Social Web Text Analytics with Mozdeh

5/17/2018
University of Wolverhampton
Mike Thelwall

Contents

1 Introduction

This book provides an overview of the data gathering and analysis capabilities of the free software Mozdeh. Detailed instructions can be found on the Mozdeh website http://mozdeh.wlv.ac.uk.

The social web has encouraged people to publish their thoughts on a previously unknown scale. These posts are widely used by business for consumer monitoring and marketing purposes, such as by targeting online adverts and altering campaigns in response to public reactions. Political parties also exploit this information to assess the resonance of campaign messages and to identify wavering voters. For social scientists, public information on the social web provides an easily accessible window into public attitudes, saving time in comparison to interviews and questionnaires and making larger scale studies possible. This book describes methods for researchers to gather and analyse this information with the free Windows software Mozdeh. The name means “good news” in Persian and derives from Mozdeh’s origins for news analysis.

This book focuses on gathering and analysing tweets and YouTube comments. These sites offer access to large numbers of posts but all methods in this book can be conducted on a standard laptop and do not require computer programming. The methods can also be applied to other sites in which users post text, although extra work is needed to import their data into Mozdeh. The book supports research into topics that are reflected in the social web, such as politics, as well as research into social web phenomena, such as the popularity of YouTube celebrities or gender differences in tweets about an issue. Whilst the Mozdeh website (http://mozdeh.wlv.ac.uk) contains step by step instructions, this book gives an overview of the purpose and rationale for the methods embedded in Mozdeh, which include the following types.

- **Gathering data**: Gathering tweets matching keywords or from a set of users is (currently) straightforward with Mozdeh, depending on continued support from Twitter and YouTube. With additional steps, texts can be imported from other sources, such as Facebook pages, TripAdvisor and Scopus.
- **Exploring topic content**: Mozdeh supports simple filtering and searching of a dataset, helping to reveal what people are posting about.
- **Describing topic content**: Mozdeh supports content analysis to characterise the key attributes (e.g., issues) of a data set, by allowing texts to be listed in random order and showing associated information (e.g., usernames, time of posting).
- **Analysing sentiment**: Mozdeh calculates the average positive and negative sentiment strength for all posts matching a query or filters, including 95% confidence intervals. This allows comparisons between different sets of texts, such as from different users, time periods or topics. In conjunction with association mining (see below) it can detect issues that attract positive or negative reactions.
- **Identifying gender differences:** Mozdeh can filter texts by author gender, allowing comparisons between males and females. In conjunction with association mining (see below) it can detect issues and language that are more common for one gender.
- **Analysing time series**: Mozdeh graphs the volume of texts gathered over time or the percentage of texts that match a given query. These time series graphs can be used to visually detect trends in volume and spikes that represent sudden events.
- **Association mining**: Mozdeh uses word association mining to detect issues that associate with topics, sentiment, gender or time periods.
- **Analysing networks**: Mozdeh draws networks from the texts to illustrate the patterns of communication between the most interactive users. These can be used to identify clusters or influential people.

1.1 *Embedding Mozdeh within a research project*

Mozdeh can play different roles within an academic research project. The following list suggests ways in which Mozdeh might fit within a research design.

- **Data collection**: Mozdeh is only used to collect social web texts.

- **Descriptive quantitative analysis**: The analysis methods of this book are systematically applied to describe the key aspects of a topic. This might be characterised as a descriptive quantitative paradigm. The exploration might start with a goal then apply all relevant methods and synthesise and summarise all information discovered that is relevant to that goal. This seems particularly suited to an exploration of a previously unknown phenomenon or for a student project. The key questions here are what, who, where, when, or how but not why (e.g., What is discussed in teenage cancer forums? Who engages in slut-shaming in YouTube? Where in the world are viewers of YouTube educational material? When did a group of YouTube stars first become popular? How do slow food communities connect to each other online?).
- **Quantitative hypothesis testing**: The analysis methods of this book are used to answer research questions or test hypotheses within a quantitative paradigm, perhaps in conjunction with other methods and data. In a well-researched area, existing theory or prior research may suggest a likely outcome for a given study that can be formulated as a binary statistical hypothesis. The key questions are typically binary, such as starting with *is*, *are* or *do*. (e.g., Do females express more positive sentiment about people with psychiatric illnesses in YouTube? Are males less popular on Twitter?).
- **Qualitative investigation support**: The analysis methods of this book are used to give one perspective as part of a holistic qualitative investigation (e.g., case study) of a phenomenon that uses multiple alternative analytical methods and data. Qualitative research may use an all-encompassing exploratory methodological approach, such as grounded theory, which guides the method to obtain data to analyse, as well as the approach to analyse it and report the findings in the form of theories. The key question for this type of research is often why (e.g., Why do goths form tight-knit communities?)

Research projects might combine elements of the above. For example, a paper about anorexia videos on YouTube might use a descriptive approach for background and then test a specific hypothesis. Similarly, a hypothesis testing paper might incorporate follow-up exploratory qualitative research to suggest a possible reason for the findings. There are some basic checks for the suitability of Mozdeh for a project.

- **Amount of data:** The analytical tools in Mozdeh are most effective on large datasets. About 10,000 texts is a minimum for a useful analysis and at least 100,000 is recommended. With fewer texts the word association mining methods are unlikely to yield useful results. Content analysis or qualitative methods are recommended for smaller datasets.
- **Research questions**: Mozdeh's analytics include tools for topic, gender, date, and sentiment so a project's research questions should be related to these.

1.2 Examples of research projects with Mozdeh

Several published academic journal articles have been based upon Mozdeh analyses. These can be found online and brief summaries are provided here of the most recent. Older articles using Mozdeh are less informative because they use methods that are no longer possible or have been improved.

- Thelwall, M. & Mas-Bleda, A. (2018). YouTube science channel video presenters and comments: Female friendly or vestiges of sexism? *Aslib Journal of Information Management*, 70(1), 28-46. The data for this article was a set comments on the videos of 50 science channels on YouTube. The data was gathered by Mozdeh using the YouTube API and the results use Mozdeh's word association and gender detection features to identify the proportion of male and female commenters on each video and the topics associating with males and females.
- Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, 21(3), 303-316. The data for this article was the comments from a set of videos matching searches for

31 popular dance styles, from ballet to jumpstyle. This paper compares comments on the videos for the different dance styles to identify differences in their online meanings. The paper includes a gender analysis and a network of topic similarity for the dance styles analysed. It also introduces and analyses a suite of analytics techniques for this type of data, the Comment Term Frequency Comparison (CTFC) social media analytics method.

1.3 Ethics

All researchers should consider the ethics of their research and ensure that they have all necessary institutional permissions before starting. There is a long and honourable tradition in the social sciences of ensuring that research involving people obtains informed consent before starting and takes additional steps, when necessary, to safeguard the wellbeing of participants. The requirements are different for social web research when the authors of the texts analysed are not directly contacted by the researcher.

Informed consent for individual projects using public data is unnecessary Research that involves human subjects, such as through interviews and questionnaires, needs ethical approval and appropriate measures to obtain informed consent. In contrast, research on public documents or performances does not: permission to research an author’s public works is unnecessary, even if publishing a damaging critical review. Texts in the public social web are clearly documents rather than humans. Because of this, explicit approval from the authors is not needed. From the author’s perspective, if they place something in the public domain then they do not have a right to complain if it is used by others for legal purposes. Authors of public texts do not have the right to privacy for these texts (Wilkinson & Thelwall, 2011). This case needs to be made to research committees to get approval to waive informed consent procedures. The May 25, 2018 changes to EU privacy laws means that researchers must check that their data source includes consent for analysis of their data. For example, this should be specified in general terms in the Twitter and YouTube terms and conditions of use.

Anonymity is essential It is important to fully anonymise users to avoid drawing attention to them. To give an extreme example, if someone had tweeted that they were having suicidal thoughts then it could be distressing if researchers subsequently drew attention to their old posts by reporting them in an academic paper, perhaps even leading to press coverage. Thus, unless the post author is a public figure (such as a UK Member of Parliament), their identity should be treated as private. They should not be named without permission and their online pseudonyms not be revealed without permission. It is also important to avoid giving information that would allow the users to be traced, such as by quoting all or part of a social web post, even if it has subsequently been deleted. Thus, when reporting research, either no quotes should be used or quotes should be extensively paraphrased so that the individual author could not be traced with it. Ideally, even the author of a modified quoted post should be unable to detect that your pseudo-quote is from them. This should be done for all posts, as a matter of principle. Even quoting an innocuous text can have negative side-effects if it draws attention to someone that values their privacy. This connects with a person’s right to be forgotten which allows individuals to request that their public information is not broadcast, such as through Google search results.

1.4 Workflow

The following steps are recommended as a general guide to help give the best results from a social web research project, especially if it is a descriptive quantitative analysis, as described above.

1. Decide on preliminary goals: What will be investigated? What could be discovered?
2. Pilot test the data collection and analysis: Collect data and analyse it on a small scale to discover what type of data can be gathered and what types of results are likely. Association mining (see below) does not work well on a small scale, however. Pilot testing is important to ensure that the main project does not suffer from a fatal flaw, resulting in wasted time.
3. Finalise goals and/or research questions: This should be driven by the results of the pilot test.

4. Construct a research design: Decide how the data will be gathered and which methods will be applied to address the research goals. This should be informed by the pilot study and directly match the goals and/or research questions. A descriptive study might run all possible analyses and then summarise the results.
5. Gather data as planned with Mozdeh or importing it into Mozdeh.
6. Analyse data as planned with Mozdeh and/or with other methods.
7. Conduct follow-up investigations to get deeper insights into the results, if relevant.

Finally, Mozdeh’s analyses work best when there are many relevant texts in the social web and it is possible to gather then without also including many false matches.

2 Data gathering

Mozdeh can gather social web texts directly from websites that share them, such as Twitter and YouTube, or can import texts that have been gathered in other ways. Both Twitter and YouTube currently share their data through an Applications Programming Interface (API) which is an information sharing technology that Mozdeh interacts with to download posts. In the future, these sites might withdraw their APIs or charge for them. In this case, the feature will be withdrawn from Mozdeh. For this reason, it is a good idea to collect data as soon as possible. Other sites might also start to offer useful APIs and these may then be added to Mozdeh.

This section gives an overview of how to gather texts with Mozdeh, focusing on theoretical issues. Step by step instructions are given on the Mozdeh website.

2.1 Tweets

Twitter allows two types of texts to be collected free via Mozdeh: recent tweets matching a query and all (up to a maximum) tweets posted by a user. It does not allow older tweets to be gathered via queries, unless they are collected from a specific named user. It also allows real-time monitoring for as many days, weeks or months as necessary, if long term data is needed. Old tweets that are not from a specific set of known users must be bought instead from a Twitter data reseller, such as Pulsar (pulsarplatform.com).

2.1.1 Query-based retrieval

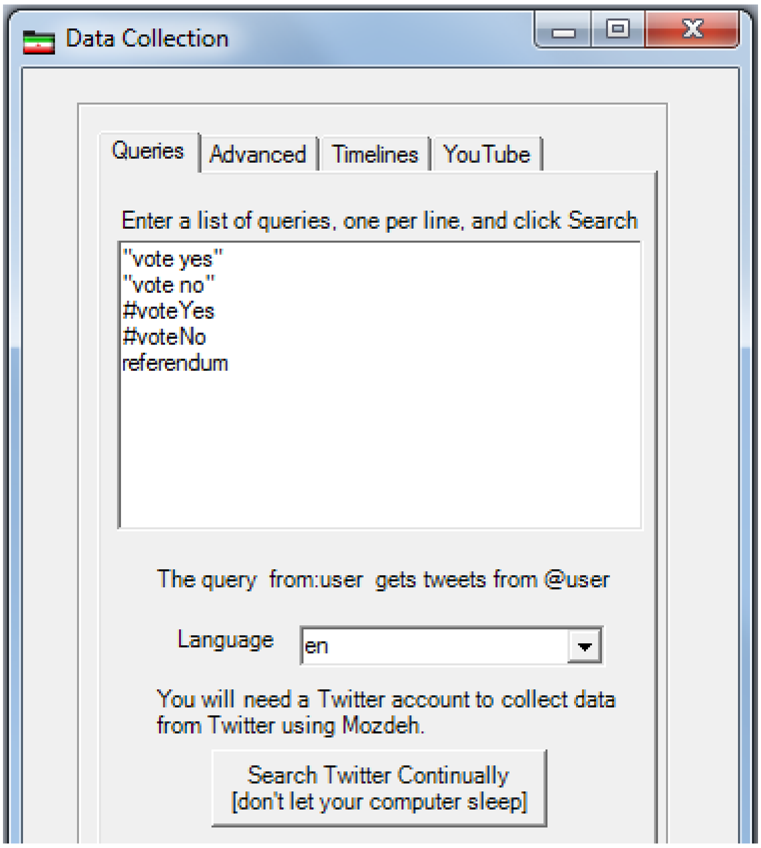
Mozdeh can gather tweets matching one or more queries. These queries can be listed in the data collection screen (Figure 2.1). When Mozdeh is started, it cycles through these queries, submits them to Twitter, via its API, and saves the matches returned. It repeats this process indefinitely to continually check for new content. If it is left on a computer that is constantly connected to the internet and does not go into sleep or hibernation modes, Mozdeh can collect data for years without stopping. There are three main limitations, however.

- Old tweets are not retrieved: The search API does not return tweets that are older than about a week even if they match the keyword search. To gather free tweets for an event, it is therefore important to start Mozdeh before the event and leave it collecting throughout the event.
- Not all tweets are retrieved: The API returns a maximum of 1% of all tweets posted at the time of data collection. This is known as the “rate limit”. Most queries generate far fewer matches than 1% of all tweets, so this is often not a problem (Thelwall, 2015). It is mainly an issue when running a large set of queries with a high cumulative total number of matching tweets or a few very popular queries, such as for a breaking major news story. Twitter do not say whether the 1% limit is a random sample.
- Some tweets are filtered out. The API seems to remove spam-like tweets as well as duplicate tweets, especially retweets. For example, if a tweet has 10,000 retweets then the API may only return a few of these identical retweets. The exact process by which tweets are filtered out is not reported by Twitter.

