Assessing Topic Representations for Gist-Forming

Eric Alexander
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI 53706
ealexand@cs.wisc.edu

Michael Gleicher
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI 53706
gleicher@cs.wisc.edu

ABSTRACT

As topic modeling has grown in popularity, tools for visualizing the process have become increasingly common. Though these tools support a variety of different tasks, they generally have a view or module that conveys the contents of an individual topic. These views support the important task of **gist-forming**: helping the user build a cohesive overall sense of the topic's semantic content that can be generalized outside the specific subset of words that are shown. There are a number of factors that affect these views, including the visual encoding used, the number of topic words included, and the quality of the topics themselves. To our knowledge, there has been no formal evaluation comparing the ways in which these factors might change users' interpretations. In a series of crowdsourced experiments, we sought to compare features of visual topic representations in their suitability for gist-forming. We found that gist-forming ability is remarkably resistant to changes in visual representation, though it deteriorates with topics of lower quality.

CCS Concepts

•**Human-centered computing** → **Empirical studies in visualization;** *Information visualization;*

Keywords

Topic model visualization; word clouds

1. INTRODUCTION

Probabilistic topic modeling is an increasingly popular method of exploring large collections of text documents. Though there are many algorithms for creating topic models, they generally treat each document as a combination of *topics*, which are themselves collections of words that co-occur within the documents. Such models support a variety of investigative tasks, from comparing groups of documents to finding temporal trends. Critical to these sorts of high level inquiries, however, are lower level tasks associated with being able to understand what individual topics are *about*.

One of the most important of these is a task we call **gist-forming**. Gist-forming is the act of a user building a general sense of the semantic content of the words contained within a topic—of grasping the overarching concept or idea that connects them all, if such a connection exists. In most models, some topics are just the result of a statistical coincidence without any meaningful association of the words. Consequently, an important task related to gist-forming is that of **topic evaluation**. When building their gist, a user must not only be concerned with what the topic is about, but also how much its words actually go together, how meaningful it is, and whether or not they can *trust* it to provide insight.

In this paper, we evaluate different factors that may affect a user's ability to form a gist from a topic. These include: the visual encoding used, which can range from lists of words to bar charts to word clouds; the number of words included in the representation of the topic; and the semantic quality or cohesiveness of the topic itself. We conducted a set of crowdsourced experiments that used concrete subtasks to explore the ways in which these factors influence the abstract tasks of gist-forming and topic evaluation. We found that some factors matter more than others. In particular, though we hypothesized that visual encoding would have the greatest effect, performance was remarkably resistant to changes in encoding. Far greater were the effects of topic quality. In the following sections, we lay out the experiments we designed to discover this effect, and describe their implications for architects of topic model visualizations.

Our contributions are as such:

- We articulate the task of gist-finding for future investigation.
- We create an experimental mechanism for assessing it, combining the subtasks of **topic naming** and **word matching**.
- We evaluate the factors of visual encoding, number of words, noise, and topic quality for their influence on this task.

2. RELATED WORKS

There are a variety of different methods for conveying topics that are employed in the literature. Word lists are by far the most common, generally ordered by frequency [3, 5], but occasionally ordered by other metrics [7]. Word clouds, now ubiquitous in online settings, are increasingly being used in

AVI '16 June 07 - 10, 2016, Bari, Italy

ISBN 978-1-4503-4131-8/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2909132.2909252

topic visualizations [10, 16, 23]. Bar charts are sometimes used, as well [1, 20]. We focused primarily on word lists and word clouds for this investigation due to their popularity.

Many comparisons of these encodings have focused primarily on word clouds, and have identified weaknesses in them as a data encoding. In one such study, Rivadeneira et al. consider tasks including search, browsing, impression formation, and word recognition [18]. Though their description of impression formation is similar to gist-forming, their experiments do not match our goals for topic representations. They determine that search and browsing are easier with a simple sorted list—understandable given the organization offered by alphabetical ordering (a finding also supported by Halvey et al. [12]). They focus on the task of recognition—recall of specific words seen in the cloud, and the ability to distinguish from similar but absent words—which is the opposite goal of a generalizable gist (described more in §3). While they penalize participants for identifying words that do not appear in the word cloud but are related, topic gists need to extend to words other than those explicitly contained in the visualization. The authors of this study also put explicit time limits of 20 seconds on all tasks, negating the possibility that some encodings might encourage longer engagement.

Other critiques, while not adding experimentation, describe the perceived strengths and weaknesses of word clouds for other tasks. Viégas and Wattenberg assert that given word clouds' broad appeal, there must be something worthwhile about them as an encoding [22]. Word clouds are admonished in a well circulated blog post titled "Word Clouds Considered Harmful," but the criticism is more about poor journalistic practice than a commentary on gist-forming capabilities [13]. Finally, Meeks discusses the use of word clouds with topic models, citing in particular their compact representation [16]. However, he does not seek to make his justification empirically.

