **@user1**:  Hello @user3 how are you?
* **@user1**:   Hello @user2 and @user3

Then there would be two arrows, one from @user1 to @user2 and a double thickness arrow from @user1 to @user3 but no arrow connecting @user4

A graphic of a user

Description automatically generated with medium confidence

**2) Co-tweeted network**.*This network illustrates which users tweeted to within the same tweet and how often.* If any tweet mentions both @user1 and @user2 then this creates a line between @user1 to @user2. If there are multiple tweets from @user1 mentioning both @user1 and @user2 then the width of the line is proportional to the number of these tweets.  
For example if @user1 - @user4 are in data set and there are three tweets:

* **@anyuser**: Hello @user3 how are you?
* **@anyuser**: Hello @user2 and @user3
* **@anyuser**: Hello @user3 and @user4

Then there would be two lines, one between @user2 and @user3 and one between @user3 and @user4.

A long silver and red hand watch

Description automatically generated with low confidence

Both types of networks can be created by clicking on the Network menu in Mozdeh and selecting the *Create networks of user connections in tweets for the Top N users: a) tweeter-@tweetee and b) @tweetee1-@tweetee2 networks* submenu item.

A screenshot of a computer error message

Description automatically generated

Here is an explanation of the questions that are asked after clicking this button.

* [you might not see this] "Select raw data file of tweets to use": Select a file in the folder that opens in the browser. This folder contains the tweets collected by Mozdeh in a simple plain text format. If there are multiple files, select the file containing “filtered” in the filename because it will have duplicate tweets removed.
* A long question about how to select the top N tweeters– answer 1 unless you are doing something fancy. This will include people that are most embedded in the network because they both send and receive tweets:
  + For tweet from-to network, select users with:
  + 1: Largest Minimum of {tweets FROM user to someone, tweets TO user from someone - @user}. *Users are ranked in order of the minimum number of tweets sent and received. So if Suki had sent 20 tweets and received 5 then her score would be 5.*
  + 2: Most tweets sent + received. *Users are ranked in order of the total number of tweets sent and received. So if Suki had sent 20 tweets and received 5 then her score would be 25.*
  + 3: Most tweets TO user (always used for co-networks). *Users are ranked in order of the total number of tweets received. If Suki had sent 20 tweets and received 5 then her score would be 5.*
  + 4: Most tweets FROM user. *Users are ranked in order of the total number of tweets sent. If Suki had sent 20 tweets and received 5 then her score would be 20.*
* “Enter maximum number of nodes for network. More than 50 makes the network too cluttered, and large numbers (e.g., Over 20,000) may result in an out of memory error when creating the network. Unless you are doing something fancy, accept the suggested answer of 50. Bigger numbers make cluttered networks and may crash the computer. Although it would be nice to make a complete network with all nodes, this is usually impractical.
* “Filename for main results”: Accept the suggestion.
* “Use raw numbers in the networks rather than scaled numbers? Click YES ONLY if exporting the network to a spreadsheet or another graphics package” Click NO if you just want to draw the network. Click YES if you want to know the exact number of connections rather than needing the network diagram.

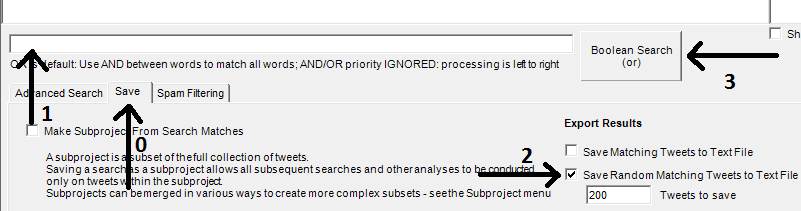
# Web text thick description analysis

This section describes a specific overall methodology for a complete project analysing a collection of posts. A *thick description* is a method of creating a detailed description of a topic, as captured by Twitter searches. It includes a content analysis, time series analysis and a word frequency analyses. This is based upon the chapter 8 of the free book below.

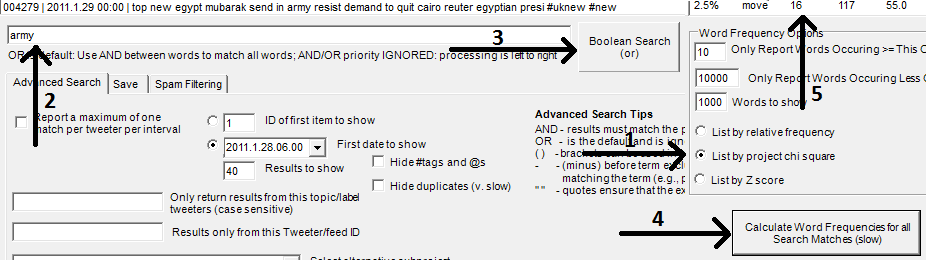
Thelwall, M. (2013). Webometrics and social web research methods [free in-progress draft copy]. University of Wolverhampton. <http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf>.

