SUPPLEMENTAL MATERIALS

VizCommender: Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations

Michael Oppermann, Robert Kincaid, and Tamara Munzner

1. Additional Materials

`stopwords.txt`	List of stop-words that are applied to the whole corpus.
`stopwords_data_fields.txt`	List of stop-words that are used to filter data fields generated by Tableau.
`triplets.csv`	Triplets: • Viz sheet specification IDs (`reference_id`, `alternative_1_id`, `alternative_2_id`) • Number of participant votes (`alternative_1_votes`, `alternative_2_votes`) • `batch_id` (triplets are divided into 3 batches of 45 triplets)
`viz_extracts.json`	JSON file with viz sheet specifications.

2. Viz-to-Viz Workbook Recommendation

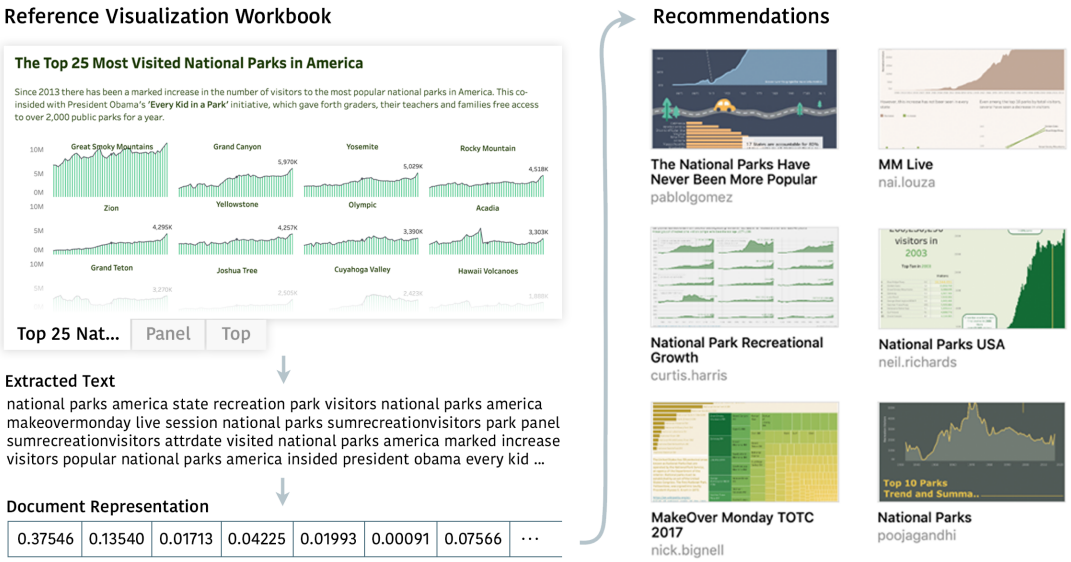


Fig. S1. Example Recommendation. Text gets extracted from a reference workbook (viz workbook specification) and an NLP model (e.g., LDA) is applied to transform the bag-of-words representation into a numeric document representation. A distance metric (e.g., Jensen-Shannon divergence) is used to compute pairwise similarity scores to other workbooks. Top results within a specified score range are selected as recommendations.

3. Visual Similarity vs. Data/Topic Similarity

After several experiments with our VizCommender interface and extensive discussions with collaborators, we decided to exclude visual encoding specifications when computing similarity between visualization workbooks. We argue that visual encoding features can add noise to the model when the task is information seeking in VizRepos.

Figure S2 shows simplified examples to illustrate some challenges with similarity perception. The pie charts in Figure S2a are perceived visually similar although they use different data and address distinct topics. In contrast, identical data may be visualized fundamentally differently, as shown in Figure S2.

We conclude that visual similarity is not an important factor for finding topically related workbooks but incorporating visual encodings to increase the diversity of recommendations and to improve the detection of duplicates would be an interesting direction for future work.

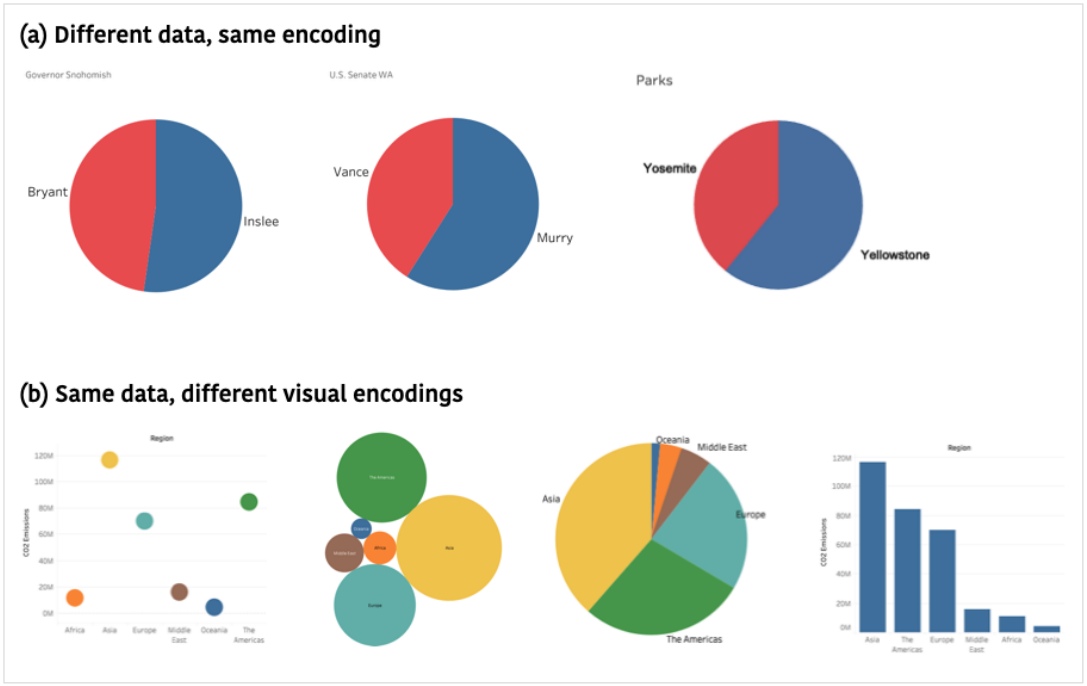


Fig. S2. Simple examples illustrating that visual similarity is typically topic independent.

4. Model Investigation

In addition to the crowdsourced study, we informally investigated each NLP model based on the Tableau Public VizRepo (sample of 18,820 workbooks) and internal corporate VizRepo (3,424 workbooks).

After computing a numeric representation of each workbook with the four models and varying hyper-parameters, our first step was to use UMAP dimensionality reduction to create two-dimensional embeddings. A visual inspection of the scatterplot projections showed that very similar workbooks are indeed grouped together, but higher level semantic clusters were not obvious and it was not possible to identify whether any of the methods performed particularly well or poorly.

For TF-IDF, we compared the representative keywords of each workbook, and for LDA and LSI, we analyzed the most important keywords describing individual topics. However, the generated topics were not always interpretable and meaningful because documents were often assigned to similar topics although we found them to be clearly distinct; this problem is a well-known issue in NLP [https://dl.acm.org/doi/10.5555/2984093.2984126, http://dx.doi.org/10.3115/v1/W14-3110].

4.1. UMAP Projections of Document Representations

4.1.1. Tableau Public VizRepo

The following UMAP projections are based on a corpus of 15,482 viz specifications (workbooks).

