Timed Tags: Theme Identification with Word Clouds

Leigh Hagestad
Stanford University Computer Science
Stanford, USA
leighh1@stanford.edu

Sarah Nader
Stanford University Computer Science
Stanford, USA
snader2@stanford.edu

ABSTRACT
Tag clouds have a low-class reputation in the data visualization community [10]. However, like their numbered-data counterpart the pie chart, they are used quite often and are frequently found online. While looked down upon in theory, word clouds persist in practice. This paper seeks to explore the use and effectiveness of tag clouds (word clouds) as a means of data visualization. Tag clouds are visual representations of (usually) the frequency count of terms or words in a given document/set of documents. This paper compares the effectiveness of tag clouds in conveying themes, ideas, or a main idea of the relevant document(s) quickly, and how effectively tag clouds enable viewers to retain such concepts or themes over time. We compare tag cloud's effectiveness along these dimensions with the effectiveness of using frequency-based lists of words.

Author Keywords
Tag clouds, word clouds, word lists, retention, TF-IDF cosine similarity

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
For those who are unfamiliar with tag clouds, they are a text data visualization method in which tags, usually single words, have their relative importance encoded in font size and/or color. Common definitions for importance in the tag cloud context are the frequency of the word in a corpus or categorization hierarchies, where larger tags represent the quantity of content items in that category. For the purpose of this article, "word clouds" and "tag clouds" will be used interchangeably, although there are differences to the data behind them [8].

In general, tag clouds have a pretty negative reputation among the data visualization community. Some even go as far as calling them the "mullets of the internet" [10]. Despite the general dismissal of tag clouds as a sensible visualization tool, tag clouds thrive on the internet (where attention spans go to die). Given this contradiction between academic perception and real-world use, we wanted to further investigate the practical utility of tag clouds. If they've made it this far, there must be some redeeming qualities to them. Based on prior literature, we hypothesize that they are indeed useful for conveying ideas, themes, or general concepts - scenarios where the quantitative encoding is not the primary goal.

A prominent use case for tag clouds is in various fields of psychology, such as personnel psychology or social psychology, where the study aims to portray qualitative data like "what is it like to be introverted?" As one personnel psychologist put it, "Although I-O psychologists really like long boring tables, the person on the street probably doesn't. I had to come up with a more engaging way to display this information if I wanted to successfully explain PP" [9]. The existence of word cloud proponents such as Stetz motivated us to attempt to quantify the utility of tag clouds compared to other representations, such as frequency-ordered word lists.

RELATED WORK

Tag Clouds and "Vernacular Visualization"
Fernanda Viegas and Martin Wattenberg succinctly and sufficiently outlines the (then) history and usage Tag Clouds [1]. The note combine past work and commentary on the uses of and cases for/against tag clouds. This note was most useful by identifying the principles of visualization design (or lack thereof) which make the visualization effective/ ineffective. It provides a holistic examination as the word cloud as a phenomenon in Web 2.0 and in the information visualization community. One of the most important themes identified was the use of tag clouds by non academics and/ or data scientists, and how the lack of precision and/or resolution provided in the visualization medium makes it effective for 'vernacular' rather than scientifically rigorous applications.

Tag Clouds for Social Signals
Marti A. Hearst and Daniela Rosner scraped opinions from blogs and articles written about tag clouds and coded the commentary for different categories about how tag clouds are appreciated, used, and perceived by those who chose to write about them and make commentary about them [4]. This method pulls a group of highly-detailed and well articulated feedback about word-clouds/tag-clouds from those individuals on the internet that have spent a great deal of time and energy thinking about them (as evidenced by their taking the time to produce an article on the matter). This study effectively identifies some well-stated advantages and use cases for word-clouds which are corroborated by their fellows. The study found that "tag clouds are primarily a visualization used to signal the existence of tags and collaborative human activity, as

opposed to a visualization used for data analysis" and determined that - in terms of their usability - tag clouds are a good visualization best used for lower-brow visualization needs. They are fun, social, dynamic, and extemporaneous. It is a good medium for drawing out higher-level patterns and frequency information about a collection of text, but is not good for lower-level, specific and precise data about frequency.

