# Explainability in Transparent Information Retrieval Systems

**Jerome Ramos, ScB,[1] Carsten Eickhoff, PhD[1,2]**
**[1]Department of Computer Science, Brown University, Providence, RI, USA**
**[2]Center for Biomedical Informatics, Brown University, Providence, RI, USA**

**ABSTRACT**

*When performing exploratory search, one must be able to quickly identify documents that may contain relevant information. Typical search engines provide users with ranked results of a query, but provide little transparency and explainability on how the results are ranked. The lack of transparency prevents users from understanding the results and can potentially lead to biased, unfair search engines. Non-transparent search engines may also hurt trust in the presented ranking. ExplainSearch was developed in order to provide users a way to understand how the ranking algorithm scores each document. A user study was performed which showed that ExplainSearch provided better explainability and transparency compared to traditional search engines and made exploratory search quicker, easier and more trustworthy.*

## 1 INTRODUCTION

Search engines are a vital part of gaining information in today's time. Information retrieval systems are often the main source of knowledge for many people. Due to the ubiquitousness of search engines, it is vital that they provide unbiased, factual data. “Unfair” search engines could lead to significant negative social and economic impact. In the era of "fake news", it is important to be able to determine which articles are factual. Additionally, the information retrieval system itself could potentially be trained to be biased because of how it is trained. Since machine learning models use datasets as a ground source of truth, a dataset that is not representative of the real world may lead to inaccurate results. For example, social media accounts with non-western names are more likely to be flagged as fraudulent because the classifiers have been trained on Western names [1]. Research has also shown that people trust more sources ranked higher (known as position bias) in the search results, but the ranking criteria may be based on signals of user satisfaction rather than signals of factual information [1].

In addition, being able to quickly find relevant knowledge is an important skill. Exploratory search allows people to discover new topics and learn key facts in a particular domain. One of the key skills in exploratory search is to be able to discern which information is interesting, relevant, and truthful. Although users are often provided a title and snippet of a document, this information is often not enough to determine if the document is relevant without having to read the entire document. Forming a query to search for can also be a difficult task if a user is unfamiliar with the terminology of the topic [2]. Thus, search strategies tend to be navigating articles through a trial-and-error approach. Belkin’s Anomalous State of Knowledge also found that most users are unable to precisely formulate the query they need [3]. In addition, deep learning systems do not answer the "why" results are ranked in the given order and practitioners want to know the underlying reasoning behind the results [1].

ExplainSearch was developed to tackle explainability and transparency in additive exact term matching retrieval models. It is a search engine application that allows user to understand the explanation behind the search results. Not only does ExplainSearch show the overall score of each result, but it also shows how each query term contributes to the overall score of the ranking. ExplainSearch provides a way to make search results transparent and explainable which allows users to find relevant information more quickly and trust that the rankings are an accurate representation of the ground truth.

## 2 BACKGROUND

Past research in search engine visualizations include Tile-Bars and HotMaps+WordBars, which explored the use of colors and shapes to provide explainability to search results [4,5]. Tile-Bars focused on showing users the document length and the density of query terms relative to the documents. HotMaps and HotMaps+WordBars displayed the term frequency of query terms over all documents returned. However, these approaches use Boolean queries and do not return scores for the documents. In addition, these interfaces do not look similar to modern day search engines and could potentially have a higher cognitive load relative to today’s tools because it would take time for users to

become familiar with a new interface. Other research has also explored using alternatives to graphs for visualizing information.However, studies have shown that users prefer bars over numbers or the absence of graphical explanations of relevance scores [2]. These alternative visualizations can be intuitive and difficult to use for first time users because of the lack of familiarity. URank provided a modern approach to providing visualizations to search engines. URank also displayed relevance scores for results. However, their user interface does not resemble many common search engines, which can be difficult to use for first time users.

Although the results of the URank show that their search engine interface had a lower cognitive load compared to traditional search engines, their sample size was quite small and it is worth exploring the combination of a traditional search engine interface and interpretable visualizations. Additionally, the focus of the URank studies was on cognitive load and search strategies, particularly query adjustments. One of the limits in their user studies was how strict to make the search tasks. They explored making both broad and focused exploration tasks. Additionally, the authors state that in the future, they would perform an online study with an incentive for finishing the task to bridge the gap between broad and focused tasks. Although the use of stacked bars for search engine explainability is not a novel approach in and of itself, ExplainSearch aims to combine familiarity with popular search engines and easily interpretable graphs. Furthermore, ExplainSearch allows users to make free form queries rather than having to select keywords, which is an unnatural way of searching.