The second and third issues above are problematic for research projects because they introduce an unknown quantity into the data source and a potential source of bias. This must be accepted and explicitly acknowledged as a limitation of the data gathering process. Although it is undesirable, all social research has imperfect data and so it is normal to have some limitations.

The queries must be carefully constructed to match as many relevant tweets and as few irrelevant tweets as possible. If the queries match many irrelevant tweets, then it will be difficult to analyse the resulting data. If many relevant tweets are not found, then this will weaken the analysis. A balance must therefore be chosen to gather as many relevant tweets as possible. For this, it is important to test the queries in the online Twitter search interface and in a pilot test with Mozdeh to look for problems. The optimal strategy will vary by research project but here are some suggestions.

- Consider focusing on a set of unambiguous hashtags or keywords, ignoring other tweets. This may be necessary if the topic is a general one.
- Consider using phrase searches in quotes to increase the precision of a query if the individual words are too general. For example, for energy saving tweets, the phrase searches “energy saving” and “saving energy” would generate far fewer false matches than the query *energy saving* without quotes.
- If keywords describing the topic generate many false matches, consider making a list of queries that combine the keywords with extra terms that would tend to only match relevant tweets. Alternatively, a list of specific phrases might be used. For example, to gather tweets relevant to European Union energy policy, the queries might be “*EU energy policy*”, “*energy policy in the EU*” but these are likely to be too specific and a more flexible query like *“energy policy” EU* might be more useful if it did not generate too many false matches.



Enter a list of queries, one per line, and click Search

```
"vote yes"
"vote no"
#voteYes
#voteNo
referendum
```

The query from:user gets tweets from @user

Language en

You will need a Twitter account to collect data from Twitter using Mozdeh.

Figure 2.1. An example of the Mozdeh data collection window with five queries.

2.1.2 User-based retrieval

Mozdeh can download the most recent tweets from a set of users, up to a limit of between 1800 and 3200 (the Timelines tab in Figure 2.0). The tweets can be very old and are not limited by the one-week restriction of the keyword search facility. There is an option to collect all tweets from the users or to ignore their retweets. User-based retrieval is suitable for research projects that are interested in a set of users (e.g., UK MPs). The main limitation is that the set of tweets downloaded can exclude the oldest ones for users that have posted more than about 1800.

2.2 YouTube comments

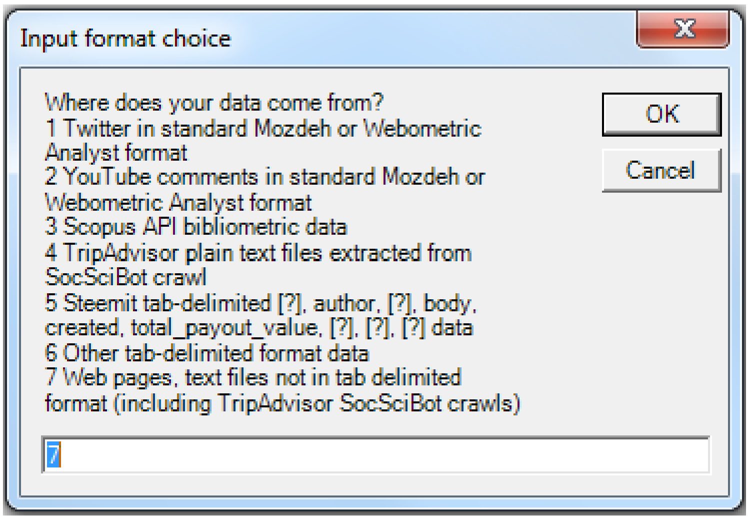
Mozdeh can download the most recent comments on YouTube videos via the YouTube API. There is a limit of about 700 comments per video. If a video has more comments, then the most recent comments will be retrieved. It is not possible to search YouTube directly for comments; the only way to get comments is by specifying videos. These videos can be specified in multiple different ways.

- **By list**: A list of YouTube video URLs or IDs can be entered into Mozdeh to download their comments.
- **By channel**: One or more YouTube channels (i.e., users in most cases) can be identified so that Mozdeh downloads the comments on the videos in the channel. If the channel has too many videos (greater than about 350) then only comments on the most recent videos will be downloaded. If not all videos are returned then it may be necessary to manually identify the remaining videos from the YouTube website and then submit them to Mozdeh as a list, as described above.
- **By query**: One or more queries can be entered to retrieve matching YouTube videos. YouTube retrieves videos by an unknown process that probably includes matching video titles and metadata as well as matching videos that are similar to other matching videos. This is likely to generate at least a few strange matches for a query. To filter out these strange matches, it is possible to insist that Mozdeh discards videos unless their titles and/or descriptions match the query submitted. If the query has too many matching videos (greater than about 350) then not all matching videos will be reported.

Searching for videos by query can be problematic if some irrelevant videos are returned. For example, a query for a song might match a video of a computer game that mentions the song as part of the game soundtrack. If this happens then it may be necessary to manually filter the list of videos (this is saved by Mozdeh: see the Analyse menu | Open raw data folders) and then run the collection again with the filtered list.

2.3 Importing TripAdvisor, Scopus, Reddit, Facebook, Web pages

Mozdeh can read some types of texts that have been collected by other programs. It can import reviews from the TripAdvisor website, as crawled by the free web crawler SocSciBot, as well as Scopus, Steemit, Reddit, Facebook and general webpages. To import this data, save it to a separate folder, start Mozdeh, enter a project name and click the **Import Data** button in the start-up wizard. After this, select the data to import from the Input format choice dialog box (Figure 2.2) and wait for Mozdeh to finish.


Input format choice

Where does your data come from?
1 Twitter in standard Mozdeh or Webometric Analyst format
2 YouTube comments in standard Mozdeh or Webometric Analyst format
3 Scopus API bibliometric data
4 TripAdvisor plain text files extracted from SocSciBot crawl
5 Steemit tab-delimited [?], author, [?], body, created, total_payout_value, [?], [?], [?] data
6 Other tab-delimited format data
7 Web pages, text files not in tab delimited format (including TripAdvisor SocSciBot crawls)

7

Figure 2.2. An example of the Mozdeh data collection window with five queries. More import options are now available.

2.4 Importing texts from other sources

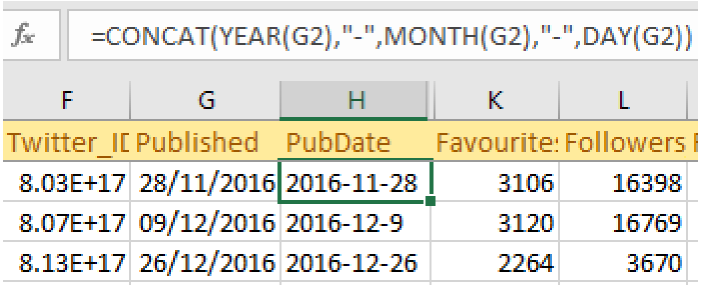
Mozdeh can read texts that have been collected by other programs if they are in a simple format that it recognises. Data must be saved into a simple format that Mozdeh can understand to be loaded via the **Import Data** button in the start-up wizard.

The date must be in tab-delimited plain text format, with the information arranged in columns. Data in other formats can be converted by loading into a spreadsheet, such as Excel, and then using the Save As menu item and choosing the text tab-delimited format. There should be one column for the text and one for the date. Optional extra columns include author name, and topic label. he date must be in a format recognised my Mozdeh. It recognises several formats but the simplest is year-month-day. For example, 2017-05-30 indicates 30 May 2017.

If using Excel and the date is in a different format in a column, then you may be able to translate it into Mozdeh format using the formula

```
=CONCAT(YEAR(G2),"-",MONTH(G2),"-",DAY(G2))
```

as in the Figure 2.3 example.



```
=CONCAT(YEAR(G2),"-",MONTH(G2),"-",DAY(G2))
```

F	G	H	K	L
Twitter_I	Published	PubDate	Favourite	Followers
8.03E+17	28/11/2016	2016-11-28	3106	16398
8.07E+17	09/12/2016	2016-12-9	3120	16769
8.13E+17	26/12/2016	2016-12-26	2264	3670

Figure 2.3. A Microsoft Excel formula for translating a date into the Mozdeh format.

2.5 Duplicate elimination

An important generic issue for data collection is whether to exclude duplicate texts. Many social web sites contain large numbers of repeated posts. These include retweets, modified tweets, copied posts and spam. Unless an analysis focuses on the details of interpersonal communications between users and needs a complete record of all posts, it is better to avoid having duplicate posts within a dataset. Mozdeh supports this with its filtering options. Posts can be filtered out if they are identical to a previous post (or if they are identical and have the same author). For Twitter, Mozdeh gives the option to remove any links and usernames before checking for duplicates. This helps to eliminate simple spam campaigns that have mostly the same text but just change URLs or the usernames.

2.6 Spam

If data gathering results in a large amount of spam and this is unavoidable then Mozdeh has features to mark and remove spam. These are described on the website (http://mozdeh.wlv.ac.uk/SpamRemoval.html). They involve marking individual posts as spam either manually or by running queries in Mozdeh to identify spam texts and marking all the results as spam.

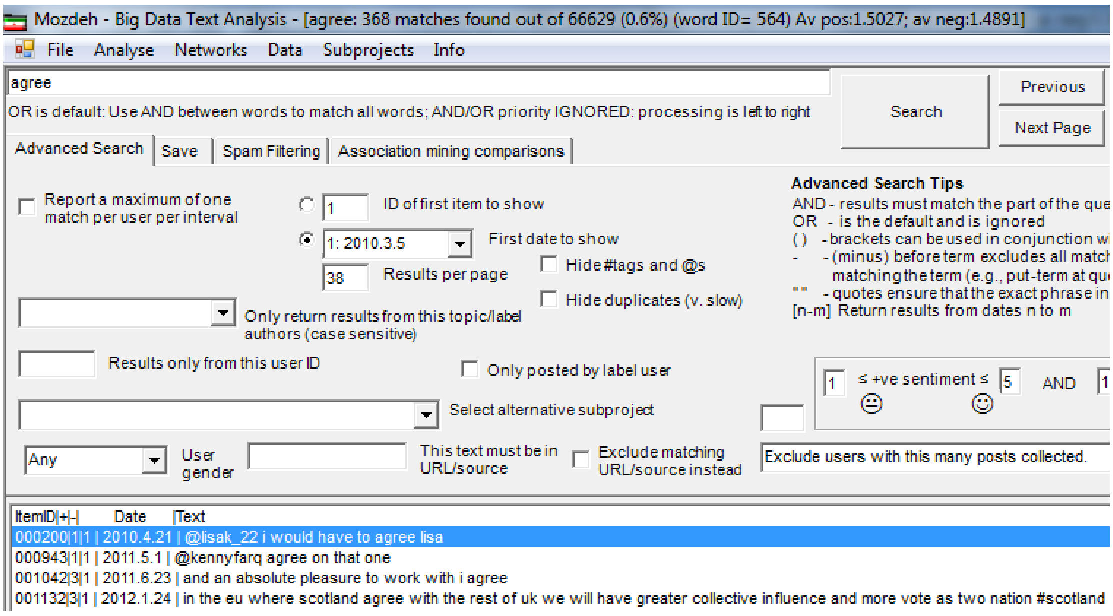
3 Searching and filtering

When Mozdeh gathers or imports texts, they are grouped together into a “project” for analysis. After the data gathering stage is complete, the project analysis can begin. The downloaded texts can be explored in Mozdeh through a combination of keyword searches, filters and sorting options. These facilities are useful to explore a topic and as part of association mining and other methods. Help is available for some of these options in Mozdeh on its website or by hovering the mouse over a button.

3.1 Keyword searches and search results

The search text box is at the top left hand corner of the main Mozdeh window after a project has been loaded (Figure 3.1). The simplest way to search for texts in Mozdeh is to enter a single keyword in this box and click the search button. The first 38 matching posts will then be listed. To see the next page of 38 matches, click the Next Page button near the Search button.

Matches are identified by Mozdeh comparing the keyword entered to all the texts and listing the texts containing the word. Before the matching process starts, Mozdeh attempts to remove terminal “s” characters to convert plural words to singular so that a search will match both versions. For example, the search agree will match tweets containing agree or agrees (even though these are not nouns in this case). Mozdeh also converts all words to lower case before matching so that the search is case-insensitive.



Mozdeh - Big Data Text Analysis - [agree: 368 matches found out of 66629 (0.6%) (word ID= 564) Av pos:1.5027; av neg:1.4891]

agree

OR is default: Use AND between words to match all words; AND/OR priority IGNORED: processing is left to right

□ Report a maximum of one match per user per interval

○ 1 ID of first item to show

● 1: 2010.3.5 First date to show

38 Results per page

□ Hide #tags and @s

□ Hide duplicates (v. slow)

Only return results from this topic/label authors (case sensitive)

Results only from this user ID

□ Only posted by label user

Select alternative subproject

Any User gender

This text must be in URL/source □ Exclude matching URL/source instead

Exclude users with this many posts collected.

