

# Improving SpaceNews' Information Retrieval System for key-word based queries

ADELAIDE SANTOS\*, RITA MENDES\*, and TIAGO RODRIGUES\*, Faculty of Engineering - University of Porto, Portugal

Allowing users to easily retrieve news articles upon keyword-based queries is an essential operation for any news-related information system. SpaceNews, a reliable space-related news agency [15], is one of the best on the topic, even though the search system they provide is lackluster. In this paper, we propose an improvement of SpaceNews' Information Retrieval System. By cleaning up a dataset composed of some of their news, and making it comply with a well-defined format, we aim to design queries that can deliver relevant results on keyword-based queries, with the help of Apache Solr [1]. These queries are tested and compared using relevant metrics like Precision at 10, Average Precision and the Precision-Recall curve, so that the ones that best fit our proposal are chosen. As a fully-fledged search system must have usability in mind, a User Interface was developed so any user can quickly retrieve information, and Solr's More Like This feature [2] was employed to facilitate the fetching of even more relevant articles, creating a final experience much more streamlined than the currently existing one.

Additional Key Words and Phrases: datasets, information processing, information retrieval, data pipeline

## ACM Reference Format:

Adelaide Santos, Rita Mendes, and Tiago Rodrigues. 2022. Improving SpaceNews' Information Retrieval System for key-word based queries. *ACM Trans. Graph.* 1, 1, Article 1 (October 2022), 14 pages. <https://doi.org/420969.420969>

## 1 INTRODUCTION

News websites are one of the biggest outlets of information in the world. With social media and the widespread use of the web, it is one of the easiest mediums people have to be up to date and know what is happening. However, for the websites to be engaging, they need to show users relevant content. When a user searches for something, it should bring up the most compatible results possible. This is what we want to achieve.

In this report, we will demonstrate how to start with a dataset of the Spacenews website [7] and create a search system that brings up relevant results every time. When exploring interesting sets of data, we came upon this possibility when searching Kaggle. It had no license restrictions (it simply stated that the copyright belongs to the publisher, Spacenews.com [7]) and it seemed quite usable. The data was also updated recently, with the newest entries coming in from June 2nd, 2022 [7]. In this dataset, the data is clean and well organized. It doesn't have a lot of columns, but each one has a lot of information, and the structured nature makes it easier to find patterns. The final objective is for the user to be able to type a normal sentence, in regular text, and relevant articles regarding the query will show up, even though the title may indicate it. We hope

\* All authors contributed equally to this research.

Authors' address: Adelaide Santos, [up201907487@up.pt](mailto:up201907487@up.pt); Rita Mendes, [up201907877@up.pt](mailto:up201907877@up.pt); Tiago Rodrigues, [up201907021@up.pt](mailto:up201907021@up.pt), Faculty of Engineering - University of Porto, Portugal.

to extract features from the text and match them against the query for relevance, getting with this compatible articles whose titles may not show it.

This paper is divided into three parts. The first, in Section 2, will start by showing the steps taken in order to prepare and clean the data, and suit it to our needs. We will highlight the pipeline produced and designed in order to make these steps reproducible with ease. Then, we will show some of the analysis done on the dataset, and some of the more interesting features found in the exploration, along with the definition of some example information needs that users might have.

In the second part, starting in Section 3, we will show the work done using Apache Solr [1], from the definition of the document model and its indexation to the evaluation of the search results, using some of the most common metrics, and checking if they answer the information needs defined previously.

The final part, Section 4, focuses on improving the features developed in the previous iterations and, with this, creating a seamless experience for the end users. We show how the User Interface was devised and what it brings to viewers of the page, how the search features were improved to deliver more accurate results overall, and how the MoreLikeThis feature from Solr [2] was used to improve the results when taking into account feedback from the user.

## 2 DATA PREPARATION AND CHARACTERIZATION

Preparing the data before processing it is essential to smooth the rough edges that can appear in datasets, creating a uniform look that is consistent with the final vision for the project. Here, we will show the steps that were taken to eliminate those inconsistencies and regularize the dataset, and some of the exploratory analysis done with the clean data.

### 2.1 Filling Missing Parameters

While exploring the data, namely using Python and the Pandas library [13], we noticed that not all of the cells were included, as we initially thought. Upon further analysis, we found out that all the missing cells belonged to the content column, meaning the article had not been parsed yet. To combat this, we created a script that would download the lost articles and add them to the database, making sure that no cell was left a void.

With this, we encountered yet another problem. Some of the articles mentioned in the dataset were no longer available and, as such, they could not be downloaded from the website. Considering this, and the fact that there were only 4 such entries, they were deemed outliers and dropped, as they could not bring any relevant information.

Whilst visiting the actual website, we saw that there were some tags and sections that showed up on almost all pages, but that

weren't present in the dataset. As they seemed to be necessary for a search system, since they act like keywords for an article, we went and fetched all tags and sections for all articles that had them, developing the dataset even further.

## 2.2 Cleaning the Data

Continuing the exploration, the next thing that stood out was the post-excerpt column, which only repeated the first paragraph of the article. Since this was duplicate information and did not add much to the set, we decided it was best to simply drop it, and focus on the remaining items, from which we could extract the same information.

The same thing happened with the tags and sections, not all articles had them. This time, instead of dropping those articles, we assumed this meant that they were broader, and not suited to be tagged, so we kept them, just with empty tags and section values. This seemed to be a recurring feature, so it may indicate that the tags and sections were added recently, and not all articles were published in that time frame.

## 2.3 Creating a data Pipeline

To make these steps reproducible, it was necessary to create a data pipeline, from which we could start out with the downloaded data set and get to the updated version. This includes all the steps related to the data filling and cleaning described in Subsections 2.1 and 2.2. Initially, the simple pipeline, modeled after the pipelines described by Sérgio Nunes [11], just tried to couple the dataset with the missing articles, writing them all into a `spacenews_filled.csv` dataset. It was a bit crude when using Beautiful Soup, a Python library for parsing HTML files [14], since it kept all the scripts the page had, as well as the HTML tags, which would take too much time to remove. But, with the help of Newspaper3k, another library made specifically for extracting articles [12], it was a breeze. With that handled, it was possible to advance to the next stage of the pipeline.

With further exploration and consideration, namely the dropping of certain parts of the dataset, the second iteration of the pipeline focused more on making slight fixes to the already existing data. It parsed the articles from the website, and then it removed all the entries that didn't have a corresponding article online. For this, we also used the Pandas Python library and its drop nulls feature. Also, it eliminated the `postexcerpt` column, which was deemed unnecessary when exploring (explained in Section 2.1).

To keep the dataset more modular, we decided to split the authors' names from the news entries. While this makes sense from a database standpoint, it also helps in reducing the amount of data that is repeated in each row. Once again, we used the Pandas library for this task.

Whilst the Newspaper3k [12] library could fetch the article and strip all of the HTML cleanly, it was not capable of fetching all the relevant data, such as the article's tags and sections. To combat this, we developed a python script that fetches that information using BeautifulSoup [14], which proved helpful for this use case.

Finally, for easier use with Apache Solr, we decided to generate a final version, in JSON format, so that it could easily be parsed by the program, and also because it has native support for lists, which CSV doesn't. This was important for the tags and sections because since

we stored them in lists, they would be converted to a string literal in CSV format, which was far from ideal when analyzing them.