The instructions below assume that you have completed the task above and have a collection of posts collected by Mozdeh to analyse.

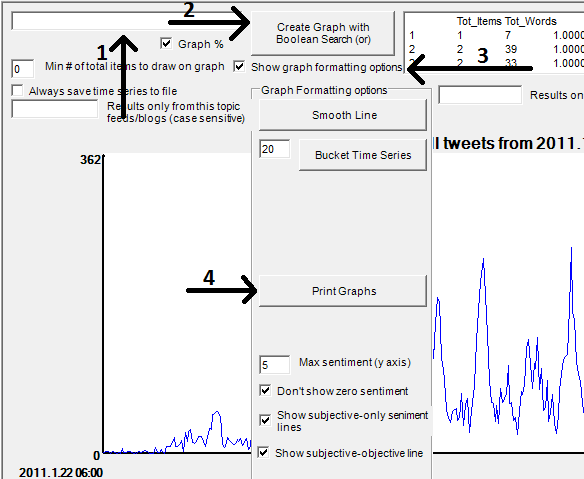
1. **Content analysis**. A content analysis needs a random sample of posts to start with. A random sample of texts can be saved by Mozdeh by clicking the Save tab (0 below), clearing the search text box (1), checking the Save Random Matching Posts to Text File (2) and clicking the Boolean Search button (3). When asked for a location to save the results to, make sure that you remember and can find the location that you chose. Hint: If you can't find it, select View all Reports Created So Far from the Analyse menu.



1. To run a content analysis, load the new text file just created into a spreadsheet. The easiest way to do this is to open a spreadsheet program, then open the text file, select and copy all the text in the text file and paste it into the spreadsheet. Now decide upon some categories for your posts, write names and descriptions for the categories and categorise the random sample according to your descriptions. See the book above or content analysis books for more information on conducting a content analysis.
2. **Relative word frequency analyses for queries within a topic**. This identifies words that are unusually common in posts that match a particular query. These words may point to important aspects of the topic discussed. This can be achieved in four steps; click the List by project chi square option (1); enter the query (2); click the Boolean Search button (3) and click the Calculate Word Frequencies for Search Matches button (4). Words that are unusually common in posts matching the query compared to posts not matching the query (as measured by the chi square statistic) will be listed in descending order in the box above (5).



1. The chi square value indicates how statistically significant the difference is between the frequency of the terms in the topic-specific collection of posts compared to all posts collected, with higher values indicating more significant results. The results can be copied to a spreadsheet or word processer for analysis. To do this, right click in the box of results on the top right of the screen, left click and click Select All from the menu that appears, then left click and click *Copy* from the menu and then paste the results into your spreadsheet or word processor document.
2. For a quick analysis of the relatively high frequency words, write a paragraph summarising the patterns that you see in them. For example, they may point to topics that are frequent for your query or sentiment words that are frequent for it (e.g., good, bad).
3. **Creating and saving a time series of *all* posts in the collection**. A time series of how often the posts in the collection occurred for each hour selected can show patterns of changes in interest over time. To create such a graph, select *Graph Time Series* from the *Analyse* menu, then click on the box in the top left hand corner (1) and clear the text so that it is blank, then click Create Graph with Boolean Search (2). To print the graph, click Show Graph Formatting Options (3) and click Print Graph (4). Before printing, select the printer options in the print dialog box and change the paper layout to Landscape or some of the graph will be missing. To return to the main search screen, select *Search* from the *Analyse* menu.



1. **Creating and saving a time series of posts in the collection *that match a query***. A time series of how often the posts in the collection match a particular query can show patterns of changes in interest related to that query (e.g., riot) over time. To create such a graph, select Graph Time Series from the Analyse menu, then click on the box in the top left hand corner (1 above) and enter your search term in the box, then click *Create Graph with Boolean Search (2)*. To print the graph, click Show Graph Formatting Options (3) and click Print Graph (4). Before printing, select the printer options in the print dialog box and change the paper layout to Landscape or some of the graph will be missing.

