Association Mining with Mozdeh

Mozdeh can automatically check for patterns of association between any query you enter and the words in all the texts in your project. It does this by:

1. Finding all texts that match your query and filters.
2. Extracting the words from all matching texts
3. Checking if these words occur more often in matching texts
4. Listing the results in decreasing order of “surprisingness” (using a chisquare statistical test of surprisingness).

These words suggest concepts, topics and linguistic styles that associate with texts matching your query and filters. In addition to the surprisingness, it is also important to consider the *coverage* – the percentage of comments matching the query (i.e., the Matches percentage column). If a term is rare (i.e., low coverage) then it may not be an important factor for your query, even if it has a strong association with it.

This is a three-step process:

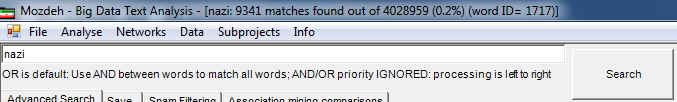
1. You chose the search and/or filter to mine associations for.
2. Mozdeh produces a list of associating terms.
3. You interpret the terms in Mozdeh’s list, rejecting those that are irrelevant and investigating the relevant ones further by reading comments containing them.

This process will be illustrated with an investigation of the Nazi controversy of the YouTube star PewDiePie. The project for this consists of four million comments extracted from videos in the PewDiePie channel.

# Basic queries

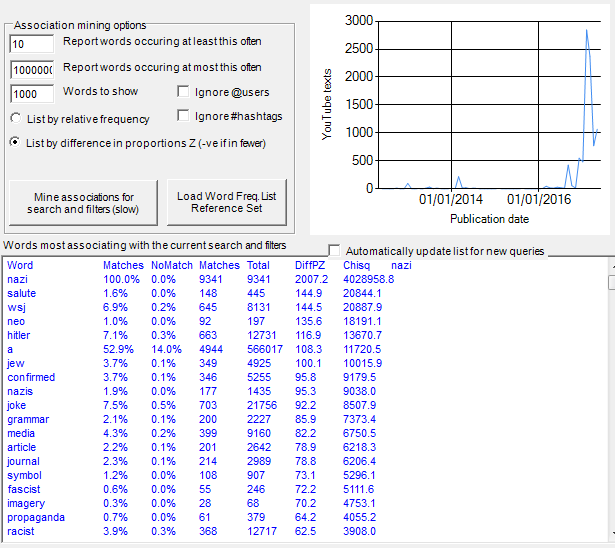
To identify associations with a basic word or phrase query, enter the query in the search box, click the **Search** button and then the **Mine associations for search and filters (slow)** button. Words that associate with the query will then be listed in the bottom right text box. The next step is to draw your own conclusions about why these words associate with your query, and whether the association is relevant to your research.

illustrate the basic process, the query nazi was run in the PewDiePie project, and the mine associations button clicked.



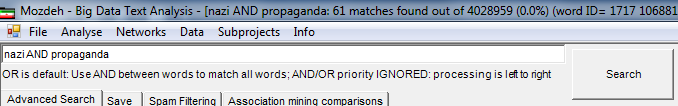
From the spike in the graph below right, it seems that the Nazi controversy generated lots of comments. The word list shows the words that associate with “Nazi” in PewDiePie comments. Some are discussed here to illustrate how you might analyse these kinds of results. Note that capital letters and plurals have been removed from the words before building the list (e.g., Nazi and Nazis are counted as the same word).

* *Nazi*: this word is unsurprisingly at the top of the list. In the list, the Matches column reports 100%, meaning that all matching YouTube comments contain the word Nazi. Similarly, 0% in the NoMatch column means that the YouTube comments not matching the Nazi query do not contain the term Nazi. This is not surprising!
* *Salute* occurs in 1.6% of the comments that contain Nazi and 0.0% of the remaining comments. Although 1.6% is a low percentage so a manual browsing of the matching comments might not notice it, it is clearly relevant to one of the incidents in which PewDiePie gave a Nazi salute. The low percentage (i.e., low coverage) nevertheless suggests that the salute itself was *not* a critical issue for many people.
* *wsj* occurs in 6.9% of the comments that contain Nazi and 0.2% of the remaining comments. This suggests that Wall Street Journal coverage of the event was important to PewDiePie commenters.
* *A* occurs in 52.9% of matching comments compared to 14/0% of the rest, probably due to phrases like “a nazi salute”, so this can be ignored.
* *Grammar* occurs in 2.1% of matching comments and 0.1% of the rest but the matches are irrelevant because they are from unrelated comments about “grammar Nazis”.