With all the transformations put together, a final pipeline, exemplified in Figure 1 was achieved. It includes the filling of missing information, the data cleaning that ensues, and the separation into different models, to keep it consistent with a relational database paradigm.

## 2.4 Dataset Characterization and Analysis

To better understand the dataset that we will be working with, we wrote an R script that performed statistical analysis on the data.

The dataset we chose contains about 17.500 news articles published from 2005 until May 2022. As we can see in Figure 2, the website has been growing in the past years, with 2021 being the year with the most published news.

We wanted to find if any authors prevailed over the rest, so we plotted the authors with more than 100 news articles sorted by the number of publications, shown in Figure 3, and noticed that Jeff Foust was the one with the most publications, topping most other authors by a large margin.

We also wanted to find if any subject prevailed over the others, so we plotted the frequency of the tags (Figure 4) and the sections (Figure 5) associated with the articles.

## 2.5 Text Analysis

In order to identify the most common words in the dataset text we made a word cloud. This is a visual representation of text data in the form of tags, whose importance is visualized by their size. The bigger the word is in the word cloud, the more frequent it is in the text.

To build the word cloud in Figure 6, the data used belongs to the content and title of each article, as these are the columns with a higher value text content. Thus, we can verify that some of the most common words in the articles are "space", "satellite", "launch" and "NASA".

These kinds of words are to be expected as common, seeing as it is a website about news from space, and so they can be inferred as keywords to the articles and can guide the search system to better results.

## 2.6 Data Domain Conceptual Model

After preparing the data, we ended up with the relations shown in the following diagram:

The main class is the **Article** class, which contains:

- **Title:** The title of the article;
- **URL:** The URL to the article page on the website;
- **Content:** The body of the article;
- **Date:** The article's publication date.

There is also an **Author** class, containing:

- **Name:** The name of the author.

The relation between **Article** and **Author** is a relation of *many to one*, as each article is published by only one author, but an author can have many published articles.

An article can have multiple tags and sections associated, so we also have a **Tag** class with:

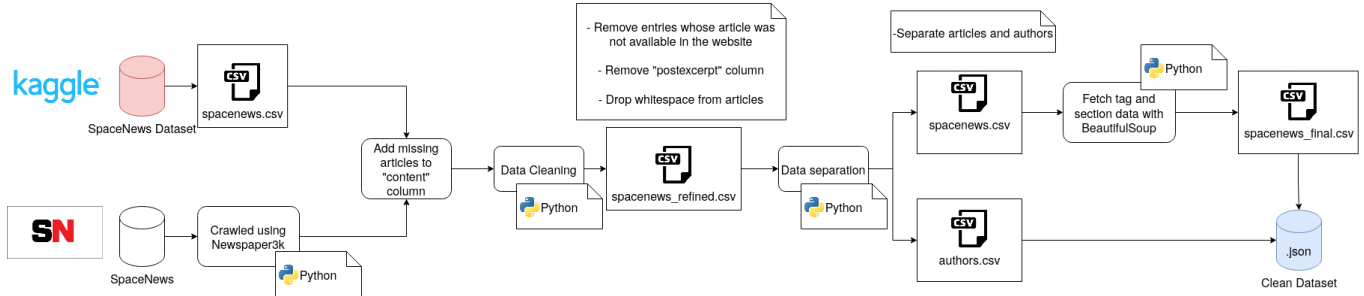


Fig. 1. Final, refined pipeline

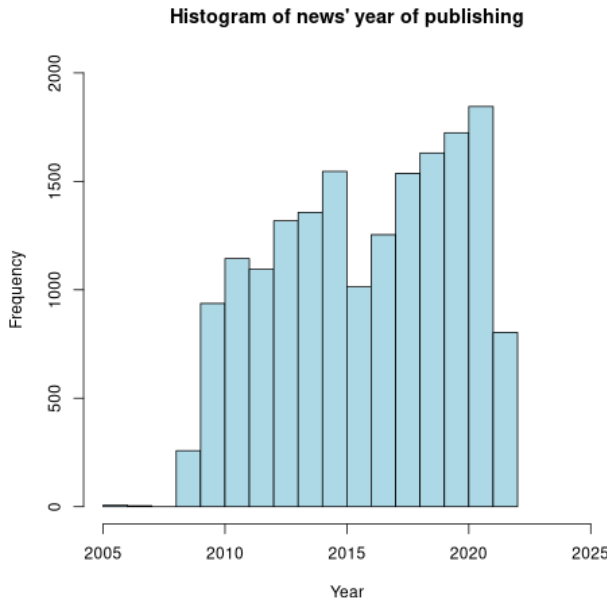


Fig. 2. Distributions of the articles based on publication year

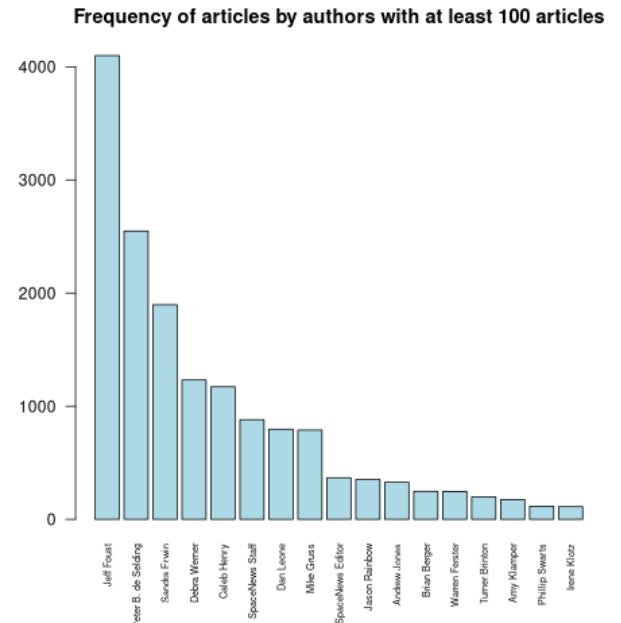


Fig. 3. Authors with the most published news articles

- **Name:** The name of the tag.
- and a **Section** class with:
- **Name:** The name of the section.
- These are *many to many* relations.