## 3 EXPLAINSEARCH

### User Interface

ExplainSearch is designed in a minimalistic fashion in order to look like other common search engines. Familiarity was an important goal to keep in mind while designing ExplainSearch to make the application feel natural and usable. The search bar appears at the top of the screen and allows users to input their queries. Once a query is made, users can see the title of the document and the abstract, which consists of the first 50 words of the document. Additionally, each title is a hyperlink that leads to the full article. Every document has a star button next to it that users can click on to bookmark documents. They can view all bookmarked documents by clicking on the hamburger menu. Every search result also comes with a stacked bar graph that shows the overall relevance score of the document as well as the constituent scores contributed by each word in the query. Users can hover over one of the bars in the graph to see the exact score of that term. A legend is also provided at the top of the screen to associate a term and its respective color in the graphs.

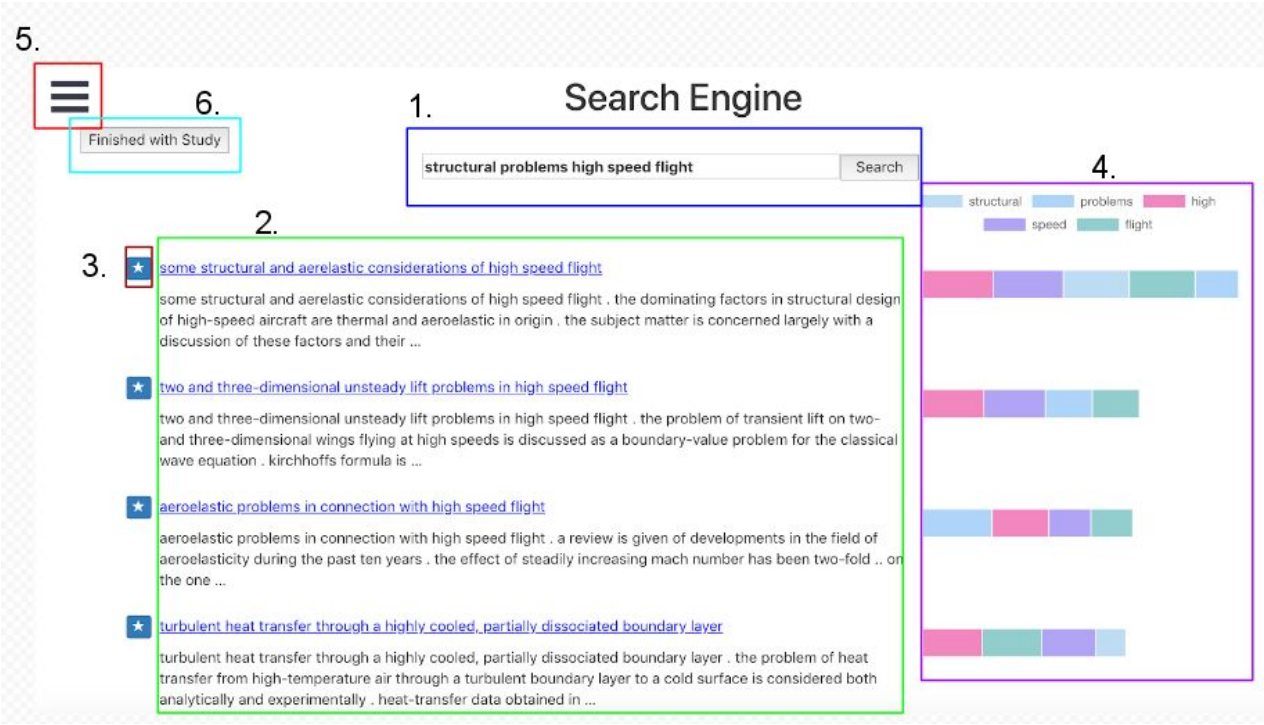


**Figure 1. ExplainSearch User Interface** 1) Searchbar to make queries 2) Search results, ranked from most relevant to least relevant 3) Star button to bookmark documents 4) Stacked bar graphs to explain how each query term contributes to the overall score of the document. 5) Hamburger Menu to show bookmarked documents and exploration task prompt 6) Button to redirect users to the quiz after they are done searching

