allourideas

download your data

We created allourideas.org to improve the way that groups collect information, and now we are making it easy for you to download your data from our site. I'm hoping that once more people have their hands on their data, we'll start to discover cool new ways to learn from it.

To further facilitate this development, we are happy to release all the data from one of our largest wiki surveys: the Washington Post's wiki survey about who had the worst year in Washington (thank you Ryan Kellett and colleagues at the Post). Please feel free to play with the data and send us your own analysis of it. We'll post the coolest things we get here on the blog. It's kind of like the Netflix Prize, but minus the million dollars.

Keep reading for detailed documentation about the data files and for links to the Washington Post's data.

The website can generate files in comma-separated values (csv) format. These csv files can then be analyzed using spreadsheet programs (e.g., Excel or OpenOffice) or statistical programs (e.g., R). You can request the csv files at the bottom of the admin page (http://www.allourideas.org/[yourwikisurvey]/admin). Then our server will create the files (this could take up to one hour) and email you a link to download them. You can only request files from the wiki surveys where you are an administrator.

There are three files that you can request: 1) a file where each record is an idea, 2) a file where each record is a vote, and 3) a file where each record is a non-vote. Here are the three files from the Washington Post's Worst year in Washington wiki survey:

- wikisurvey_727_ideas_2014-04-05T18_21_17Z.csv
- wikisurvey_727_votes_2014-04-05T18_21_13Z.csv
- wikisurvey_727_nonvotes_2014-04-05T18_21_19Z.csv

[Note, these files were regenerated on April 5, 2014 after our most recent improvements to the csv files. If you would like the version of the files we originally posted, you can find them here: ideas, votes, non_notes]

Below is the documentation for each type of file.

1) *Idea file (wikisurvey_[ID#]_ideas_[time].csv)*

This file has one record for each idea, and the time in the filename is the UTC time that the file was requested in ISO-8601 format.

- **Wiki survey ID**: ID number for your wiki survey. Every row in your file will have the same number. It is also the number that will be in the name of all of the csv files from this wiki survey.
- **Idea ID**: ID number for this idea (unique).
- **Idea Text**: Text of the idea (e.g., "Free ice cream all the time").
- **Wins**: Number of times that this idea has won a pairwise comparison.
- **Losses**: Number of times that this idea has lost a pairwise comparison.
- **Times involved in Cant Decide**: Number of times this idea was in a contest where "I Can't Decide" was selected.
- **Score**: Estimated probability that this idea will beat a randomly chosen other idea from this wiki survey for a randomly chosen session. The score will always be between 0 and 100.
- **User Submitted**: TRUE/FALSE to indicate whether the idea was uploaded by a user. Ideas uploaded from the creation page

(http://www.allourideas.org/questions/new) are not user submitted. Ideas submitted from the voting page (e.g., http://www.allourideas.org/planyc_example/) are user submitted.

- **Session ID**: ID number of the session in which this idea was added.
- **Created at**: Indicates when this record was created in the database. So, all seed ideas will have approximately the same "Created at" time. All timestamps represent time UTC time.
- **Last Activity**: Indicates the last time that this record was updated.
- **Active**: TRUE/FALSE to indicate whether the idea is active (i.e., in the pool of ideas that can be shown to users) at the time the csv file was created.
- **Appearances on Left**: Number of times that this idea appeared on the left of the pair. Note that appearance just refers to this idea appearing on the screen. Not all appearances result in a vote. This would occur, for example, i the voter closes her browser while this idea is on the screen.
- **Appearances on Right**: Number of times that this idea appeared on the right of the pair. See note about Appearances on Left for more information.
- **Info:** records the "info" parameter passed into the url (more information) for the session where this idea was uploaded. If no parameter was passed, this value will be NA.

2) *Vote file (wikisurvey_[ID#]_votes_[time].csv)*

This file has one record for each vote, and the time in the filename is the UTC time that the file was requested in ISO-8061 format.

- **Vote ID:** ID number for this vote (unique).
- **Session ID**: ID number for the session in which this vote was cast.
- **Wikisurvey ID**: ID number for this wikisurvey. This will be the same for each row in the file.
- **Winner ID**: Idea ID for winner of this vote.
- **Winner Text**: Text of winner of this vote.