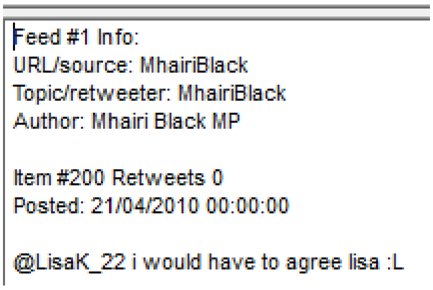
1 ≤ +ve sentiment ≤ 5 AND 1

Advanced Search Tips
AND - results must match the part of the que
OR - is the default and is ignored
( ) - brackets can be used in conjunction w
- - (minus) before term excludes all match
matching the term (e.g., put-term at qu
"" - quotes ensure that the exact phrase in
[n-m] Return results from dates n to m

```
ItemID|+|-|      Date      |Text
000200|1|1 | 2010.4.21 | @lisak_22 i would have to agree lisa
000943|1|1 | 2011.5.1 | @kennyfarq agree on that one
001042|3|1 | 2011.6.23 | and an absolute pleasure to work with i agree
001132|3|1 | 2012.1.24 | in the eu where scotland agree with the rest of uk we will have greater collective influence and more vote as two nation #scotland
```

Figure 3.1. Search results (bottom of screen) for the query *agree* for a set of Scottish MPs’ tweets.

Posts matching a query are displayed towards the bottom of the screen in the same lower case depluralised form that is used for searching (Figure 1). The list also gives the date of the post and its positive and negative sentiment score. To see the original tweet and extra information, such as its author name and retweet count, click the search result (Figure 3.2).



Feed #1 Info:
URL/source: MhairiBlack
Topic/retweeter: MhairiBlack
Author: Mhairi Black MP

Item #200 Retweets 0
Posted: 21/04/2010 00:00:00

@LisaK_22 i would have to agree lisa :L

Figure 3.2. Information about a Scottish MP’s tweet displayed by Mozdeh when a search result is clicked.

3.2 Advanced searches

Searches can be more complicated than individual keywords. If a list of keywords is entered, then this is treated as a Boolean OR search and texts are returned that contain any of the keywords. For example, the query *agree disagree* would match texts containing the term agree, the term disagree or both terms.

AND: To search for texts containing all terms in a list, enter the word AND between them in capital letters (including spaces on either side of this term). The query *agree AND disagree* would therefore only match texts that contained both of *agree*, *disagree*.

Phrase search: To search for texts containing a specific phrase, put standard (not smart) quotes around the phrase. For example, *"happy birthday"* matches texts that contain both *happy* and *birthday*, with the first immediately preceding the second.

Exclude term or phrase: To exclude texts from the results that mention a given term, add the term to the end of a query, preceded by a minus sign – with no gap between the minus sign and the term to be excluded. For example, the query *jaguar -car* would match texts containing the term *jaguar* but not *car*. This can also be used for phrases. So, the query *happy AND birthday -"happy birthday"* would match texts that contain both *happy* and *birthday* but not the phrase *happy birthday*. The matches might therefore include *happy first birthday*.

Date specific searches: To ensure that all matches contain results from a specified date range, enter the start and end number of the dates in square brackets at the end of the query, separated by spaces. For example, to get matches for the query “cat feet towel” from the tenth to fifteenth date only the query would be *"cat feet towel" [10 15]*. The numbers of the dates in the project can be found in the *First date to show* dropdown box. For the project in Figure 3.3, [10 15] means July to December 2016.

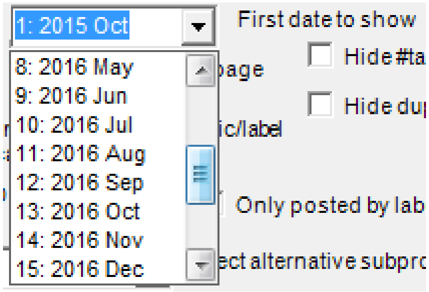


Figure 3.3. Numbered dates, as displayed in the *First date to show* box. These are needed for date-specific queries.

3.3 Complex searches

Brackets can be used to build more complex searches. Searches are normally processed from left to right. So, the query *happy birthday AND mum* would first find texts that contained either happy or birthday and then restrict this set to those that also contained *mum*. The matching texts would therefore all contain mum and would all contain either happy or birthday. To find texts that contained the word *happy* or both *birthday* and *mum*, add brackets round the second half of the query to get *happy (birthday AND mum)*. The brackets ensure that the search segment *birthday AND mum* is processed first. Thus, Mozdeh would first find texts containing both *birthday* and *mum* and would then add in texts containing *happy*, generating a lot more matches.

3.4 Filters

Mozdeh has filters for sentiment, author gender, retweet count, label, user ID, and author posting frequency. These filters work in conjunction with the keyword search facility. If a filter is set and the Search button is clicked, then all the results displayed and processed will match the filter. These filters are cumulative. If multiple filters are selected, they will all be applied.

Sentiment: As described in the sentiment analysis chapter, the sentiment filters can be used to set the minimum or maximum positive or negative sentiment of the texts returned.

Author gender: As described in the author gender chapter, the author gender filter ensures that the posts are authored by the selected gender (male, female, male or female, none). Since gender is deduced from author names, when possible, the category “none” means that Mozdeh has not been able to guess the user’s gender rather that they are genderless or third gender.

Retweet count: This can be used to set the minimum or maximum retweet count of search matches. The right-hand box is usually pre-filled with the highest value in the project.

Label: The label dropdown box lists information about the information used to gather the posts, such as the Twitter search queries or the YouTube channel IDs. To filter using this information, select a value from the dropdown box. To select multiple values at the same time, list them in the box, separating them with a pipe symbol |. Wildcards (*) are also allowed.

User ID: This can be used to ensure that all posts returned are from a single user. To find the numerical ID of that user click on any of their tweets and look for the number at the top right hand corner of the information box (Figure 3.2).

Author posting frequency: This can be used to set the minimum or maximum number of posts from an author for them to be included in the search matches. If the project allows a maximum of one post per user, then this filter is useless.

3.5 Search results order

The search results (i.e., the list at the bottom of Figure 3.1) is displayed in ascending date order (oldest first) but this can be changed to help explore the data. This change can be made using the *Sort by* drop-down box at the top middle of the screen (Figure 3.4). The *Search* button must be clicked after changing the search order to see the results. The results can be sorted ascending or descending by date, retweet count, or positive/negative sentiment.

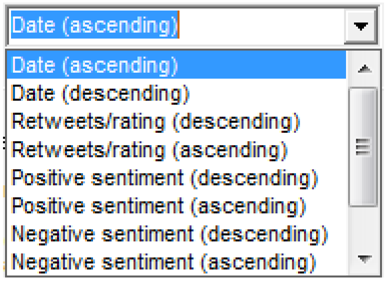

Date (ascending)
Date (ascending)
Date (descending)
Retweets/rating (descending)
Retweets/rating (ascending)
Positive sentiment (descending)
Positive sentiment (ascending)
Negative sentiment (descending)
Negative sentiment (ascending)

Figure 3.4. The *Sort by* drop-down box showing some of the sort options.

There is a *random* sort order that changes each time the *Search* button is clicked but not when the *Next Page* and *Previous* navigation buttons are clicked. Random sorting aids exploration of the data and provides a random sample for content analyses. Sorting the texts randomly reduces the risk that browsing them will lead to incorrect conclusions because, for example, an unusual topic was discussed on the date when the first tweets were collected.

4 Sentiment

Sentiment analysis is the use of a computer program to detect and classify subjective content. Mozdeh incorporates the program SentiStrength to estimate the strength of positive and negative sentiment in texts. SentiStrength works best for English but can also classify sentiment in a few other languages. Sentiment analysis is useful to identify issues that attract positive or negative comments as well as to compare the attitudes of different groups, such as males and females.

4.1 SentiStrength

SentiStrength detects sentiment using a lexicon of sentiment terms and a set of extra rules. It allocates each text two scores between 1 (no positive sentiment) and 5 (very strong positive sentiment) and the same for negative sentiment, from -1 (no negative sentiment) and -5 (very strong negative sentiment). Thus, a score of -1,1 indicates that there is no positive sentiment and no negative sentiment and -3,5 indicates very strong positive sentiment and moderate negative sentiment.

SentiStrength's lexicon is a manually curated list of over 3,000 terms and term stems, each annotated with a positive or negative sentiment strength. For example, love has a value of +3 so the sentence, "I love you" would score -1, 3. The extra rules include negation (e.g., *not happy* becomes negative because of *not*), boosting (e.g., *very happy* is more positive than *happy* because of the booster word *very*). Sentiment can also be detected through emoticons :) emphasis spelling (e.g., *haaaaapy* is more positive than *happy*) and excess punctuation!!! There is a list of sentiment-related phrases (e.g., "shock horror") to detect idioms. Full details and an evaluation of the accuracy of the methods are published elsewhere (Thelwall, Buckley, & Paltoglou, 2012).

4.2 Mozdeh sentiment outputs

Mozdeh classifies all the texts in a project for positive and negative sentiment, reporting the scores alongside the texts when they are listed on screen. This information can be used to sort or filter the texts, to mine sentiment associations and to calculate the average sentiment of a set of texts. It is also used to produce a bar chart and a bubble chart for each set of search results (Figure 4.1). The bubble chart cross references the positive and negative sentiment scores to show how they associate with each other.

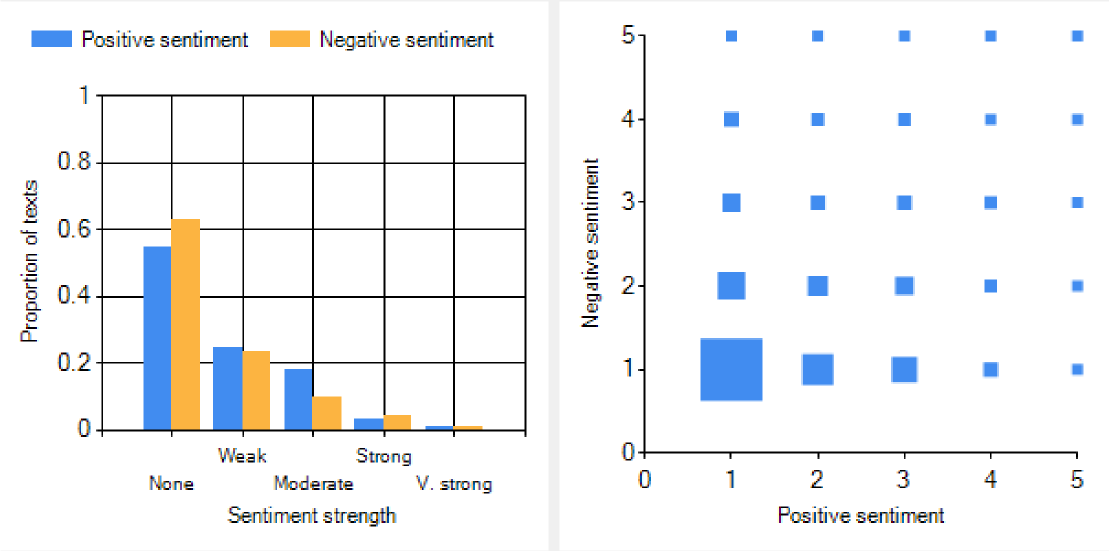


Figure 4.1. The bar and bubble charts for sentiment produced by Mozdeh for a search.

Sorting: To sort texts by sentiment, select one of the four options for positive or negative sentiment ascending or descending in the *Sort by* drop-down box and then click the Search button. This is useful to browse the most or least positive or negative texts gathered.

Filtering: To filter texts by sentiment, change the minimum and/or maximum positive and/or negative sentiment scores in the sentiment filter box and click *Search*. This ensures that only texts with the specified range of sentiments are shown. For example, changing the minimum positive sentiment from 1 (the default) to 3 (moderate positive sentiment) ensures that all texts displayed have at least moderate positivity. This section also has a few shortcut buttons: + sets 3 (moderate) as the minimum positive sentiment and -2(mild) as the maximum negative sentiment so that all matching texts are positive overall. There is a similar – button for negative sentiment.

Averages: Whenever a search is run in Mozdeh (i.e., by clicking *Search*), it calculates the positive and negative sentiment strength of each matching text and reports the overall average alongside a 95% confidence interval, as in the example below.

Pos 1.6955 (1.6862, 1.7048)
Neg 1.4089 (1.4003, 1.4175)

The averages are useful for comparing between different sets of texts. For example, the above results are for male-authored texts in a project. For female-authored texts, the results differ.

Pos 1.8960 (1.8804, 1.9116)
Neg 1.3816 (1.3777, 1.4056)

For positive sentiment, the female average above is higher and the difference is statistically significant since the confidence intervals (1.6862, 1.7048) and (1.8804, 1.9116) do not overlap. We can therefore be reasonably sure that females tend to tweet more positive sentiment in this project. For negative sentiment, the male average is higher but the confidence intervals overlap slightly so the statistical evidence of an underlying gender difference is much weaker. Strictly speaking, if the confidence interval overlap is small, as in this case, then the difference may still be statistically significant but for simplicity it is easier to ignore this.

Association mining: If a sentiment filter is set up and the *Mine associations* button is clicked then Mozdeh will report a list of words that associate with the specified sentiment range. Some of these words will be explicit sentiment words (e.g., happy, sad) that are of interest primarily for the type of sentiment expressed. The non-sentiment words may be more interesting because they suggest the topics that attract the sentiment specified. See the Association mining chapter for more information.

5 Gender

Mozdeh can guess the gender of a post author a from their username. It does this by splitting the username into two parts, interpreting the first part as their first (given) name and then checking if this name is used almost exclusively by one gender. If it is, then the author is assumed to be from that gender. If not, then the author gender is left unassigned.

The list of gendered first names is taken from the 1992 census of the USA and the 1022 most common male-dominated and 3938 most common female-dominated names from that list are used. A first name is included if it is used by one gender at least 90% of the time. The USA is a cosmopolitan country and so this list should work well for all English-speaking countries and should also be reasonable for most Western nations. The list also includes a few gender-specific titles (e.g., Mr, Ms) for people that incorporate them within their usernames.

On Twitter and YouTube, this method seems to give few false matches and to give a gender to about 30% of users. The remaining people tend to have names that are not derived from their legal names.