Graphical Perception and Presentation Techniques

Halvey & Keane measured the selection time across different variations of word lists and and tag clouds [2]. Their variations include: alphabetical cloud, cloud, alphabetical horizontal list, horizontal list, alphabetical vertical list, vertical list. The user task was to complete 24 different selections from a randomly presented visualization of 10 countries (which were also randomly selected from a list of 60 country names). An alphabetical cloud (2.94s) has similar selection time to an alphabetical horizontal list (2.887s) and alphabetical vertical list (2.892s). However, a cloud has the highest selection time (3.409s) with a vertical list the second highest at 3.241s. After these results, they decided to investigate further the effect of font size on tag cloud selection. They found that the larger the font used, the quicker the selection.

Contrary to the Havley & Keane study, Hassan-Montero proposes that alphabetization is not effective in tag clouds [3]. Instead, he proposes an improvement by clustering words based on similarity. Their study draws from the 'del.icio.us' bookmarking tool to produce a "sample containing 218,063 URLs tagged with 242,349 tags by 111,234 users." With this dataset, they create a tool that calculates tag similarity and produces a corresponding tag cloud. Although they did not test the effectiveness of this new visualization with real users, they devise an equation to evaluate it. The equation measures coverage and overlapping, where "coverage is the number of resources that have been described by at least one of the 95 selected tags. Overlapping is the relative co-occurrence between tags."

Perhaps most relevant to our work is the study on the perception of word/tag clouds as a means for interpreting information [6]. "[Their] note describes two studies to evaluate the effectiveness of differently constructed tagclouds for the various tasks they can be used to support, including searching, browsing, impression formation and recognition." In this note they examined the effectiveness of tag clouds when varying font weight, color, and size, as well as the placement features of sorting, clustering, and spatial layout.

However, while the study was generally well constructed and well-formed, we were concerned about the significance of their results, given the low number of participants used in their study; the first experiment had 13 participants, the second had 11 participants - some of which participated in the first study [6]. Furthermore, a great deal of the work in this study focuses on precise recall rather than impression formation and retention. We sought to formulate a study which would both increase the robustness of their results and conclusions as well as expand upon their findings in both scope and scale by examining impression formation and retention from the 'gist' of the text and draw inferences about the base corpus from two different forms of information presentation.

METHODS

In order to compare the effectiveness of word clouds versus word lists, we selected a reputable paper that contains word clouds as the source of our stimuli. After transcribing the tag cloud into a frequency-ordered word list, we created and distributed Qualtrics surveys on Mechanical Turk with different permutations of the variables being studied. The two main tasks being tested are immediate recall and short-term retention. We accomplish this by asking participants to identify the corresponding age group of the visualizations as well as overall themes, ideas, and values presented in the visualization. Processing the participants' responses involved averaging accuracies and calculating TF-IDF cosine similarity between the responses and word cloud corpus.

Data Source

Building off the social sciences scenario, we decided to focus on a somewhat recent study on Facebook statuses as they relate to personality, gender, and age. The study " 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers, who also took standard personality tests, and found striking variations in language with personality, gender, and age" [7]. As this study has been cited multiple times and employs tag clouds as its main visualization strategy, it became an obvious choice for stimuli in our survey. For our purposes, we utilized the age tag clouds, which consisted of four age buckets: 13 to 18, 19 to 22, 23 to 29, and 30 to 65. The gender word clouds seemed to obvious and had a 50% chance at guessing, while the personality clouds seemed to require prior knowledge about personality traits. Thus, the age categories was a happy medium in that it was easily relatable yet offered more than two options.

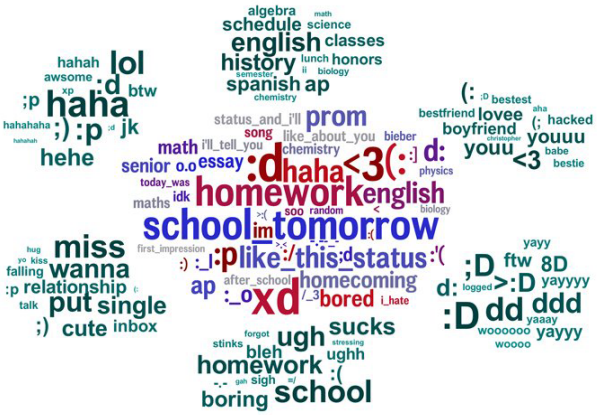

Figure 1. Tag cloud representing the words, phrases, and topics most distinguishing subjects aged 13 to 18. [7]

Pilot Survey

Before launching the survey to a larger audience, it was necessary to test the quality of the questions and responses to ensure well-formed data. In this initial prototype, if you will, we incorporated all four age categories. Participants would be randomly presented with either a word cloud or word list, representing one of the four age groups. We created the word lists manually by transcribing the words in the tag clouds in order of font size (see Figure 2). Furthermore, the participant would either be presented the survey on the same page or be taken to a two-minute and thirty second distractor task, which was a YouTube video from the popular show, "How It's Made," demonstrating how frozen cookie dough is made. To ensure that participants watched the video in its entirety, we employed the Qualtrics feature that hides the advance button until two minutes and thirty seconds have passed.