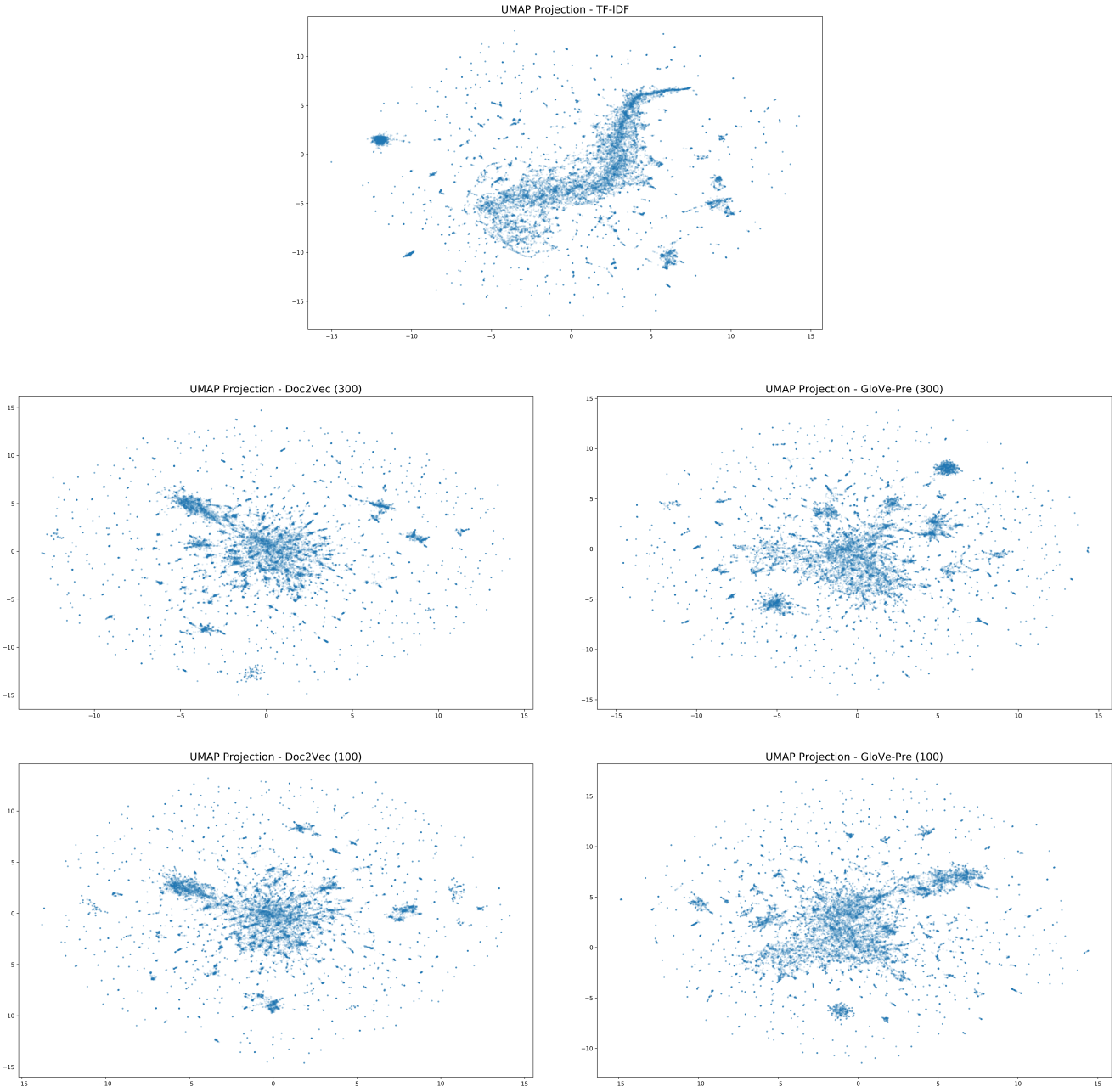


Fig. S3. UMAP applied to **TF-IDF** and **Doc2Vec** (100 and 300 dimensions) and **GloVe** (100 and 300 dimensions) document vectors.

Fig. S4. UMAP applied to **LSI** (15, 30, 75, and 150 dimensions) and **LDA** (15, 30, 75, and 150 topics) document vectors.

4.1.2. Tableau Corporate VizRepo

The following UMAP projections are based on a corpus of 15,482 viz specifications (workbooks).

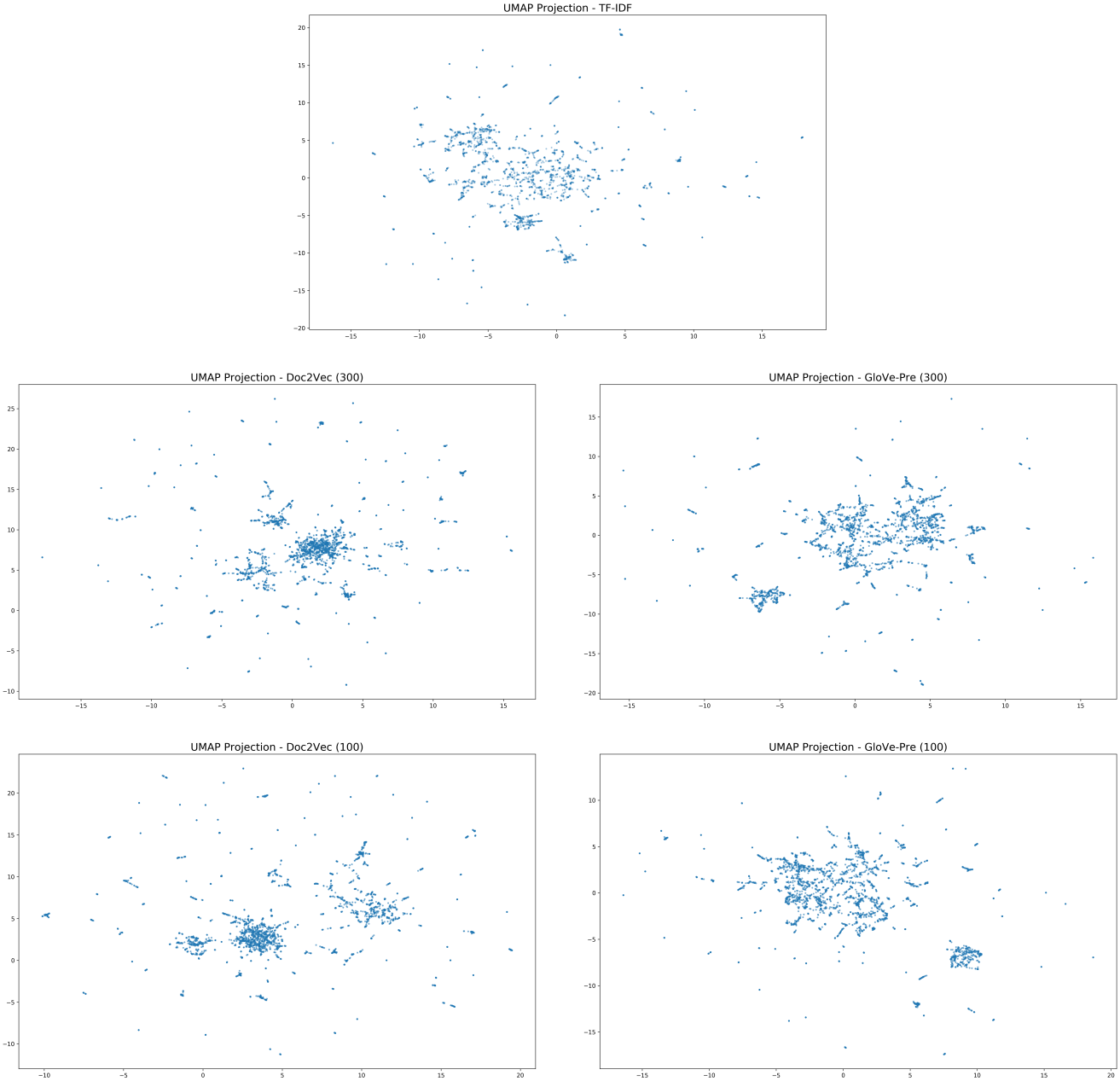


Fig. S5. UMAP applied to **TF-IDF** and **Doc2Vec** (100 and 300 dimensions) and **GloVe** (100 and 300 dimensions) document vectors.

Fig. S6. UMAP applied to **LSI** (15, 30, 75, and 150 dimensions) and **LDA** (15, 30, 75, and 150 topics) document vectors.

4.2. Keyword Probing

We use an approach that we call *keyword probing* to analyze the dominant topic of workbooks to better understand if document representations capture underlying topics.

For instance, we investigate workbooks from the Tableau Public sample VizRepo that contain the term "superstore" (220 out of 18,820 workbooks). Since we know that those workbooks are highly related, if they are assigned to a broad range of topics, it is an indicator that the model does not capture the content well enough.

The results for the *superstore* example are shown in Figure S7. We try multiple NLP models with 15, 30, 75, and 150 topics. LDA is based on topic distributions and we can choose the topic with the highest probability as the dominant topic. For LSI (150 dim.), Doc2Vec (100 dim.), and GloVe (pre-trained; 100 dim.), we first compute the workbook vectors and then apply k-means clustering to identify the most dominant topic. LDA with 15 and 30 topics perform best because nearly all *superstore* workbooks are assigned to the same topic. With a larger number of topics, and across all models, those workbooks are distributed over many topics/clusters.