If a dataset is mostly authored from other parts of the world, then another gender list may need to be substituted. To do this, replace the files *Female Names.txt* and *Male Names.txt* in the Moz_data folder with appropriate plain text lower case lists. If this is done after collecting the data, delete the file *genderinfo.txt* in the project folder and restart Mozdeh.

There is no gender output from Mozdeh but the gender filter allows separate searches for males and females. There is also support for gender based word association mining.

5.1 Sentiment and gender

A simple way to compare males and females is to assess whether they differ in the average strength of positive and negative sentiment in all the texts collected or for a set of filters and/or a keyword search. This can be done by setting the search terms and/or filters and then selecting a gender and clicking the search button. The average positive and negative sentiment strengths are automatically displayed in the box towards the bottom right of the screen, together with confidence intervals. If a second search is run for the other gender then the confidence intervals can be used to assess whether there is a statistically significant difference between the genders in average sentiment, as illustrated in the sentiment chapter.

5.2 Sentiment and association mining

As discussed in the association mining chapter, if a gender filter is set (male or female) then the *Mine associations* button can be clicked to discover words that associate with that gender. This can reveal gender-specific topics and communication styles. It works by finding words that are more common in texts matching the search and/or filter than in the remaining texts. For example, if the filter *Male* is set then it will find words that are more common in male users' texts than in the remaining texts. The remaining texts are likely to include many male-authored texts amongst the approximately 70% for which the author gender could not be guessed. The test would be more powerful if the comparison was made directly between the male authored texts and the female authored texts, ignoring the unknown 70%. Three strategies can be used to achieve this. The third option is likely to be the fastest if the Mine associations feature is to be used extensively for a project.

- Use the *Association mining comparisons* tab instead of the *Mine associations* button and check the *Compare male vs. female for each query* option.
- Select the gender filter *Male or Female*, clear all other filters and the query, and in the Save tab, select *Make subproject from search matches*. After clicking *Search*, a new subproject will be created that excludes all users without a gender. The *Mine associations* button will be more powerful when this subproject without ungendered users is selected.
- Select the gender filter *Male or Female*, clear all other filters and the query, and in the Save tab, select *Make new project from search matches*. After clicking *Search*, a new project will be created that excludes all users without a gender. Close Mozdeh and start the new project. The *Mine association button* will be more powerful when this new project without ungendered users is loaded.

6 Time

Many social web texts include an exact date and time of posting. If a collection of texts spans a long period, then it can be used to detect changes over time. Mozdeh supports three types of time analysis: trend detection, spike detection and difference detection. The first two are based on visual inspections of time series graphs.

6.1 Visual trend detection

Mozdeh has two features for producing time series graphs and this section discusses the first. Whenever a search is run a time series volume graph is drawn on the right of the screen to show the volume of texts matching the search over time. To see a time series graph for an entire project, run a blank search (Figure 6.1).

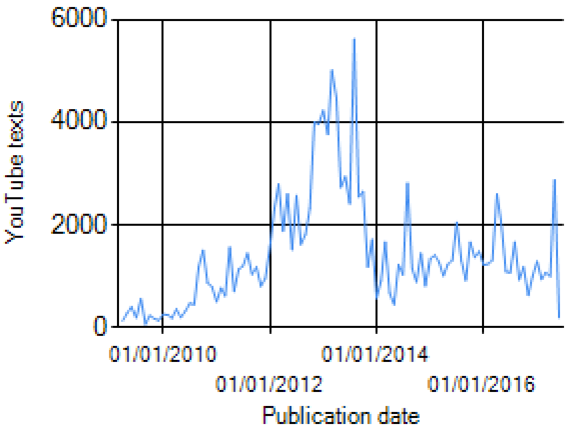


Figure 6.1. An example of the time series graph produced by Mozdeh whenever a query is run.

Trend detection means visually inspecting a graph to identify overall trends, such as increasing volume, decreasing volume. Figure 6.1 shows a gradual increase followed by a period of high activity and then a steady lower level of activity. The overall trend may be useful background information about interest in the topic itself or, for YouTube, the period when the videos analysed were most watched.

6.2 Proportion trends and spikes

Larger graphs can be produced using the special graph window (select *Graph time series* from Mozdeh’s *Analyse* menu). This draws a much larger graph and can illustrate the percentage (rather than number) of tweets matching a query. The y axis in Figure 6.2 illustrates this with the percentage of all texts in this data set of comments on YouTube science videos from the Sixty Symbols channel containing the term *interesting*. There is a slight overall decreasing trend in this graph, suggesting a small decrease in the proportion of commenters that apparently found the videos to be interesting. The line in the graph is not very smooth but this is probably due to the relatively small amount of data rather than any underlying pattern of jaggedness.

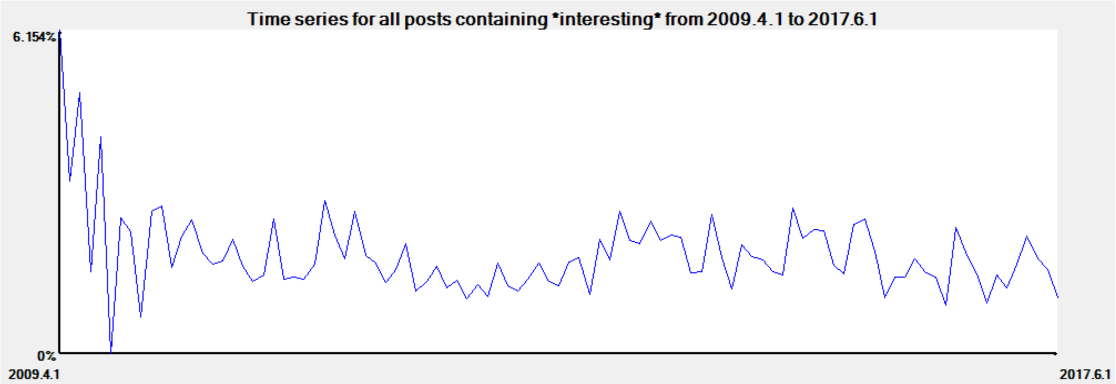


Figure 6.2. A large time series graph produced by Mozdeh to show the proportion of texts containing the word *interesting* over time.

Spike detection means visually inspecting a graph for sudden increases in volume that indicate an event occurring. Figure 6.3 shows a few small spikes and one very large spike, from August 2014, corresponding to 322 YouTube comments containing the term atom – about 11% of all posts from this channel in this month. Clicking on the spike in Mozdeh gives a list of texts from the date of the spike. These can be explored to find out what caused the spike. In this case the cause was the 5 August 2014 release of the Sixty Symbols video, "Do atoms ever touch". It clearly attracted a very high number of viewers as soon as it was launched – presumably these were from Sixty Symbols followers who would watch the video as soon as it was released, rather than casual visitors or school classes.

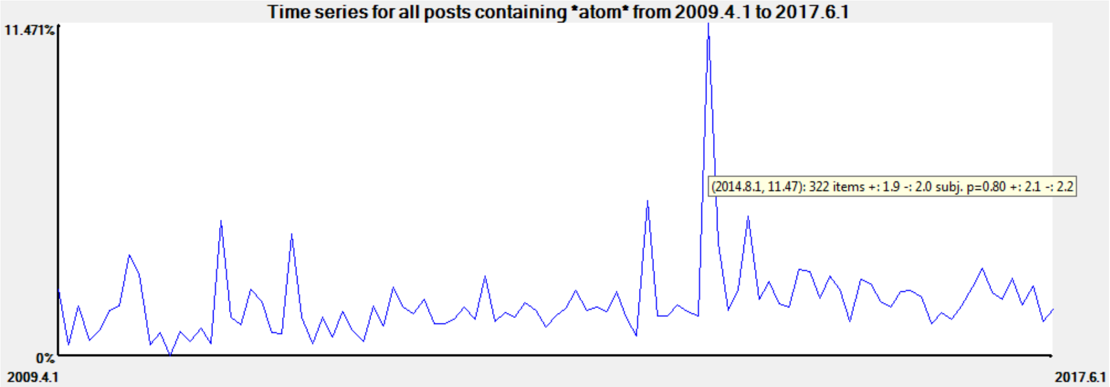


Figure 6.3. A large time series graph for the proportion of YouTube comments from the Sixty Symbols channel containing the term atom. Hovering the cursor on the spike has produced a bubble of extra information about it. This project is organised by month so each date corresponds to an entire month.

6.3 Automated spike detection

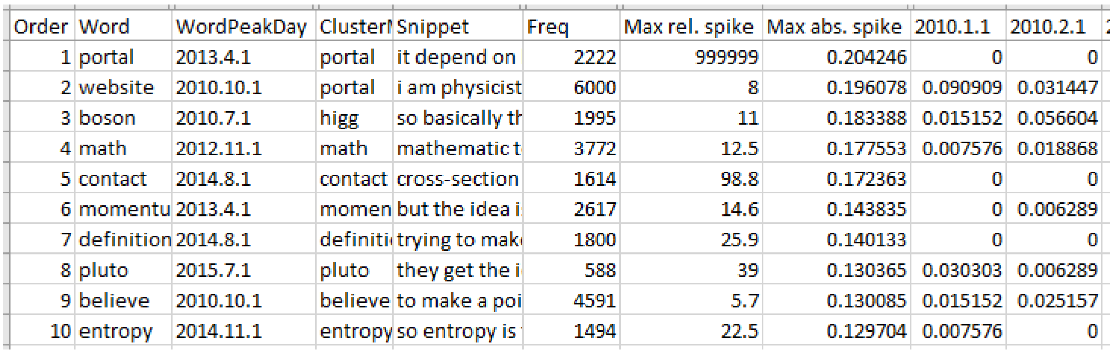
Mozdeh has an advanced feature for automatically detecting spikes. This can be useful if a project has hundreds of thousands of texts and manual explorations are not enough. To automatically detect spikes, select *Time Series Scanning* from the *Analyse* menu and then click *Make time series file for all words matching conditions*. This will attempt to find the 1000 words that generate the biggest spikes, when drawn on a graph. This method has previously been used to detect news stories in a set of general tweets (Thelwall, Buckley, & Paltoglou, 2011).

Spikes are measured in two ways; for absolute and relative height. The absolute height of a spike is the proportion of texts containing the word on the day of the spike subtract the average proportion of texts containing the word for all previous days. For example, if *wedding* occurs in 20% of texts from March 3 and an average of 1% of all texts from previous days then the absolute spike size would be 0.20-0.01=0.19. The relative spike size divides the absolute spike size by the average of all texts from previous days. In this case the relative spike size would be 0.19/0.01=19. The importance of a spike is statistically better reflected by a chi-square statistic but absolute spike size seems to work best in practice, so this is used.

Mozdeh produces the following files from time series scanning.

- ***proportion of Feeds containing word.info.log** Summarises the options used to detect spikes. This is a reminder of the options if different ones are tried.
- ***proportion of Feeds containing word.txt** Lists the maximum absolute and relative spike sizes for all words in the project matching the criteria. Includes the proportion of posts containing the word at each time or date. Does not include the chi-square statistic. Avoid this file unless comprehensive information is needed.
- ***proportion of Feeds containing word.sorted.log** Ignore this file.
- ***proportion of Feeds containing word.clusters.log** Ignore this file.
- ***proportion of Feeds containing word.clusters+time.txt** This is the main results file and is described below.

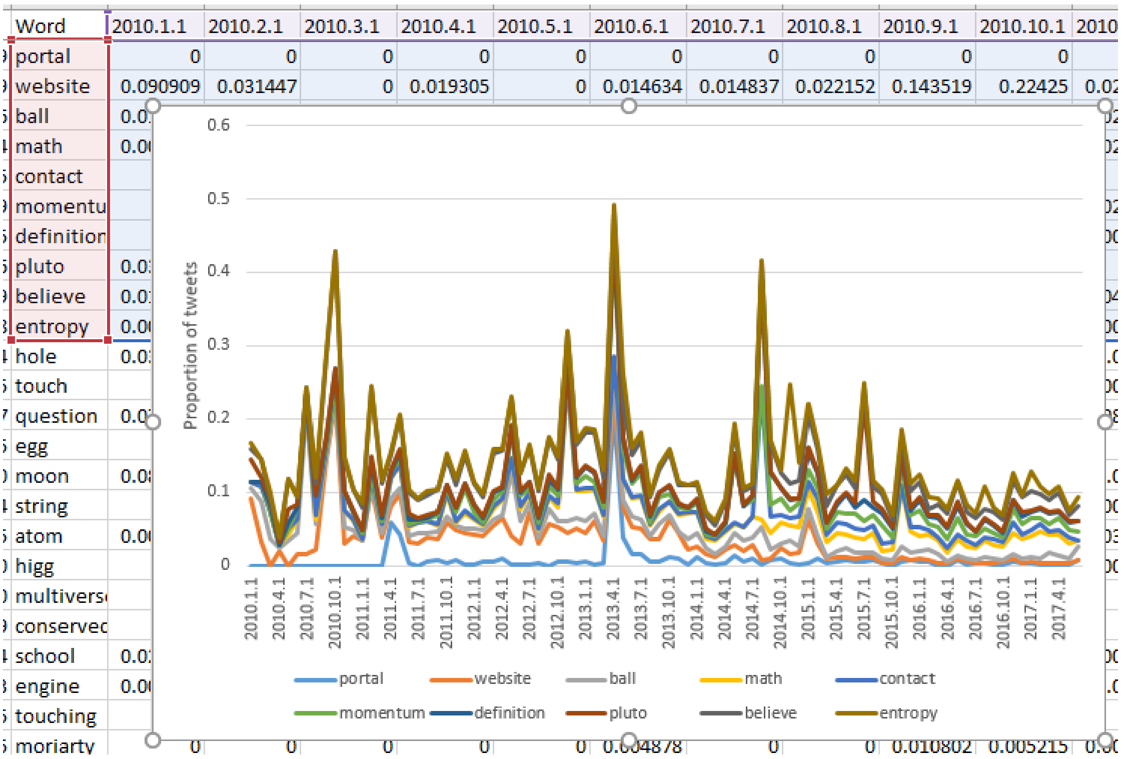
The **clusters+time** file is in text (tab delimited) format that can be loaded into a spreadsheet like Excel to view (Figure 6.4). It lists the spike terms in descending order of absolute spike size. The day/time of the spike is also reported. In the example, the word causing the biggest spike is *portal*, which has an absolute spike size of about 0.20, during April 2013. This file also includes a snippet – a text containing the word – to give some quick context. The cluster column estimates whether the term is usually used in conjunction with another term. This helps to identify phrases. In the example, the words portal and website are in the same cluster (portal) so tend to occur in the same texts. The clustering often does not work very well.