The survey asked the following questions about the user group being represented in the word cloud/list:

- What gender do you believe the people who wrote the posts from the first page are?
- What themes or values appear most important to the people who wrote posts from the first page?
- What is the average age of the people who wrote the posts from the first page?

As an incentive for Turkers to not, for lack of a better word, troll the survey, we also included a caveat: "Note: In order

semester	class	drunk	papers	shit	classes	wanna
fuck	science	sober	week	fuck	semester	dont
21st	exam	hungover	exams	fucked	class	join
campus	lecture	drunken	assignments	fucking	college	hang
classes	professor	hangover	tests	damn	schedule	kinda
shit	psychology	drink	projects	ass	summer	chill
fucking	psych	drinking	due	fucking	registered	somethin
studying	test	bar	homework	bullshit	taking	badly
in the library	notes	fucked	finals	bitch	online	dun
fuckin	studying	slightly	finished	shitty	fall	Lemme
apartment	philosophy	alcohol	finish	pissed	spring	takers
exam	final	wasted	semester	dude	credits	hangout
go to class	semester	stoned	weeks	fucks	register	lend
roommate	prof	tipsy	classes	bitches	signed	jogging
uni	history	hung	writing	shits		skip
midterm						
lab						
study						
hit me up						
wanna						
haha						
finals						
fucked						
college						
fml						
im						
professor						
dorm						
lectures						
last final						
bitches						
this paper						
interview						
<3						
damn						
go home						
page paper						
roomie						
misses						
drunk						
roommates						
hahaha						
wat						
tattoo						
done_with						
cant						
ko						
na						
ka						
sem						

Figure 2. Frequency-ordered word list depicting the 19-22 age group. This is one of 8 potential stimuli that was presented to participants. The list was presented as an image so that browser formatting was not an issue.

Figure 3. Revised tag cloud representing the words, phrases, and topics most distinguishing subjects aged 23-29, now in black and white. [7]

to receive payment, your response should be similar other respondent's responses by at least 75%." This strategy was adapted from a similar one used in prior research where Turkers were asked to label medical terms in patient-authored posts [5]. After launching the survey on Mechanical Turk, we had approximately 300 total participants across all survey variations.

From the pilot results, we learned that including all four age categories was rather unnecessary. Randomizing which age visualization was portrayed provided no interesting results, as the accuracy remained relatively constant across age buckets. Moreover, from the in-class feedback, we were given the wise advice to make the tag clouds black and white, since the use of color in the original tag clouds could be an unfair advantage over the word lists.

Another key finding from the pilot survey was that presenting the survey on the same page as the visualization was ineffective, since the participant could simply scroll back up to the image to answer the questions. Thus moving forward, we decided to always present the survey on a separate page.

Revised Survey

As aforementioned, one major change in the next round of surveys was to eliminate color from the word cloud visualization. This was completed by manipulating the colors in Photoshop (see Figure 3). We also focused our study to only present the visualization variations for one age group - the 23-29 year old bucket, since age identification across all buckets wasn't one of the main goals of our study.

This particular age group was selected because the younger two were too obvious to identify. The 13-18 bucket contained a plethora of emoticons (as demonstrated in Figure 1), and the 19-22 bucket had 21st as one of the most visually prominent terms (as well as other key words such as "college," "fuck," and "beer"). The 23-29 age group was selected over the 30-65 group because it was less ambiguous to identify, as indicated from the average accuracy results in our pilot survey.

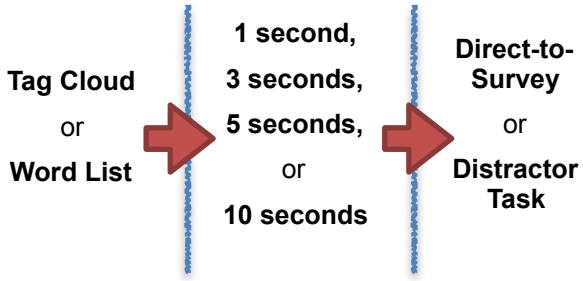

Figure 4. The different variables contained with the survey, resulting in 16 different variations.