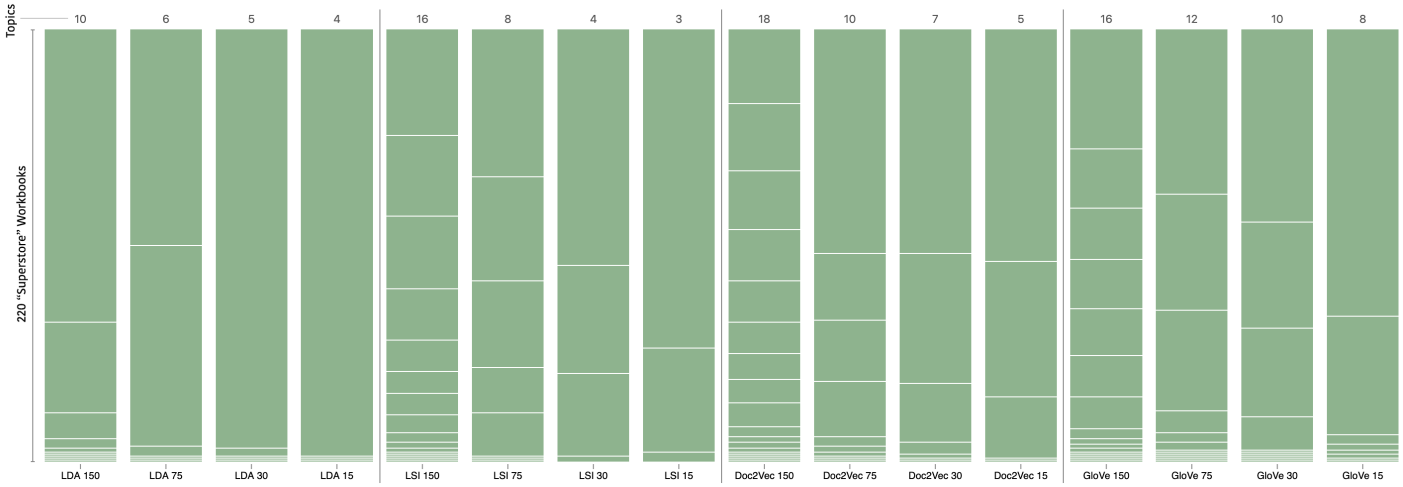


Fig. S7. Model comparison via keyword probing. Distribution of *superstore* workbooks into different topics/clusters. Each green block represents one topic and the height indicates how many *superstore* workbooks are assigned.

5. Crowdsourced Study

5.1. Experiment Interface

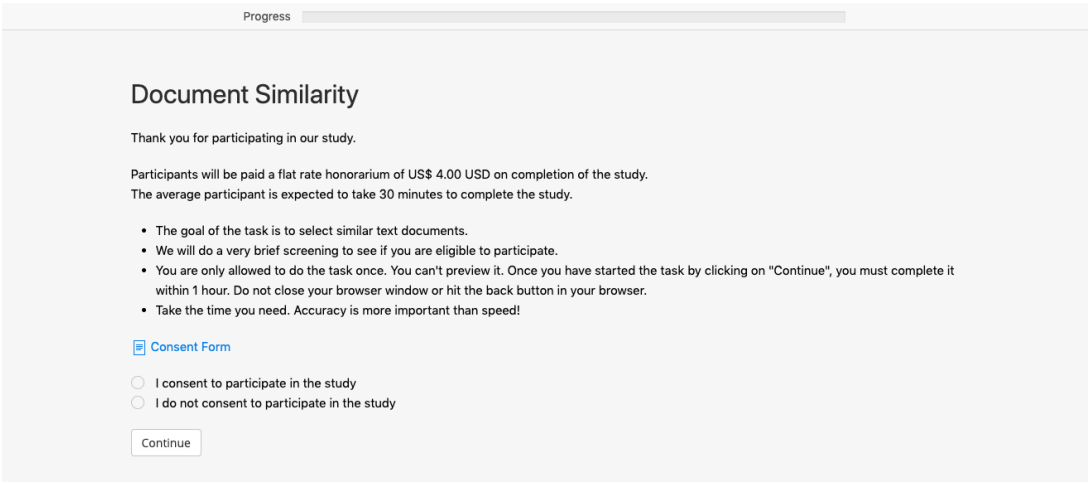


Fig. S8. Obtaining informed consent

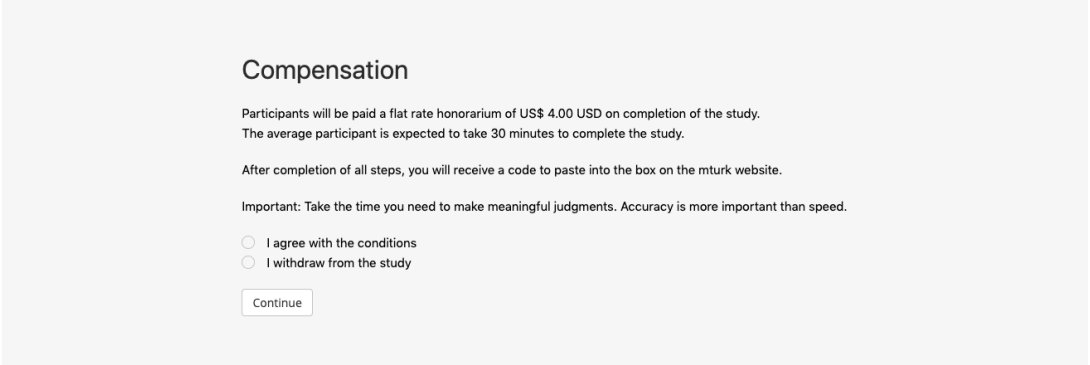


Fig. S9. Additional note that all steps not be completed.

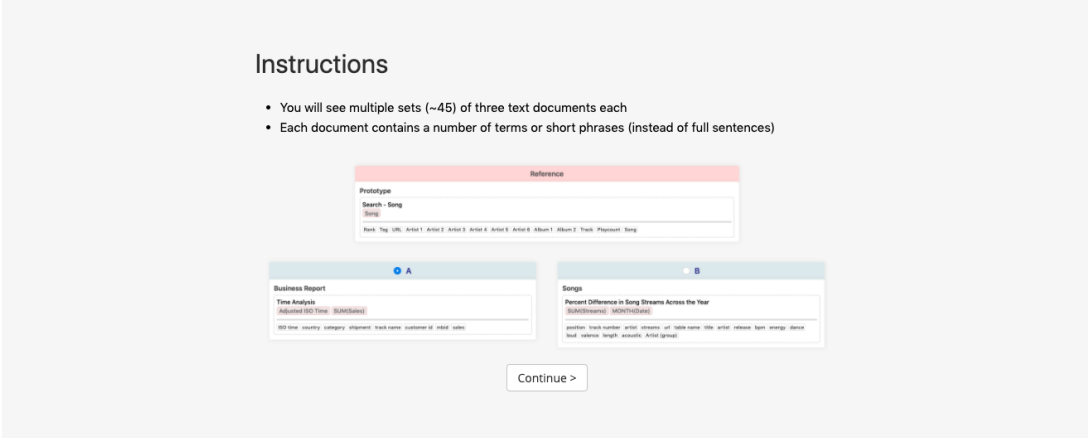


Fig. S10. Interface instructions.