## Investigating the results

If the associating words do not make sense then they can be investigated by adding them to the search with the keyword AND, reading the matching comments. For example, tin investigate “propaganda” in the above results, change the query from nazi to nazi AND propaganda (tip: first uncheck the Automatically update list for new queries box above the associating words list if you don’t want it to change). Reading the matching comments in the list box underneath (not shown here), most either accuse PewDiePie of producing Nazi propaganda or mock these claims.

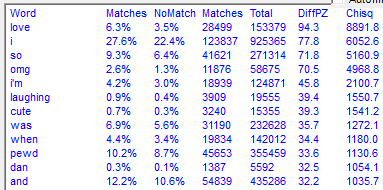


# Gender

To find out 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 the **Mine associations for search and filters (slow)** button.



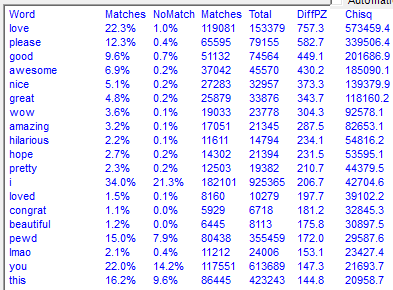
Words associating with (i.e., tending to occur in) female commenters’ posts will be displayed in the bottom left association mining box. Commenters’ genders are guessed from their first names. For example, any commenter with first name Maria would be assumed to be female, whereas those with a first name of Jacob would be assumed to be male. Most commenters will have an unknown gender because this rule works only about a third of the time. The associating word list therefore compares words in comments from known female commenters against comments from male or unknown gender commenters.



* *Love* occurs in 6.3% of the comments from known females compared to 3.5% of the male or unknown gender comments, a very strong gender association. This suggests either that females love PewDiePie more, or his activities, or are more expressive in YouTube posts.
* *I* occurs in 27.6% of the comments from known females compared to 22.4% of the male or unknown gender comments, a very strong gender association. This suggests either that females express a personal perspective more than do males.
* *pewd* occurs in 10.2% of the comments from known females compared to 8.7% of the male or unknown gender comments, a very strong gender association. This suggests either that females tend to address PewDiePie directly or that they use this name abbreviation more frequently.

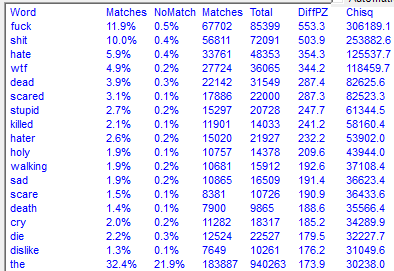
# Sentiment

To find out issues that attract predominantly positive (or negative) comments, enter a blank query, click the + button in the sentiment box and click **Mine associations for search and filters (slow)**. Words associating with positive sentiment (i.e., occurring most in positive sentiment posts) will be displayed in the bottom left association mining box. Comment sentiment is estimated with a 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.



* *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.

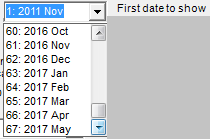
For negative associations, click the - button in the sentiment box and click **Mine associations for search and filters (slow)**.

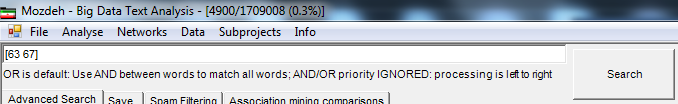


* Swear words are often used to express negative sentiment.
* *Walking* occurs in 1.9% of the negative comments compared to 0.1% of the rest. This is due to discussions of the TV series Walking Dead, which falsely triggers negativity due to the presence of the word dead in the title. This can be ignored.
* *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”).
* *People (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.

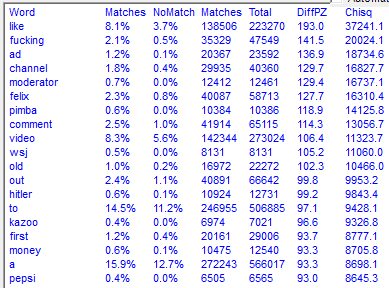
# Time

To find out issues that attract more interest in one time 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 2017 the date range to enter as a search would be [63 67].





Finally, click **Mine associations for search and filters (slow)**. Words associating with the specified date range (i.e., 2017 in the above example) will be displayed in the bottom left association mining box.



* *like* occurs in 8.1% of the 2017 comments compared to 3.7% of the earlier comments. This is due discussions in 2017 about why PewDiePie is popular. See the comparisons section below for information about how this was discovere3d.
* *ad* occurs in 1.2% of the 2017 comments compared to 0.1% of the rest. Browsing the results of the query *ad AND like [63 67]* suggests that this is due to discussions of some of PewDiePie’s sponsors withdrawing, and apparently increased comment spam.
* *Moderator* occurs in 0.7% of the negative comments compared to 0.0% of the rest. Browsing the results of the query *moderator AND like [63 67],* this is due to a PewDiePie’s call for comment moderators, with many applications taking the form of comments starting with “I would like to be a moderator because…”.