3. GENERAL EXPERIMENTAL DESIGN

Having a good "gist" of a topic is an abstract concept, making it difficult to quantitatively measure. It is important that the gist derived by a user is generalizable beyond just the specific words that they see in the representation. This is because given the size of most topics, any representation will necessarily display only a subset of the words that the topic contains. As mentioned in §2, generalizability is almost the opposite of the "recognition" goal used in [18], which penalized participants for recalling words that were similar to the words shown, but not actually present. It is also important that the user be able to form this general sense *quickly*, though this issue is more subtle. On one hand, being able to form an accurate idea at a quick glance is valuable, but so too might be a visualization that encourages longer linger time and engagement [14].

To evaluate the abstract task of gist-forming, we developed an experimental procedure that combines two concrete tasks, topic naming and word matching, with measures of participant confidence.

Topic naming.

For the topic naming task, participants are presented with a representation of a topic and asked to provide a name that captures the essence of the topic as closely as possible. This is meant to induce the process of gist-forming, as a user cannot create a properly descriptive name without first coming up with a cohesive idea of what the topic is about. This task was inspired by watching a group of collaborators in literature studies build an understanding of a model through the process of creating names for each topic. Given the subtlety of meaning that can be contained in a name, however, *correctness* of these names is often subjective and difficult to accurately measure.

Word matching.

Trying to evaluate topic names for some concept of correctness is difficult, given the wide variety of valid names that may fit. For this reason, we pair the naming task with a word matching task to evaluate the robustness of participants' gists, similar to the word intrusion task introduced in [6]. In it, we hold out a small number of high-ranking words from each topic that is presented. We then show participants these words, mixed with a roughly equal number of highly ranked words from unrelated topics, and for each word ask them to decide if it could have come from the represented topic. This allows us to compute objective accuracy measurements to assess how well a participant's concept of a topic can be generalized to other words that they might see in the context of that topic.

Confidence.

To measure how hard these tasks seemed for participants, as well as the amount of trust that they put into their own gists, we also had them report their confidence in their answers. These took the form of scores from 1 to 7 for both their confidence that their name fully captured the nature of the topic and their confidence in whether or not each word belonged (see Figure 1). It is worth noting that we are not always looking for complete confidence. Due to the probabilistic nature of topic models, there are times when users should be cautious in the conclusions they draw. This is especially true for lower quality topics.

In a pretesting phase, we asked for both user confidence as well as their opinion of the topic's quality. However, these two questions were strongly correlated enough to seem redundant, and we wanted to avoid having to explain to participants precisely what topic quality means, which is why we only asked for their confidence in these experiments.

In addition to measuring word matching accuracy and these two kinds of confidence (confidence in topic name and confidence in having matched the correct words), we also measured the time that participants spent with each topic representation. Though we gave them unlimited time to complete each question, we wanted to see whether or not linger time differed across factors or correlated with any of our other measures.

Topic contents and representations differed across experiments, as will be described in §4 and §5, but all used the same basic experimental setup. After giving consent, participants were shown a brief tutorial explaining the experimental task and the different encodings they might see. They were then presented with a succession of stimuli of the form shown in Figure 1 (in random order). On a single web page, they were asked to look at the topic representation that was shown, input a name for the topic and a number indicating their confidence in that name (from 1 to 7). They then were presented with a selection of words not contained in the

representation and asked if they seemed to belong with the topic. This binary choice ("yes" or "no") was accompanied with a confidence score (also from 1 to 7). After a participant had completed all of their allotted questions, they were asked to input demographic data and any comments.

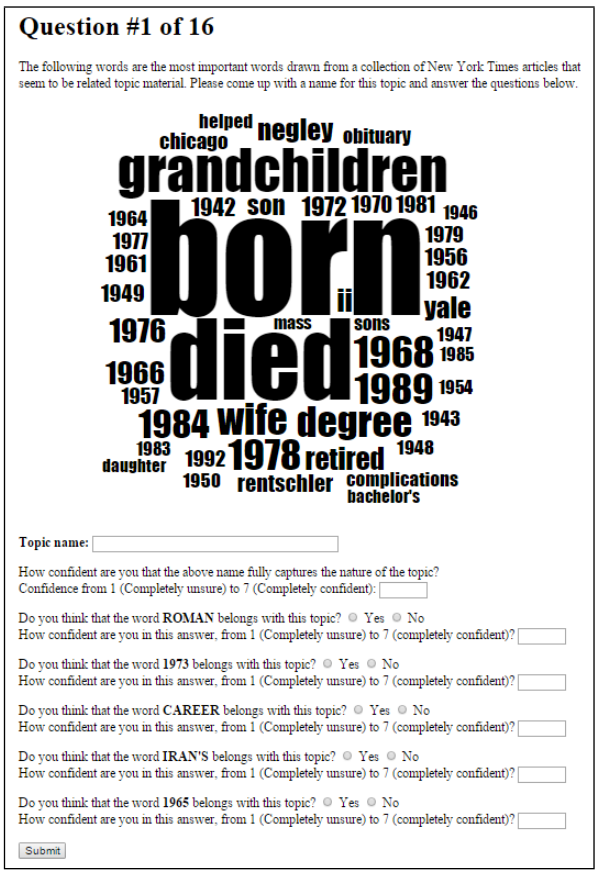

Figure 1: This is an example of a stimulus that might have been presented to a participant. This particular representation is in the word cloud category.