- **Loser ID**: Idea ID for loser of this vote.
- **Loser Text**: Text of loser of this vote.
- **Prompt ID**: ID number of prompt (e.g., “free beer” vs “free ice cream”). Note that this is left/right sensitive: (“free beer” vs “free ice cream”) has a different prompt ID from (“free ice cream” vs “free beer”).
- **Appearance ID**: ID number for this appearance. In early wiki surveys appearance IDs might not be unique because in very rare situations multiple votes were cast on the same appearance ID.
- **Left Choice ID**: ID number of idea on the left of the prompt.
- **Right Choice ID**: ID number of idea on the right of the prompt.
- **Created at**: When this record was created in the database. All timestamps represent UTC time.
- **Updated at**: When this record was last updated in the database.
- **Response Time (s)**: Response time for this vote as measured on the client side (in seconds). Beware that measuring response time using javascript is not exact, and we recommend that you treat this measurement with some caution.
- **Missing Response Time Explanation**: reason that response time is missing (if it is missing).
- **Valid**: TRUE/FALSE indicating whether this record is valid. This is part of our infrastructure for handling attempted gaming of the voting process. For your analysis, we recommend only using records where valid==TRUE. Please emai if you have questions.
- **Hashed IP Address**: a cryptographic hash of the IP address of the session. From this value you can see if two sessions are from the same IP address, but not what that IP address actually is.
- **URL Alias:** the string that appears after “www.allourideas.org/” for this wiki survey. For example, for “www.allourideas.org/studentgovernment” the url alias is “studentgovernment.”
- **User agent string:** this string records the browser and operating system of

the computer used for this vote. For more information here is the Wikipedia page on user agent strings.

- **Referring url:** this is the url from which this session originated. For example, if the session originated after the person clicked on a link to your wiki survey from http://www.nytimes.com/example, then the referring url would be http://www.nytimes.com/example. For more information here is the Wikipedia page on HTTP referrers. Other possible values are "DIRECT_VISIT" if the url is typed in directly, and "REFERRER_NOT_FOUND" if we are not able to locate the referrer.
- **Widget**: TRUE/FALSE value that records whether the vote was recorded at the widget interface (e.g., http://widget.allourideas.org/[url]). If FALSE the vote was recorded at the main interface (e.g., http://www.allourideas.org/[url]).
- **Info:** records the "info" parameter passed into the url (more information). If no parameter was passed, this value will be NA.
- **City:** "locality" field in call to Google Geocoding API based on lat-lon inferred from IP address. This field is only made available under a data protection agreement.
- **State:** "administrative_area_level_1″ field in call to Google Geocoding API based on lat-lon inferred from IP address. This field is only made available under a data protection agreement.
- **Country:** "country" field in call to Google Geocoding API based on lat-lon inferred from IP address. This field is only made available under a data protection agreement.

3) Non-vote file (wikisurvey_[ID#]_non_votes_[time].csv)

This file has one record for each non-vote. There are three kinds of non-votes: "Bounce," "Stopped_Voting_Or_Skipping," and "Skip". A "Bounce" is recorded when a session begins but no vote or skip occurs (i.e., someone visits the site, but does not participate). A "Stopped_Voting_Or_Skipping" is recorded when a prompt (e.g., "Free ice cream" vs "free beer") is shown to a voter and no response of any kind is returned. This occurs most frequently when the voters closes her browser. A “Skip” is recorded when the voter clicks “I can’t decide." The time in the filename is the UTC time that the file was requested in ISO-8601 format.

- **Record Type**: "Bounce,” “Stopped_Voting_Or_Skipping,” and “Skip” are the only valid values. These are defined above.
- **Skip ID**: ID number for this skip. If row is not a skip, this value will be NA.
- **Appearance ID**: ID number of this appearance.
- **Session ID**: ID number of the session in which this action occurred.
- **Wikisurvey ID**: ID number for this wiki survey. This should be the same for all rows in the file.
- **Left Choice ID**: ID of idea appearing on left of the prompt.
- **Left Choice Text**: text of idea appearing on the left.
- **Right Choice ID**: ID of idea appearing on the right of the prompt.
- **Right Choice Text**: text of idea appearing on the left.
- **Prompt ID**: ID number of prompt (e.g., “free beer” vs “free ice cream”). Note that this is left/right sensitive: (“free beer” vs “free ice cream”) has a different prompt ID from (“free ice cream” vs “free beer”).
- **Reason**: If the Record Type is “Skip” this indicates why the voter could not decide. For “Bounces” and “Stopped_Voting_Or_Skipping” this field should always be NA.
- **Created at**: When this record was created in the database. All timestamps represent UTC time.
- **Updated at**: When this record was last updated in the database.
- **Response Time (s)**: Response time for this vote as measured on the client side (in seconds). Beware that measuring response time using javascript is not exact, and we recommend that you treat this measurement with some caution. For “Bounces” and “Stopped_Voting_Or_Skipping” this field should always be NA.