In addition to changing the range of stimuli presented, we also altered the way they were presented. In the initial study, we realized that not only could participants scroll back and forth in some versions, but they could also stare at the stimuli for however long they wanted. This introduced potential confounding factors in our results, so we decided to control the duration of time for which participants could view the visualization. The exposure times were 1, 3, 5, or 10 seconds long. We included very short exposure times in hopes of gaining insight into the pre-processing differences between word lists and tag clouds. We were also interested in observing the overall effect of exposure time on trend identification.

Similar to the initial survey, the participants would be directed either to the survey or a distractor task after seeing the stimulus. There were no changes to the actual questions in the survey. As this is a lot of variables to keep track of (essentially a 2 by 2 by 4), we present a flow chart of the survey flow in Figure 4.

Because we realized participants in the 1 or 3 second condition may not be prepared for the flash of a visual, we added the following introduction text on the first page to the survey: "The following page will show collections of words taken from many Facebook posts. The image will only be shown for 1 second, so please pay attention. You will not be able to return to the page once it is shown. Please examine the information shown in the image and think about the type of users who might generate posts using these types of words. You will be asked several questions about these images in a few moments."

Once again, we launched the survey on Mechanical Turk, this time limiting the pool of participants to countries where English is the primary language in an attempt to reduce confounding factors. In the end, there was approximately 25 participants for each survey, resulting in over 400 participants total.

Analysis

Before conducting our analysis on the responses, we manually filtered out malformed inputs. We defined malformed inputs as responses that left either of the age-related questions or responses that contained content related to chocolate chip cookies. Surprisingly, about 1-2 responses per survey type mentioned chocolate chip cookies in some manner. We believe that this is either due to trolling, misunderstanding the question, or being in the 1 second exposure group.

We separate our analysis into two metrics: accuracy and TF-IDF scores. We calculated accuracy by counting how many participants per survey were able to correctly identify the age group and dividing the sum by total number of valid responses in that bucket. This gave us a sense of average accuracy per condition.

Term frequency–inverse document frequency (TF-IDF) cosine similarity was calculated for each valid response using a python script that we wrote. In our case, the corpus was derived from the tag cloud and the query was each participant's free-form response. We used the popular Porter Stemming Algorithm to stem words in the corpus and queries and also stripped out common stop words as well as all punctuation and capitalization. We also normalized all scores by dividing by the maximum score within each condition. Therefore, all scores for responses within a condition were contained in the range 0.0 to 1.0. Although TF-IDF is typically used for information retrieval and text mining, we felt its application was relevant to our study because it provided a quantitative metric for how similar participants' responses were to the text in the stimuli.

Once all the comparative metrics were calculated, we ran ANOVA and T-test examinations on the TF-IDF results, which we shall discuss in the next section.

RESULTS

Accuracy

From Table 1, we see that the worse performance, with an average accuracy of 27%, resulted from a 1-second exposure to the word list without the distractor task. The best performance, 70% accuracy, resulted from a 10 second exposure to the tag cloud without the distractor task. Aside from those two outliers, the average accuracies across conditions were actually quite similar.

TF-IDF Cosine Score

From Table 2, we see that the worst performance (TF-IDF score of 0.427) resulted from a 5-second exposure to the word list without a distractor task. The best performance (TF-IDF score of 0.765) resulted from a 10-second

Average Age-Identification Accuracy

Test type	1 s	3s	5s	10s
List, Distractor	0.5	0.526	0.625	0.65
List, Direct	0.269	0.526	0.538	0.615
Cloud, Distractor	0.44	0.541	0.606	0.625
Cloud, Direct	0.391	0.615	0.571	0.7

Table 1. Average accuracy in identifying the correct age bucket, across all variables

Average TF-IDF Cosine Similarity

Test type	1 s	3s	5s	10s
List, Distractor	0.530	0.441	0.505	0.416
List, Direct	0.519	0.516	0.427	0.409
Cloud, Distractor	0.574	0.529	0.505	0.633
Cloud, Direct	0.614	0.494	0.598	0.765

Table 2. Average TF-IDF cosine similarity scores between the free-text responses and stimuli corpus, across all variables

exposure to the tag cloud without a distractor task. Similar to the accuracy results, the scores across conditions are actually pretty close together. Ultimately, the T-tests between the visualization types proved insignificant for all exposure times. Although ANOVA analysis agreed with this conclusion for most time exposures, the 10-second exposure time was actually ANOVA significant.