3.1 Stimuli

The topics we used for these experiments were drawn from a model built with 100 topics on a collection of New York Times articles from 2006 [19]. The model was built with the Latent Dirichlet Allocation (LDA) algorithm [3] as implemented in the Gensim Python library [17]. The topic representations were created with the D3 visualization library for Javascript [4] and Jason Davies' d3-cloud extension [11], employing an Archimedean spiral technique that places the largest words towards the center of the visualization.

3.2 Participants

Over the course of four experiments, we recruited 111 participants using Amazon's Mechanical Turk framework, specifically restricted to native English speakers residing in North America with at least a 95% approval rating. These participants ranged in age from 19 to 65 (with a mean of 33) and were made up of 64 males and 47 females. We paid participants $2.00 for their time.

4. COMPARING VISUAL ENCODINGS

The first factor impacting gist-forming that we evaluated was the designer's choice of visual encoding. In particular, we compared the two encodings most commonly used in practice: word lists and word clouds. Word lists are a subset of the words in a topic—the most frequent ones—displayed in descending order by frequency (see Figure 2 for examples of this technique). Word clouds are pictures displaying weighted lists of words, with weight—in this case, frequency within the topic—encoded using font-size (see Figure 1 for an example). We decided upon these two encodings both for their ubiquity in the literature and for their distinct difference in appearance. We felt this would make them most likely to expose differences in user performance.

Word clouds are an often polarizing visualization technique [13]. While many of the arguments against them are more about how they are used than the encoding itself, it has been empirically shown that word clouds are poor at helping users do tasks like recall or searching for a particular word [18]. However, we hypothesized that they may be particularly suited to the task of gist-forming. They fit a large amount of data (which is to say, many words) into a compact space [16]. With proper layouts and sizing, they can quickly direct the user's eye to the most important words of the visualization [15]. Their aesthetics may increase engagement and time spent with the visualization [21]. Finally, their popularity and ubiquity mean that most users are already equipped to interpret them.

Given these strengths, we hypothesized that:

- User accuracy would be at least as good when using word clouds as when using word lists.
- Users would take longer with word clouds (i.e., longer linger time).
- Users would *prefer* word clouds to word lists.

4.1 Experiment 1A: Good topics

For our stimuli, we hand-selected 16 topics from our New York Times model (see §3.1) that seemed to be highly cohesive, so as to avoid floor effects. In addition to our subjective impressions, we also confirmed that these topics scored highly on the Uniform Distribution ranking and Vacuous Semantic Distribution ranking proposed in [2]. Using a within-subjects design, we presented each participant with 16 stimuli—8 word clouds and 8 word lists—each containing the top 50 most frequent words from their respective topics. The order of the stimuli was randomized, as were which representations were paired with which topics.

We recruited 23 participants (13 male, 10 female) on Amazon's Mechanical Turk with ages ranging from 19 to 47 (with an average of 31).

4.1.1 Results

We ran a series of two-way analyses of variance (ANOVAs) to look for effects of representation and word ranking on the measures of accuracy, word confidence, name confidence, and time taken. We saw no effects of representation on accuracy ($F(1, 154) = 0.13$, $p = 0.72$), name confidence ($F(1, 22) = 1.23$, $p = 0.28$), or time taken ($F(1, 22) = 3.26$, $p = 0.08$). We did see a significant effect of representation on the user's confidence for their individual word decisions ($F(1, 154) = 5.62$, $p = 0.02$), but the effect size was small

"Good" topics			"Mediocre" topics		
air	bar	oil	island	images	report
force	beer	energy	long	years	department
side	drink	power	cat	work	agency
plane	ice	environmental	sound	photographs	officials
crash	cocktail	plant	animal	history	office
aircraft	made	gas	shore	early	investigation
safety	back	plants	bear	american	information
bags	glass	water	animals	ago	federal
pounds	coffee	fuel	deer	modern	government
vehicle	drinking	natural	people	native	general

Figure 2: Examples of "good" topics and "mediocre" topics from our model built on New York Times articles.

(a difference in means of .19 on an integer scale from 1 to 7—see Figure 4). Accuracy across the two representations was exceptionally close: when presented with a word cloud representation, participants correctly identified associated words at a rate of 0.866 as compared to a rate of 0.87 when presented with word lists. Despite resulting in such similar performance, nearly two thirds of the participants (14 of 22) expressed a preference for word clouds over word lists, generally citing reasons such as they were "easier to read."

We did see a significant effect of a word's ranking within the topic on both the participants' accuracy at correctly matching it ($F(3, 154) = 7.97$, $p < 0.0001$), as well as their confidence in said matches ($F(3, 154) = 48.17$, $p < 0.0001$). Accuracy and confidence went down the further down the topic's ranking a selected word was drawn from.