Order	Word	WordPeakDay	Cluster	Snippet	Freq	Max rel. spike	Max abs. spike	2010.1.1	2010.2.1
1	portal	2013.4.1	portal	it depend on	2222	999999	0.204246	0	0
2	website	2010.10.1	portal	i am physicist	6000	8	0.196078	0.090909	0.031447
3	boson	2010.7.1	higg	so basically th	1995	11	0.183388	0.015152	0.056604
4	math	2012.11.1	math	mathematic t	3772	12.5	0.177553	0.007576	0.018868
5	contact	2014.8.1	contact	cross-section	1614	98.8	0.172363	0	0
6	momentu	2013.4.1	momen	but the idea i	2617	14.6	0.143835	0	0.006289
7	definition	2014.8.1	definiti	trying to mak	1800	25.9	0.140133	0	0
8	pluto	2015.7.1	pluto	they get the i	588	39	0.130365	0.030303	0.006289
9	believe	2010.10.1	believe	to make a poi	4591	5.7	0.130085	0.015152	0.025157
10	entropy	2014.11.1	entropy	so entropy is	1494	22.5	0.129704	0.007576	0

Figure 6.4. A section of a **word.clusters+time.txt** file copied into a spreadsheet. Some columns have been hidden.

To graph a spike, either enter the keyword as a query in Mozdeh to get a small time series graph, enter the keyword in the Graph window to see a larger time series graph or create a graph in the spreadsheet from the data in the time columns (Figure 6.5).



Word	2010.1.1	2010.2.1	2010.3.1	2010.4.1	2010.5.1	2010.6.1	2010.7.1	2010.8.1	2010.9.1	2010.10.1	2010
portal	0	0	0	0	0	0	0	0	0	0	
website	0.090909	0.031447	0	0.019305	0	0.014634	0.014837	0.022152	0.143519	0.22425	0.02
ball	0.0										02
math	0.0										02
contact											
momentu											02
definition											00
pluto	0.0										
believe	0.0										04
entropy	0.0										00
hole	0.0										.0
touch											00
question	0.0										08
egg											
moon	0.0										.0
string											00
atom	0.0										03
higg											00
multivers											
conservec											
school	0.0										00
engine	0.0										.0
touching											
moriarty	0	0	0	0	0	0.004878	0	0	0.010802	0.005215	0.0

Figure 6.5. Time series graphs created in a spreadsheet from the **word.clusters+time.txt** file. Some columns have been hidden.

7 Association mining

Mozdeh can automatically check for associations between a query keywords and/or filters in the main search screen and the words in all the texts of a project. It does this by:

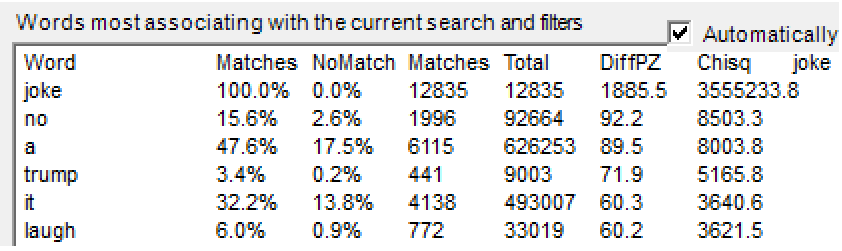
1. Finding all texts that match the query and/or filters.
2. Extracting the words from all matching texts.
3. Checking if these words occur more often in matching texts than in the remaining texts.
4. Listing the results in decreasing order of difference significance, using difference in proportions and chi-square statistical tests. The chi-square value is the most appropriate for statistical tests.
5. Further analysing the statistically significant terms to identify their causes.

This is achieved in the search window with the *Mine associations* button.

7.1 Search terms or Boolean queries

If a query is entered in the top left-hand corner search box and the Mine Associations button is clicked then a word association list will be displayed in the bottom right hand corner box (see Figure 1). The words at the top of the results list displayed on the screen are those that are most likely to have a genuine association with the query, including concepts, topics and linguistic styles.

For example, in a project containing a set of female comedians' YouTube channels, a keyword search for joke, one of the most strongly associating terms is *no* (Figure 7.1). Exploring the project (with the query *joke AND no*) shows that the phrase "no joke" was common, with commenters claiming that their comments were serious (perhaps ironically).



Words most associating with the current search and filters ☑ Automatically

Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	joke
joke	100.0%	0.0%	12835	12835	1885.5	3555233.8	
no	15.6%	2.6%	1996	92664	92.2	8503.3	
a	47.6%	17.5%	6115	626253	89.5	8003.8	
trump	3.4%	0.2%	441	9003	71.9	5165.8	
it	32.2%	13.8%	4138	493007	60.3	3640.6	
laugh	6.0%	0.9%	772	33019	60.2	3621.5	

Figure 7.1. Association mining results for the query *joke* in a project of comments on videos in a set of female comedians' YouTube channels.

The word association results report the percentage of texts that match a query and contain the listed word (the Matches column in Figure 7.1). They also report the percentage of texts not matching the query that contain the term. In the example, the term no occurs in 15.6% of comments matching the query joke and in 2.6% of comments not matching it.

The matches percentage should be considered when assessing the importance of an association. For example, whilst Trump associates with joke (Figure 7.1), this term only occurs in 3.4% of tweets.

7.2 Statistical association tests

The chi-square values reported in the association mining box cannot be interpreted at face value for statistical hypothesis tests. This is because when many statistical tests are run at the same time, the chances of drawing a false conclusion from at least one of them is very high. This is the statistical problem of familywise error rates. This problem occurs with Mozdeh when it calculates many chi-square values at the same time, leading to a high chance that at least one of the chi-square values is misleadingly high.

To give a concrete example, (and using approximate language) there is only a 5% chance of one of Mozdeh's chi-square values being higher than 3.841 if there wasn't really an association. Thus, if Mozdeh reported only one word in its association box and that word had a chi-squared higher than 3.841 then you would only have a 5% chance of being incorrect if you concluded that there was an association. In contrast, if Mozdeh reported 100 chi-squared values and one of them was 3.841 then you could not safely conclude that this word had a genuine association. This is because Mozdeh has run 100 tests and so it is very likely one or more of them had a chi-square value of at least 3.841 by accident.

To reduce the risk of falsely believing that a term is significant when examining multiple chi-square values, Mozdeh uses the Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure. This effectively tests all the words at once and reports the significant terms using a single test. This controls the risk of false positives from running 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.
- Three stars *** is significant at the (familywise) 0.1% level.

As illustrated in Table 7.1, higher chi-square values are needed for a term to be significant when the Benjamini-Hochberg test is run. For example, whilst the term *appreciation* is significant at the 0.1% level, the term *children* is not even significant at the 5% level, despite its chi-square value 16.7 being above the critical value 3.841 for a single test. This term is not significant because Mozdeh conducted so many plausible tests (47,009) that the threshold of 3.841 is no longer enough to guard against having at least one false result from this huge number of tests. In contrast, even with 47,009 tests, a chi-square value of 266.6 is extremely unlikely to occur by chance if there is not a genuine underlying association and so it is reasonable to assume that the term *appreciation* has a genuine association.

On a 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 generate a statistically significant result. This is normally the number of words that occur in at least 2 or 3 different posts.

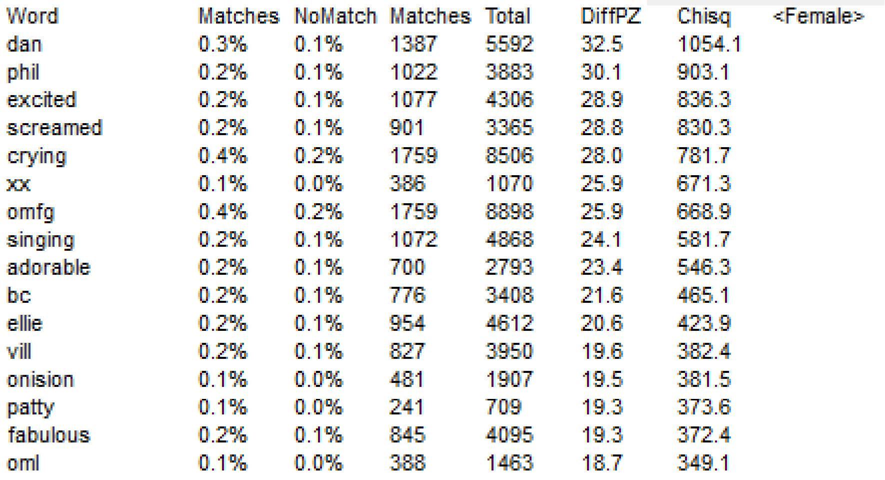
The example below illustrates the star system. At the 5% and 1% levels there is statistically significant evidence of associations for 11 terms. At the 0.1% level there is statistically significant evidence of associations for 10 terms (excluding *footage* from the previous set).

Table 7.1. The top 20 word association mining results for the term facepalm within a corpus of 55,634 comments on YouTube videos from official museum channels. A total of 47,009 words occurred in at least three different texts and could potentially have generated a statistically significant result, so the Benjamini-Hochberg correction is based on this figure.

Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	Sig (47009 tests)
appreciation	25.0%	0.0%	1	13	32.7	266.6	***
expressing	25.0%	0.0%	1	15	30.4	230.9	***
attacking	25.0%	0.0%	1	24	24	144	***
flie	25.0%	0.1%	1	29	21.9	119	***
progress	25.0%	0.1%	1	31	21.1	111.2	***
guessing	25.0%	0.1%	1	31	21.1	111.2	***
hmm	25.0%	0.1%	1	34	20.2	101.3	***
thread	25.0%	0.1%	1	36	19.6	95.7	***
reasonable	25.0%	0.1%	1	36	19.6	95.7	***
insult	25.0%	0.1%	1	38	19.1	90.6	***
k	25.0%	0.1%	1	46	17.3	74.7	***
constant	25.0%	0.1%	1	47	17.2	73.1	***
arguing	25.0%	0.1%	1	48	17	71.5	***
argue	25.0%	0.1%	1	58	15.4	59	***
pussy	25.0%	0.1%	1	67	14.3	51	***
apart	25.0%	0.2%	1	90	12.4	37.7	***
section	25.0%	0.2%	1	92	12.2	36.9	***
footage	25.0%	0.2%	1	120	10.7	28	**
children	25.0%	0.4%	1	197	8.3	16.7	
cute	25.0%	0.4%	1	229	7.7	14.3	

7.3 Gender

Mozdeh can detect terms that associate with male or female authors. To discover what males or females comment about more often than each other, enter a blank query, select a gender from the *User gender* drop-down box and click *Mine associations*. Words associating with the selected gender authors’ posts will be displayed in the association mining box. As explained in the Gender chapter, author genders are guessed from their first names. For example, selecting a female author filter and clicking *Mine associations* would compare texts known to be female-authored against texts from male or unknown gender authors (Figure 7.1).

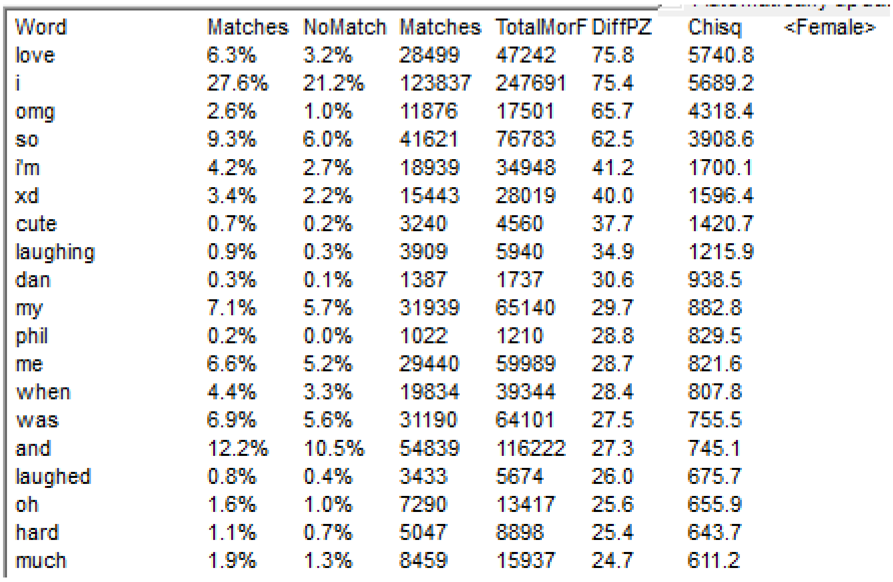


Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	<Female>
dan	0.3%	0.1%	1387	5592	32.5	1054.1	
phil	0.2%	0.1%	1022	3883	30.1	903.1	
excited	0.2%	0.1%	1077	4306	28.9	836.3	
screamed	0.2%	0.1%	901	3365	28.8	830.3	
crying	0.4%	0.2%	1759	8506	28.0	781.7	
xx	0.1%	0.0%	386	1070	25.9	671.3	
omfg	0.4%	0.2%	1759	8898	25.9	668.9	
singing	0.2%	0.1%	1072	4868	24.1	581.7	
adorable	0.2%	0.1%	700	2793	23.4	546.3	
bc	0.2%	0.1%	776	3408	21.6	465.1	
ellie	0.2%	0.1%	954	4612	20.6	423.9	
vill	0.2%	0.1%	827	3950	19.6	382.4	
onision	0.1%	0.0%	481	1907	19.5	381.5	
patty	0.1%	0.0%	241	709	19.3	373.6	
fabulous	0.2%	0.1%	845	4095	19.3	372.4	
oml	0.1%	0.0%	388	1463	18.7	349.1	

Dan
Phil
excited
screamed
crying

Figure 7.2. Association mining results for the gender filter female in a project of comments on videos in the PewDiePie YouTube channel.