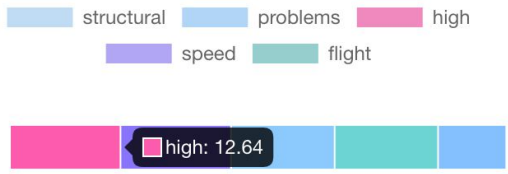


**Figure 2. Hovering over the bar reveals the score of that word**

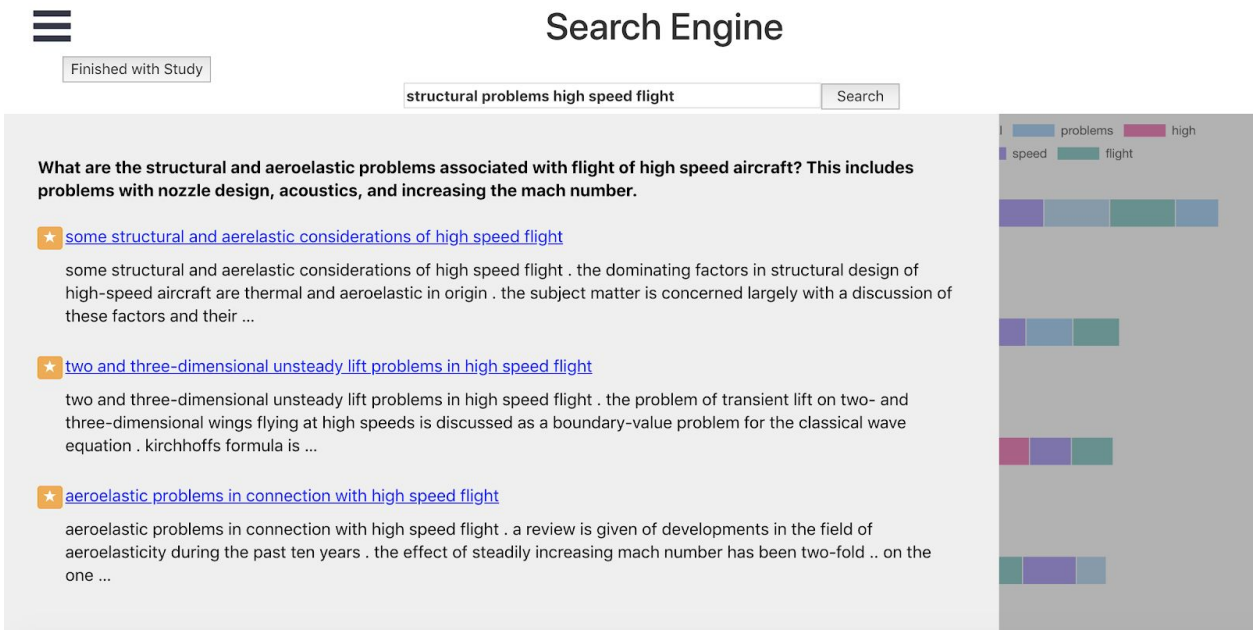


**Figure 3. Viewing the exploration task and the bookmarked documents in the hamburger menu**

**Ranking Explanation**

The words of the bar graph are sorted by magnitude of the query term score. Thus, larger scores appear on the left side of the graph and smaller scores appear on the right side. This makes it easy to scan the bar graphs and quickly determine the most relevant words in the document.

The ranking function used in ExplainSearch is the Okapi BM25 algorithm, which ranks matching documents according to their relevance to a given search query [6]. For every document $D$ and a query $Q$, with words $q_1,q_2,...,q_n$, the score of $D$ is given by:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i)\frac{f(q_i,D)(k_1+1)}{f(q_i,D)k_1(1-b+b\frac{|D|}{avgdl})}$$

where $f(q_i,D)$ is the frequency of term i in document $D$, $|D|$ is the number of the words in the document, and *avgdl* is the average document length in the collection. $k_1$ is a discount factor for repeated term mentions and $b$ is a parameter for document length normalization. Additionally, the *IDF* function used in the Okapi BM25 algorithm is the inverse document frequency, which is given by:

$$IDF(q_i) = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

where $N$ is the total number of documents in the collection and $n(q_i)$ is the number of documents containing $q_i$.