4.1.2 Discussion

It seems as though for topics as good as the ones selected for this experiment, the difference in visual representation is too small to matter. Though we hypothesized that word clouds would have at least as high performance as word lists, such consistency across all of our measures is surprising. These results suggest the question of whether participants are deriving the same interpretations across representations or if the topics are so good (having been selected for their coherence) that we are seeing an accuracy ceiling regardless of the representations' differences.

The effect of a word's ranking within a topic is encouraging to see, as it reinforces the use of such ranking schemes for creating topic representations (see Figure 6). Still, without having seen an explicit difference across conditions, more experiments were needed to ensure that the experimental mechanism was sufficiently sensitive to find differences when they exist.

4.2 Experiment 1B: Mediocre topics

To be sure that the similar performance we saw across visual encodings in the first experiment was not simply a factor of having picked the best topics possible, we ran a second experiment with a set of lower quality topics. While topics from the first experiment were selected to be as coherent as possible, these topics were selected to be "mediocre," in that they still seemed to show some level of semantic cohesion (i.e., they were not junk topics) but the connection between the words was harder to grasp. Though this was a subjective selection, they were also ensured to be topics that scored lower using objective topic rankings [2].

For this experiment, we used a within-subjects design identical to that in §4.1, with each participant seeing 8 word list stimuli and 8 word cloud stimuli, each containing a topic's top 50 words. However, we substituted the original 16 "good" topics for these new 16 "mediocre" topics.

We recruited 28 participants (19 male, 9 female) with ages ranging from 21 to 65 (with a mean of 34).

4.2.1 Results

We ran a series of two-way ANOVAs to look for effects of representation and word ranking on accuracy, confidence, and time taken. There was once again no significant effect of representation type on accuracy ($F(1, 189) = 0.09$, $p = 0.76$), name confidence ($F(1, 27) = 1.85$, $p = 0.19$), or time taken ($F(1, 27) = 3.98$, $p = 0.056$). There was also no significant effect on word matching confidence ($F(1, 189) = 0.12$, $p = 0.73$). However, the overall accuracy with the new topics was lower than that measured in the first experiment, with a mean of 0.694 as compared to 0.868. A word's ranking was a significant factor in participant's confidence in their word matching ($F(3, 189) = 30.53$, $p < 0.0001$), though not in their accuracy ($F(3, 189) = 1.58$, $p = 0.20$). Participants favored word clouds over word lists at an even higher rate than before (21 of 28).

4.2.2 Discussion

Where we might have expected the lower quality topics to expose differences between the representations, performance remained steady across the two conditions. This seems to indicate that users' gist-forming abilities are, in fact, resistant to this shift in visual representation. However, the overall change in accuracy indicates a dramatic difference between topics of good quality and topics of mediocre quality.

This difference becomes particularly apparent when we look at the data from the two experiments together. It is worth noting that this was a sequence of two experiments, and not a proper single between-subjects experiment, as we ran the two experiments on different days. As the participant pool of Mechanical Turkers can vary, this is a potential source of variance that is unaccounted for when making this comparison, although the experiments were run at similar times each day. Furthermore, our measurements for these experiments were consistent with successive experiments as described in §5.1 and §5.2.

We ran a series of two-way ANOVAs with the combined data to look for effects of representation and topic quality. We saw main effects of topic quality on accuracy ($F(1, 49) = 49.37$, $p < 0.0001$), word confidence ($F(1, 49) = 10.85$, $p = 0.002$), and name confidence ($F(1, 49) = 19.14$, $p < 0.0001$), each of which go down for topics of lower quality. There were no other main or interaction effects.