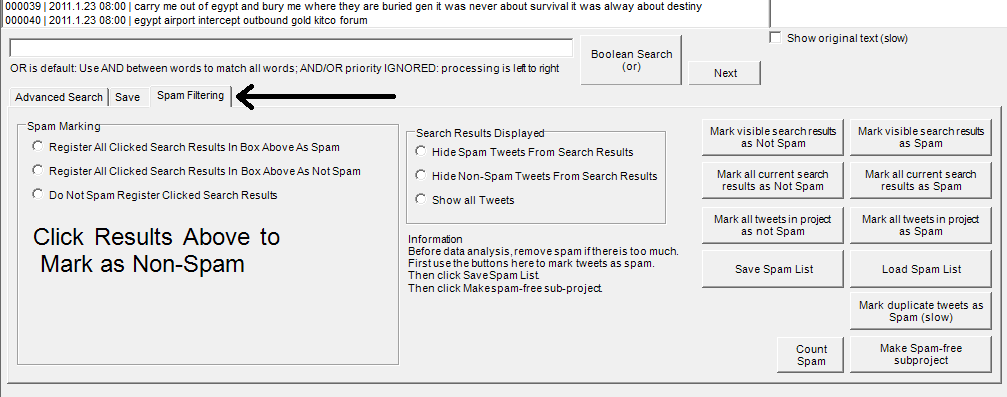
# Marking and Removing Spam

Mozdeh has built in procedures to help you to mark and remove spam content from a set of posts that has already been collected. This is important because otherwise your analyses are likely to reflect spam rather than genuine users.

## Getting started

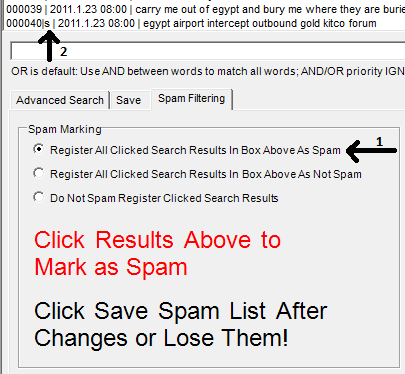
Collect the posts, load them into Mozdeh, allow Mozdeh to index them and then go to the start screen. Run a blank search (i.e., clear the search box) and then click the Spam Filtering tab below the search box.

This gives lots of options for marking and processing spam.



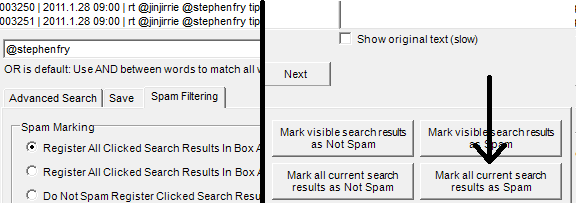
## Marking individual items as spam

To mark individual posts as spam, first click the spam marking option Register All Clicked Search Results in Box Above as Spam (marked by arrow 1 below) and then click any spam result in the list box above it. When you click a spam result, it will be marked with "s" in the results box (as shown by arrow 2 above). You can click as many results as you like and if you navigate to other pages then all your results will be remembered. If you are marking a lot of results as Spam then please click the Save Spam List button occasionally in case of a program crash. Browse your data by repeatedly clicking the next button and when you see a spam item, click on it to mark it. If you accidentally click the wrong item, click the option below arrow 1 above and click your item again to clear it.



## Marking search results as spam

If you identify a pattern, such as that all texts containing "@stephenfry" are spam then you can specify this by clicking the Mark all current search results as Spam after running the search, as shown below.

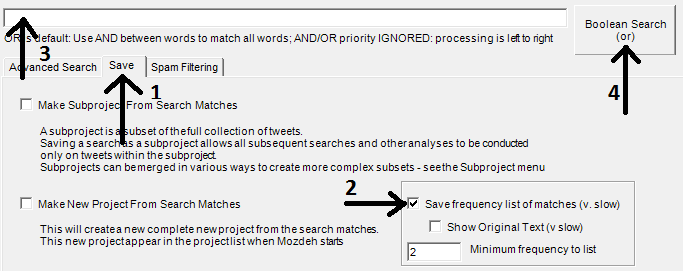


## Marking search results as not spam

It is possible to mark all search results as not spam. For example, you might want to mark all search results containing the word holiday as spam except those containing the word riot. To do this, first mark all search results for holiday as spam, as described above. Then run a query for holiday AND riot (to match all texts containing both words) and click the button Mark all current search results as Not Spam button.

## Marking duplicates as spam

Before marking a list of duplicates as spam, make a list of the duplicates in case they are needed for future analyses. To do this, click the Save tab (1 below), then click the Save frequency list of matches button (2), clear the search text box (3) and run this blank search by clicking the Boolean Search button (4).



If you are using any automated analyses, such as trend detection or the word frequency table, click the Mark duplicate posts as spam button. This will mark all duplicate texts as spam so that there will never be two identical texts in the results.

## Working with the spam-free texts

To work with spam-free texts, it is best to create a new project after excluding all spam. To do this after completing the spam identification, click the Make new Spam free project button, close Mozdeh and then reopen the new project.

# Old Twitter/x instructions

These no longer work but are included for historical reasons.