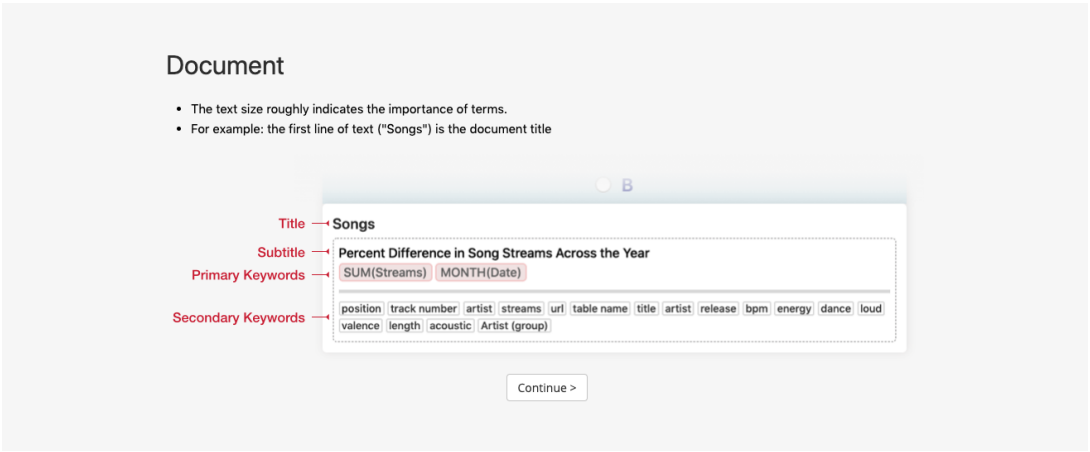


Fig. S11. Description of text document structure

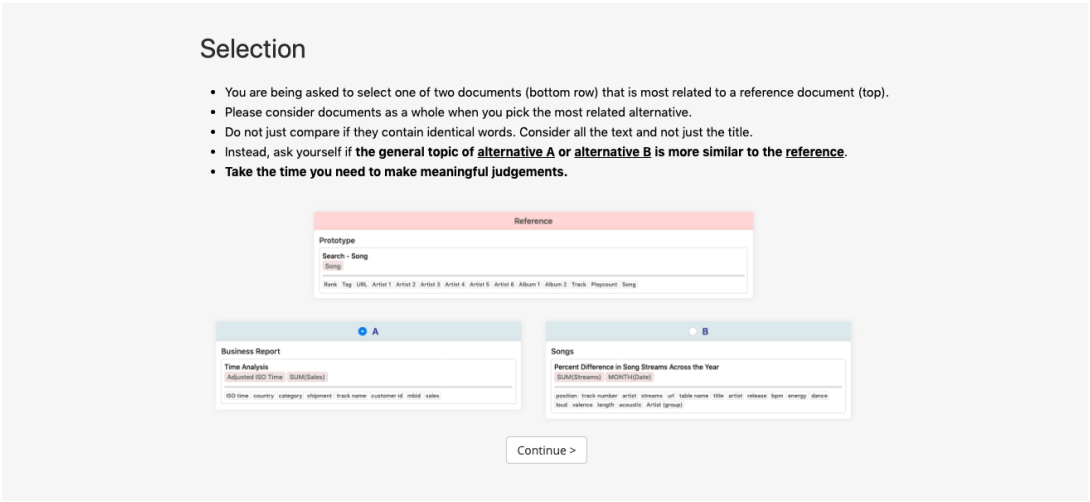


Fig. S12. Task description.

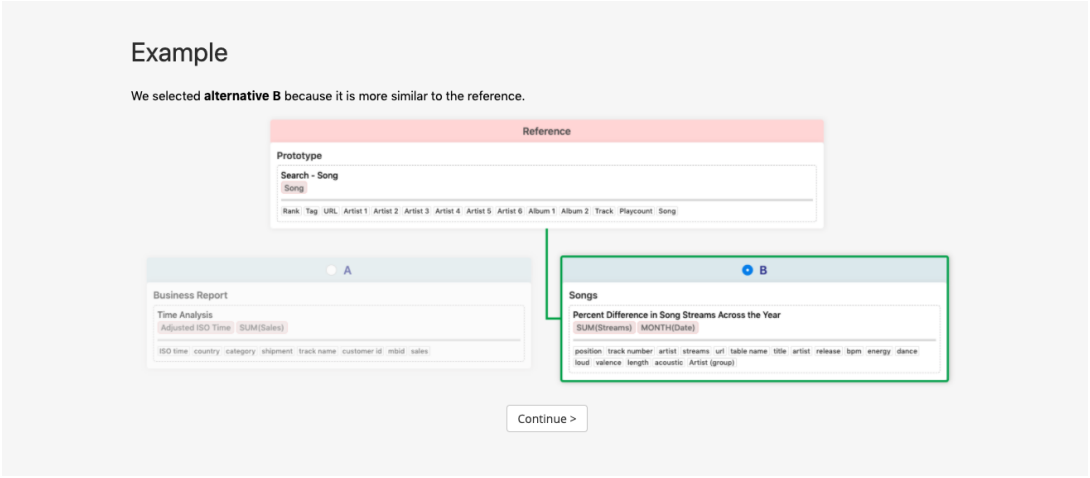


Fig. S13. Example selection.

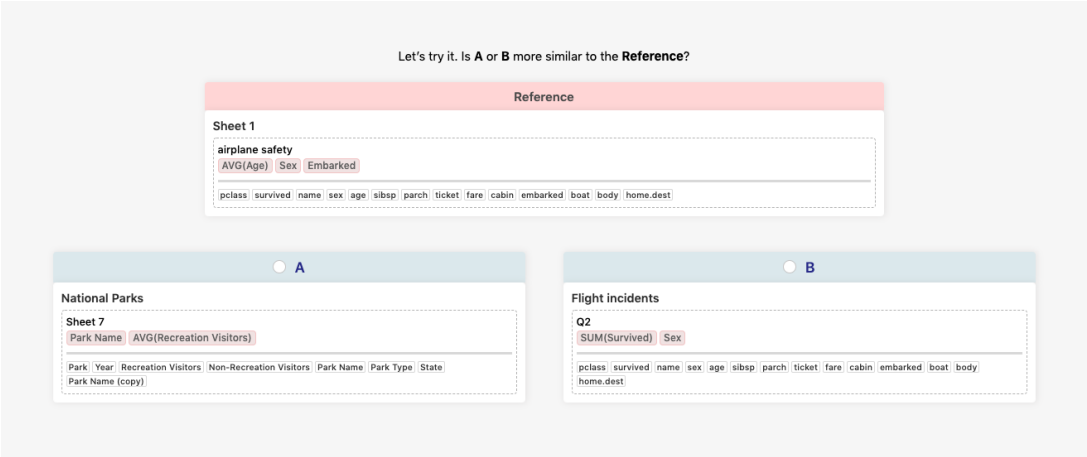


Fig. S14. Screening test.

5.2. Demographics

Participants were asked to fill out a post-experiment questionnaire which asked about their gender, age, educational background, and their experience during the study.

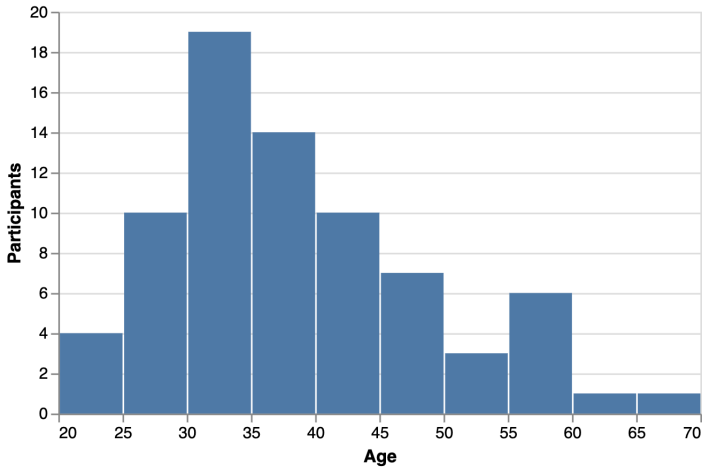


Fig. S15. Age distribution. Mean=38; Median=36

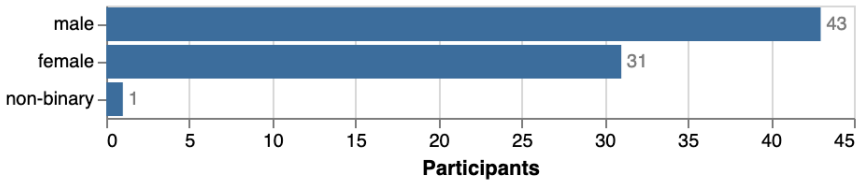


Fig. S16. Gender distribution.

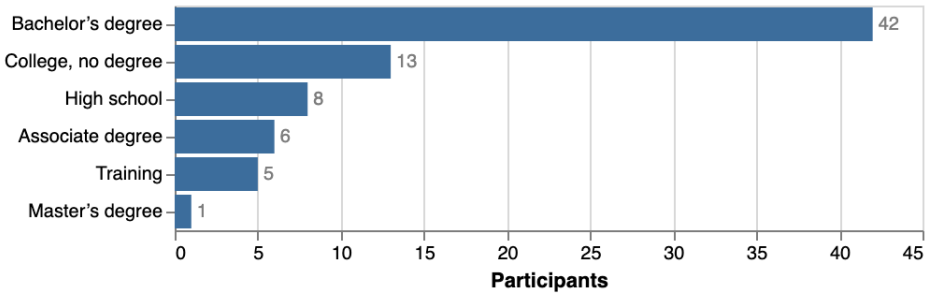


Fig. S17. Educational background.