DISCUSSION

Overall, any exposure for 5 seconds or less doesn't really have an effect on either age identification or trend extraction. This is probably because especially in the 1 or 3 second case, that the stimuli is presented much too quickly for the viewer to glean anything meaningful. This especially evident in the 1-second word list condition, where the age identification accuracy is essentially due to chance.

In general, age identification was not affected by the visualization type. As demonstrated by Figure 6, average accuracy beyond a 1-second exposure fluctuates very little as time increases, visualization changes, and distractor task toggles.

For theme identification, as measured by TF-IDF, there appears to be a general trend of tag clouds out-performing word lists (see Figure 7). While ANOVA deemed most of these differences insignificant, the disparity in the 10-second exposure time is significant, which is the key finding of our study. There is a distinct inverse relationship between the visualization type and TF-IDF, thereby making a case for the utility of word clouds for conveying qualitative information. While tag clouds did not perform significantly better in a simple task like selecting the correct age group from a list of choices, there was notable improvement for a more complex task such as extracting and summarizing themes from a visualization. This also supports the popularity of tag clouds in more "everyday" mediums such as blogs and websites, as it's easier to extract high level themes in this representation.

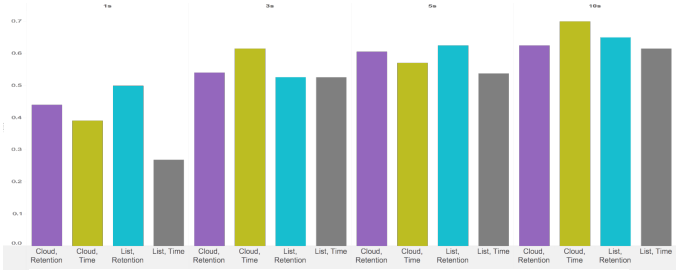

Figure 6. Age-Identification accuracy across all survey variations. Purple & green: tag clouds. Blue & grey: word list.

Shortcomings and Limitations

Although our study in total had over 400 participants, that still only averages out to about 25 participants per survey type. After filtering out the malformed responses, that results in about 22 participants per survey, which really isn't that large of a sample size to run be conclusive.

Regarding our use of TF-IDF cosine similarity, our study's corpus only had 131 unique terms, and the average response length was 7.5 unique terms. Compared to the large corpuses that TF-IDF is typically used for, our corpus seems trivial. In an ideal situation, we would have had access to the data behind the word clouds to use as our corpus, offering more diversity and opportunity to match terms when calculating the TF-IDF per response. That also would make the scores a more accurate representation of similarity.

Furthermore, the TF-IDF calculation introduced a bias for shorter responses. For example, a response that only says "office" would receive a higher score than a response that says "office, wedding, lemonade," even though the latter contains a higher quantity of relevant terms. This is because TF-IDF would penalize the irrelevant term rather than rewarding the additional relevant terms. While we understood this trade-off going into the study, it would be interesting to re-work the metric to accommodate these cases, especially since responses were so short that these differences could have significant impact.

Another limitation of TF-IDF is that it did not incorporate the weighting of the tag clouds, meaning the largest word in the tag cloud and the smallest word would result in the same score.

FUTURE WORK

The most pivotal next steps for a study of this nature would be expand the scope of the project along all dimensions: number of visualization types, time period of recall and

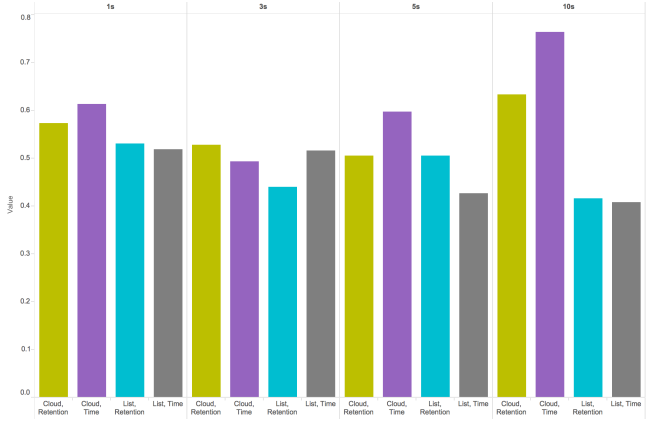

Figure 7. TF-IDF scores across all survey variations. Purple & green: tag clouds. Blue & grey: word list.

response rate, as well as number and diversity of participants.