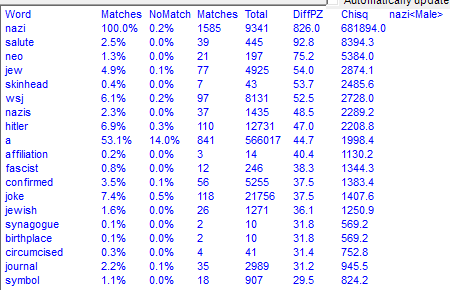
# Word/filter combinations

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

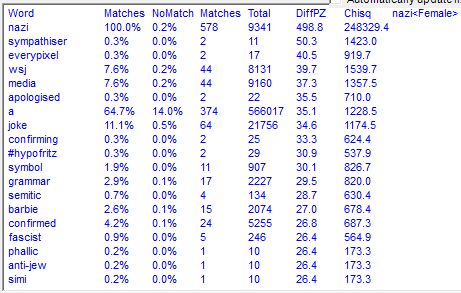
To illustrate, to mine associations for male comments mentioning the word Nazi, enter Nazi as the query and select male as the gender (see composite image below and note the top-right indicator of the search and gender filter), then click **Mine associations for search and filters (slow)**.



This gives the results below.



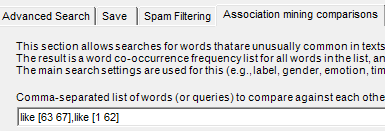
Changing gender to female and running again:



* *sympathiser* is a more significant term for females than for males but this is based on only 2 matching comments (see the second Matches column in the female set) so it is not important.
* *salute* is a more significant term for males than for females.

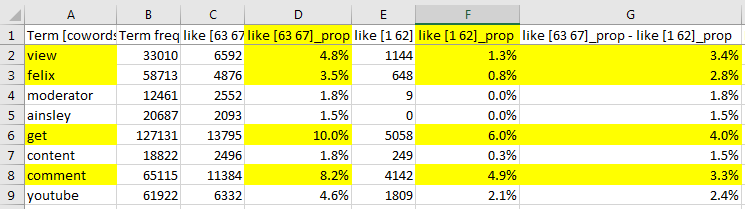
# Comparisons

Sometimes it is helpful to compare two queries and/or filters against each other rather than one query against the remaining texts. For example, in one of the examples above it was shown that like occurred more in 2017 than in comments posted before this year. To find out why, we could compare comments containing like from 2017 with comments containing like from before, ignoring all comments not containing like. To do this, select the **Association mining comparisons** tab and enter the two searches in the new text box, separated by a comma, as below, and click the **Compare words matching the above queries (Slow)** button. Enter a file for the results (they are not displayed on the screen).



When the file is ready, open it and copy (Ctrl-A, Ctrl-C) and paste (Ctrl-V) its contents into a spreadsheet to see the results. They can be sorted in increasing or decreasing order of difference in proportions z (one of the columns) as a convenient way to find words associated with the earlier or later time 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, comment* 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* in particular.



# Tips

Here are some tips that may help to improve the performance of some of the methods described above.

**Refining queries to eliminate irrelevant matches**. If your query matches some irrelevant comments then you may be able to change the query to solve this problem in two ways.

*Increased specificity*: Adding an extra term with AND can make a search more specific. For example, suppose that you are interested in comments about Nazi salutes and your 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 you are interested in comments about Nazi salutes and your 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*:



# Method limitations

The following limitations must be considered when interpreting the results.

* The chisquared 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 increases the likely error rate; b) the data should be independent, whereas commenters are likely to copy each other and there may be some spam.
* Apart from the issues discussed below, using the starred results is safe because they protect the “familywise error rate” with much stricter requirements for a positive result. <https://en.wikipedia.org/wiki/Family-wise_error_rate#Hochberg's_step-up_procedure>
* 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 video with Nazi imagery by commenting that it is “disgusting fascist propaganda”, avoiding the term Nazi. Another might reply to this comment with “I agree, it is terrible”, using “it” instead of any specific topic word.

# The chisquared test

The chisquared statistic is from the following 4x4 table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Texts containing word | Texts not containing word | Total texts |
| Texts matching query | A=Match | B= | A+B |
| Texts not matching | C | D | C+D |
| Total texts | A+C | B+D | A+B+C+D |

For example, considering the wsj word in the first set of results shown above, and using the fact that the total number of comments in the project is 4028959, some of the cells can be completed (if you like maths puzzles you will be able to fill in the rest).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Texts containing wsj | Texts not containing wsj | Total texts |
| Texts matching nazi query | A=645 | B | A+B=9341 |
| Texts not matching nazi query | C | D | C+D |
| Total texts | A+C=8131 | B+D | A+B+C+D=4028959 |