1. Start Mozdeh, enter a project name, click New Project.
2. Enter your queries in Twitter part of the Data Collection screen (e.g., those below to look for posts about some of the most vulnerable people in any society, those seeking asylum), select en (English) as the language and click the *Search Twitter Once* button for a quick test.

A screenshot of a computer

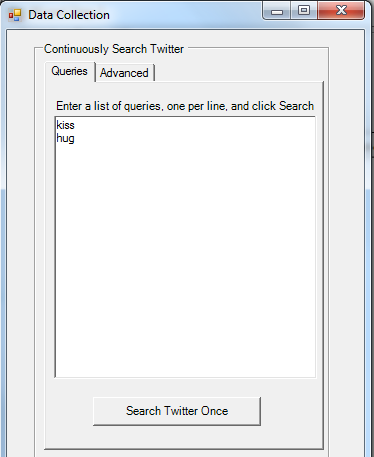
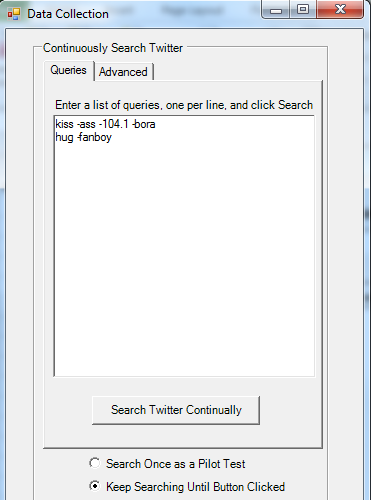
Description automatically generated

1. You will be taken to a web page asking you to logon to Twitter. This will give you a pin number to enter into Mozdeh that gives Mozdeh permission to collect data for you (but not permission to tamper with your account, so this is safe to do).
2. Mozdeh will now collect posts matching your queries. When it has finished, it will ask a series of questions. Click OK or give the suggested answer to these questions **but make sure the *hours* option is checked rather than the *days* option** (the other answers don’t matter at this stage) and then you should get the main search screen.

## Generating a full-scale set of queries

Having completed the pilot test in section 1, you should have an idea how Mozdeh works and what it can do. The rest of this document describes how to use Mozdeh's main features in more detail for use in a full-scale research project.

The first stage of a full scale Twitter analysis project is to generate an effective set of queries to match relevant posts. This is a key step of the project that must be carried out carefully to ensure that all relevant queries are included, that queries matching too many irrelevant posts are rejected, and that queries matching some irrelevant posts are refined to eliminate most irrelevant posts. To refine a query to exclude relevant results add *–term* to the end of the query, where *term* matches many of the irrelevant results. For example, a study to investigate how the words *kiss* and *hug* are used in Twitter to express affection started with these two terms as queries (below left). After pilot testing the queries, some terms were subtracted from them to remove lots of unwanted matches (below right). For example, subtracting *ass* removed many instances of the phrase "kiss my ass", removing *104.1* excluded references to the Kiss 104.1 radio station and subtracting *bora* removed lots of spam related to a TV meme at the time of checking. The final queries used were **kiss –ass –bora -104.1** and **hug –fanboy**.

 -> 

The process to build a useful set of queries is:

1. Brainstorm a set of potentially relevant queries
2. Test them in <https://twitter.com/search-home> for relevant results.
3. Check each query for irrelevant content and remove if almost all content is irrelevant, or refine the query by adding *–term* to exclude irrelevant matches if there is some irrelevant content.
4. Check for additional queries to add to the set.
5. Repeat from a. until the set of queries seems to be stable and satisfactory.

Once this is complete, the queries can be used to gather the data for analysis. The above steps are described below.

### Brainstorm a set of potentially relevant queries

Create a list of queries that you believe should match relevant posts. The queries can be keywords or phrases that describe your topic and, as far as possible, do not also describe irrelevant topics. Enclose phrases in quotes and save the queries in a plain text file. It is particularly important to use a plain text file (e.g., in Windows Notepad) because if you have quotes in a word processor document then it will change the type of quotes from straight to smart, which the Twitter Search will not recognise.

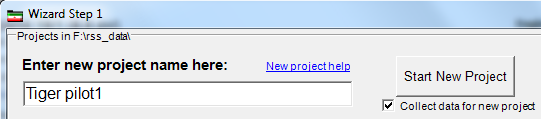
Test each query but submitting it to Twitter Search (<https://twitter.com/search-home>) and checking the results for false matches. Queries can be modified by adding extra terms specific to the topic or by subtracting terms matching irrelevant topics. For example, if the initial search was *tiger* for the big cat then the refinement might be *tiger cat* or even *tiger –woods*, both of which should get a higher percentage of relevant queries than the initial query.

Try the list of [search operators](https://dev.twitter.com/docs/using-search) at <https://dev.twitter.com/docs/using-search> too - and beware that the web search results may not work in the same way as Mozdeh search results. For example, searching for a username returns some of the users posts in the results online but seems not to in Mozdeh - the *from:user* command is needed instead.