5.3. Judging Similarity

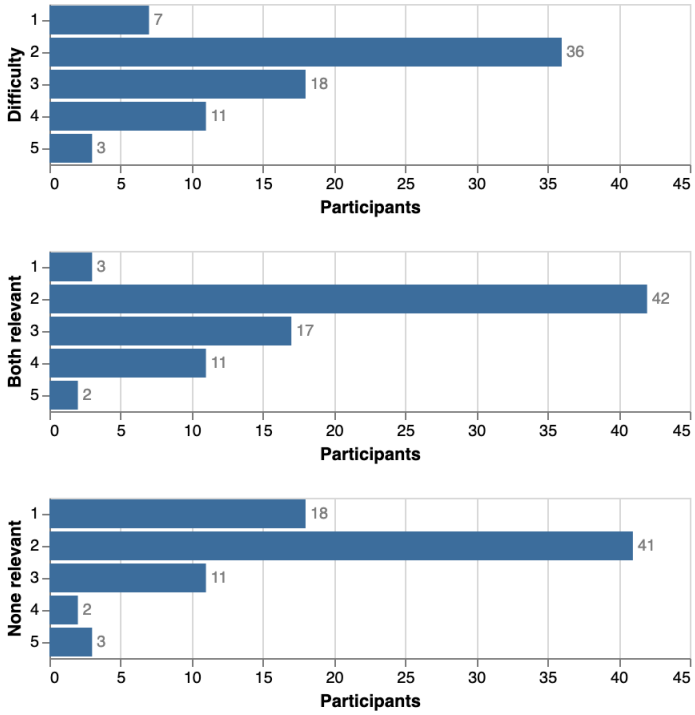
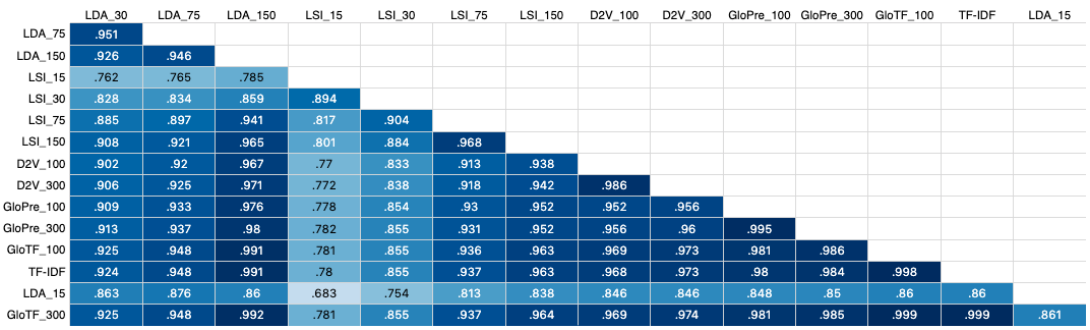


Fig. S18. Experience ratings on 5-point Likert scale. Task difficulty from 1 (=very easy) to 5 (=very difficult). Number of times participants thought both alternatives were equally related to the reference (1=Never to 5=Very often). Number of times participants thought none of the alternatives were related to the reference (1=Never to 5=Very often).

5.4. Agreement Between Model Predictions and Human Judgements

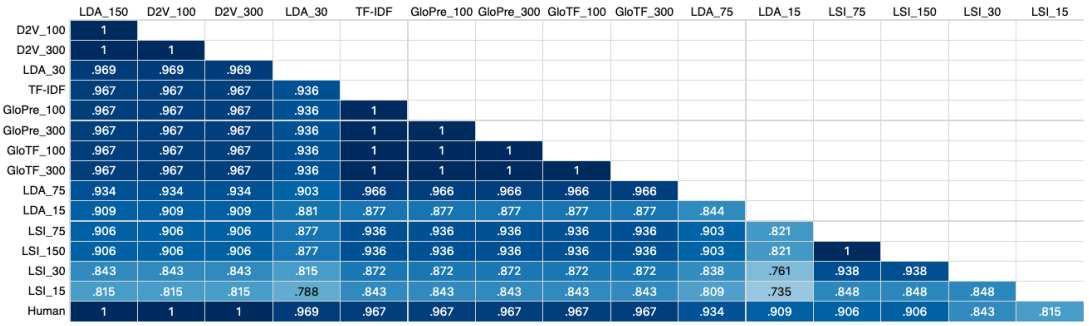
We use Fleiss' Kappa κ to quantify the chance-corrected agreement between models and human judgements.

Model abbreviations: *D2V*= newly trained Doc2Vec model; *GloPre* = Pre-trained GloVe model; *GloTF* = Pre-trained GloVe model that we further trained on viz specifications. Numbers next to the model abbreviations indicate the number of topics/dimensions.



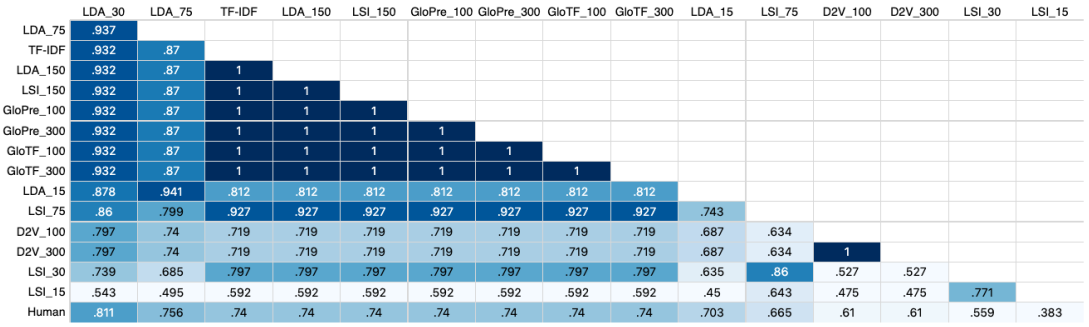
	LDA_30	LDA_75	LDA_150	LSI_15	LSI_30	LSI_75	LSI_150	D2V_100	D2V_300	GloPre_100	GloPre_300	GloTF_100	TF-IDF	LDA_15
LDA_75	.951													
LDA_150	.926	.946												
LSI_15	.762	.765	.785											
LSI_30	.828	.834	.859	.894										
LSI_75	.885	.897	.941	.817	.904									
LSI_150	.908	.921	.965	.801	.884	.968								
D2V_100	.902	.92	.967	.77	.833	.913	.938							
D2V_300	.906	.925	.971	.772	.838	.918	.942	.986						
GloPre_100	.909	.933	.976	.778	.854	.93	.952	.952	.956					
GloPre_300	.913	.937	.98	.782	.855	.931	.952	.956	.96	.995				
GloTF_100	.925	.948	.991	.781	.855	.936	.963	.969	.973	.981	.986			
TF-IDF	.924	.948	.991	.78	.855	.937	.963	.968	.973	.98	.984	.998		
LDA_15	.863	.876	.86	.683	.754	.813	.838	.846	.846	.848	.85	.86	.86	
GloTF_300	.925	.948	.992	.781	.855	.937	.964	.969	.974	.981	.985	.999	.999	.861

Fig. S19. Agreement between model predictions and human judgements for 135 triplets.



	LDA_150	D2V_100	D2V_300	LDA_30	TF-IDF	GloPre_100	GloPre_300	GloTF_100	GloTF_300	LDA_75	LDA_15	LSI_75	LSI_150	LSI_30	LSI_15
D2V_100	1														
D2V_300	1	1													
LDA_30	.969	.969	.969												
TF-IDF	.967	.967	.967	.936											
GloPre_100	.967	.967	.967	.936	1										
GloPre_300	.967	.967	.967	.936	1	1									
GloTF_100	.967	.967	.967	.936	1	1	1								
GloTF_300	.967	.967	.967	.936	1	1	1	1							
LDA_75	.934	.934	.934	.903	.966	.966	.966	.966	.966						
LDA_15	.909	.909	.909	.881	.877	.877	.877	.877	.877	.844					
LSI_75	.906	.906	.906	.877	.936	.936	.936	.936	.936	.903	.821				
LSI_150	.906	.906	.906	.877	.936	.936	.936	.936	.936	.903	.821	1			
LSI_30	.843	.843	.843	.815	.872	.872	.872	.872	.872	.838	.761	.938	.938		
LSI_15	.815	.815	.815	.788	.843	.843	.843	.843	.843	.809	.735	.848	.848	.848	
Human	1	1	1	.969	.967	.967	.967	.967	.967	.934	.909	.906	.906	.843	.815

Fig. S20. Agreement between model predictions and human judgements for 92 triplets with higher consensus (at least 80% agreement between all participants and agreement between the majority vote and the expert's gold standard)



	LDA_30	LDA_75	TF-IDF	LDA_150	LSI_150	GloPre_100	GloPre_300	GloTF_100	GloTF_300	LDA_15	LSI_75	D2V_100	D2V_300	LSI_30	LSI_15
LDA_75	.937														
TF-IDF	.932	.87													
LDA_150	.932	.87	1												
LSI_150	.932	.87	1	1											
GloPre_100	.932	.87	1	1	1										
GloPre_300	.932	.87	1	1	1	1									
GloTF_100	.932	.87	1	1	1	1	1								
GloTF_300	.932	.87	1	1	1	1	1	1							
LDA_15	.878	.941	.812	.812	.812	.812	.812	.812	.812						
LSI_75	.86	.799	.927	.927	.927	.927	.927	.927	.927	.743					
D2V_100	.797	.74	.719	.719	.719	.719	.719	.719	.719	.687	.634				
D2V_300	.797	.74	.719	.719	.719	.719	.719	.719	.719	.687	.634	1			
LSI_30	.739	.685	.797	.797	.797	.797	.797	.797	.797	.635	.86	.527	.527		
LSI_15	.543	.495	.592	.592	.592	.592	.592	.592	.592	.45	.643	.475	.475	.771	
Human	.811	.756	.74	.74	.74	.74	.74	.74	.74	.703	.665	.61	.61	.559	.383

Fig. S21. Agreement between model predictions and human judgements for 43 triplets with lower consensus (less than 80% agreement between participants or disagreement between the majority vote and the expert's gold standard.)