## 2.7 Setting the Information Needs

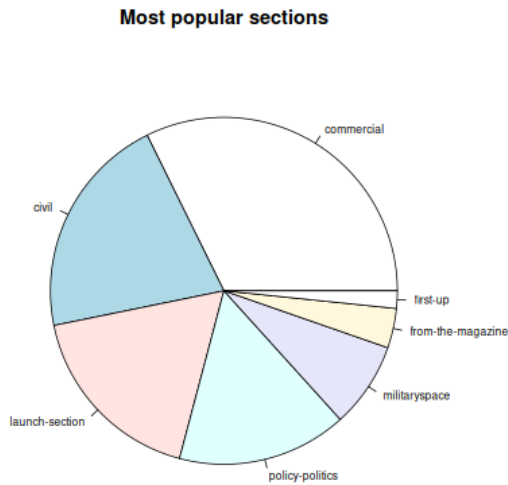
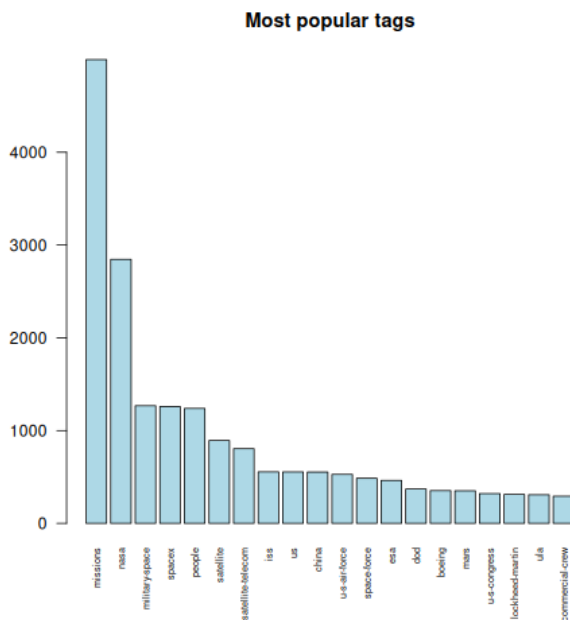
When analyzing an Information Retrieval System, it is important to consider the possible searches that the users will perform and the underlying scenarios. For this, we can consider Information Needs, which encapsulate a possible search task to be executed that satisfies a user's need. Here are some interesting ones that pertain to our search system:

- What were the latest launches by SpaceX?
- What articles did Jeff Foust publish in 2021?
- Are there launches from French Guiana?
- What are the news regarding astronauts working in the ISS?

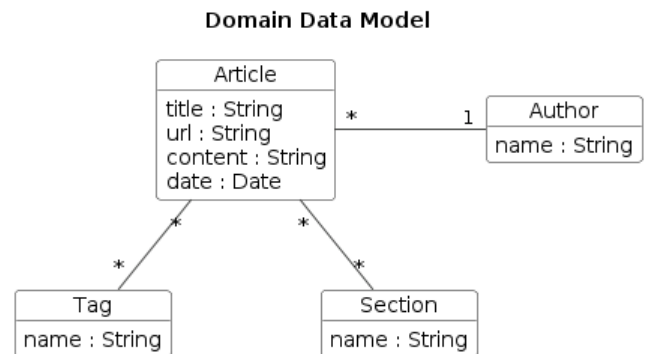
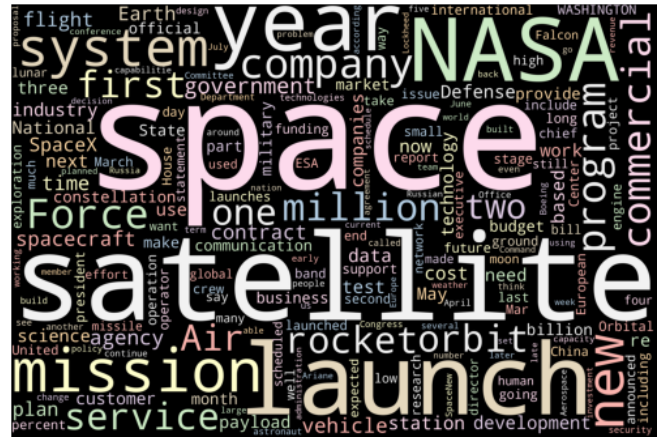
## 3 INFORMATION RETRIEVAL

According to Manning et al., Information Retrieval can be defined as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [9]. This job is now widespread and done automatically within search systems, where the most common task is retrieving a file, in a given format, through the use of a search query. The system matches said query against all files and sorts them according to relevance, leaving the user with, hopefully, the most significant results first.

In this section, we will demonstrate how we used an Information Retrieval tool, Apache Solr [1], to develop a search system for our project. We will first describe the document model that was created and its indexation so that Solr could know how to handle each field properly. Afterward, we will be answering some of the information needs described in Subsection 2.7 with plain and modified queries,



and showing the results of each one. Finally, we will evaluate said results with some of the most common measures, such as Average Precision and Precision@10, and interpret their meaning.



### 3.1 Creating the Document Model

In Subsection 2.3, it was hinted that the CSV file created originally was not the best fit for working with our data, especially with Apache Solr. First and most importantly, CSV does not have a way to store lists natively, it simply converts them into string format. This is problematic with the tags and sections that were retrieved later on, as Solr would have trouble parsing them. Also, the separation that we tried to achieve when separating the authors, to make the dataset look like a relational database was far from ideal as well, since Solr expects the information for a full document to be self-contained, like in a NoSQL database.

This meant that we had to make some changes to adapt the dataset for Solr. The tweaks that were made were:

- Removing the authorID and substituting back with the author's name.
- Converting the CSV file into JSON and transforming the string lists into actual lists.
- Transforming the date fields into %YYYY-%mm-%dd format and appending "T00:00:00Z" to it, so that it conforms to Solr's specification.

With this, our dataset looked more like a NoSQL database, but it fits much better into the Solr model, as each entry was its isolated document, and contained all information within it.

### 3.2 Indexing the Documents

After determining the model for each document, it was time to index it appropriately. In Solr, this is done by creating types for each attribute of the document, and specifying which of them should be indexed in the first place.

Our first tasks were determining what the user should be able to search for, and defining which of the attributes should be indexed. It was decided that all but the URL should be indexed, as it did not make sense, according to our information needs, for a user to search for a URL, and also because we could not extract any relevant information from it when designing queries. All other fields were indexed as they provide valuable content and can aid in answering information needs.

With the fields to be indexed defined, it was now necessary to assign them types. A type in Solr is defined with analyzers, which describes how the text in the database and the query should be broken down to be indexed, as the original values are kept intact. Each type is composed of a tokenizer and several filters, and it is applied to both the dataset and the query. For more details, it is possible to check the documentation present in their website [8].

For the title and content fields, we created a custom type, `articleText`, that has the following features:

- `StandardTokenizerFactory`, which splits the text into tokens.
- `ASCIIFoldingFilterFactory`, which converts non-ASCII characters into their ASCII equivalent, if possible, like turning ç into c.
- `LowerCaseFilterFactory`, which turns everything into lower case.
- `EnglishMinimalStemFilterFactory`, which converts all words into their minimal stem, so something like "dogs" would turn into "dog".
- `PorterStemFilterFactory`, which applies the Porter algorithm for stemming, so that tokens like "jump", "jumping" and "jumped" would all match to "jump", hopefully leading to better results regarding temporal actions

Ideally, these transformations make it easier for Solr to match with relevant articles, whilst ignoring some of the minor issues with the text.

For all other fields, we used the default `text_en` type included with Solr, first because they were not as relevant as the two main ones, so we wouldn't need to customize it as much, and second because the filters that it applies seem to function well with the remaining fields, like the authors, the tags and the sections. It is also worth mentioning that both tags and sections were considered multi-valued fields, meaning they represent a collection and not just a single item. This is important so that Solr can make the distinction when indexing them. A full description of the types and indexes can be found in Table 1.

| Field    | Type        | Indexed | Multi-Valued |
|----------|-------------|---------|--------------|
| title    | articleText | Yes     | No           |
| url      | text_en     | No      | No           |
| content  | articleText | Yes     | No           |
| author   | text_en     | Yes     | No           |
| date     | date        | Yes     | No           |
| tags     | text_en     | Yes     | Yes          |
| sections | text_en     | Yes     | Yes          |

Table 1. Field types and respective indexation

### 3.3 Answering the Information Needs

When putting such a system up to the test, it is key to evaluate it by seeing if it answers the underlying information needs of each query. To do this, we will use the needs mentioned in Subsection 2.7, and check what results will the system bring up. We will show the top 10 results that the system brings up, even though more might have appeared. We will first try the regular version of each query and then try and boost some fields or terms so that better results can be seen. The results that are considered relevant will also be highlighted in the different tables.

All queries, regular and boosted, will then be evaluated in their Precision at 10, Average Precision and their Precision-Recall curve. Precision at 10 was used instead of the standard Precision since most users are only interested in the first few results that show up, instead of the overall performance [10]. Average Precision on the other hand, is used to complement Precision at 10, as it does not depend on Recall. In order to calculate recall appropriately for the queries, we would have to check all the dataset for relevancy, which is not feasible, so we will not be using this metric.

#### 3.3.1 Information Need 1 - Search for possible alien planets

For the first information need, we intend to find some news related to possible alien planets. The chosen query was `alien planets`, as relevant articles should have these words on them. When boosting, the documents where this information appears in the title are seen as more favorable, as they are more likely to be relevant for the user. The Solr configuration is presented in Table 2 and the results are displayed in Table 3 and Figure 8.

| Regular              | Boosted                |
|----------------------|------------------------|
| q.op: OR             | q.op: OR               |
| qf: title<br>content | qf: title^3<br>content |

Table 2. Parameters used for configuring the 'Alien Planets' Query

#### 3.3.2 Information Need 2 - Articles authored by Jeff Foust in 2021

The second need tries to find out what articles Jeff Foust published in 2021. For this, we can once again take advantage of Solr's ability to filter results, and only include the ones published in 2021. With the boosted query, the author field was favored, as we want to find the main author of the article. The chosen query was `Jeff Foust`,

| Rank | Regular  | R | Boosted  | R |
|------|--|---|--|---|
| 1    | Despite Early Success, Kepler Far from Finding Another Earth           | Y | 'Orphan' Alien Planets May Be Common                                   | Y |
| 2    | Kepler Points to 50 Billion Planets in the Milky Way                   | Y | Alien Planet Has Strange, Methane-free Atmosphere                      | Y |
| 3    | NASA's Kepler Space Telescope Confirms Alien Planet in Habitable Zone  | Y | Astronomer Stands By Alien Planet Discovery Amid Doubts                | Y |
| 4    | 'Orphan' Alien Planets May Be Common                                   | Y | NASA's Kepler Space Telescope Confirms Alien Planet in Habitable Zone  | Y |
| 5    | Kepler Finds First Earth-sized Planet in Habitable Zone                | Y | Study Suggests Earth-size Alien Planets May Be Surprisingly Common     | Y |
| 6    | Study says Rogue Worlds May Outnumber Planets with Suns                | Y | Hobbled Kepler May Resume Alien World Search                           | Y |
| 7    | Light from Alien Super-Earth Seen for 1st Time                         | Y | Light from Alien Super-Earth Seen for 1st Time                         | Y |
| 8    | Alien Planet Has Strange, Methane-free Atmosphere                      | Y | NASA's Kepler Craft Begins New Search for Alien Worlds                 | Y |
| 9    | SETI's Allen Telescope Array Resumes Listening Duties                  | N | Newly Discovered Planets Could Support Life                            | Y |
| 10   | For First Time, Alien Planet's True Color Revealed: 'Deep Cobalt Blue' | Y | For First Time, Alien Planet's True Color Revealed: 'Deep Cobalt Blue' | Y |
| P@10 | 0.9  |   | 1  |   |
| AvgP | 0.927234   |   | 0.991536   |   |

Table 3. Results for the 'Alien Planets' Query

as it should be enough to search for the author's name. The Solr configuration can be seen in Table 4 and the results are available in Table 5 and Figure 9.

It is important to highlight that, in this query, we could actually get the most relevant results possible, since we could actually determine all of the articles written by Jeff Foust in this time period by looking at the dataset. Nevertheless, the results from Solr could be, and in fact were, very distinct.

We can see a clear difference between the regular and boosted queries in the beginning. With the author field favored, Solr could determine the articles authored by Jeff Foust with a lot more precision, and keep the recall high, maintaining a perfect curve during the tests. In this case, it means that the query actually did retrieve all of the articles within that time period, so it behaved as best as it possibly could.

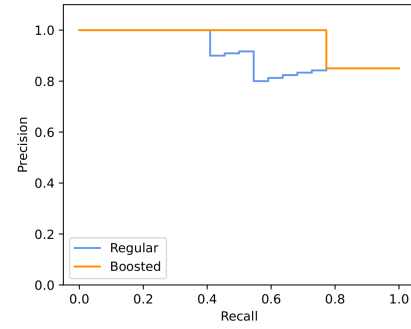


Fig. 8. Precision-Recall curves for the regular and boosted 'Alien Planets' queries

| Regular   | Boosted   |
|---|---|
| q.op: AND   | q.op: AND   |
| fq: date[2021-01-01T00:00:00Z TO 2021-12-31T00:00:00Z]; | fq: date[2021-01-01T00:00:00Z TO 2021-12-31T00:00:00Z]; |
| qf: title<br>content<br>author                          | qf: title<br>content<br>author^5                        |

Table 4. Parameters used for configuring the 'Jeff Foust' Query

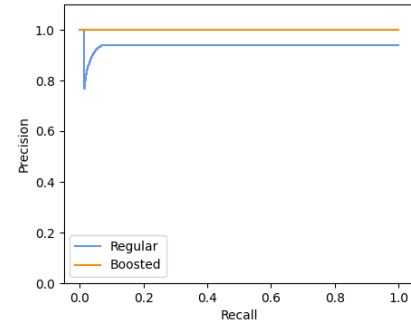


Fig. 9. Precision-Recall curves for the regular and boosted 'Jeff Foust' queries

### 3.3.3 Information Need 3 - Launches leaving from French Guiana

For the third information need, we intended to find news related to launches from French Guiana. For this, we should focus more on keyword based searching. The chosen query was "French Guiana Launch", as relevant articles should have these words on them. For the boosted query, the title was given the biggest boost, as it usually contains the most important keywords. The sections and tags attributes were also boosted, as they are the website's own mechanism of associating articles with relevant keywords. We also boosted the "Launch" term in order to show that we are more focused on results about launches than other news related to that Space Centre. The Solr configuration can be seen in Table 6 and the results are available in Table 7 and Figures 10

| Rank | Regular   | R | Boosted  | R |
|------|---|---|--|---|
| 1    | Foust Forward   Will Jeff Bezos kick-start Blue Origin? Does he need to?                                    | Y | White House commits to ISS extension   | Y |
| 2    | Foust Forward   The other human space-flight race   | Y | Firefly halts launch preparations after federal government seeks divestment of foreign ownership | Y |
| 3    | Foust Forward   The sky isn't falling (yet)   | Y | JWST begins sunshield deployment   | Y |
| 4    | Foust Forward   Dmitry in Dubai: Rogozin grabs the spotlight at the International Astronautical Congress    | Y | Virgin Orbit raises far less than expected from SPAC merger                                      | Y |
| 5    | Foust Forward   A schedule better suited for Artemis  | Y | FAA delays completion of...ip environmental review   | Y |
| 6    | Foust Forward   The space community could use some Perseverance   | Y | Virgin Orbit investing in startups as SPAC merger wraps up                                       | Y |
| 7    | Foust Forward   Inspiration and resilience in commercial human spaceflight                                  | Y | Ariane 5 launches NASA's James Webb Space Telescope  | Y |
| 8    | Foust Forward   The missing element of the first National Space Council meeting of the Biden administration | Y | ESA moving ahead on new Copernicus missions despite lack of U.K. agreement                       | Y |
| 9    | Foust Forward   Will the National Space Council remain effective in the Biden administration?               | Y | Crypto entrepreneur to go to space on New Shepard  | Y |
| 10   | Foust Forward   What senators should ask future NASA administrator Bill Nelson                              | Y | JWST launch marks only the start of a risky deployment process                                   | Y |
| P@10 | 1   |   | 1  |   |
| AvgP | 1   |   | 1  |   |

Table 5. Results for the 'Jeff Foust' Query

### 3.3.4 Information Need 4 - News regarding astronauts working in the ISS

For the fourth information need, we intend to find news related to the astronauts working in the International Space Station. For this, we also figured that keyword searching would be the best bet. As such, we used the query "Astronaut working ISS". When boosting, we boosted matches on the title and tags, as those are the places where keywords usually occur. We did not boost sections, as these words were not broad enough to make for a section of the website.

| Regular  | Boosted   |
|--|---|
| q: French Guiana Launch                            | q: French Guiana Launch^10                                |
| q.op: AND  | q.op: AND   |
| qf: title<br>content<br>author<br>sections<br>tags | qf: title^10<br>content<br>author<br>sections^5<br>tags^5 |

Table 6. Parameters used for configuring the 'French Guiana Launch' Query

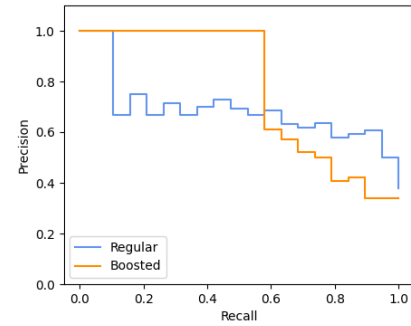


Fig. 10. Precision-Recall curve for the regular and boosted 'French Guiana Launch' queries

The Solr configuration can be seen in Table 8 and the results are available in Table 9 and Figure 11.

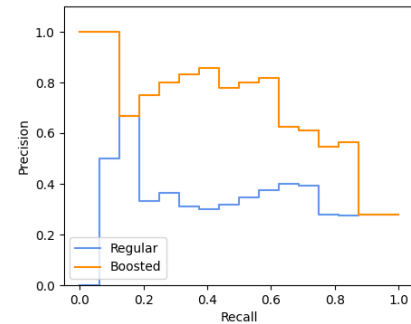


Fig. 11. Precision-Recall curve for the regular and boosted 'Astronaut working ISS' queries

### 3.4 Evaluating the Queries

Overall, the results are promising. All queries successfully achieved a high level on all metrics, and some even have perfect results. The boosts seem to be effective as well, improving the Precision and keeping the Average Precision steady on all queries.

| Rank | Regular  | R | Boosted   | R |
|------|--|---|---|---|
| 1    | Russia halts Soyuz launches from French Guiana                             | Y | Russia halts Soyuz launches from French Guiana                            | Y |
| 2    | Arianespace suspends French Guiana launches amid coronavirus response      | Y | Arianespace suspends French Guiana launches amid coronavirus response     | Y |
| 3    | French space agency pledges 10-million-euro boost to French Guiana economy | N | Soyuz launches French reconnaissance satellite in final 2020 launch       | Y |
| 4    | Airbus Ships Measat-3b to French Guiana Launch Site                        | Y | Airbus Ships Measat-3b to French Guiana Launch Site                       | Y |
| 5    | Soyuz launches French reconnaissance satellite in final 2020 launch        | Y | Arianespace launches for first time since French Guiana protests ended    | Y |
| 6    | Launch Activity Hits 20-year High in 2014                                  | N | French Guiana accord sets stage for Arianespace to resume launches        | Y |
| 7    | Arianespace launches for first time since French Guiana protests ended     | Y | Soyuz Rocket Launches Second Batch of O3b Satellites from French Guiana   | Y |
| 8    | French Guiana accord sets stage for Arianespace to resume launches         | Y | Soyuz Launches French Pleiades Imaging Satellite                          | Y |
| 9    | Eutelsat satellite returned to factory as French Guiana unrest continues   | N | Le Gall confident French Guiana launches will resume "in the coming days" | Y |
| 10   | Guiana Space Center launches to resume in June                             | Y | Guiana Space Center launches to resume in June                            | Y |
| P@10 | 0.7  |   | 1   |   |
| AvgP | 0.625  |   | 0.625   |   |

Table 7. Results for the 'French Guiana Launch' Query

In query 2, for example, the results are indeed improved and they show that having a well-defined document model can be a great help in determining the best results. This query was more targeted at the mechanism Solr uses to determine relevancy. Although the overall score was better on the boosted query, the regular one still managed to show some good first articles, which is the most important feature of a search system.

The first query shows that boosting drives different results upwards, and often they are more relevant. Overall, the results on this query were as expected, and seeing both the regular and Boosted PR curves in Figure 8 at 1 up until the 0.4 recall mark indicates that

| Regular  | Boosted   |
|--|---|
| q.op: AND  | q.op: AND   |
| qf: title<br>content<br>author<br>sections<br>tags | qf: title^10<br>content<br>author<br>sections<br>tags^5 |

Table 8. Parameters used for configuring the 'Astronaut working ISS' Query

| Rank | Regular   | R | Boosted   | R |
|------|---|---|---|---|
| 1    | Saber Astronautics to work with Axiom to bring Australian astronauts to space station | N | Demo-2 astronauts get to work on ISS  | Y |
| 2    | Demo-2 astronauts get to work on ISS  | Y | Pace of work put strain on private astronaut mission to ISS                           | Y |
| 3    | Pace of work put strain on private astronaut mission to ISS                           | Y | Saber Astronautics to work with Axiom to bring Australian astronauts to space station | N |
| 4    | Astronauts Repair Space Station Satellite Deployer                                    | Y | Soyuz launches Japanese private astronauts to ISS                                     | Y |
| 5    | NASA still working with Russia on ISS seat barter agreement                           | N | Soyuz spacecraft set to launch astronauts to ISS                                      | Y |
| 6    | Japan to recruit first new astronauts in 13 years to support Artemis program          | N | NASA astronaut may have extended stay on ISS  | Y |
| 7    | Virgin Galactic to work with NASA on private orbital spaceflight experiences          | N | Astronaut preparing for ISS mission with reduced crew                                 | Y |
| 8    | NASA working with cosmetics company on space station commercialization                | N | Two NASA astronauts to get extended ISS stays   | Y |
| 9    | Scientists want NASA and ESA to work together on a Europa lander mission              | N | NASA astronaut still baffled by removal from ISS mission                              | N |
| 10   | Private astronaut mission cleared for launch  | Y | ISS operations remain normal ahead of private astronaut mission                       | Y |
| P@10 | 0.4   |   | 0.8   |   |
| AvgP | 0.559   |   | 0.558   |   |

Table 9. Results for the 'Astronaut working ISS' Query

good results do show up at the beginning, which is key since most users only look at the first few results.



In the last two queries we can see some interesting results. For the third one, the precision-recall curves shown in Figure 10 present a sharp difference, with the boosted query far outperforming the first one, showing that term and field boosting do improve results. Meanwhile, the curves in Figure 11 show a stark difference between regular and boosted queries. Since only fields were boosted, we can infer that boosting fields is advantageous in various scenarios, meaning they are a sound option to use in all queries.

Whilst all queries are satisfactory, work can be done on improving the results. For example, the sharp drop seen in Figures 10 and 11, indicates that documents stop being relevant as the list continues. After investigating the issue, it appears that most of the problems seem to stem from the definition of relevant instead of the search system itself. Since we cannot feasibly have information about all the relevant documents in the dataset for each query, having these approximate results is a good indication that the system is going in the right direction.

## 4 DEVELOPING A SEARCH SYSTEM

Creating a complete search system does not simply entail cleaning the data and setting up a service to respond to queries. Even though those steps are key, the result needs to be refined to be a suitable answer.

Here, we will show the improvements done to create the best overall experience possible, for the general user. We start by describing the User Interface we developed, which was one of the biggest problems with the SpaceNews' service [15]. Next, we highlight the tweaks made in Apache Solr to make it as general as possible, whilst maintaining the relevancy of the results. Finally, we will outline what we did with Solr's MoreLikeThis tool [2], a feature that was used to implement a relevance feedback loop and generate more and more relevant results, based on user opinion.

### 4.1 Building a User Interface

To facilitate interaction with users, we decided it was vital to develop an easy-to-use User Interface. All the visual components were developed using Vue [6] and TailwindCSS [5], whilst the requests were made using Node.js [4] and ExpressJS [3] as a proxy.

The app was made with simplicity in mind and, as such, the home page only presents a search bar, much like most popular search engines. The search results show up as a query is submitted, and they are paginated to not create a bottleneck. Every time the user scrolls to the bottom of the page, more articles are fetched. In each search result, a "More like this" button can be found, which the user can take advantage of to select even more results based on the article he found relevant.

The only limitation of the interface is the text content of each article appears as a single block. This is due to the dataset itself not containing any newline characters in the content and, as such, it is impossible to have them appear in the article.

To make the life of the front end easier, we developed a server in ExpressJS that would act as a proxy to Solr. This server abstracted away all the intricate details of the Solr API, which doesn't possess the cleanest syntax, and forwards all the results to the front end, making the API much easier to use.

### 4.2 Improving Information Retrieval

In Subsection 3.3, several possible configurations were explored for creating queries with Solr. Although some were extremely successful, they were also very tailored to the specific query done. In a real-world case, making adjustments to the queries based on user input is not practical. With this in mind, we decided to apply the boosts on a more general level, taking into account the most common targets that users have when searching for articles, like the title and content. The best solution for this was field boosts, which focused on searching in the correct places by giving them more importance.

To prevent trying to guess the context of the query, the same boosts were applied to all queries. These boosts, shown in Table 10, were chosen based on the most common things users search for, based on general knowledge.

| Field    | Boost |
|----------|-------|
| title    | 10    |
| content  | 5     |
| tags     | 3     |
| sections | 3     |
| author   | 1     |

Table 10. Boosts applied to each field in the document model

Even though these may not be the best boosts for some queries, like Query 2 in Subsection 3.3.2, they work in the generality of cases and keep the relevancy high, which is more important when dealing with a large-scale system like this one. As it will be seen in section 4.3, these parameters provide similar results to the previously described ones, albeit with some exceptions when taking into account other features.

### 4.3 More Like This

One key feature that was missing from the original search system was the fact that there was no way for the user to tell the system what articles were relevant for a certain query, nor for the user to search for more similar articles based on results. This would be a great feature to have when trying to gather similar information in a certain topic.

To make this a reality in Solr, we implemented a form of "relevance feedback" in our system. We made use of Solr's MoreLikeThis [2], a feature that would generate new, ideally more accurate, results based on a result the user considered relevant, by searching for other items similar to that one in the dataset.

In order to create a fair baseline against this feature, we tested the results given by MoreLikeThis by comparing them to the system with the boosts described in Subsection 4.2. With this, we hope to see if the MoreLikeThis results can be considered a better alternative to the original ones.

To collect the MoreLikeThis results, we needed to determine a relevant article within the original query results from which to run the query. For this, we assumed that the first result was always relevant, and that the user would run that every time. Even though this may skew some results, like we will see in Subsection 4.3.1, it

should apply to the generality of cases, as the system shows good precision at 1 values in all previously ran tests, like the ones in Subsection 3.3.

#### 4.3.1 Answering Information Need 1 - Launches leaving from French Guiana

For our testing of MoreLikeThis we decided it was good to test it with already seen Information Needs. This way, not only could we have a base comparison against the new system, but also with the old, tailored one.

The first chosen one was the Information Need described in Subsection 3.3.3 and it gave some really unexpected results, as for the first time the query that would theoretically perform the best, had the worst result, by far. The results are available in Table 11 and Figure 12.

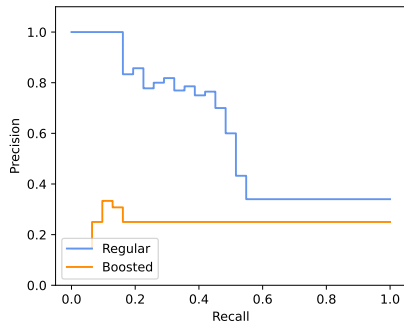


Fig. 12. Precision-Recall curves for the regular and MoreLikeThis 'French Guiana Launch' queries

#### 4.3.2 Answering Information Need 2 - Searching for possible alien planets

Considering the second test the Information Need from Subsection 3.3.4, we can start to see some more realistic results from MoreLikeThis, when applied to broader terms. Although the scores still indicate a curious turn of events, they seem to improve over the previous query. The results for this query can be found in Table 12 and Figure 13.

#### 4.3.3 Improved system evaluation

While initially we hoped the results would show that MoreLikeThis made a clear difference in improving the search system, the reality was much unforeseen. Having an outcome like the one from Query 1 show that not always can the system function well. Even though the results seem precise in the regular query at the beginning, as shown in Figure 12, they drop sharply as the recall levels advance. The error there can be attributed to the way MoreLikeThis works fundamentally, and the consequences the first result in the list had.

Since the first article talked about Russia extensively, in the context of launching from French Guiana, the system considered it was a highly relevant part of the article, and focused heavily on the Russia aspect, leaving behind the initial subject. Even though

| Rank | Regular  | R | MoreLikeThis   | R |
|------|--|---|--|---|
| 1    | Russia halts Soyuz launches from French Guiana                               | Y | Russia-Ukraine war raises questions for upcoming OneWeb launches                           | Y |
| 2    | Arianespace suspends French Guiana launches amid coronavirus response        | Y | The ending of an era in international space cooperation                                    | N |
| 3    | Airbus Ships Measat-3b to French Guiana Launch Site                          | Y | ESA suspends work with Russia on ExoMars mission   | N |
| 4    | Arianespace launches for first time since French Guiana protests ended       | Y | ESA says it's "very unlikely" ExoMars will launch this year                                | N |
| 5    | French Guiana accord sets stage for Arianeespace to resume launches          | Y | Russia looks to China for collaboration in space but faces isolation over Ukraine invasion | N |
| 6    | French space agency pledges 10-million-euro boost to French Guiana economy   | N | U.S. and Europe say space cooperation with Russia not affected yet by Ukraine crisis       | N |
| 7    | Soyuz Rocket Launches Second Batch of O3b Satellites from French Guiana      | Y | Rogozin puts poison-pill conditions on OneWeb Soyuz launch                                 | N |
| 8    | Le Gall confident French Guiana launches will resume "in the coming days"    | Y | Arianespace assessing impact of crewed Soyuz failure on satellite-launching variant        | Y |
| 9    | Eutelsat satellite returned to factory as French Guiana unrest continues     | N | ESA weighs options for replacing Soyuz launches  | Y |
| 10   | Ariane 5 launch facing further delays as a general strike hits French Guiana | Y | Rogozin delays decision on space station future  | Y |
| P@10 | 0.8  |   | 0.4  |   |
| AvgP | 0.634  |   | 0.625  |   |

Table 11. Results for the MoreLikeThis 'French Guiana Launch' Query

some of the results seemed relevant, especially towards the end, the initial, more important ones, deviate substantially from the original subject.

In the second query, although results do improve, they still seem lackluster when compared to the ones in Subsection 3.3. The fact that the precision is kept high in the beginning of Figure 13 is promising, but it does not keep up to what was expected of the system. One

| Rank | Regular   | R | MoreLikeThis   | R |
|------|---|---|--|---|
| 1    | 'Orphan' Alien Planets May Be Common                                  | Y | 'Orphan' Alien Planets May Be Common                               | Y |
| 2    | Alien Planet Has Strange, Methane-free Atmosphere                     | Y | Study says Rogue Worlds May Outnumber Planets with Suns            | Y |
| 3    | Astronomer Stands By Alien Planet Discovery Amid Doubts               | Y | Light from Alien Super-Earth Seen for 1st Time                     | Y |
| 4    | NASA's Kepler Space Telescope Confirms Alien Planet in Habitable Zone | Y | Alien Planet Has Strange, Methane-free Atmosphere                  | Y |
| 5    | Study Suggests Earth-size Alien Planets May Be Surprisingly Common    | Y | Study Suggests Earth-size Alien Planets May Be Surprisingly Common | Y |
| 6    | Hobbled Kepler May Resume Alien World Search                          | Y | Despite Early Success, Kepler Far from Finding Another Earth       | Y |
| 7    | Newly Discovered Planets Could Support Life                           | Y | CNES Space Telescope Finds Familiar Exoplanet                      | Y |
| 8    | Kepler Points to 50 Billion Planets in the Milky Way                  | Y | Kepler Points to 50 Billion Planets in the Milky Way               | Y |
| 9    | Light from Alien Super-Earth Seen for 1st Time                        | Y | Planet-Hunting Kepler Spacecraft Suffers Major Failure             | Y |
| 10   | NASA's Kepler Craft Begins New Search for Alien Worlds                | Y | Scientists Think Kepler Could Locate Habitable Exomoons            | N |
| P@10 | 1   |   | 0.8  |   |
| AvgP | 0.778   |   | 0.869  |   |

Table 12. Results for the MoreLikeThis 'Alien Planets' Query

favorable fact about MoreLikeThis is the fact that, although many articles become irrelevant, the relevant ones are articles that were not found in the original results.

Overall, the results indicate that MoreLikeThis works well as a tool to surface more relevant articles that have fallen in priority and, as such, it works tremendously well as a companion to the solid base system that already exists, and that most users will end up only using anyway. For those that wish to extend or go deep on a specific topic, having a service like this can aid in bringing up relevant, newer information on the topic.

## 5 CONCLUSION

This paper addressed three key steps in developing a full information retrieval system. First, a set of data was picked and cleaned to form a pattern and create an organized structure, using a reproducible pipeline for all this. Then, a search tool was used to explore this collection, and answer some common information needs that a user

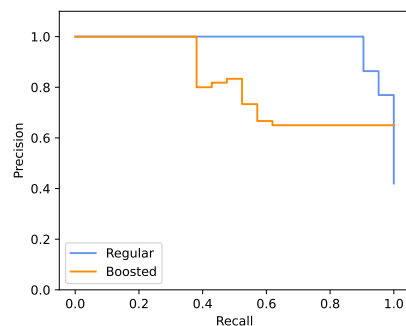


Fig. 13. Precision-Recall curves for the regular and MoreLikeThis 'Alien Planets' queries

would have, in this context. Finally, the modifications were fixed to bring the most relevant results possible to the most amount of queries, along with a companion service that can bring aid for those that wish to go deep in a single topic.


In the future, further improvements to the ranking system could be employed, namely using a form of classification model that would learn from user patterns and order the results not according to pre-determined algorithms, but by using data and Machine Learning as principal feature in determining relevancy. Nonetheless, the resulting evaluation that was performed shows that these alterations were effective in most cases, with known exceptions, making this system a complete and solid search tool for completing our goal and creating an apt replacement for the SpaceNews' current Information Retrieval system.

## REFERENCES

- [1] Apache Solr dev team. 2022. *Apache Solr*. Apache Software Foundation. Retrieved November 2nd, 2022 from <https://solr.apache.org/>
- [2] Apache Solr dev team. 2022. *Apache Solr MoreLikeThis*. Apache Software Foundation. Retrieved December 6th, 2022 from [https://solr.apache.org/guide/8\\_8/morelikethis.html](https://solr.apache.org/guide/8_8/morelikethis.html)
- [3] Express.js development team. 2022. *ExpressJS*. Express. Retrieved December 7th, 2022 from <https://expressjs.com/>
- [4] Node.js development team. 2022. *Node*. Node. Retrieved December 1st, 2022 from <https://nodejs.org/en/>
- [5] TailwindCSS development team. 2022. *TailwindCSS*. TailwindCSS. Retrieved December 9th, 2022 from <https://tailwindcss.com/>
- [6] Vue.js development team. 2022. *Vue*. Vue.js. Retrieved December 9th, 2022 from <https://vuejs.org/>
- [7] Patrick Fleith. 2022. *SpaceNews Dataset*. Kaggle. Retrieved October 1st, 2022 from <https://www.kaggle.com/datasets/patrickfleith/space-news-dataset>
- [8] Apache Software Foundation. 2019. *Understanding Analyzers, Tokenizers, and Filters*. Apache Software Foundation. Retrieved November 10th, 2022 from [https://solr.apache.org/guide/8\\_1/understanding-analyzers-tokenizers-and-filters.html#for-more-information](https://solr.apache.org/guide/8_1/understanding-analyzers-tokenizers-and-filters.html#for-more-information)
- [9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge. <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- [10] Sérgio Nunes. 2022. *Evaluation in Information Retrieval*.
- [11] Sérgio Nunes. 2022. *Lecture notes in Information Processing and Retrieval*.
- [12] Lucas Ou-Yang. 2022. *Newspaper3k*. Newspaper3k. Retrieved October 7th, 2022 from <https://newspaper.readthedocs.io/en/latest/>
- [13] The pandas development team. 2022. *Pandas*. Pandas. Retrieved October 6th, 2022 from <https://pandas.pydata.org/>
- [14] Leonard Richardson. 2022. *Beautiful Soup*. BeautifulSoup. Retrieved October 4th, 2022 from <https://www.crummy.com/software/BeautifulSoup/>
- [15] The SpaceNews Team. 2022. *SpaceNews*. SpaceNews. Retrieved December 10th, 2022 from <https://spacenews.com/>

## A HOME PAGE



 To infinity, and beyond!

© 2022 PRI Group 33 . All Rights Reserved.

## B SEARCH RESULTS PAGE



---

Showing 10 of 3645 results

**SpaceX vs. the world**  
 Perhaps Elon Musk decided his appearance before the National Academies needed an X factor. When it was time fo...  
 Jeff Foust  
 December 18, 2021  
 blue-origin nasa spacex from-the-magazine
 [More like this](#)

**SpaceX launches SiriusXM satellite**  
 Updated 10:20 a.m. Eastern with Maxar announcement of post-launch contact with SXM-8. WASHINGTON — SpaceX...  
 Jeff Foust  
 June 6, 2021  
 falcon-9 maxar-technologies siriusxm spacex commercial launch-section
 [More like this](#)

**SpaceX launches Starlink satellites**  
 WASHINGTON — SpaceX launched another set of Starlink satellites April 28, its first since the FCC approved a modifi...  
 Jeff Foust  
 April 29, 2021  
 falcon-9 spacex starlink commercial launch-section
 [More like this](#)

**SpaceX launches Turksat 5A**  
 WASHINGTON — A SpaceX Falcon 9 launched a Turkish communications satellite Jan. 7 to start what may be the bus...  
 Jeff Foust  
 January 7, 2021  
 airbus-defence-and-space falcon-9 spacex turksat commercial launch-section
 [More like this](#)

**SpaceX launches SiriusXM satellite**  
 WASHINGTON — A SpaceX Falcon 9 launched a new spacecraft for satellite radio company SiriusXM Dec. 13 as the c...  
 Jeff Foust  
 December 13, 2020  
 falcon-9 siriusxm spacex commercial launch-section
 [More like this](#)

**SpaceX launches EchoStar 23**  
 WASHINGTON — A SpaceX Falcon 9 launched the EchoStar 23 satellite early March 16 on the rare mission that did n...  
 Jeff Foust  
 March 16, 2017  
 launch-section
 [More like this](#)

**Winds postpone SpaceX launch**  
 WASHINGTON — High winds in the upper atmosphere have led SpaceX to postpone an already delayed Falcon 9 lau...  
 Jeff Foust  
 March 1, 2016  
 falcon-9 ses spacex launch-section
 [More like this](#)

**SpaceX Confirms Google Investment**  
 SpaceX confirmed Jan. 20 that Google is taking a stake in the Hawthorne, California, firm, joining Fidelity in a \$1 billi...  
 Brian Berger  
 January 20, 2015  
 google spacex commercial
 [More like this](#)

**SpaceX heeds Ukraine's Starlink SOS**  
 This story was updated March 1 at 1:55 p.m. EST WASHINGTON — SpaceX CEO Elon Musk said Saturday that he's se...  
 Brian Berger  
 February 28, 2022  
 russia spacex starlink ukraine europe
 [More like this](#)

**SpaceX to launch Turksat 6A**  
 KIHAEI, Hawaii — Turksat will launch its first domestically built communications satellite on a SpaceX Falcon 9, the Tu...  
 Jeff Foust  
 September 18, 2021  
 falcon-9 spacex turksat commercial launch-section
 [More like this](#)

C ARTICLE PAGE



Q SpaceX

## SpaceX launches Starlink satellites

Jeff Foust - April 29, 2021

[falcon-9](#) [spacex](#) [starlink](#) [commercial](#) [launch-section](#)

WASHINGTON — SpaceX launched another set of Starlink satellites April 28, its first since the FCC approved a modification that allows the company to operate more satellites in lower orbits. The Falcon 9 rocket lifted off from Space Launch Complex 40 at Cape Canaveral Space Force Station in Florida at 11:44 p.m. Eastern. The rocket's upper stage deployed its payload of 60 Starlink satellites into low Earth orbit nearly 65 minutes later. The launch took place a day after the Federal Communications Commission approved SpaceX's request to modify its Starlink constellation. The modification will move 2,814 satellites originally approved for launch in orbits of 1,100 to 1,300 kilometers to orbits of 540 to 570 kilometers, similar to the 550-kilometer orbits used by existing Starlink satellites. SpaceX did not mention the FCC's decision in its webcast. However, it did discuss how it chose lower orbits for spaceflight safety, ensuring that satellites will deorbit within several years of the end of their lives. It also mentioned its work with the 18th Space Control Squadron, sharing data on the orbits of Starlink satellites for collision avoidance activities, as well as a recent agreement with NASA to coordinate maneuvers between Starlink and NASA spacecraft in low Earth orbit. "We are extremely proud of our efforts to not only provide internet access to the disconnected, but also ensure space remains a place where human spaceflight continues to grow," Jessie Anderson, host of the webcast, said. With this launch, SpaceX has now placed 1,505 Starlink satellites into orbit, of which 1,434 remain in orbit. The company was approaching its previous authorization of 1,584 satellites in 550-kilometer orbits when the FCC approved its license modification to allow more satellites in those lower orbits. The Falcon 9's first stage landed on a droneship in the Atlantic about eight and a half minutes after launch. The booster completed its seventh flight, which included launches of a GPS 3 satellite, the Turksat 5A communications satellite and five Starlink missions. SpaceX has been using the Starlink launches to push the limits of reusability of the Falcon 9 first stage. "There doesn't seem to be any obvious limit to the reusability of the vehicle," Elon Musk, chief executive of SpaceX, said at an April 23 NASA press conference after the Crew-2 launch. "We do intend to fly the Falcon 9 booster until we see some kind of a failure with the Starlink missions, have that be a life-leader." Musk's comments came after the first launch of a reused Falcon 9 first stage on a crewed mission. The Crew-2 launch used the same first stage that flew the Crew-1 mission the previous November. Musk said he and NASA have discussed what the optimal number of launches of a booster might be. "Do you want to be on a brand-new booster?" he asked. "You probably don't want to be on the life leader for a crewed mission, but it's probably good to have a flight or two under its belt." He suggested a "couple of flights" might be best for a booster launching a crewed mission. "It's a hard problem for a rocket," he said of reusability. SpaceX also used the launch to honor Michael Collins, the Apollo 11 astronaut who died earlier that day at the age of 90. "Godspeed Apollo 11's Michael Collins," the SpaceX launch director said as the rocket lifted off. "May the pursuit of exploration live on."

© 2022 PRI Group 33. All Rights Reserved.