### Pilot the queries by gathering matching posts

First, Mozdeh must be installed on the computer that will run the analysis. To do this, download Mozdeh by following the instructions here <http://mozdeh.wlv.ac.uk/installation.html> to save it onto your computer (Windows only, sorry, and use at your own risk). Start Mozdeh and follow the instructions about selecting a folder in which to store your data. Mozdeh will ask to create a folder called rss\_data on your computer. This folder will be initially empty but you will eventually populate this with new subfolders, one for each Twitter project.

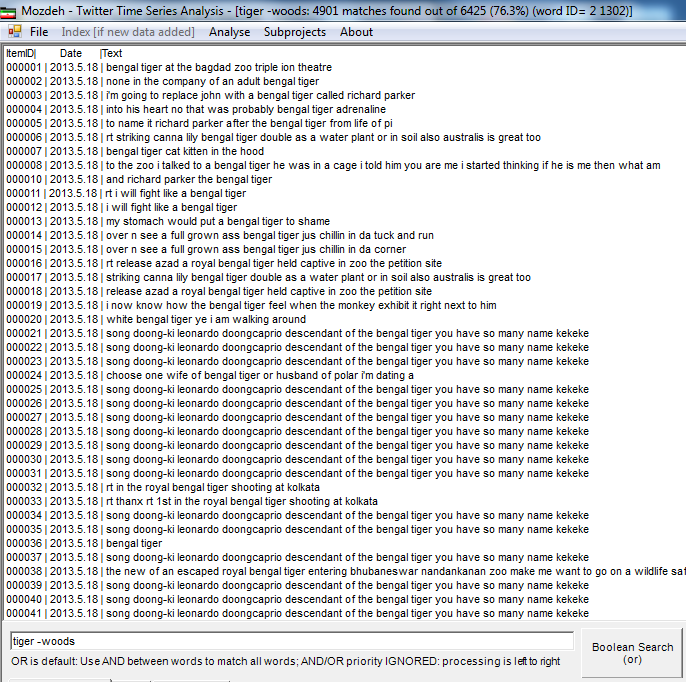
Mozdeh can help with piloting the original queries. Start Mozdeh, enter a name for the pilot test (e.g., Tiger pilot1) and click the Start New Project button.



Then enter the queries in the Data Collection screen and click the Search Twitter Once button. You will be taken to a web page asking you to logon to Twitter. This will give you a pin number to enter into Mozdeh that gives Mozdeh permission to search on your behalf (but not permission to tamper with your account, so this is safe).

Mozdeh will now collect posts matching your queries. When it has finished, it will ask a series of questions – please click OK or give the suggested answer to these questions and then you should get the main search screen.

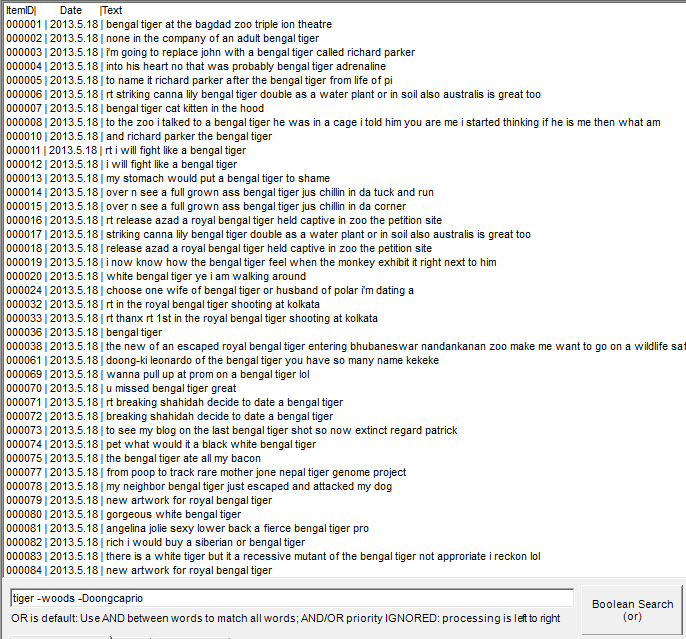
At the main search screen, enter the first query and then click Boolean Search to see posts that match that search. NB. The search button may be at the top of the screen rather than at the bottom as shown below.



The above screen shows the first page of results for the query tiger –woods. Some of the results also match some of the other queries (e.g., "bengal tiger"). Note that the query results are a bit strange – they reflect the words indexed by Mozdeh and the form in which they are indexed, which is not always the same as the original text. For instance, plural words are converted to singular versions. The original versions can be read by checking the *Show original text (slow)* option and clicking a text in the search results list.