The Okapi BM25 algorithm was chosen as the ranking function because it is a widely used algorithm, the default ranking algorithm for Apache Lucene, and it is simple to compute the importance of each word in the query. Additionally, the algorithm provides better search results than using the term-frequency approach and unlike the term-frequency approach, the overall score of BM25 is not drastically affected by the use of stop words such as "the" or "and".

**Implementation**

The client of ExplainSearch was developed in ReactJS. The web server API developed in Node.js and primarily handles routing as well as storing metrics in a MongoDB server. Finally, the Java Spring Boot Framework as well as the Lucene library was used to parse through the document collection, compute the BM25 score of each document, and send the relevance scores to the client through an API.

## 4 USER STUDY

**Experiment Setup**

The user studies were conducted on Mechanical Turk (MTurk), a crowdsourcing platform where users can perform various tasks for pay. These users were separated into three distinct experimental conditions: 1) Titles-Only: a search engine that only shows the titles; 2) Titles+Abstract: a search engine that shows the titles and first 50 words of the document; 3) ExplainSearch: the titles, and first 50 words of the documents. The two control group search engines contained the same functionality as ExplainSearch, except Titles+Abstract does not include the bar graphs and Titles-Only does not include the bar graphs and the abstracts. Each group had a total of 50 users perform an open-ended exploratory task.