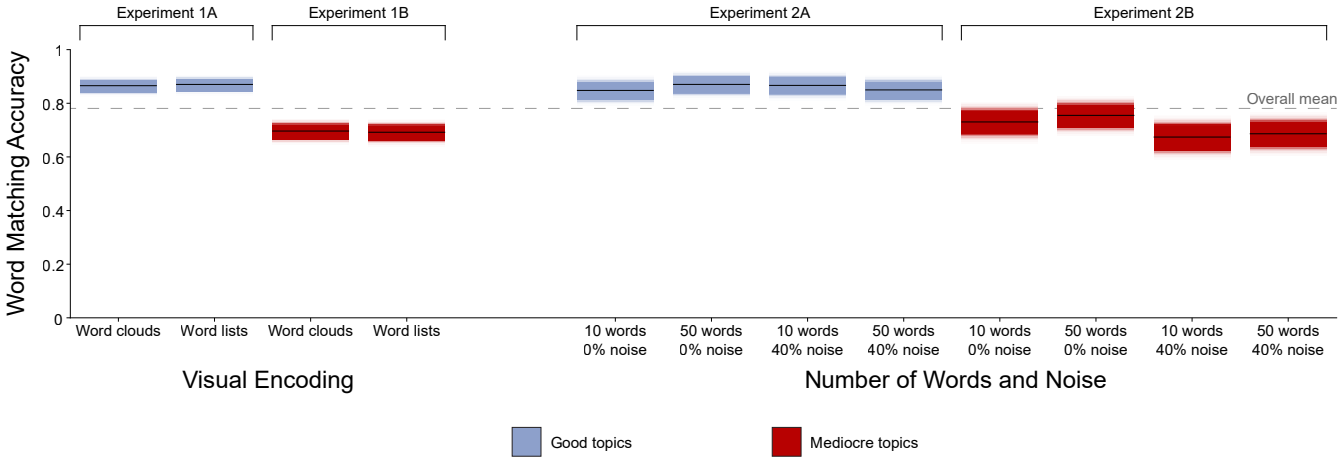

Figure 3: The effects of representation features on word matching accuracy, as gradient plots [9]. Fully opaque colored regions represent a 95% confidence interval and an alpha gradient extends to the 100% confidence interval. We saw no significant effects of visual encoding or number of words. There was no effect of noise with good quality topics, and only a small effect with lower quality topics However, with the data combined as described in §4.2.2 and §5.2.2, we did see significant differences between topics of good and mediocre quality.

5. NUMBER OF WORDS AND NOISE

In our next set of experiments, we evaluated two different factors for their effects on gist-forming. The first was the number of words that the user is shown from the topic. In practice, this number can range anywhere from two or three to hundreds. On one hand, seeing more words would seem to be beneficial, as it gives the user more data to form their gist. This seems especially important given our observation in §4 that lower ranked words are harder to match with their topic. On the other hand, more words could be overwhelming to the user.

In addition to number of words, we also examined the effect of injecting noise into the topic words. For this factor, we replaced words in a topic representation with "noise words" that did not appear in that topic. We hypothesized that this mechanism might be a way of simulating topics of poorer quality in a way that we could quantifiably measure as the percentage of each topic made up of noise. When choosing noise words for a particular topic, we selected words that had no significant probability in the topic's distribution, but did appear at the same position in a *different* topic in the model. This ensured that while the noise words shared no relation with the topic at hand, they did not stand out as completely obscure from the rest of the corpus.

For these factors, we hypothesized that:

- Providing more words would improve participant accuracy (though possibly hurt confidence).
- Introducing noise would decrease both participant accuracy and confidence.
- The improved accuracy with more words would be *more* pronounced with noisy topics.

5.1 Experiment 2A: Good topics

For the first experiment exploring these factors, we used the same high-quality topics as in §4.1. We created a within-subjects design to look for any main or interaction effects between the two factors. We used two levels for the number of words factor (10 words and 50 words) and two levels for the noise factor (0% noise and 40% noise). Each participant again saw 16 stimuli, 4 from each combination of levels. All stimuli used the word list encoding. We collected responses from 20 participants (11 male, 9 female) with ages ranging from 23 to 48 (with a mean of 33).

5.1.1 Results

We ran a series of two-way ANOVAs to look for effects of noise, number of words, and word ranking. The number of words factor exhibited no main effects on accuracy ($F(1, 284) = 0.63$, $p = 0.43$), word confidence ($F(1, 284) = 0.06$, $p = 0.81$), or name confidence ($F(1, 57) = 0.18$, $p = 0.68$). After excluding outliers that appear to have been instances of the participant leaving the computer for extended periods of time, participants did spend significantly longer on stimuli with 50 words ($F(1, 57) = 4.17$, $p = 0.04$), but this is to be expected given the longer time it would take to read.

While the presence of noise seemed to decrease participants' confidence in their topic names ($F(1, 57) = 45.56$, $p < 0.0001$), it had no discernible effect on either their accuracy ($F(1, 284) = 0.56$, $p = 0.45$) or their confidence when asked whether or not new words went with the topic ($F(1, 284) = 0.0001$, $p = 0.99$).

Once again, a new word's ranking within the topic had a significant effect on participant's accuracy ($F(3, 284) = 3.42$, $p = 0.02$) and confidence ($F(3, 284) = 21.85$, $p < 0.0001$) when matching it to its topic.

5.1.2 Discussion

We were surprised not to see an effect of the number of words included on either accuracy or confidence. The difference in magnitude from 10 to 50 words is drastic, with the latter group receiving five times as much information to work with as the former. The longer times spent on the questions with more words seem to indicate that participants were *looking at* the extra words, and yet the extra data offered no benefit for the word matching task.