	LDA_30	LDA_75	LDA_150	LSI_15	LSI_30	LSI_75	LSI_150	D2V_100	D2V_300	GloPre_100	GloPre_300	GloTF_100	TF-IDF	LDA_15
LDA_75	.951													
LDA_150	.926	.946												
LSI_15	.762	.765	.785											
LSI_30	.828	.834	.859	.894										
LSI_75	.885	.897	.941	.817	.904									
LSI_150	.908	.921	.965	.801	.884	.968								
D2V_100	.902	.92	.967	.77	.833	.913	.938							
D2V_300	.906	.925	.971	.772	.838	.918	.942	.986						
GloPre_100	.909	.933	.976	.778	.854	.93	.952	.952	.956					
GloPre_300	.913	.937	.98	.782	.855	.931	.952	.956	.96	.995				
GloTF_100	.925	.948	.991	.781	.855	.936	.963	.969	.973	.981	.986			
TF-IDF	.924	.948	.991	.78	.855	.937	.963	.968	.973	.98	.984	.998		
LDA_15	.863	.876	.86	.683	.754	.813	.838	.846	.846	.848	.85	.86	.86	
GloTF_300	.925	.948	.992	.781	.855	.937	.964	.969	.974	.981	.985	.999	.999	.861

Fig. S22. Within-model agreement for all 4,211 generated triplets.

5.5. Confidence of Model Decisions

For the triplet judgement task, models produce predictions by independently computing the pairwise similarity scores between the reference document and one of the alternatives. The model compares the two scores and the alternative with the higher score is deemed to be more similar.

The delta between two scores is an indicator for the confidence of model predictions. Are both scores very close, and in fact, both alternatives are relevant or is there a significant difference?

We normalized the pairwise similarity scores for each model and compared the difference.

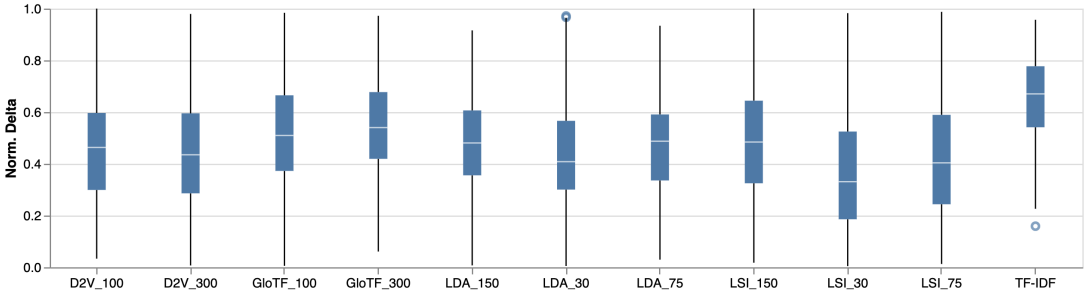


Fig. S23. Model comparison based on normalized similarity score deltas. The median difference between pairs of similarity scores is slightly higher for TF-IDF. All other models are in a similar range.

6. Study Analysis Tool

We implemented an interactive visual analysis tool to both clean the crowdsourced data and to better understand the results and the implications for the model selection.

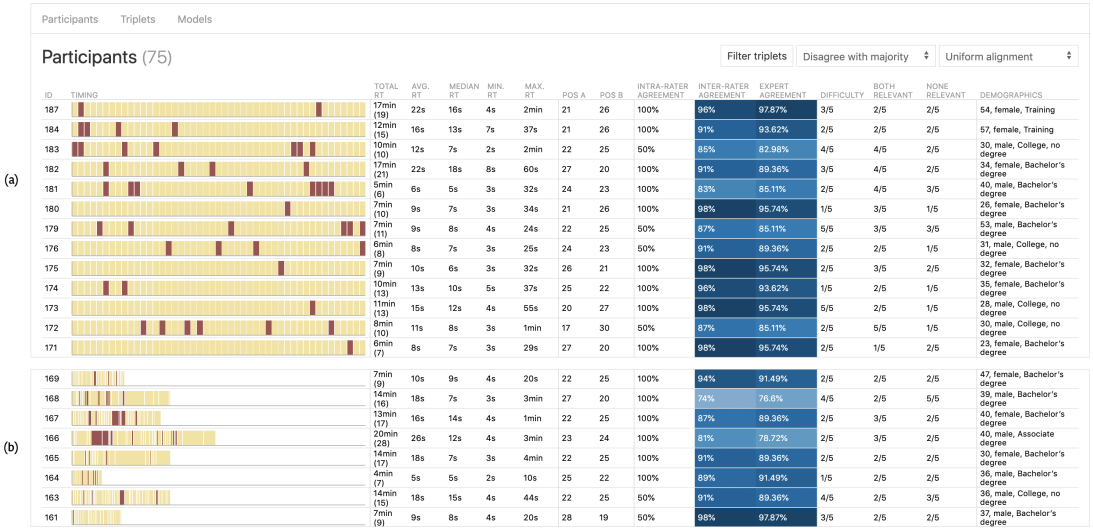


Fig. S24. Our custom analysis tool provides an overview of all MTurk participants. The responses of each participant are displayed along a timeline, (a) either laid out uniformly or (b) by reaction time, and color-coded based on the agreement with the gold standard or majority vote.

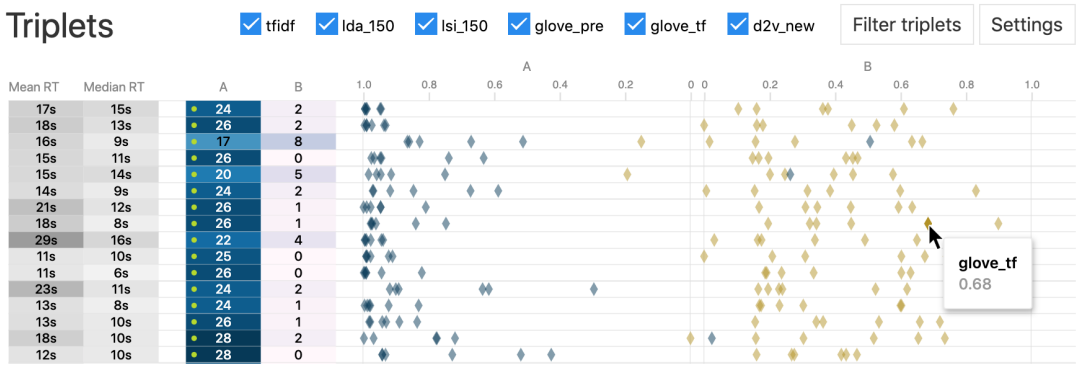


Fig. S25. A separate view in the analysis tool allows us to compare model predictions with human judgements for individual triplets. Each row corresponds to one triplet (out of 135). The first two columns show the reaction time followed by two columns that summarize the number of human votes that each alternative received. The yellow circle inside the blue boxes indicates which alternative the expert annotator has picked. Model predictions are shown using diamond symbols. For each triplet, models produce one similarity score for alternative A and one score for alternative B. We have pre-selected 6 models which leads to 12 symbols per triplet. Similarity scores are between 0-1 and the values are encoded through the x-position. The alternative with the higher score is considered more similar and shown in blue; otherwise in gold. For these example triplets, we can see that nearly all models agree with the human majority that alternative A is more similar to the reference. There are only three triplets where one model predicted that alternative B is more similar (blue diamonds in column B).