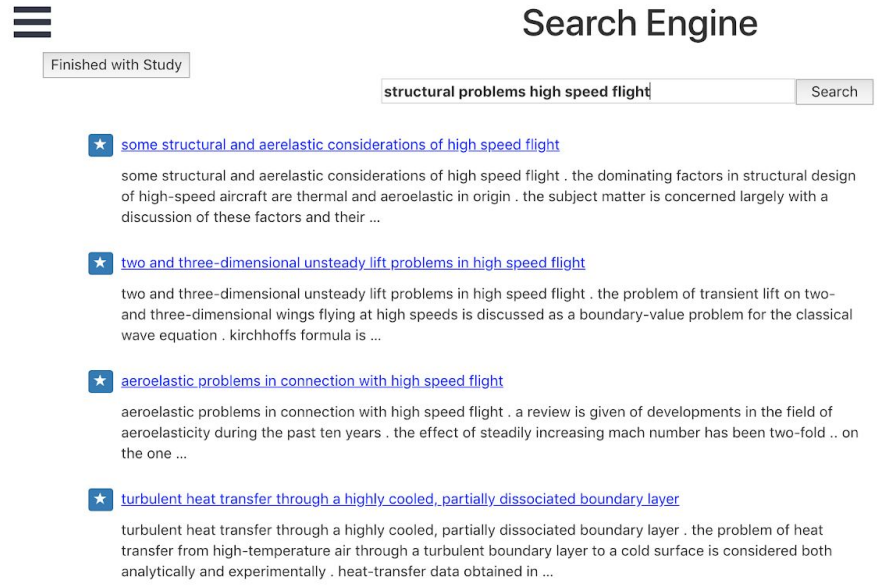


**Figure 4. Search Engine with only the titles and abstracts given**

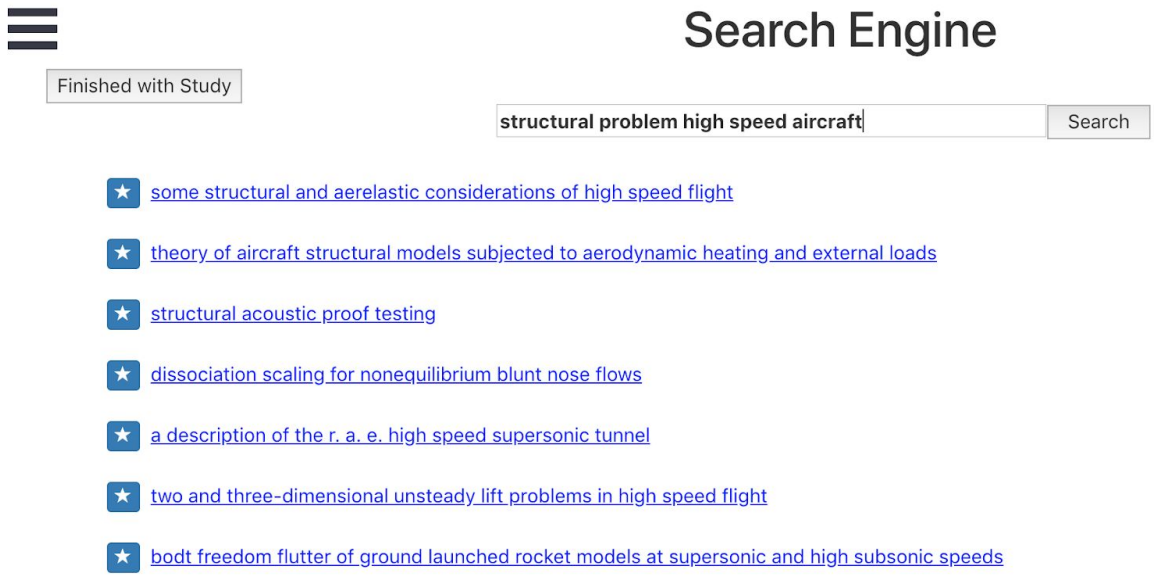


**Figure 5. Search Engine with only the titles returned**

The document collection chosen was the Cranfield Collection [8], a collection of research documents on aerospace engineering. A relatively niche topic was chosen because the focus of the study was whether users could quickly gain knowledge on an unfamiliar topic. Users were tasked with an exploratory search, where they attempted to answer the question: "What are the structural and aeroelastic problems associated with flight of high speed aircraft?". This prompt was chosen from the set of queries in the Cranfield Collection. The users used their respective search engines to perform queries and learn more about the topic. Once they were satisfied with their research, they were redirected to a multiple choice quiz, where they answered questions related to the given prompt.

Due to the low barrier of entry into MTurk, this quiz was designed to filter out submissions that did not make a serious attempt at the experiment. Additionally, users were instructed that if they did not make any queries, view, or bookmarked any documents, they would not be compensated and their data was filtered out of the final data. If a user failed the test, they are redirected back to the search engine to perform research again on the topic and are given three more attempts to pass the quiz. The reasoning behind this is that users would be given another chance to perform more research on the prompt and, in particular, look up information about the questions that they did not know the answer to previously. Once the user passes they quiz, they are given a survey to describe their experiences with their respective search engine. If the user failed the quiz four times, then their data would not be included in the final results.

**Metrics Collected**

The queries the user made, as well as which documents they view and bookmarked were recorded. The time they took to complete the task was also recorded. The exit survey given to users at the end of the experiment asks eight questions about the simplicity, transparency, trustworthiness, and explainability of the search engine they used. The responses were on a five point likert scale that ranged from "strongly disagree" to "strongly agree". The NASA Task Load Index was also used to measure the load of the experiment. Users were also given the option to provide free-form feedback to comment on their experience with the search engine.

## 5 RESULTS

A total of 50 users completed the task for each of the three experimental conditions. Users filled out Hart and Staveland's NASA Task Load Index (TLX) at the end of the survey to assess how they felt using their respective search engine. For each question, users indicated a number between 1 and 20, where 1 indicates very low and 20 indicates very high. The performance of all three groups was relatively even, meaning users felt that they were successful in completing the task. The temporal demand of ExplainSearch was lower than both of the alternative search engines. Mental demand for ExplainSearch was also lower, especially when compared to Titles+Abstract users.

**Table 1. Average NASA Task Load Index for each Group**

| Workload Type | ExplainSearch | Titles+Abstract | Titles-Only |
|---|---|---|---|
| Mental Demand | 14.08 | 15.86 | 14.60 |
| Physical Demand | 3.90 | 3.42 | 5.04 |
| Temporal Demand | 5.14 | 6.64 | 8.02 |
| Performance | 13.76 | 13.30 | 13.41 |
| Effort | 14.26 | 15.60 | 14.10 |
| Frustration | 7.42 | 10.42 | 9.30 |

In addition, users were asked questions about the transparency, simplicity, trustworthiness, and explainability of their search engine. Responses ranged from strongly-disagree to strongly agree. The responses are shown in the tables below. The p-values between each search engine was calculated using the Mann-Whitney U Test, which tests if it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. At a significance level of p=0.05, it can be said that there is a significant difference between the responses of ExplainSearch and Titles-Only for all questions. Additionally, for all but two survey questions, there was a significant difference between ExplainSearch and Titles-Only. There was only a single question with a significant difference in responses between Titles+Abstract and Titles-Only.