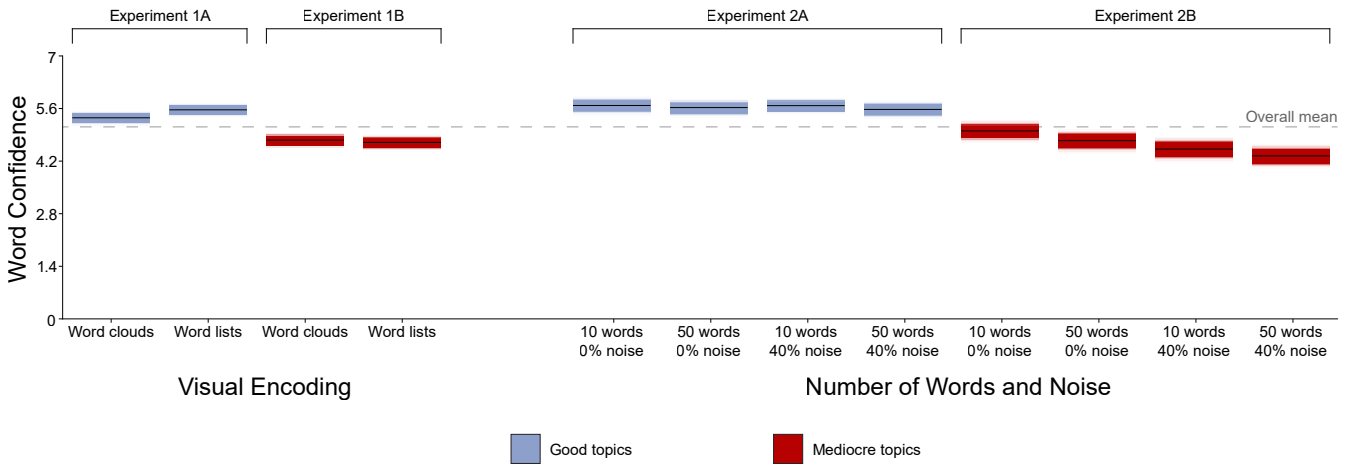

Figure 4: The effects of representation features on word matching confidence, as gradient plots [9]. Fully opaque colored regions represent a 95% confidence interval and an alpha gradient extends to the 100% confidence interval. We see significant effects of noise and extra words in Experiment 2B (see §5.2), as well as a significant effect of topic quality with combined data as described in §4.2.2 and §5.2.2.

The presence of noise words within the stimuli seemed to create a *perception* of difficulty without actually affecting performance in the word matching task. This is surprising, as we had expected introducing noise to be a way of artificially making the task harder, but participants appeared to be fully adept at *seeing through* the noise.

It is interesting to note that participants' overall accuracy (0.859) nearly matched that of the first visual encodings experiment that used the same "good" topics (0.868), further reinforcing the resistance of the gist-forming task to changes in presentation.

5.2 Experiment 2B: Mediocre topics

As in §4.2, our next step was to ensure that the consistency in accuracy we observed was not the result of ceiling effects associated with the high-quality topics. We ran an experiment using the same experimental design but switching out the high-quality topics for "mediocre" ones. This was again a within-subjects design, presenting each subject with 16 word list stimuli, 4 of each combination of noise levels (0% and 40%) and number of words (10 and 50). We recruited 38 participants (20 male, 18 female) with ages ranging from 23 to 60 (with a mean of 34).

5.2.1 *Results*

We ran a series of two-way ANOVAs to look for effects of number of words, noise, and word ranking. Once again, the number of words factor showed no significant effect on accuracy ($F(1, 238) = 2.18$, $p = 0.14$) or participants' confidence in their word matches ($F(1, 238) = 1.48$, $p = 0.23$). We did see a significant (though small) effect indicating that participants' confidence in their names ($F(1, 48) = 6.32$, $p = 0.02$) dropped in the 50-words condition (see Figure 5).

Introducing noise to the stimuli once again lowered participants' confidence in their names ($F(1, 48) = 32.94$, $p < 0.0001$) and in their word matches ($F(1, 238) = 30.45$, $p < 0.0001$).We also saw an effect of noise on accuracy that was not present with good topics, in which accuracy was slightly lower in the noisy case ($F(1, 238) = 4.68$, $p = 0.03$). However, the size of this effect was small ($M_{0\%} = 0.71$, $SD_{0\%} = 0.22$, $M_{40\%} = 0.65$, $SD_{40\%} = 0.23$).

No interaction effects between noise and number of words were observed. There were no effects to be observed on time taken to answer each question.

As in the previous experiments, a word's ranking within the topic had a significant effect on the participant's ability to match it to the representation ($F(3, 238) = 6.44$, $p = 0.0003$) and their confidence in said match ($F(3, 238) = 8.59$, $p < 0.0001$).

5.2.2 *Discussion*

Once again, our hypothesis for improved performance with more words was not substantiated. Confidence with more words actually went down—possibly indicating that participants were overwhelmed by the extra information (see Figures 4 and 5). Noise once again introduced uncertainty in the participants' responses without negatively affecting their word matching accuracy.

The overall accuracy for the word matching task was 0.711, down from 0.859 in the first experiment looking at these factors. Looking at the combined data from these two experiments reinforces this trend (though it must be done with the same caveats as described in 4.2.2). With the combined data, we ran a series of two-way ANOVAs looking for the effects of topic quality with number of words and noise. Upon doing this, we are able to see a main effect of topic quality on accuracy ($F(1, 35) = 86.58$, $p < 0.0001$), word confidence ($F(1, 35) = 16.14$, $p = 0.0003$), and name confidence ($F(1, 35) = 16.91$, $p = 0.0002$), each of which go down with the worse topics.