The gender association results are weakened by the presence of authors with unknown genders. To increase their power, first create a subproject with only male or female commenters in and then apply the association mining again (see the Gender chapter for more information). As shown in Figure 7.3 the results have overlaps but are statistically much more powerful (i.e., higher chi-square values).



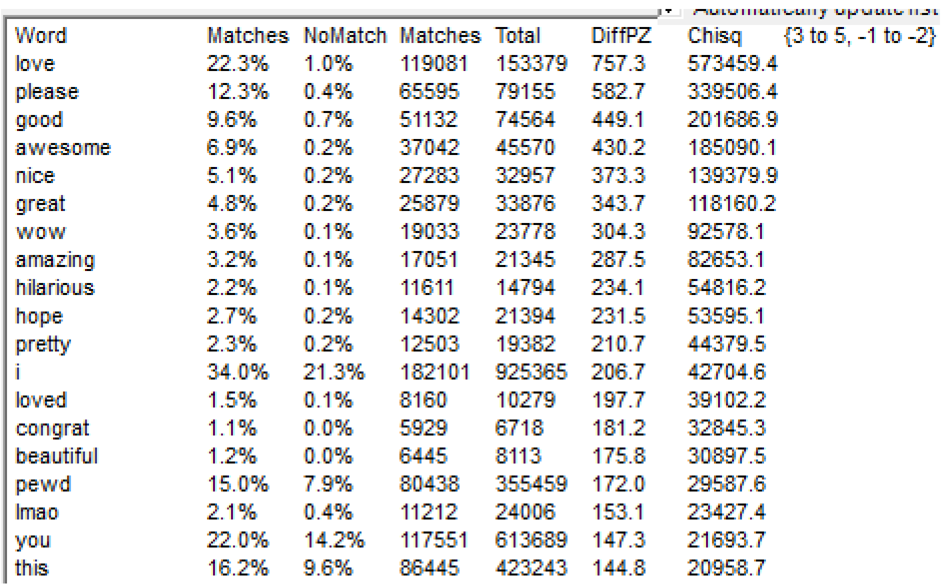
Word	Matches	NoMatch	Matches	TotalMorF	DiffPZ	Chisq	<Female>
love	6.3%	3.2%	28499	47242	75.8	5740.8	
i	27.6%	21.2%	123837	247691	75.4	5689.2	
omg	2.6%	1.0%	11876	17501	65.7	4318.4	
so	9.3%	6.0%	41621	76783	62.5	3908.6	
i'm	4.2%	2.7%	18939	34948	41.2	1700.1	
xd	3.4%	2.2%	15443	28019	40.0	1596.4	
cute	0.7%	0.2%	3240	4560	37.7	1420.7	
laughing	0.9%	0.3%	3909	5940	34.9	1215.9	
dan	0.3%	0.1%	1387	1737	30.6	938.5	
my	7.1%	5.7%	31939	65140	29.7	882.8	
phil	0.2%	0.0%	1022	1210	28.8	829.5	
me	6.6%	5.2%	29440	59989	28.7	821.6	
when	4.4%	3.3%	19834	39344	28.4	807.8	
was	6.9%	5.6%	31190	64101	27.5	755.5	
and	12.2%	10.5%	54839	116222	27.3	745.1	
laughed	0.8%	0.4%	3433	5674	26.0	675.7	
oh	1.6%	1.0%	7290	13417	25.6	655.9	
hard	1.1%	0.7%	5047	8898	25.4	643.7	
much	1.9%	1.3%	8459	15937	24.7	611.2	

Love
I
OMG
XD
cute

Figure 7.3. Association mining results for the gender filter female in a project of comments on videos in the PewDiePie YouTube channel, after selecting a subproject containing only texts authored by known male or female commenters.

7.4 Sentiment

Issues that attract predominantly positive (or negative) comments can be found by entering a blank query, clicking the + button in the sentiment box and then clicking **Mine associations**. Words associating with positive sentiment (i.e., occurring most in positive texts) will be displayed. Comment sentiment is estimated with the sentiment analysis program, SentiStrength, that is built in to Mozdeh. Terms that describe positive sentiment are to be expected so the remaining words are more interesting. The + button sets the sentiment to be at least 3 out of 5 for positivity (i.e., at least moderately positive) and either -1 or -2 for negativity (i.e., mild negative sentiment or none).



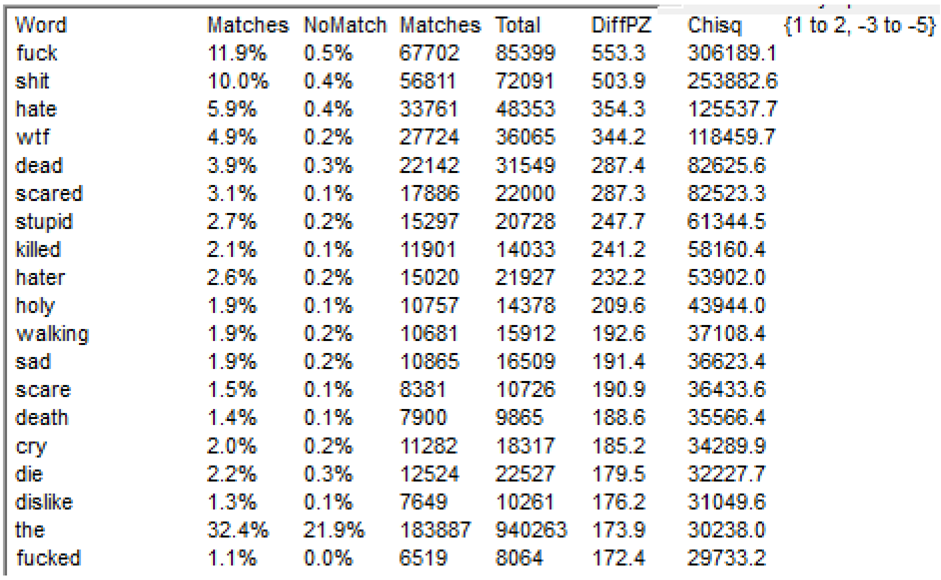
Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	{3 to 5, -1 to -2}
love	22.3%	1.0%	119081	153379	757.3	573459.4	
please	12.3%	0.4%	65595	79155	582.7	339506.4	
good	9.6%	0.7%	51132	74564	449.1	201686.9	
awesome	6.9%	0.2%	37042	45570	430.2	185090.1	
nice	5.1%	0.2%	27283	32957	373.3	139379.9	
great	4.8%	0.2%	25879	33876	343.7	118160.2	
wow	3.6%	0.1%	19033	23778	304.3	92578.1	
amazing	3.2%	0.1%	17051	21345	287.5	82653.1	
hilarious	2.2%	0.1%	11611	14794	234.1	54816.2	
hope	2.7%	0.2%	14302	21394	231.5	53595.1	
pretty	2.3%	0.2%	12503	19382	210.7	44379.5	
i	34.0%	21.3%	182101	925365	206.7	42704.6	
loved	1.5%	0.1%	8160	10279	197.7	39102.2	
congrat	1.1%	0.0%	5929	6718	181.2	32845.3	
beautiful	1.2%	0.0%	6445	8113	175.8	30897.5	
pewd	15.0%	7.9%	80438	355459	172.0	29587.6	
lmao	2.1%	0.4%	11212	24006	153.1	23427.4	
you	22.0%	14.2%	117551	613689	147.3	21693.7	
this	16.2%	9.6%	86445	423243	144.8	20958.7	

Figure 7.4. Terms associating with positive sentiment for comments on videos from PewDiePie.

- *Love* occurs in 22.3% of the positive comments compared to 1.0% of the remaining comments. This term describes positive sentiment so is not particularly surprising but indicates that sentiment is often expressed with this strong term.
- *Hilarious* occurs in 2.2% of the positive comments compared to 0.1% of the rest. Together with lmao, these terms suggest that humour is commented on positively but the low percentages suggest that humour is *not* the main cause of positivity.
- *Game* (not shown above, scrolled further down the list) occurs in 6.6% of the positive comments compared to 3.4% of the rest. This suggest that the games played by PewDiePie attract positive comments.

For negative associations, click the - button in the sentiment box and click **Mine associations.** This button sets the negative sentiment to be -3 to -5 for negativity (i.e., at least moderately negative) and either 1 or 2 for positivity (i.e., mild positive sentiment or none).

.



Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	{1 to 2, -3 to -5}
fuck	11.9%	0.5%	67702	85399	553.3	306189.1	
shit	10.0%	0.4%	56811	72091	503.9	253882.6	
hate	5.9%	0.4%	33761	48353	354.3	125537.7	
wtf	4.9%	0.2%	27724	36065	344.2	118459.7	
dead	3.9%	0.3%	22142	31549	287.4	82625.6	
scared	3.1%	0.1%	17886	22000	287.3	82523.3	
stupid	2.7%	0.2%	15297	20728	247.7	61344.5	
killed	2.1%	0.1%	11901	14033	241.2	58160.4	
hater	2.6%	0.2%	15020	21927	232.2	53902.0	
holy	1.9%	0.1%	10757	14378	209.6	43944.0	
walking	1.9%	0.2%	10681	15912	192.6	37108.4	
sad	1.9%	0.2%	10865	16509	191.4	36623.4	
scare	1.5%	0.1%	8381	10726	190.9	36433.6	
death	1.4%	0.1%	7900	9865	188.6	35566.4	
cry	2.0%	0.2%	11282	18317	185.2	34289.9	
die	2.2%	0.3%	12524	22527	179.5	32227.7	
dislike	1.3%	0.1%	7649	10261	176.2	31049.6	
the	32.4%	21.9%	183887	940263	173.9	30238.0	
fucked	1.1%	0.0%	6519	8064	172.4	29733.2	

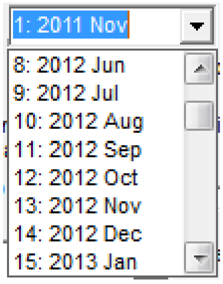
Figure 7.5. Terms associating with negative sentiment for comments on videos from PewDiePie.

- Swear words are often used to express negative sentiment.

- *Walking* occurs in 1.9% of the negative comments compared to 0.1% of the rest. This is due to discussions of the TV series Walking Dead, which falsely triggers negativity due to the presence of the word dead in the title. This can be ignored.
- *People* (not shown above, scrolled further down the list) occurs in 4.4% of the negative comments compared to 1.9% of the rest. This strange result occurs because commenters sometimes generalise when disagreeing with a behaviour (e.g., "[] is one of the annoying people that…", "people are [at war with] PewDiePie").
- *media* (not shown above, scrolled further down the list) occurs in 0.6% of the negative comments compared to 0.2% of the rest. This is due to a combination of criticism of media coverage of PewDiePie and comments on negative stories about PewDiePie in the mass media.

7.5 Time

To find out issues that attract more interest during one period than another, choose a date range, identify the number of the start and end of the period in the *First date to show* dropdown box and enter the first and last numbers in square brackets in the search box. In the example below, to check for important words from before 2013 the date range query would be [1 14].

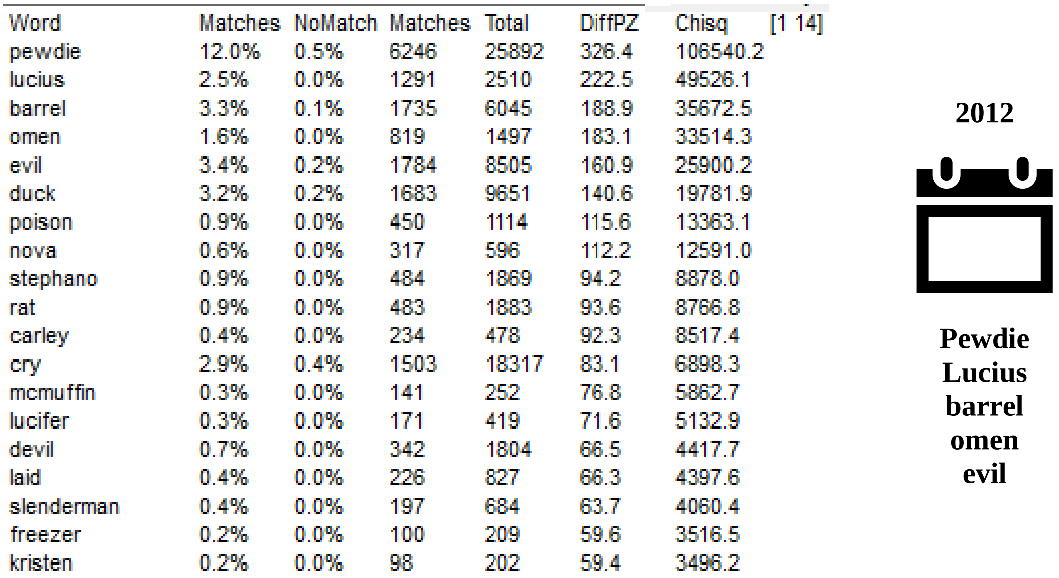


1: 2011 Nov

8: 2012 Jun
9: 2012 Jul
10: 2012 Aug
11: 2012 Sep
12: 2012 Oct
13: 2012 Nov
14: 2012 Dec
15: 2013 Jan

Figure 7.6. Dates and associated numbers displayed in the First date to show dropdown box.