**Table 2. Survey Responses for ExplainSearch Users**

| Question | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The search engine was simple to use | 1 | 2 | 6 | 31 | 10 |
| Transparency in search results made it easier to find relevant documents | 1 | 2 | 10 | 22 | 15 |
| Finding relevant results to searches was intuitive | 0 | 5 | 10 | 29 | 6 |
| The results from the search engine are an accurate representation of the truth | 2 | 1 | 11 | 28 | 8 |
| The application clearly explains the rankings of the results | 1 | 4 | 8 | 25 | 12 |
| Transparency in search results helped to find relevant documents quickly | 1 | 4 | 6 | 30 | 9 |
| It is clear why a certain result is ranked higher than another results | 1 | 6 | 6 | 23 | 14 |
| The rankings of the search results corresponds to how trustworthy they were | 1 | 11 | 13 | 21 | 4 |

**Table 2. Survey Responses for Titles+Abstract Users**

| Question | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The search engine was simple to use | 4 | 7 | 10 | 23 | 6 |
| Transparency in search results made it easier to find relevant documents | 0 | 11 | 11 | 24 | 4 |
| Finding relevant results to searches was intuitive | 2 | 10 | 16 | 20 | 2 |
| The results from the search engine are an accurate representation of the truth | 0 | 4 | 22 | 20 | 4 |
| The application clearly explains the rankings of the results | 10 | 15 | 14 | 9 | 2 |
| Transparency in search results helped to find relevant documents quickly | 1 | 8 | 14 | 25 | 2 |
| It is clear why a certain result is ranked higher than another results | 11 | 13 | 11 | 10 | 5 |
| The rankings of the search results corresponds to how trustworthy they were | 5 | 16 | 24 | 4 | 1 |

**Table 3. Survey Responses for Titles-Only Users**

| Question | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The search engine was simple to use | 1 | 5 | 12 | 22 | 10 |
| Transparency in search results made it easier to find relevant documents | 2 | 9 | 13 | 19 | 7 |
| Finding relevant results to searches was intuitive | 2 | 10 | 13 | 21 | 4 |
| The results from the search engine are an accurate representation of the truth | 3 | 1 | 13 | 26 | 7 |
| The application clearly explains the rankings of the results | 8 | 15 | 14 | 9 | 4 |
| Transparency in search results helped to find relevant documents quickly | 2 | 10 | 14 | 16 | 8 |
| It is clear why a certain result is ranked higher than another results | 5 | 17 | 10 | 15 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| The rankings of the search results corresponds to how trustworthy they were | 4 | 11 | 22 | 9 | 4 |

**Table 4. P Values between search engines for each survey question. P Values less than 0.05 are highlighted**

| Question | ExplainSearch and Titles+Abstract | ExplainSearch and Titles-Only | Titles+Abstract and Titles-Only |
|---|---|---|---|
| The search engine was simple to use | **0.005126** | 0.089919 | 0.112866 |
| Transparency in search results made it easier to find relevant documents | **0.001960** | **0.003467** | 0.489840 |
| Finding relevant results to searches was intuitive | **0.002350** | **0.016368** | 0.291470 |
| The results from the search engine are an accurate representation of the truth | **0.010598** | 0.263157 | 0.051996 |
| The application clearly explains the rankings of the results | **0.000000** | **0.000001** | 0.266837 |
| Transparency in search results helped to find relevant documents quickly | **0.002346** | **0.008428** | 0.458097 |
| It is clear why a certain result is ranked higher than another results | **0.000005** | **0.000015** | 0.213351 |
| The rankings of the search results corresponds to how trustworthy they were | **0.000121** | **0.032027** | **0.033363** |

## 6 DISCUSSION

ExplainSearch received more positive feedback compared to Titles+Abstract and Titles-Only. The p-values between ExplainSearch and Titles+Abstract indicate that there is a significant differences in how users feel about the two engines. Additionally, the p-values for all questions are extremely low, showing that there was much greater positive feedback for ExplainSearch compared to Titles+Abstract. ExplainSearch also had significant p-values for six of the eight questions, meaning that overall, ExplainSearch performed better than Titles-Only. Interestingly, Titles-Only performed better than Titles+Abstract even though the only difference between the two was the exclusion of the abstract. Seven of the eight questions had a p-value greater than .05 for Titles+Abstract and Titles-Only, indicating that, in general, users felt quite similarly about these search engines.