In these experiments, we see again that topic quality has a dramatic effect on both accuracy and confidence. We also find that noise turns out not to be a good way of simulating poor topics. Introducing noise to good topics did not result in a decrease in accuracy, while replacing good topics with mediocre topics resulted in a substantial decrease in accuracy. However, it is still very interesting that participants' gists were able to withstand that level of manipulation.

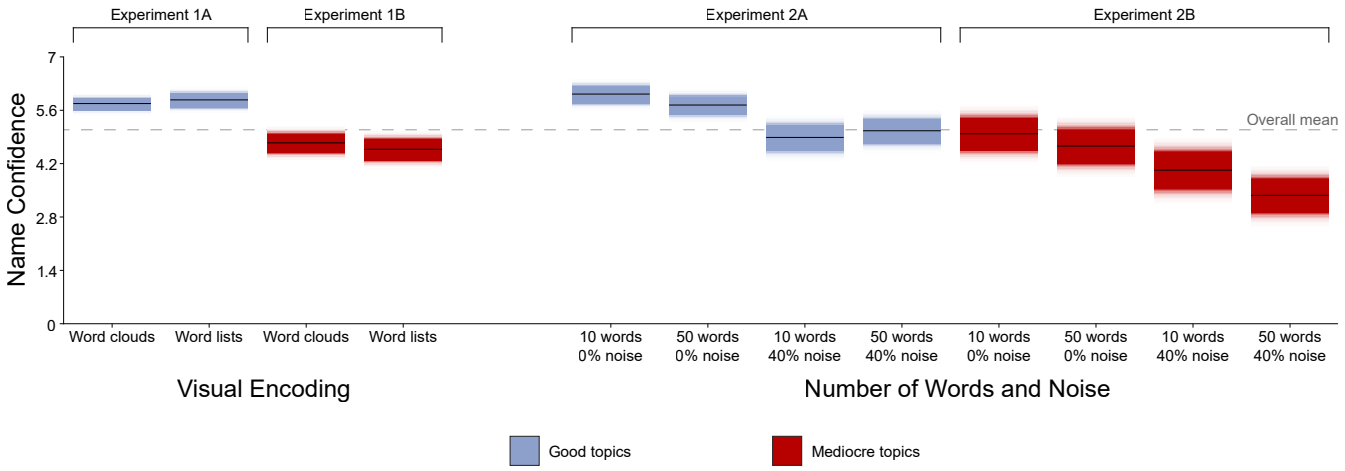

Figure 5: The effects of representation features on topic name confidence, as gradient plots [9]. Fully opaque colored regions represent a 95% confidence interval and an alpha gradient extends to the 100% confidence interval. While there was no effect of visual encoding, noise resulted in significantly lower confidence, as did more words with mediocre topics. After combining the data from experiments together as described in §4.2.2 and §5.2.2, topic quality showed a significant effect on confidence, as well.

6. DISCUSSION

We are able to offer several takeaways from this exploration of gist-forming that are relevant to designers of topic modeling tools and visualizations. Counter to our expectations, gist-forming seems to be quite robust to changes in the visual encoding used to convey topics. The robustness of user performance across encoding, combined with participants' preference for word clouds over word lists, may indicate that word clouds are suitable to use for topic interpretation tasks, despite their documented poor performance in helping users with other tasks such as search and recall. It is possible that other encodings (e.g., bar charts) may differ in ways not captured by the pairwise experiments described here, but we believe word clouds and word lists are representative of the literature.

Gist-forming also appears to be resistant to changes in the number of words shown, which did not affect accuracy in either §5.1 or §5.2 and showed only a minor effect on confidence in §5.2. The drop in confidence seen with more words on worse topics may suggest that number of words can be used by designers as a method of tempering the tendency in some users to make overly broad generalizations about what topic trends may mean.

Similarly, the drop in user confidence for both mediocre and noisy topics is beneficial for the task of topic evaluation. These lower quality topics are instances when one would *want* user confidence to go down—for users to form their interpretations with a grain of salt rather than making sweeping claims based on tenuous connections. As users seem to be able to differentiate between topics of different quality, designers may be able to leave the task of topic evaluation largely in their hands.

Finally, it is clear that the factor that has the greatest effect on the gist-forming task is topic quality. While creating good visualizations can help users achieve many new insights, this finding reinforces the need to help them arrive at good models. Tools that incorporate users into the *training* process are crucial to this effort.