After clicking *Mine associations*, words associating with the specified date range (i.e., 2017 in the above example) will be displayed.



Word	Matches	NoMatch	Matches	Total	DiffPZ	Chisq	[1 14]
pewdie	12.0%	0.5%	6246	25892	326.4	106540.2	
lucius	2.5%	0.0%	1291	2510	222.5	49526.1	
barrel	3.3%	0.1%	1735	6045	188.9	35672.5	
omen	1.6%	0.0%	819	1497	183.1	33514.3	
evil	3.4%	0.2%	1784	8505	160.9	25900.2	
duck	3.2%	0.2%	1683	9651	140.6	19781.9	
poison	0.9%	0.0%	450	1114	115.6	13363.1	
nova	0.6%	0.0%	317	596	112.2	12591.0	
stephano	0.9%	0.0%	484	1869	94.2	8878.0	
rat	0.9%	0.0%	483	1883	93.6	8766.8	
carley	0.4%	0.0%	234	478	92.3	8517.4	
cry	2.9%	0.4%	1503	18317	83.1	6898.3	
mcmuffin	0.3%	0.0%	141	252	76.8	5862.7	
lucifer	0.3%	0.0%	171	419	71.6	5132.9	
devil	0.7%	0.0%	342	1804	66.5	4417.7	
laid	0.4%	0.0%	226	827	66.3	4397.6	
slenderman	0.4%	0.0%	197	684	63.7	4060.4	
freezer	0.2%	0.0%	100	209	59.6	3516.5	
kristen	0.2%	0.0%	98	202	59.4	3496.2	

2012

Pewdie
Lucius
barrel
omen
evil

Figure 7.7. Terms associating with comments from 2011 and 2012 on PewDiePie videos.

- *Pewdie* occurs in 12.0% of the 2011-12 comments compared to 0.5% of the later comments, signifying a greater focus on the channel owner in the early years.
- *Lucius* occurs in 2.5% of the 2011-12 comments compared to 0.0% of the rest. Browsing comments reveals that this is the name of a game that PewDiePie played

and posted videos of in 2012. Other terms in the list also associate with this game (e.g., barrel, evil), so it was clearly an important part of the channel in the early days even though it is now forgotten.

- *Duck* occurs in 3.2% of the 2011-12 comments compared to 0.2% of the rest. Browsing the comments shows that this is another game.

7.6 Word/filter combinations

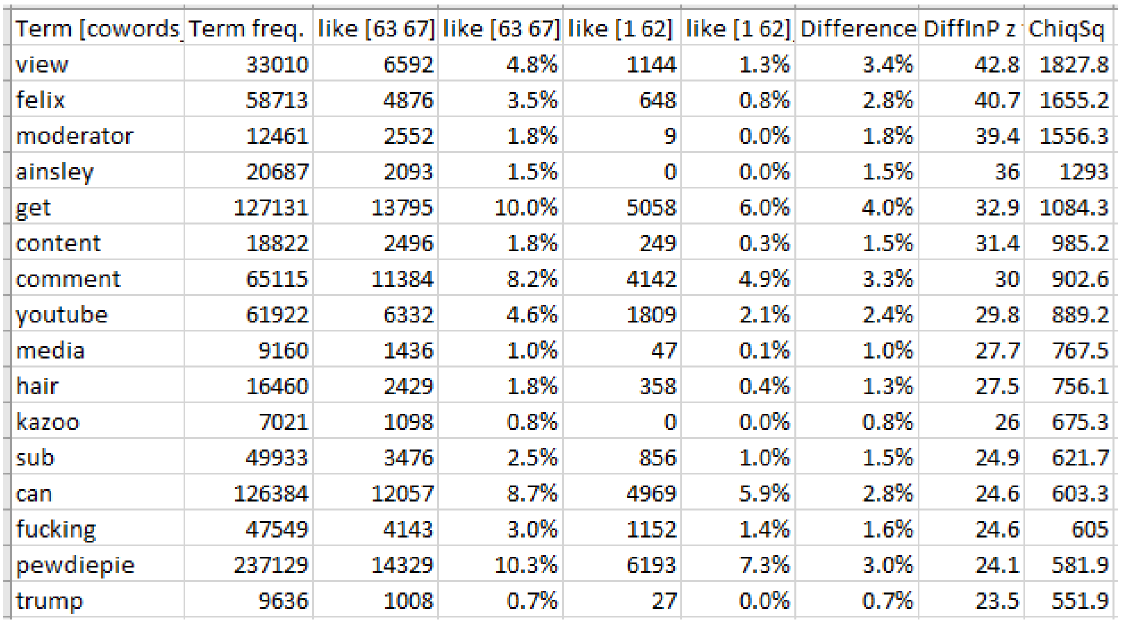
Entering a query and a gender/sentiment/time filter together gives words that associate with the combination compared to the remaining posts. It is usually better to use the comparisons method described below (Association mining comparisons tab) for this.

7.7 Comparing multiple queries and/or filters

Queries and/or filters can be compared against *each other* rather than against the remaining texts. For example, comments containing *like* from 2017 could be compared with comments containing *like* from before, ignoring all comments not containing *like*. This may show why PewDiePie was liked in 2017 compared to earlier dates. This can be achieved with the *Association mining comparisons* tab by entering the two searches separated by a comma. This produces similar results to the *Mine associations* button but saved to a plain text file rather than displayed on screen.

When the file is ready, it can be loaded into a spreadsheet to view. It can be sorted in increasing or decreasing order of difference in proportions z (the DiffInP z column) as a convenient way to find words associated with the earlier or later period. The comments have been copied to Excel in the example below, sorted, formatted as percentages, and important cells highlighted in yellow. This shows that *view*, *comment*, *felix*, *get* and other terms occur more in 2017 than before. For instance, in the spreadsheet below, *view* occurred in 4.8% of comments from 2017 containing *like*, but in only 1.3% of comments from before 2017 containing *like*.

Searching for *view AND like [63 67]* shows lots of discussion about why PewDiePie is very popular and gets many Likes, views and comments. Felix is PewDiePie's first name, associating with a shift to call him this rather than PewDiePie (e.g., "I like you now Felix") rather than associating with *like*.



Term [cowords	Term freq.	like [63 67]	like [63 67]	like [1 62]	like [1 62]	Difference	DiffInP z	ChiqSq
view	33010	6592	4.8%	1144	1.3%	3.4%	42.8	1827.8
felix	58713	4876	3.5%	648	0.8%	2.8%	40.7	1655.2
moderator	12461	2552	1.8%	9	0.0%	1.8%	39.4	1556.3
ainsley	20687	2093	1.5%	0	0.0%	1.5%	36	1293
get	127131	13795	10.0%	5058	6.0%	4.0%	32.9	1084.3
content	18822	2496	1.8%	249	0.3%	1.5%	31.4	985.2
comment	65115	11384	8.2%	4142	4.9%	3.3%	30	902.6
youtube	61922	6332	4.6%	1809	2.1%	2.4%	29.8	889.2
media	9160	1436	1.0%	47	0.1%	1.0%	27.7	767.5
hair	16460	2429	1.8%	358	0.4%	1.3%	27.5	756.1
kazoo	7021	1098	0.8%	0	0.0%	0.8%	26	675.3
sub	49933	3476	2.5%	856	1.0%	1.5%	24.9	621.7
can	126384	12057	8.7%	4969	5.9%	2.8%	24.6	603.3
fucking	47549	4143	3.0%	1152	1.4%	1.6%	24.6	605
pewdiepie	237129	14329	10.3%	6193	7.3%	3.0%	24.1	581.9
trump	9636	1008	0.7%	27	0.0%	0.7%	23.5	551.9

Figure 7.8. A comparison of terms associating with comments containing like from 2017 with comments containing like from before 2017 on PewDiePie videos. The proportions columns have been formatted as percentages in the spreadsheet.

When looking for statistically significant evidence of association then it is important to protect against false positive test results generated by testing multiple words at the same time. To help with this, Mozdeh runs a single (familywise) Benjamini-Hochberg (Benjamini & Hochberg, 1995) test on all the terms and marks them for significance level using the standard star rating scheme.

- One star * is significant at the (familywise) 5% level.
- Two stars ** is significant at the (familywise) 1% level.
- Three stars *** is significant at the (familywise) 0.1% level.

7.8 Analysis of individual significant terms

Individual terms that are statistically significant in the word association results can be investigated to find out their cause. This can be achieved in one or more of the following ways.

Key Word in Context (KWIC) investigation: This involves reading a random sample of texts containing the term and matching the specified query (e.g., gender, sentiment, time or text query) to discover the typical context in which the term is used. For example, reading a random sample of comments from 2012 on PewDiePie videos containing the term *barrel* would reveal that they are mostly about the videogame Barrels. Thus, the cause of this term was extensive commenting on his videos about Barrels in 2012 (and much less commenting in subsequent years on this game). The KWIC method is a standard technique in corpus linguistics for discovering the typical meanings of individual words from the context in which they are used. Texts can be displayed in random order using the Random option in the search results order drop-down box.

Word association investigation: This involves running an additional word association analysis using the original query/filters and adding the significant term to be investigated. For example, to investigate barrels in this way, it would be entered as a query on top of the original filter (year 2012) and then the word association mining button clicked. This generates a list of terms that associate with the significant term that give context into how it was used. For barrels, these terms include game and play, which would point to Barrels being associated with playing a game.

Both KWIC and the word association investigation have limitations. KWIC is time consuming and important context may be missed. The word association investigation does not give as rich context as KWIC but may point to trends that may be overlooked when reading the texts. Thus, it is best to use both methods to get the most comprehensive information.

7.9 Analysis of sets of significant terms

If there is a large set of statistically significant terms then it is useful to identify the key themes behind them to report a manageable amount of information rather than a long list of terms and context about them. For this, each term must first be investigated individually, as above, and then labelled with a theme that generalises their underlying cause. For example, for the PewDiePie 2012 results, both barrel and evil might be labelled with the theme *Barrels game*. At a more general level they might be labelled with the broader theme *Computer games*. After all the terms are labelled with a theme, the themes can be reported as the research discoveries instead of the individual terms.

The process of labelling terms is like the grounded theory labelling process in that it involves reflecting about what the key factor underlying the term is and generalising it as far as is reasonable for the research goals. As with grounded theory, it is important to repeatedly check the themes and revise them as necessary so that they are coherent and meaningful. This may involve ignoring aspects that are not relevant (e.g., assigning some terms to Spam or Other themes) and comparing within themes to make sure that they are coherent and between themes to make sure that they are usefully distinct. This stage may involve extensively revising the themes multiple times until they present a useful perspective about the data. This is a necessarily subjective and qualitative interpretation stage.

7.10 Refining queries to eliminate irrelevant matches

If a query matches some irrelevant texts then changing the query may give more precise results, improving the association mining accuracy or power.

Increased specificity: Adding an extra term with AND can make a search more specific. For example, suppose that someone is interested in comments about Nazi salutes and their query *salute* generated some matches about other salutes. Modifying the query to *salute AND nazi* should eliminate this problem.

Increased specificity: Searching for a quoted phrase can also make a search more specific. For example, suppose that someone is interested in comments about Nazi salutes but the query *salute* generated some matches about other salutes. Modifying the query to “nazi *salute”* should eliminate this problem.

Removing irrelevant matches: removing search matches by specifying terms that they must *not* contain using the “-” command helps when there is a group of irrelevant matches. For example, the query *nazi* has some matches for the phrase *grammar nazi*, that is irrelevant to the main controversy. Changing the query to *nazi -grammar* ensures that comments containing the term *grammar* no longer match. The minus sign must be immediately before the query term. Multiple exclusions are allowed, so discussions of the Nazi zombies game can also be excluded with *nazi -zombie -grammar*.

7.11 Method limitations

The following limitations must be considered when interpreting association mining results.

- The chi-square values should be interpreted as indicative rather than conclusive because their reporting violates two assumptions required for them to be valid: a) Only a single test should be carried out because multiple tests greatly increase the overall (familywise) error rate; b) the data should be independent, whereas commenters are likely to copy each other and there may be some spam. This can be largely ignored when using the familywise error rates (starred results) because these account for the first issue.
- False positives can occur if the query terms are ambiguous – so Nazi in the example above could refer to the political party or over-strict grammar checking.
- False negatives will occur when someone discusses something without explicitly mentioning it. For example, someone might criticise a political video by commenting “Disgusting fascist propaganda!”, without mentioning its creator or the name of the political party that produced it. A comment might also report, “I agree, it is terrible”, using “it” instead of a specific topic word to reference it (anaphora).

7.12 Statistical background: The chi-square test

The chi-square statistic used for word associations is derived from a 2x2 table (Table 7.1; see also http://www.ling.upenn.edu/~clight/chisquared.htm). For each word in the list, Mozdeh completes this table and then uses the appropriate statistical formula to deduce the chi-square value for the table. The larger the chi-square value, the more evidence there is to support the conclusion that the term associates with the query. For a single test, a value above 3.841 would be enough to reject the hypothesis that there was no underlying association, at the 0.05 level. Because of the multiple tests used, however (see: method limitations), the chi-square values should be interpreted as guides rather than rigorous tests.

Table 7.1. A 2x2 table used to calculate the chi-square values reported in the word association mining results.

	Texts containing word	**Texts not containing word**	**Total texts**
Texts matching query	A	B	**A+B**
Texts not matching query	C	D	**C+D**
Total texts	**A+C**	**B+D**	**A+B+C+D**

For example, considering the *kazoo* word in Figure 7.7, from a project with 4028959 comments with 908322 from 2017, some of the cells can be completed and the rest can be deduced from them (Table 7.2).

Table 7.2. A 2x2 table used to calculate the chi-square values reported in the word association mining results, completed for *kazoo*.