Reading the results, some themes are evident. For example Bengal tigers are used to represent courage or fierceness and there is a strange repeated post about song doong-ki. The full original post is, "RT @WeLoveDara: Song Doong-Ki, Leonardo Doongcaprio, descendant of the bengal tiger... You have so many names Dadoongie! Kekeke." Web investigations of this suggest that it is a funny comment on a cat called Dadoong owned by a member of South Korean pop group 2NE1. This is only relevant to cultural associations with tigers rather than living tigers and is quite prevalent so it makes sense to develop a strategy to remove it as spam. A simple strategy would be to add –doongcaprio to each keyword query since this unique word would surgically remove all irrelevant posts.

We can simulate this by entering tiger -woods –Doongcaprio and clicking the Boolean Search button to see the results.



Although a lot of the results in this first seem irrelevant there doesn't seem to be a pattern that would help to identify them. To get help to identify possible refining terms, click the Calculate Word Frequencies for all Search Matches (slow) button. This produces a list of words that occur at least 10 times in all pages of results for the query (i.e., tiger -woods –Doongcaprio) and lists them in descending order of relative frequency. The relative frequency is the percentage of search matches that contain the term out of all posts containing the term. For example, a score of 100% indicates that a word only occurs in search matches and never occurs in any other posts. The start of this list is copied below. Some of the terms are strange and probably refer to a single re-posted post, such as the top one. The general words, such as *coming*, can be ignored but any unusual words with a lot of matches, such as *pleaaaase, bubbly,*and*isono*, should be investigated for possible spam or being otherwise irrelevant (so add them as –term to the query). If a term is relevant then it can be ignored unless it may be useful as a query on its own to add to a revised query list. Two of the words in the table below fall into this category: panthera and tigris – investigating these would reveal that *Panthera tigris* is the Latin name for tiger, so "panthera tigris" makes sense as a new query to add. In contrast, *conservationist* is relevant to tigers but it is not worth having a query for *conservationist* because most matches will not be about tigers. Also in contrast, pleaaaase occurs in only in a common irrelevant repost so the term -pleaaaase could be added to the original query to eliminate it but the number of matches is so low that it might not be worth it if there are lots of other queries. Some of the words in the list below refer to one particular incident, a fatal tiger attack (*broke, maul, fatal*), and if this is judged irrelevant then these terms could also be subtracted from the queries.

|  |  |  |  |
| --- | --- | --- | --- |
| Rel freq | Word | Matches | Total |
| 100.0% | pleaaaaase | 14 | 14 |
| 100.0% | coming | 20 | 20 |
| 100.0% | described | 12 | 12 |
| 100.0% | bubbly | 49 | 49 |
| 100.0% | confident | 25 | 25 |
| 100.0% | won't | 27 | 27 |
| 100.0% | isono | 10 | 10 |
| 100.0% | though | 13 | 13 |
| 100.0% | amur | 18 | 18 |
| 100.0% | stripe | 10 | 10 |
| 100.0% | sweet | 24 | 24 |
| 100.0% | conservationist | 12 | 12 |
| 100.0% | broke | 96 | 96 |
| 100.0% | rule | 81 | 81 |
| 100.0% | panthera | 18 | 18 |
| 100.0% | police | 88 | 88 |
| 100.0% | maul | 13 | 13 |
| 100.0% | reason | 14 | 14 |
| 100.0% | treated | 16 | 16 |
| 100.0% | end | 14 | 14 |
| 100.0% | tigris | 10 | 10 |
| 100.0% | petition | 15 | 15 |
| 100.0% | fatal | 31 | 31 |
| 100.0% | prey | 12 | 12 |
| 100.0% | power | 11 | 11 |
| 100.0% | every | 10 | 10 |
| 100.0% | r | 10 | 10 |
| 100.0% | balm | 17 | 17 |

The above steps should be repeated with each query to get a new and revised set of queries. This revised set should then be pilot tested again in the same way, paying particular attention to any new queries. Once the pilot test results are good enough, with most of the matches being relevant, then the main data collection can start. The main data collection is the same as for the pilot test, except that the option Keep searching until button clicked should be clicked first, and when the data collection process is finished then the stop button should be clicked to start the analysis.

# FAQ

## Can Mozdeh collect all tweets relevant to a topic that is in the future or less than a week old?

No it can't collect tweets, except perhaps if you have paid to access the Twitter/X API (not tested). If you collected tweets with Mozdeh in the past then it should still work with them except that it will not be able to detect user nationalities (this requires API calls).