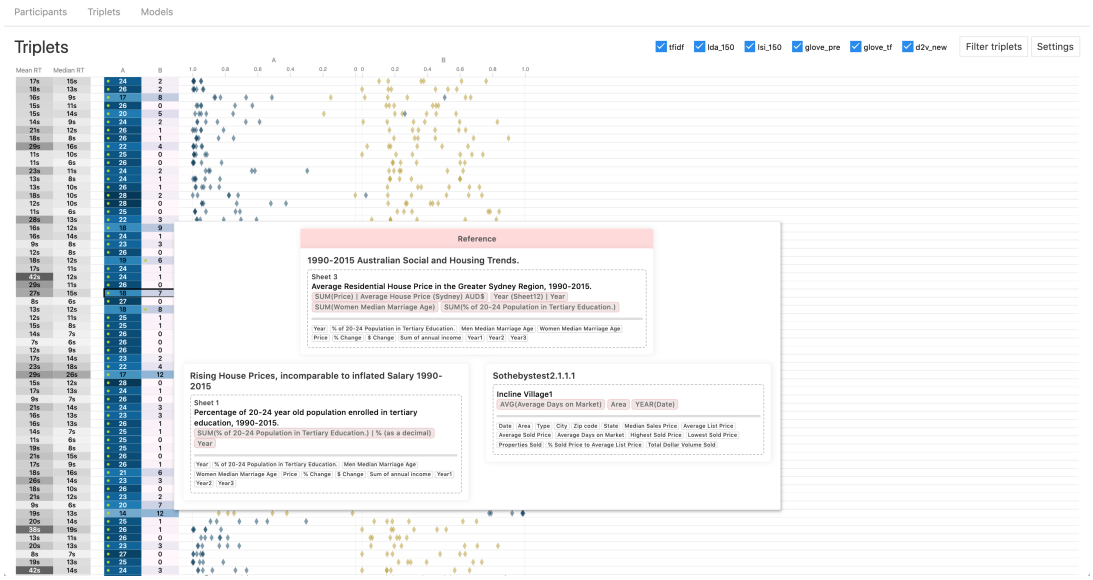


Fig. S26. Hovering over rows shows a preview of the triplet the way that it was presented to participants. The alternative that received more votes is shown on the left.

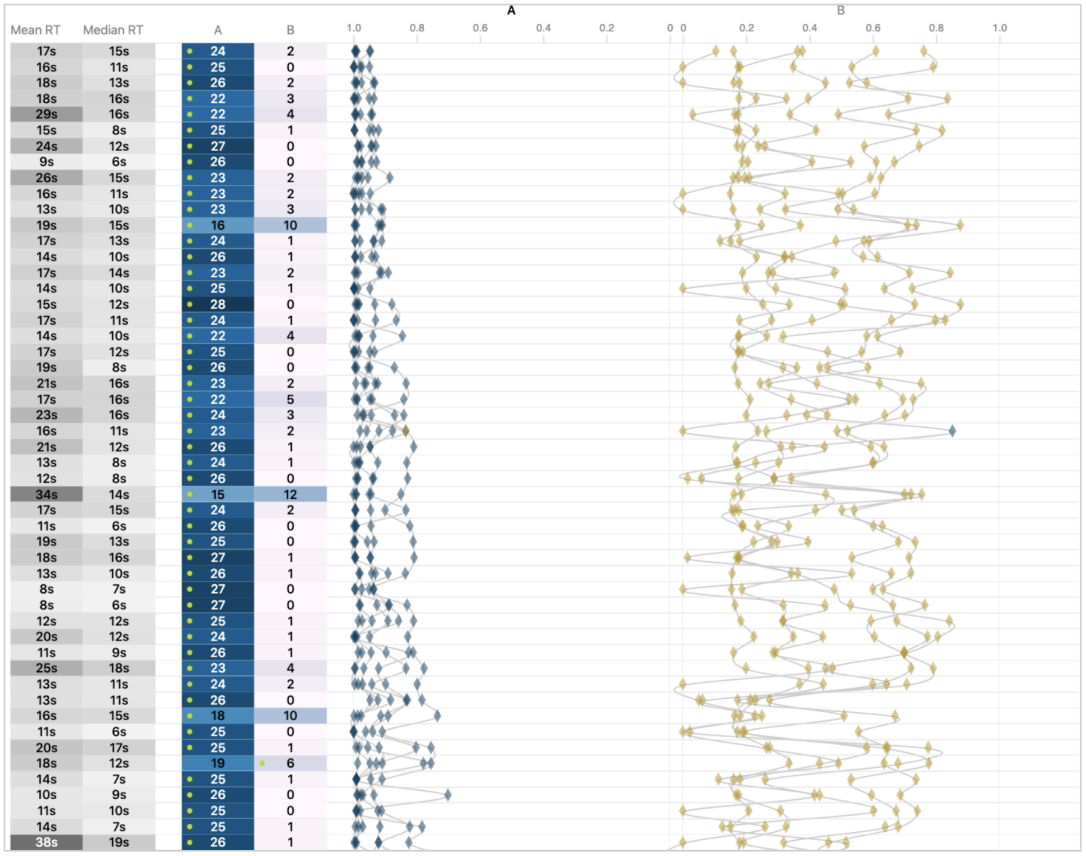


Fig. S27. Predictions that belong to a model instance can be connected with lines and all columns can be sorted to identify patterns. In this screenshot, triplets are sorted based on the standard deviation of model predictions (column *A*). All models computed high similarity scores for alternative A and diverging but lower scores for alternative B. The number of human votes mostly align with the models but in some cases human consensus is not as clear (e.g., triplet with 16 participant votes for alternative A and 10 votes for alternative B).

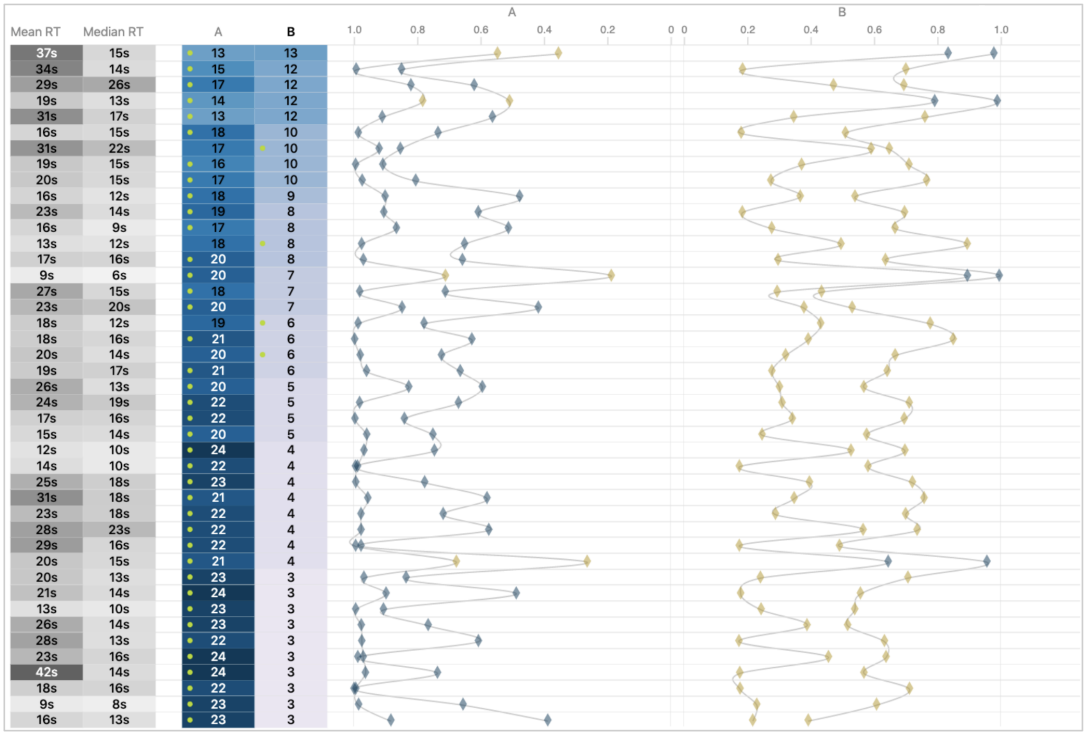


Fig. S28. Triplets can be filtered based on low or high human consensus. Model instances can be shown or hidden interactively. In this example, the triplets are sorted based on the number of participant votes alternative B received. Triplets with none or low human consensus are shown at the top. LDA and GloVe-TF are selected and all other model predictions are hidden.