	Texts containing kazoo	Texts not containing kazoo	Total texts
Texts matching the 2017 query	A=6974	B	A+B=7021
Texts not from 2017	C	D	C+D
Total texts	A+C=908322	B+D	A+B+C+D=4028959

8 Networks

8.1 Networks of connections between users

Mozdeh can draw networks of the interactions between users for some types of text. It constructs these networks from the texts by identifying when one user mentions the name of another user in their post. It then draws a network where the nodes are users and there is an arrow from one node to another when the first node's author mentions the second node's author. Arrow thicknesses are proportional to the number of texts in which a mention occurs. This only works if messages mention usernames in them and if the usernames mentioned are the same as the names recorded for text authors in the Mozdeh project. If this does not occur, then the networks drawn from a project will not contain arrows.

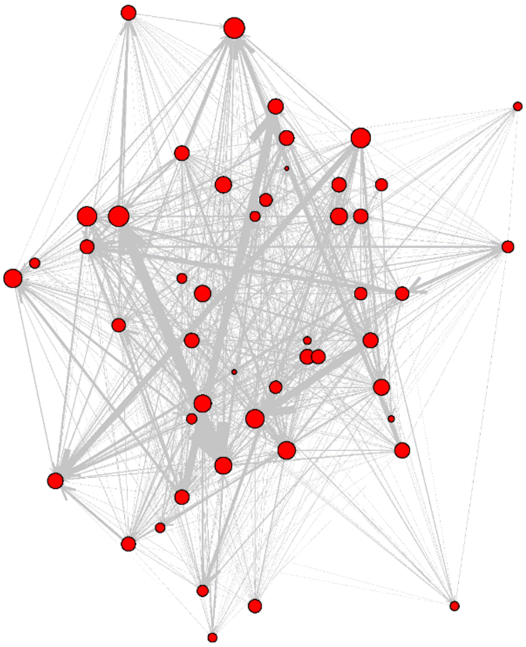


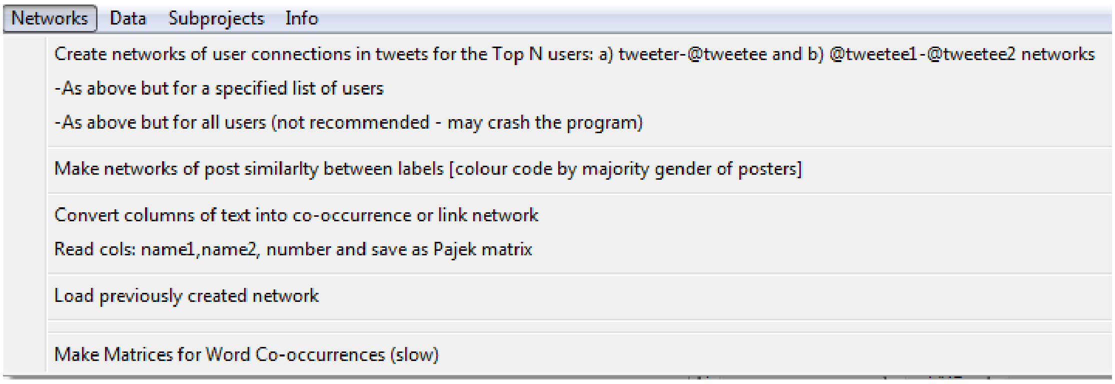
Figure 8.1. A Twitter network of the 50 nodes in a project with the most messages to or from them. *Node names are hidden to preserve anonymity*. Thicker lines indicate a greater number of messages from the arrow source to the arrow target.

When Mozdeh creates a network is often plots a selection of 50 users or texts rather than all users or texts because a diagram with large numbers of nodes tends to be too cluttered to be informative. Mozdeh arranges the nodes to help to reveal patterns but they can be adjusted to make the patterns or node names clearer. This can be manually overridden by dragging nodes in the network.

Networks can reveal the interaction patterns of the most prolific users in a project, including who they interact with and how much.

8.2 Networks of label (query) similarity

Mozdeh has many different options for ways of creating networks (see the Network menu, Figure 8.2). For the "Make networks of post similarity" option, the nodes are the labels in the dataset and lines are drawn between labels based on how similar the text associated with them. The labels here are those given by Mozdeh to each text in a project. These are usually the queries used to obtain the text. For example, if a Twitter project had data gathered from ten queries then the network would have ten nodes, one for each query, and there would be a line between nodes when the queries had generated similar tweets.



Networks Data Subprojects Info

Create networks of user connections in tweets for the Top N users: a) tweeter-@tweetee and b) @tweetee1-@tweetee2 networks

-As above but for a specified list of users

-As above but for all users (not recommended - may crash the program)

Make networks of post similarlty between labels [colour code by majority gender of posters]

Convert columns of text into co-occurrence or link network

Read cols: name1,name2, number and save as Pajek matrix

Load previously created network

Make Matrices for Word Co-occurrences (slow)

Figure 8.2. The Mozdeh Network menu with the main network creation options.

Mozdeh creates multiple networks from the same data, each network using different similarity thresholds. The size of a node is proportional to the number of posts for the label and nodes are colour coded by the majority gender of the post author (Figure 8.3).

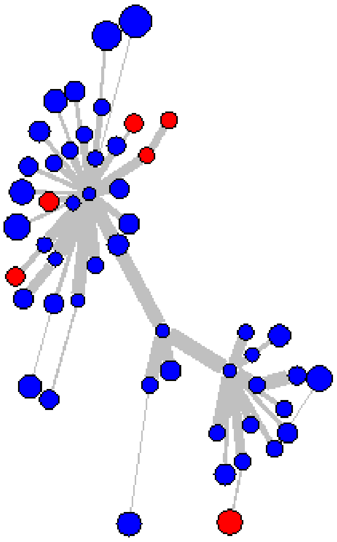


Figure 8.3. A YouTube network of the comments on videos in museum channels. *Channel names are hidden to preserve anonymity*. Each node is a YouTube museum channel. Thicker lines indicate lines indicate that the two channels have more similar content. Node sizes are proportional to the number of comments on the channel. Red: mostly female commenters; Blue: mostly male commenters.

8.3 Saving and printing networks

Networks are automatically saved by Mozdeh when created. To see the files, select View all Reports from the Analyse menu.

Network diagrams can be copied into Word by taking a screenshot of the image and then editing it in Microsoft Paint. Alternatively, the Print menu can be used to print a high-quality copy of the network to any format supported by the host computer, such as PDF or TIFF.

9 Saving and changing projects

Recall that when data is gathered by Mozdeh or imported then it is grouped together for analysis in a collection known as a “project”. A user can have multiple projects and Mozdeh always processes them separately. Mozdeh has options for restricting the texts analysed by creating subprojects (Figure 9.1), for creating new projects and for exporting a subset of texts in the project to a plain text file. These project management options can be useful when a project needs to be filtered before it is analysed.

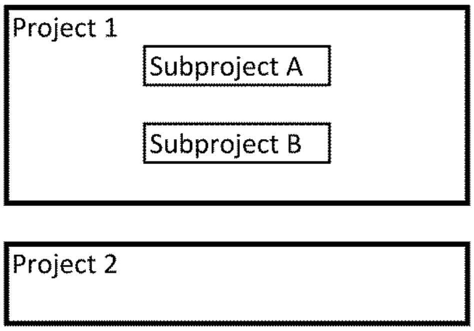


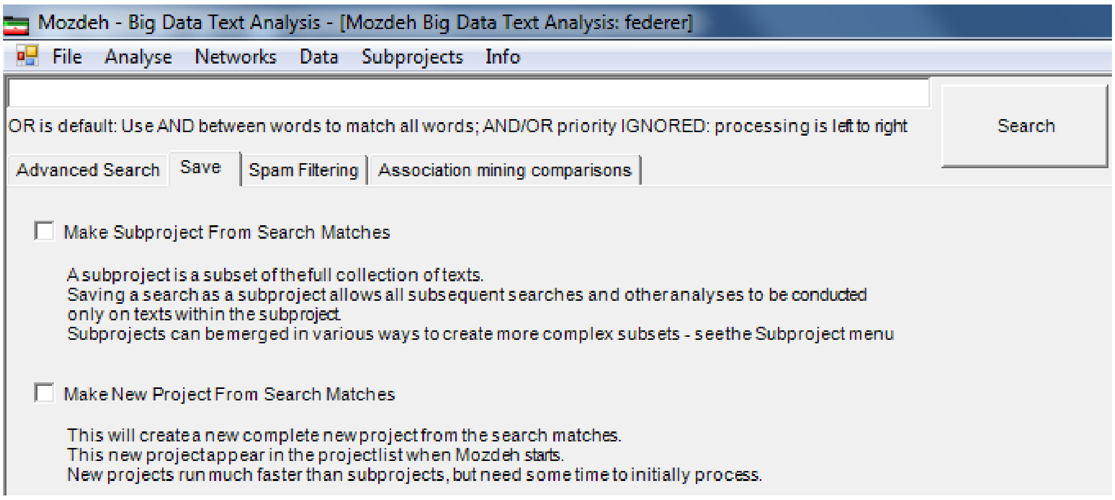
Figure 9.1. Mozdeh organises its data in separate projects. Within a project, a set of texts can be selected and labelled as a subproject. Texts within a subproject can be analysed separately, ignoring the rest of the texts in a project for most purposes.

9.1 Subprojects

The subproject feature allows Mozdeh to ignore some of the texts in a project and only process the remainder. This is achieved by registering the texts to keep as a “subproject” and loading this subproject. A subproject is saved by Mozdeh as a list of the texts that should be processed. Subprojects can be created either from a query and/or filters, or by combining existing subprojects.

Subprojects can be useful to “remember” complex queries or to refine the scope of the project for association mining. For example, a subproject could be made from all posts with authors that have been assigned a gender. This would make the word association mining more powerful for males and females because they would only be compared against the other gender and not against the texts with unknown genders.

A subproject can be made from a query and/or filters by ticking the *Make subproject* checkbox in the *Save* tab in the main search window. Subprojects can be combined in various ways (e.g., merged, inverted) using *Manage subprojects* in the *Subprojects* menu.


Mozdeh - Big Data Text Analysis - [Mozdeh Big Data Text Analysis: federer]
File
Analyse
Networks
Data
Subprojects
Info
Search
Advanced Search
Save
Spam Filtering
Association mining comparisons

OR is default: Use AND between words to match all words; AND/OR priority IGNORED: processing is left to right

□ Make Subproject From Search Matches

A subproject is a subset of the full collection of texts.
Saving a search as a subproject allows all subsequent searches and other analyses to be conducted only on texts within the subproject
Subprojects can be merged in various ways to create more complex subsets - see the Subproject menu

□ Make New Project From Search Matches

This will create a new complete new project from the search matches.
This new project appear in the project list when Mozdeh starts.
New projects run much faster than subprojects, but need some time to initially process.

Figure 9.2. The new project and subproject creation options in the Save tab.

The main disadvantage of subprojects is that they slow Mozdeh down considerably. If a subproject will be used extensively then it may be better to create a new project instead. Although a new project is slow to load for the first time, creating it can save time in the long run.

9.2 New projects

Mozdeh can create new projects from the texts matching a query and/or filters by clicking the *Make new project* option in the *Save* tab. This can be useful as follows.

- When a project contains many irrelevant or Spam texts that need to be filtered out (see: http://mozdeh.wlv.ac.uk/SpamRemoval.html).
- When subsets of a project need to be analysed separately (e.g., users that have been assigned a gender) and faster speed is needed than with subprojects.

New projects are listed in the Startup wizard when Mozdeh is next started.

9.3 Raw data and exporting text files

The raw data from a Mozdeh project is stored in a plain text (tab delimited) format file inside a folder called raw_data. This can be found by selecting *Open raw data folder* from the *Analyse* menu. If it is not too large, it can be loaded into a spreadsheet to be viewed. This is the complete set of information known by Mozdeh about the data, except for sentiment and gender.

The texts matching a query can be saved into a separate file is used by setting up the query and/or filters and checking the *Save matching texts to text file* option in the search window *Save* tab. This saves a limited set of information with each text and is less comprehensive than the raw data plain text file. It can be opened in a spreadsheet.

10 Summary

As described above, Mozdeh can be used to gather tweets and YouTube comments as well as being able to import texts from other sources. It can analyse these texts for gender, sentiment, topic and time. Detailed instructions for the methods are available in the Mozdeh website (http://mozdeh.wlv.ac.uk) as well as a frequently asked questions list.

As mentioned in the introduction, Mozdeh can be used to support different research styles. For a descriptive quantitative-led analysis, any or all the methods can be combined to describe aspects of the data, including gender, time and sentiment. The sentiment analysis methods also support quantitative hypothesis testing with the provision of average sentiment strength scores and confidence intervals. The methods can also play a supporting role in a qualitative-led analysis by providing insights into a dataset to support other methods (e.g., interviews, ethnography).

The most powerful feature of Mozdeh is its word association analysis. This can be used as the basis for research projects that investigate differences between genders, sentiments, topics or time periods. This inv

Mozdeh is likely to evolve in the future as new social web sites appear that offer free text data that it can exploit. It was originally created in 2006 for a different type of data (RSS feeds: Thelwall & Prabowo, 2007) and has since added new data sources and functions. At some stage, Twitter and YouTube may retire their free services. If this happens and would be a disaster for a long running project, then its functions can still be used on texts imported from other sources, such as commercial data providers.

11 References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 57(1), 289–300.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163-173.

Thelwall, M. & Prabowo, R. (2007). Identifying and characterising public science-related fears from RSS feeds. Journal of the American Society for Information Science and Technology, 58(3), 379-390.

Wilkinson, D. & Thelwall, M. (2011). Researching personal information on the public Web: Methods and ethics, Social Science Computer Review, 29(4), 387-401.

Thelwall, M. (2015). Evaluating the comprehensiveness of Twitter Search API results: A four step method. Cybermetrics,18-19, p1. http://www.scit.wlv.ac.uk/~cm1993/papers/TwitterComprehensiveness.pdf