## What happens to the data if Mozdeh crashes or if there is a power cut?

If Mozdeh crashes or there is a power cut then the data is not lost but will be saved in a file called something like TwitterSearches\_Tweets.txt and stored within a subfolder of the project folder within moz\_data called raw data. So the full path for the file might be something like c:\moz\_data\SNP MP test\raw data\TwitterSearches\_Tweets.txt. This will contain all of the data except perhaps a few tweets from the last few minutes before the crash. If you restart Mozdeh and select the project then it will process this file as a normal project. See below if you want to [merge projects](file:///C:\Users\Mike%20Thelwall\Dropbox\My%20CS3%20Web%20sites\Mozdeh\faq.html#merge) from two or more TwitterSearches\_Tweets.txt files due to Mozdeh crashes or power cuts. You can only merge projects after the data collection has finished.

## Can I merge two projects?

To create a single combined Mozdeh project for more than one set of posts collected by Mozdeh, then download [Webometric Analyst](http://lexiurl.wlv.ac.uk/), start it and close the startup Wizard. From the main search interface, select the Text menu, the Merge files submenu and the option Merge any number of text files (simple consecutive merge, no checking). When asked, reply Yes to the question about ignoring header lines after the first one. These files are inside the raw\_data subfolder of the main project folders in moz\_data. So the full path for the file might be something like c:\moz\_data\SNP MP test\raw data\TwitterSearches\_Tweets.txt. Select all the different Mozdeh TwitterSearches\_Tweets.txt files, one at a time, and choose any name for the merged file. One the merged file is ready, export it back to Mozdeh using the Webometric Analyst button Convert Twitter Files to Mozdeh Format in the Twitter tab on the main interface. After this, start the new project in Mozdeh and it will process it.

## Does Mozdeh continue collecting data if the computer/laptop goes into Sleep or Hibernate mode?

No. When the computer goes into sleep (or hibernate) mode, all programs stop running, including Mozdeh, so it would not be able to collect any data until it is brought out of sleep/hibernate mode. When Mozdeh is woken from sleep it can carry on as normal without a warning that it was asleep.

Mozdeh will work with the screen turned off, so it is safe to configure your computer power management settings to switch the display off after 30 minutes of inactivity as long “never” is set as the time period before going into sleep/hibernation mode (The “Put the Computer to sleep:” Power Option setting).

## Can Mozdeh find YouTube comments with timestamps?

In the search screen, check "Timestamps only" and click Search. This returns only comments containing something that looks like a timetamp, such as 5.43. Timestamps are not shown in the search results list but can be seen for individual results by clicking on them.

## Can I create a subset of my existing texts based on date?

It is tricky to create a subproject with a specified date range but with this one you can create a new project with the old data and select a range.

1. Start Mozdeh, enter a new project name and then click Import Data.
2. Select 1 as the type of data to import and leave all the settings unchanged in the next dialog box except the start and end day.
   * To enter the start day click the first yellow box and select a data in the calendar.
   * To enter the end day click the first yellow box and select a data in the calendar.
3. Copy the UserNames\_Timelines.txt or TwitterSearches\_Tweets.txt file from its original folder (called something like C:\rss\_data\OriginalProjectName\raw data) into a new empty Windows folder and point Mozdeh to this new folder when it asks.

## Can I create networks of texts from a given date range?

For this one, you will need to use the original project, **not** a date filtered new project. After opening the original project, select "Make new raw data file with date restrictions" from the data menu and choose the filtered raw data file and the date range you want. This will save a date-restricted copy of the raw data file in the same place as the original one.  
Now when you select the network creation option select the new filtered raw data file rather than the original one and you should get a network with data only from these dates.

## Does Mozdeh store the Tweet URLs?

The original link is not saved anywhere, but if the tweet is still live then you can find it by adding the tweet Entry ID (second column of raw data file) to the end of the URL [https://twitter.com/statuses/](https://twitter.com/statuses/862336612591689733) and it should redirect the standard URL. For example,  
<https://twitter.com/statuses/862336612591689733> redirects to <https://twitter.com/CamOpenAccess/status/862336612591689733>

## Does Mozdeh store the gender/sentiment/country information with the texts?

The gender information is only in the interface version, sorry. Once it has been examined in the interface, the gender information is saved in a file called genderinfo.txt in the main project folder, which matches up with the ID numbers, in case this helps. The same is true for sentiment in the "Item and Feed IDs" folder. There is a "Add country code" option in the File menu of Mozdeh, which might be useful.

## Can I preview a Mozdeh project while it is collecting data?

Yes - click the Make copy of project button towards the bottom of the data collection screen. You will need to start a second copy of Mozdeh, which can run at the same time as the first one, and open the project copy to see it.

## Can I create a copy of my project with different time slicing (day/hour/month)?

Yes but it is a bit tricky.