7. CONCLUSION

In this paper, we have described our process of using the concrete tasks of topic naming and word matching to assess the abstract task of gist-forming. By measuring accuracy and user confidence, we are able to show that the gist-forming process is remarkably resistant to changes in visual encoding and number of words, but not to dips in topic quality. There is still much more to be learned about the gist-forming process. As future work, we are interested in looking at comparisons of other static topic encodings, as well as more interactive forms of topic conveyance like Termite [7] and parallel tag clouds [8]. We believe gist-forming is important to understand, as the user's understanding of a topic's contents is the foundation upon which their semantic claims must be made.

8. ACKNOWLEDGMENTS

This work was supported in part by NSF award IIS-1162037 and a grant from the Andrew W. Mellon Foundation.

9. REFERENCES

[1] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 173–182. IEEE, 2014.

[2] L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer, 2009.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE TVCG*, 2011.

[5] A. Chaney and D. Blei. Visualizing topic models. In *Proc. AAAI on Weblogs and Social Media*, 2012.

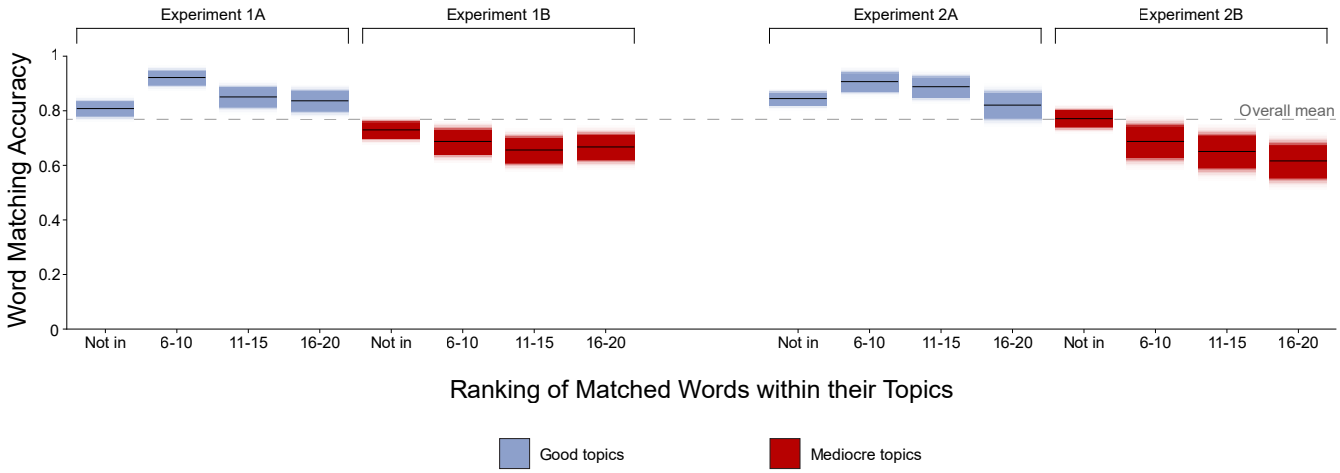

Figure 6: The effects of word ranking on accuracy, as gradient plots [9]. Fully opaque colored regions represent a 95% confidence interval and an alpha gradient extends to the 100% confidence interval. When selecting words to be matched with the topic representation, we drew from three different sections of the topic rankings: the 6-10 ranked words, the 11-15 ranked words, and the 16-20 ranked words. Here, we plot participant accuracy by these word groups (along with words that were drawn from outside of the topic).

[6] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[7] J. Chuang, C. Manning, and J. Heer. Termite: visualization techniques for assessing textual topic models. In *Proc. Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[8] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of the IEEE Symp. on Visual Analytics Science and Technology (VAST)*, 2009.

[9] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, dec 2014. IEEE Vis Conference, InfoVis track, to appear.

[10] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE TVCG*, 17(12):2412–2421, 2011.

[11] J. Davies. d3-cloud. https://github.com/jasondavies/d3-cloud, 2015.

[12] M. J. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314. ACM, 2007.

[13] J. Harris. Word clouds considered harmful, blog, http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/, 2011.

[14] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 160–160. IEEE, 2008.

[15] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-Computer Interaction–INTERACT 2009*, pages 392–404. Springer, 2009.

[16] E. Meeks. Using word clouds for topic modeling results, blog, https://dhs.stanford.edu/algorithmic-literacy/using-word-clouds-for-topic-modeling-results/, 2012.

[17] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[18] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 995–998. ACM, 2007.

[19] E. Sandhaus. The New York Times Annotated Corpus LDC2008T19. DVD. Philadelphia: Linguistic Data Consortium, 2008.

[20] A. J. Torget, R. Mihalcea, J. Christensen, and G. McGhee. Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers. 2011.

[21] T. van der Geest and R. van Dongelen. What is beautiful is useful-visual appeal and expected information quality. In *Professional Communication Conference, 2009. IPCC 2009. IEEE International*, pages 1–5. IEEE, 2009.

[22] F. B. Viégas and M. Wattenberg. Timelines tag clouds and the case for vernacular visualization. *interactions*, 15(4):49–52, 2008.

[23] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proc. ACM Knowledge discovery and data mining*, pages 153–162. ACM, 2010.