7. Proof-of-Concept: VizCommender

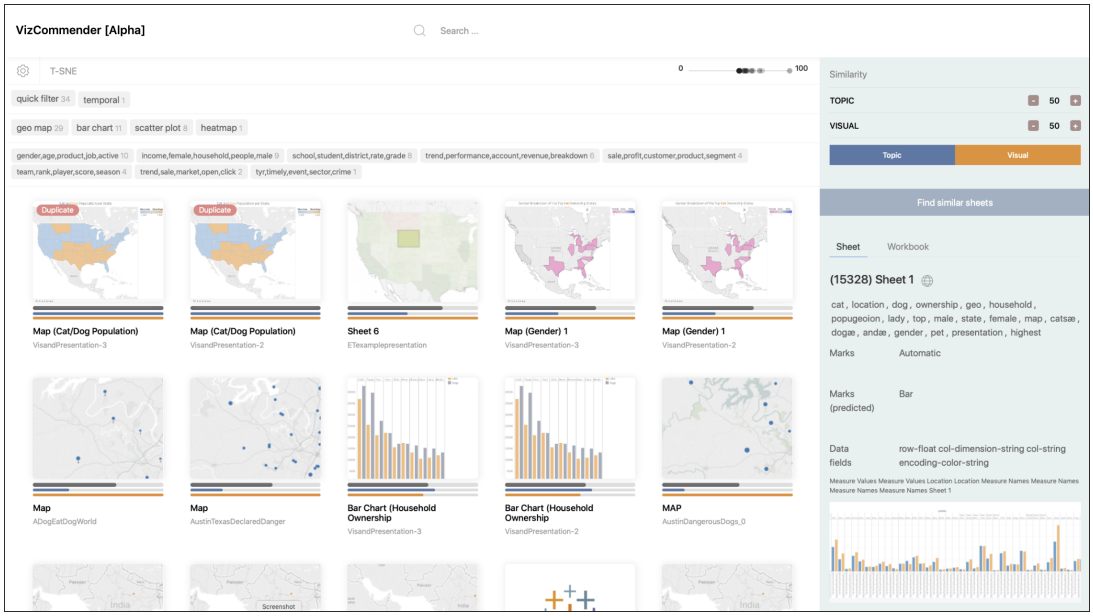


Fig. S29. Early VizCommender prototype that incorporated both visual and topic features. The interface allowed us to change the weights of these features (see right column) and observe how recommendations change. Later prototypes focused on text-based similarity because investigations and discussions with collaborators indicated that specific visual encodings are less relevant for information seeking in vis repositories.

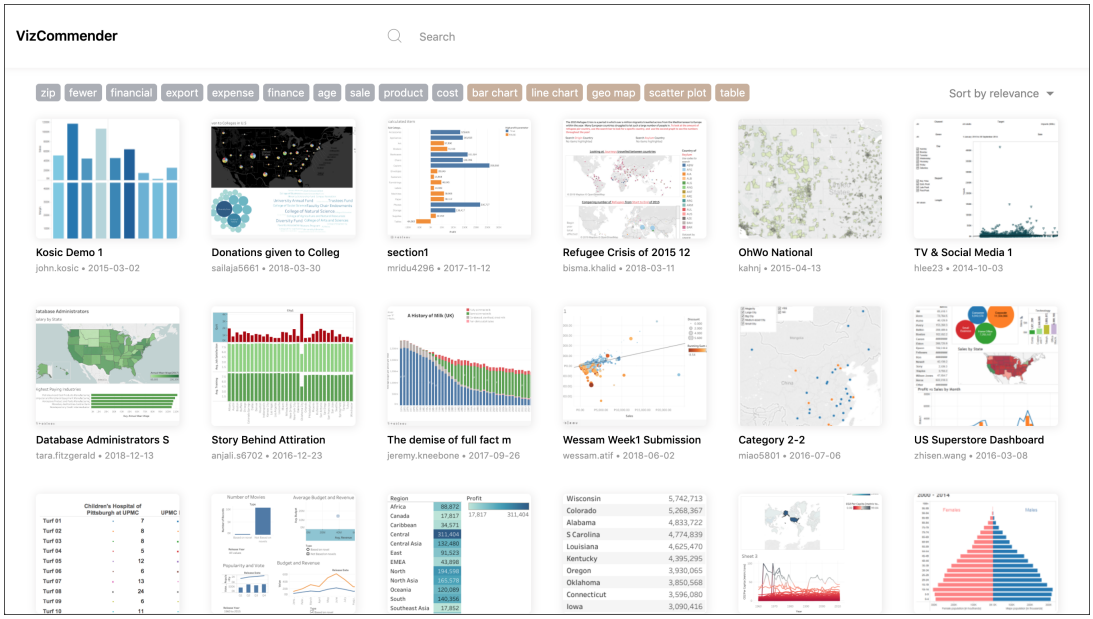


Fig. S30. VizCommender Overview.

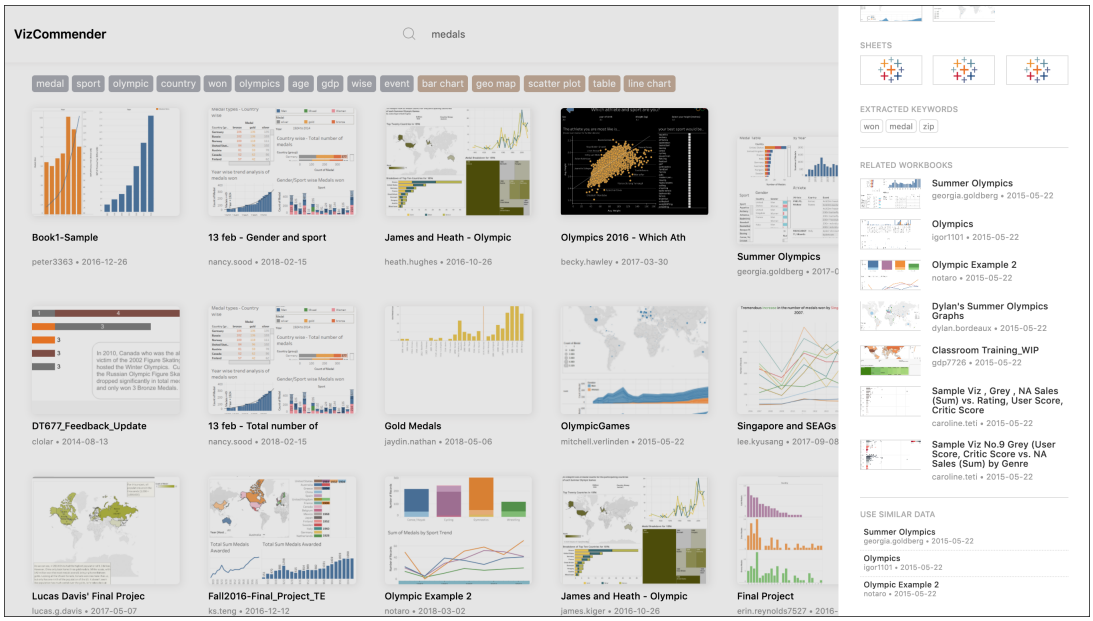


Fig. S31. VizCommender Overview. Screenshot shows results for the search query "medal". A workbook titled "OlympicGames" is selected and the quick view on the right side provides further details and recommendations. The similarity model identified related workbooks and several workbooks that use similar data sources.

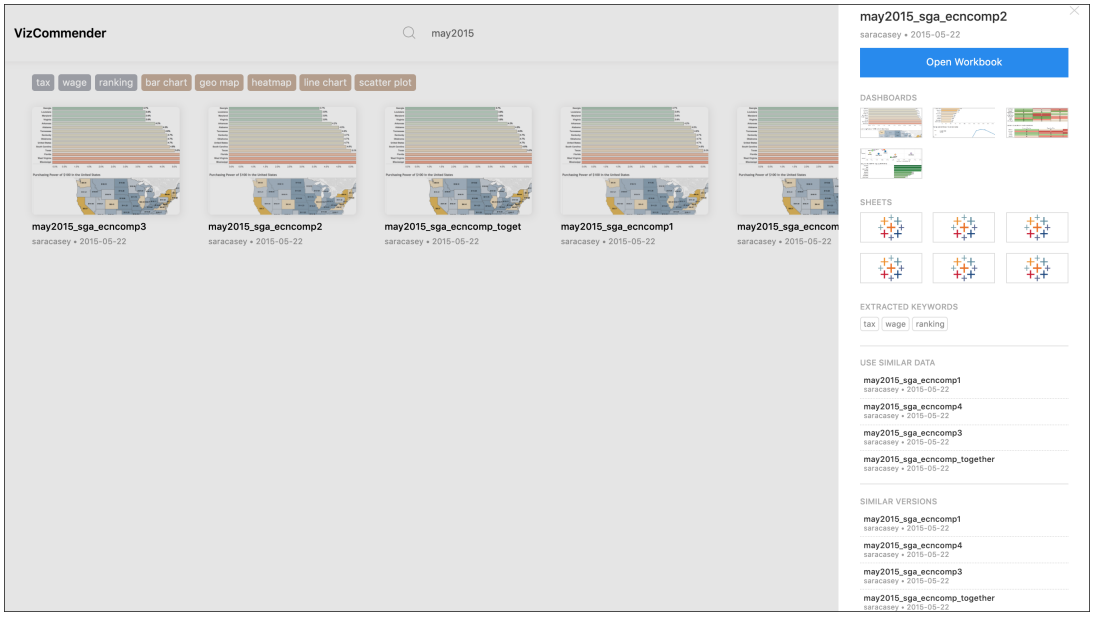


Fig. S32. VizCommender Overview. Five near-duplicate workbooks are shown and one is selected. Both similarity facets ("use similar data" and "similar versions") correctly reveal the relationship between the workbooks in the quick view sidebar. Future work could investigate how the viz-to-viz similarity measure can be used to bundle near-duplicates and to show only one representative in the interface.