- **Missing Response Time Explanation**: If response time is missing, an explanation of why.
- **Valid**: TRUE/FALSE indicating whether this record is valid. This is part of our infrastructure for handling attempted gaming of the voting process. For your analysis, we recommend only using records where valid==TRUE. Please emai if you have questions.
- **Hashed IP Address**: a cryptographic hash of the IP address of the session. From this value you can see if two sessions are from the same IP address, but not what that IP address actually is.
- **URL Alias**: the string that appears after "www.allourideas.org/" for this wiki survey. For example, for "www.allourideas.org/studentgovernment" the url alias is "studentgovernment."
- **User agent string:** this string records the browser and operating system of the computer used for this vote. For more information here is the Wikipedia page on user agent strings.
- **Referring url:** this is the url from which this session originated. For example, if the session originated after the person clicked on a link to your wiki survey from http://www.nytimes.com/example, then the referring url would be http://www.nytimes.com/example. Fore more information here is the Wikipedia page on HTTP referrers. Other possible values are "DIRECT_VISIT" if the url is typed in directly, and "REFERRER_NOT_FOUND" if we are not able to locate the referrer.
- **Widget**: TRUE/FALSE value that records whether the vote was recorded at the widget interface (e.g., http://widget.allourideas.org/[url]). If FALSE the vote was recorded at the main interface (e.g., http://www.allourideas.org/[url]).
- **Info:** records the "info" parameter passed into the url (more information). If no parameter was passed, this value will be NA.

FAQ:

1) Is it easy to read these files into R?

Yup. We use R so you can be sure that these files play nicely with R. To read the votes files into R, just use this syntax:

```
votes <- read.csv(“wikisurvey_[ID#]_votes_[time].csv”, header=TRUE, sep = “,”, dec=“.”);
```

For the ideas and non-votes, just adjust the above as appropriate.

2) How can we handle the timestamps in R?

The code below will turn the timestamps in the POSIXct data type. Then you can do nice things like time2 - time1 (etc).

```
as.POSIXct(as.character(votes[,“Created.at”]), tz=“UTC”, format=“%Y-%m-%dT%H:%M:%S+00:00”);
```

3) Why is my data file different from the layout described above?

We are constantly improving the data files by adding more information. Therefore, if you would like the most recent data file please return to your wiki survey and re-generate it. A full list of changes to the files is at the end of this page.

4) What is all that funny stuff in my filename?

Your filename includes the UTC time that your file was requested in ISO-8601 format. This helps you keep track of the possibly many files that you might have from your wiki survey. Also, we add a random four-digit string to the end of your filename in order to make it more difficult for someone to else to download it. We ask your browser to strip this four-digit string from the filename when you download it, but that might not work with all browsers. The code that generates the csv filenames is here.

5) What character encoding are you using?

All the files are in UTF-8. We know that we users from all over the world, and UTF-8 should minimize internationalization challenges.

6) Can you explain more about the timestamps?

The timestamp for a vote, skip, and I can't decide is when the response happens, while the timestamp for the bounce is when the appearance was created.

7) Are the appearance IDs ordered?

Sort of. If you sort by appearance ID within session, then you will recover the correct order of appearances within that session. However, if you sort by appearance ID globally, then the order may not be correct because of how we deal with session timeouts. More specifically, the appearance ID for a response coming after a session timeout is created when the response is received not when the appearance is created.

CHANGELOG:

Votes file:

- User agent string and referring url added on March 29, 2011.
- Changed "tracking" to "info" on July 24, 2011.
- Converted times from Pacific Time (US) to UTC time on March 21, 2012.
- Improved session handling deployed Tuesday, March 18th, 2014 at 9am EDT (1300 UTC).
- New file names deployed on April 1, 2014.
- Some geographic information (city, state, country) made available under a data protection agreement on June 30, 2016.

Non-vote file:

- Before March 29, 2011 “Bounce” and “Stopped_Voting_Or_Clicking” were both called “Orphaned Appearance”.
- User agent string and referring url added on March 29, 2011.
- Changed “tracking” to “info” on July 24, 2011.
- Converted times from Pacific Time (US) to UTC time on March 21, 2012.
- Improved session handling deployed Tuesday, March 18th, 2014 at 9am EDT (1300 UTC).
- New file names deployed on April 1, 2014.

Ideas file:

- Added info parameter on July 24, 2011.
- Converted times from Pacific Time (US) to UTC time on March 21, 2012.
- Improved session handling deployed Tuesday, March 18th, 2014 at 9am EDT (1300 UTC).
- New file names deployed on April 1, 2014.

---

bathroomideasgod liked this

i-zygzak liked this

serviceautobucuresti liked this

e-assistant-blog liked this

allourideas posted this

NEWER POST OLDER POST