* Find the TwitterSearches\_Tweets\_AllFiltered.txt file or TwitterSearches\_Tweets.txt file or YouTube file in the “raw data” folder inside the Moz\_data folder for your project. Copy this one file (only, not both files) into a new folder that does not contain any other files.
* Start a new copy of Mozdeh
* Enter a new project name and click the Import Data button (not the Start New Project button).
* Enter 1 as the option in the next dialog box (Twitter) or 2 (YouTube).
* When asked, select the new folder that you have created with a copy of the raw data file.
* After answering the other questions, this should generate a new project with the selected time grouping.

## Why is ****REPLY @user**** in the Mozdeh tweet but not in the original tweet?

If a tweet is recorded by Twitter as a reply to @user1 but @user1 is not mentioned in the tweet then Mozdeh adds **REPLY @user1** to the start of the tweet at data collection time so that these replies can be included in network analyses.

## Why do I see #NAME in my data?

If you view YouTube video comments in Excel then any line (comment or videoID) starting with a minus sign is interpreted as a "bad formula name" by Excel. To get round this problem, start a new copy of Excel, Right click in the top left hand corner of a worksheet, select format and Text. This converts all cells of the worksheet to expect text and not try to convert anything into a formula. If you copy and paste your Mozdeh data into here then it should no longer produce #NAME anywhere. If you still get #NAME then it is possible that (a) you have previously saved the file with Excel or (b) your computer is configured to process text files through Excel in some way, even though it looks like you are not using Excel. For problem (a) you would have to re-collect the data, but for (b) you might have to try a different computer.

A screenshot of a computer

Description automatically generated -> A screenshot of a computer

Description automatically generated

## Why does Mozdeh get truncated retweets, limited to 140 characters?

In the 20 June 2020 upgrade, Mozdeh should gather full tweets in all cases. It previously truncated **re**tweets.

## Why does my YouTube key not work or why has it stopped working?

YouTube keys suddenly stopping working seems to occur a lot and may be caused by a YouTube glitch. Try logging on to the [Google Developer platform](https://developers.google.com/youtube/v3/getting-started), creating a new project, adding the YouTube Data API v3 to the new project and generating credentials for it (in that order). This new key might work. This almost always works for me.

## Can I find which country users are from?

This was possible for Twitter but not for YouTube or other sources.

Versions of Mozdeh from July 2020 onwards can detect the country of Twitter users (see the menu option: Advanced|Get countries of Twitter users...). It does this by retrieving the location information from Twitter (may require a Twitter logon) and matching it against a list of country names in English, and a list of major city names in English. For example, a user location of Wolverhampton, UK would map to UK and a user location of Beijing would match to China. But a user location of My Home, 中国, or Wombourne (small village) would map to None.

## Can I find nonbinary users?

This was possible for Twitter but not for YouTube or other sources becuase nonbinary identities cannot be guessed from names and pronouns are not systematically recorded in YouTube and Reddit.

Mozdeh could detect nonbinary Twitter users (see the menu option: Advanced|Identify nonbinary, male, female...; you many need to activate Advanced| Get countries of Twitter users... first). It did this by retrieving the user self-description information from Twitter for all users in the current project (may require a Twitter logon) and searching the display name and self-description fields. Any user reporting they/them pronouns in either field and not she/her or he/him is categorised as nonbinary.

## Can I see user self-descriptions?

This is possible for Twitter but not for YouTube or other sources.

Versions of Mozdeh from July 2020 onwards can show user descriptions when clicking on a search result. This is only possible after loading the countries of search results (this also downloads the description information). To load countries, see the menu option: Advanced|Get countries of Twitter users.... It does this by retrieving the user self-description information from Twitter for all users in the current project and reporting the self-description fields. After loading countries, load the self-descriptions (see the menu option: Advanced|Report user descriptions...).

## Can I count word frequencies per user instead of per text (e.g., for Twitter timeline data)?

Yes but you will need a new project, importing your old data into it and selecting the option, Merge all texts from the same user. Here is how it works for Twitter timelines.

Copy the UserNames\_Timelines.txt file into a new folder where it is on its own.  
Start Mozdeh, enter a new project name and click the Import data button.  
Enter number 1 (Twitter) for the data type and click OK.  
Browse for the folder containing one file with all the tweeters' tweets and click OK.  
Select Merge all texts from the same user at the bottom of the massive dialog box and click OK.  
Click OK for all the other dialog boxes to accept the results.

This should give a new project in which each doctor has one "megatweet" consisting of all their tweets merged into one. The file vocabulary\_items.txt in the project folder should then report the number of users tweeting each word at least once.

## Why am I getting a proxy server error?

If you get an error message about proxy permissions when running Mozdeh from a work computer, then please either get your network administrator to allow Mozdeh to access the internet or run it from a non-work computer, such as from home, if you can.

## Why do I always get a file busy error when running Mozdeh?

If you get a file access error and are running Mozdeh from a network drive, the cloud or storage not on your computer, please try running it from a USB stick attached to your computer because network delays can cause it problems.