Looking at the responses to the NASA Task Load Index, ExplainSearch had a lower mental demand, temporal demand, and frustration compared to the other two search engines. Additionally, ExplainSearch users reported that they felt that they performed better on the task, though only marginally. Users of ExplainSearch felt that the search engine did a better job at explaining the rankings of the results, which in turn made them more confident in the trustworthiness of the results. Additionally, transparency in the documents made it easier for them to find relevant documents. Interestingly, users of ExplainSearch felt that finding relevant results was more intuitive and reported lower temporal demand even though explainable visualizations are not a common feature in most search engines.

One difficulty of the user study phase was creating an experiment that is simple enough for users to complete while still being able to collect meaningful data. Due to the low entry barrier of Mechanical Turk, many users rush to complete tasks as fast as possible rather than making a serious attempt at the task. Thus, it is vital that a task be designed to dissuade cheating in order to collect meaningful data. Although the multiple choice quiz given in

between the experiment and survey filtered out much of the non-serious attempts at the experiments, there was still cases where users were able to successfully complete the quiz without making a single query. Thus, attempts without any queries made, documents viewed, and documents bookmarked were filtered out from the final results. However, this may not be a catch-all method and there may still be cases where users performed the task non-seriously, which may dilute the overall result data set.

One experimentation with the design was to ask users to bookmark relevant documents and make the goal of the task to bookmark relevant documents. For each query given in the Cranfield collection, there is a list of query-document pairs which indicate which queries match with which documents. Once a user gets above a threshold of relevant documents bookmarked, they would successfully complete the task. However, this proved to also yield a great amount of unusable data as MTurk users would simply search the given query and bookmark several of the top documents, which would oftentimes be enough to select the majority of relevant documents.

Future work would include performing the study with document collections of a different subject to see if the complexity of the topic changes how users respond the ExplainSearch. Additionally, further trials would be conducted with users who are experienced in the subject area. It would be interesting to see how subject experts interact with ExplainSearch compared to users who have no experience in the subject. Providing more color-coded highlighting code also increase the speed of interpretability and make it easier to identify which articles have certain keywords. Along the same lines, better snippet generation could also potentially make it easier to find relevant documents more quickly.

Another important direction to explore would be modifying the experiment so that it can quickly gather useful data from crowdsourcing platforms while still being a simple task for users to complete. Additionally, it could be interesting to intentionally change the rankings of the documents and insert non-relevant documents in the results. This experiment would test whether user still felt that ExplainSearch was trustworthy, even though it is outputting tampered data and incorrect relevance scores. Perhaps the most interesting work in the future would be to provide explainability for soft matching retrieval models. Incorporating more state of the art ranking algorithms could potentially lead to more trustworthiness in the result because the new algorithm would yield better search results. However, deep learning models often act as a “black box” and it is much more difficult to explain the results of a deep learning model compared to an additive exact term matching retrieval model.

## 7 CONCLUSION

ExplainSearch is a transparent search engine that allows users to understand why the rankings of the result. By providing explainability and transparency, users were able to complete the exploratory task more quickly and were less frustrated with the task compared to users using traditional search engines. They were confident that the search results were an accurate representation of the truth and felt confident that it was clear why a certain document was ranked higher than another document. Users of ExplainSearch also better understood the rankings of the results, which helped them find relevant documents faster. This study shows that transparent and explainable search engines can speed up exploratory tasks and provide more confidence in the trustworthiness of the results.

## References


1. J. Allan., J. Arguello, L. Azzopardi, P. Bailey, T. Baldwin, et. al, Research Frontiers in Information Retrieval; Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)
2. C. di Sciascio, V. Sabol, E. Veas, “uRank: Visual analytics approach for search result exploration”, 2015 IEEE Conference on Visual Analytics Science and Technology (VAST)
3. N.K. Belkin, "Anomalous states of knowledge as a basis for information retrieval". The Canadian Journal of Information Science, 5, 1980, pages 133-143
4. M. Hearst, “TileBars: Visualization of Term Distribution Information in Full Text Information Access”, Proceedings of CHI ’95, Denver, CO, May 1995
5. Orland Hoeber, Daniel Schroeder, Michael Brooks, "Real-World User Evaluations of a Visual and Interactive Web Search Interface", Information Visualisation 2009 13th International Conference, pp. 119-126, 2009.
6. Stephen E. Robertson; Steve Walker; Susan Jones; Micheline Hancock-Beaulieu & Mike Gatford (November 1994). Okapi at TREC-3. Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA