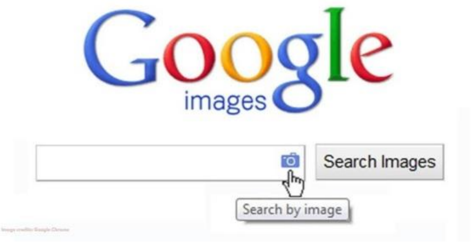
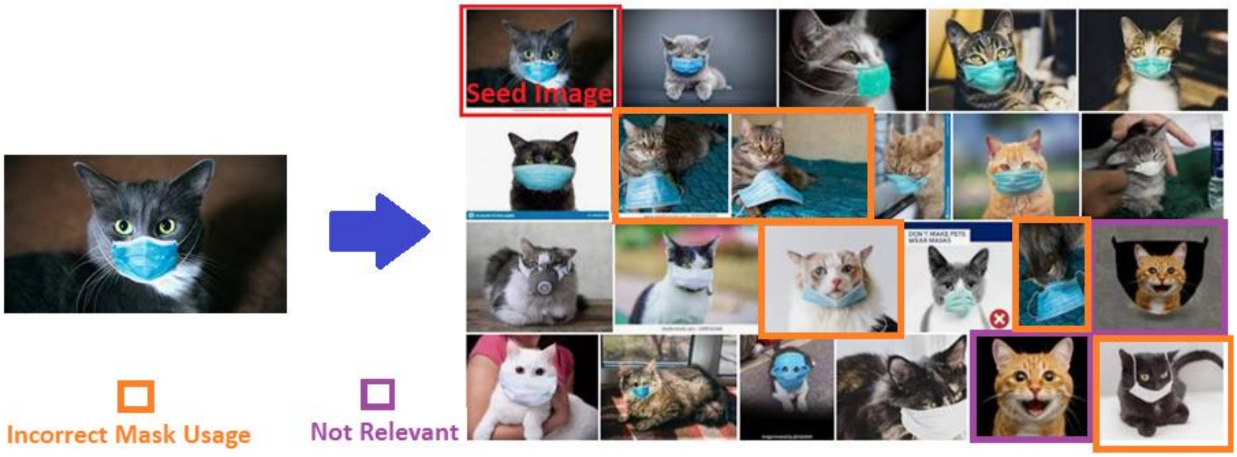
Guidelines for using the "Query-by-Document Tool"

The “Query-by-Document Tool” provided on CEPIMA’s webpage allows users to submit documents and find similar ones in the Scopus abstract and citation database. The principle can be compared to the “Query-by-Image” tool, with its most prominent implementation by Google:





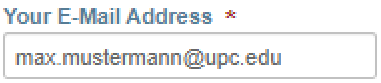
The “Query-by-Document” concept does exactly the same, but using documents instead of images. This document briefly outlines some guidelines on how to use the tool.

Requirements:

- Research Question. What kind of information are you looking for?
- Seed documents. Which are the documents that you want to find similar documents to?

Follow these simple steps to find documents that are relevant to your seed documents:

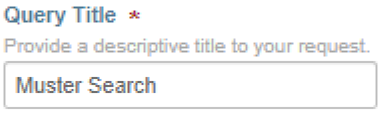
1. Visit the webpage (Link)
2. Enter your E-Mail Address



Your E-Mail Address *

max.mustermann@upc.edu

3. Provide a descriptive title to your request.
   (**Note**: Please avoid special characters such as “#”, “/”, “-“, “_” … spaces are fine)

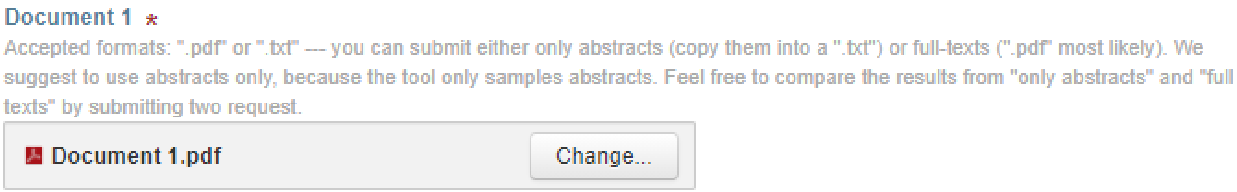


Query Title *

Provide a descriptive title to your request.

Muster Search

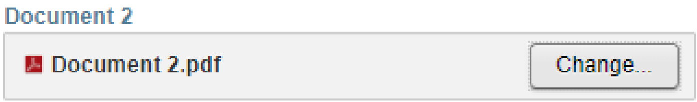
4. Upload your documents (“.pdf” or “.txt”). See description for decision between “only abstracts” and “full-text documents”



Document 1 *

Accepted formats: ".pdf" or ".txt" --- you can submit either only abstracts (copy them into a ".txt") or full-texts (".pdf" most likely). We suggest to use abstracts only, because the tool only samples abstracts. Feel free to compare the results from "only abstracts" and "full texts" by submitting two request.

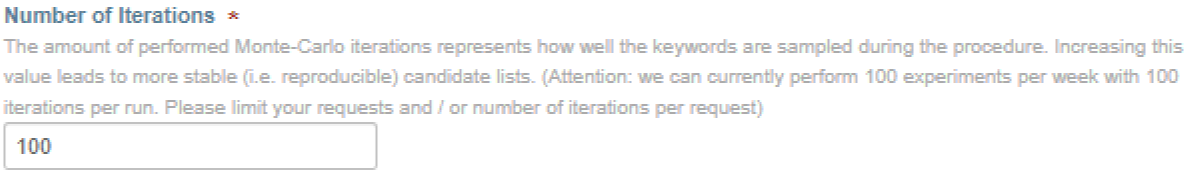
Document 1.pdf



Document 2

Document 2.pdf

5. Choose a number of iterations. Start with 100 to get an idea of how fast your seed documents converge to a stable list.

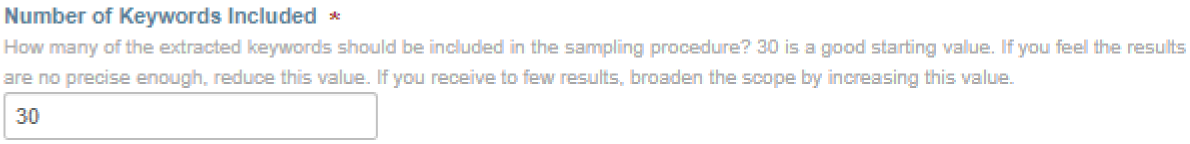


Number of Iterations *

The amount of performed Monte-Carlo iterations represents how well the keywords are sampled during the procedure. Increasing this value leads to more stable (i.e. reproducible) candidate lists. (Attention: we can currently perform 100 experiments per week with 100 iterations per run. Please limit your requests and / or number of iterations per request)

100

6. Choose a number of keywords. Start with 30 to receive the list of keywords and adjust this parameter to tune the results of your search.

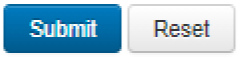


Number of Keywords Included *

How many of the extracted keywords should be included in the sampling procedure? 30 is a good starting value. If you feel the results are no precise enough, reduce this value. If you receive to few results, broaden the scope by increasing this value.

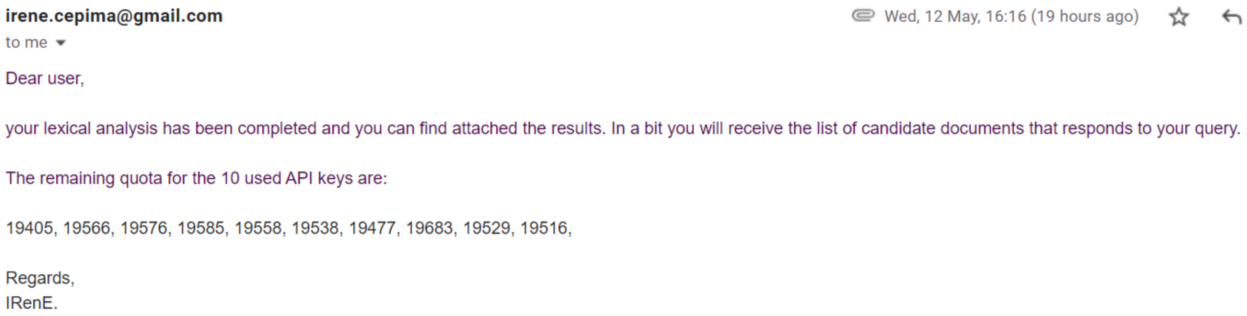
30

7. Submit your request. Congratulations, you are done!



1st Mail:

Now it is time to wait for the system to supply you with your requested information. Within a few minutes you should receive a mail with the following information:



irene.cepima@gmail.com

to me

Wed, 12 May, 16:16 (19 hours ago)

Dear user,

your lexical analysis has been completed and you can find attached the results. In a bit you will receive the list of candidate documents that responds to your query.

The remaining quota for the 10 used API keys are:

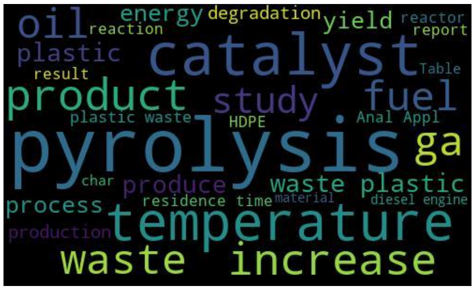
19405, 19566, 19576, 19585, 19558, 19538, 19477, 19683, 19529, 19516,

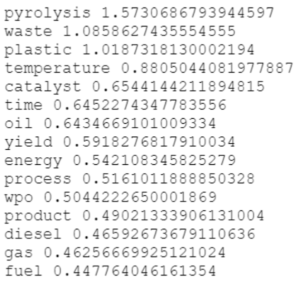
Regards,
IRenE.

This is the notification that your request has arrived and is being processed. If you do not receive this message, it might be the case that (1) the system has crashed and is offline or (2) other queries are being processed at the moment. In any case, if you receive the "Thank you" screen after submission, we will take care of your request and you will receive your results within 24 hours.

The ten numbers shown in the message inform you about the amount of request that can be performed in this week. As a rule of thumb: 100 iterations take 1,000 queries to Scopus, so these numbers will reduce by 1,000. If they approach 0, the system will be offline until the next week.

The 1st mail will contain a wordcloud with the extracted keywords and a list with all the extracted keywords with their associated relevance. The top N_{KW} (default: 30) keywords are now used in constructing queries to the Scopus® database.



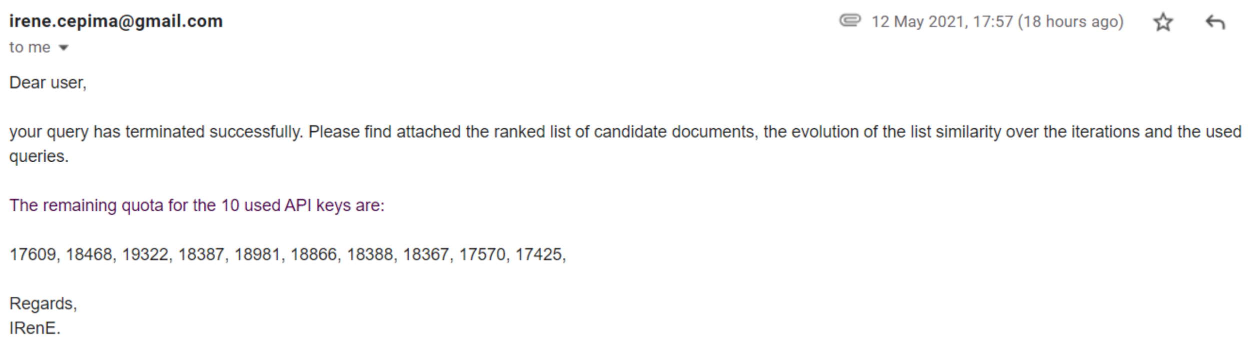


```
pyrolysis 1.5730686793944597
waste 1.0858627435554555
plastic 1.0187318130002194
temperature 0.8805044081977887
catalyst 0.6544144211894815
time 0.6452274347783556
oil 0.6434669101009334
yield 0.5918276817910034
energy 0.542108345825279
process 0.5161011888850328
wpo 0.5044222650001869
product 0.4902133390613100 4
diesel 0.4659267367911063 6
gas 0.4625666992512102 4
fuel 0.447764046161354
```

After receiving this mail, it takes then about $N_{it} \cdot 0.1$ minutes until you receive the 2nd mail with the resulting candidate list.

2nd Mail:

The second mail will look like this:



irene.cepima@gmail.com

to me

12 May 2021, 17:57 (18 hours ago)

Dear user,

your query has terminated successfully. Please find attached the ranked list of candidate documents, the evolution of the list similarity over the iterations and the used queries.

The remaining quota for the 10 used API keys are:

17609, 18468, 19322, 18387, 18981, 18866, 18388, 18367, 17570, 17425,

Regards,
IRenE.

It comes with three attachments: “ranked_candidates.xlsx”, “similarity_evolution.xlsx” and “sampled_keywords.xlsx”.

ranked_candidates.xlsx: This file contains the ranked list of candidate documents responding to your query. The list is sorted by column E “document-frequency”, that is the amount of times the document appeared in the N_{it} performed iterations. The document is characterized by its digital object identifier (DOI) and title.

similarity_evolution.xlsx: This file contains a list with an indicator of how stationary the list is. In each iteration, the updated list is compared to the previous list. Column C compares the top 1,000, column B the top 100, and column A the top 10 documents. Similarity is herein defined as:

$$Similarity(TOP\ N) = \frac{number\ of\ identical\ documents\ in\ both\ lists}{N}$$

When all three similarity indicators approach 1 it is a good indicator that N_{it} has been chosen appropriately high.

sampled_keywords.xlsx: This file contains the keywords that have been used in each iteration.