In terms of expanding the number and type of visualizations, it would be highly interesting to continue to conduct this experiment with visualizations beyond word clouds and frequency-sorted word lists. More numerically-oriented graphical representations of word counts, in-text visualizations, or completely orthogonal forms of visualization would be important for comparison. Conversely, further probing into qualitative methods could also prove fruitful. An example of this would be to include broader swath of questions and ask participants to make inferences about content and authorship of post.

Furthermore, we would consider building upon the findings discussed in the 'Prior Work' section, to incorporate those principles of word-cloud design, comparing results with those found using word clouds whose visualization might reflect changes in layout, color, font weight, or size, etc. In such a case, we could better determine how changes in those graphical features affect retention and recall of themes, ideas, and main concepts of texts.

Additionally, the metrics we used for measuring trend identification could be further improved by utilizing a data source with a richer corpus. As mentioned before, TF-IDF is often used for information retrieval with multiple documents and a large corpus. Future work should either refine the metric used for measuring trend identification or aim for a larger corpus to query responses.

Next, it would be important to relax the bounds of the periods of recall and retention participants are tested with. Rather than presenting a participant with exposure for one, three, five, or ten seconds, we might, for example, test the pre-cortal processing of word clouds versus other forms of text visualization, or - alternatively - could ask users a hour, day, or week later to recall information visualized in a word cloud. While our results regarding the 10 second exposure to tag clouds was significant, it does seem like somewhat of an anomaly in the context of our data. Thus, further work could also attempt to replicate the results presented here.

Finally, we are eager to consider how the findings discussed in this paper might extrapolate to a broader and/or more diverse group of participants. Furthermore, we would be interested to consider the level of education and/or familiarity with statistics, data visualization, or data processing that each participant might have. In being able to distinguish the levels of data familiarity amongst participants, we might be able to confirm prior studies relating to 'vernacular' data visualizations and better distinguish how effectively word clouds might convey information to those with lesser backgrounds in science or data interpretation and/or analysis.

ACKNOWLEDGMENTS

We'd like to thank the CS448B teaching team for a great quarter. This has been a fun journey through data visualization, and we thoroughly enjoyed spending our last quarter at Stanford with you all.

REFERENCES

1. Fernanda B. Viégas and Martin Wattenberg. "Tag Clouds and the Case for Vernacular Visualization". ACM Interactions (July 2008): n. pag. Web.
2. Halvey, Martin and Mark T. Keane. "An assessment of tag presentation techniques." WWW 2007: 1313-1314.
3. Hassan-Montero, Y., & Herrero-Solana, V., "Improving tagclouds as visual information retrieval interfaces," Proc. InfoSciT2006.
4. Hearst, Marti A. and Daniela Rosner. "Tag Clouds: Data Analysis Tool Or Social Signaller?". Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008) (2008): n. pag. Web.
5. Identifying Medical Terms in Patient-Authored Text: A Crowdsourcing-based Approach. Diana MacLean, Jeffrey Heer. J Am Med Inform Assoc, 2013. PDF (963.2 KB) | Journal | Software
6. Rivadeneira, A. W. et al. "Getting Our Head In The Clouds". Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07 (2007): n. pag. Web. 21 May 2016.
7. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791
8. Scott Bateman, Carl Gutwin, and Miguel Nacenta. 2008. Seeing things in the clouds: the effect of visual features on tag cloud selections. In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (HT '08). ACM, New York, NY, USA, 193-202. DOI=http://dx.doi.org/10.1145/1379092.1379130
9. Stetz, Thomas A. (2012). Personnel psychology in 75 Words (or less): A word cloud example. TIP: The Industrial-Organizational Psychologist. 49(4): 27-35. Retrieved from http://www.siop.org/tip/apr12/04stetz.aspx
10. Zeldman, Jeffrey. "Tag Clouds Are the New Mullets." Jeffrey Zeldman Presents The Daily Report. 19 Apr. 2005. Web. 08 May 2016.