SIANI Master – Data Science in Engineering

Course Work 2023/2024

Scientific journal recommender to submit a publication

Aim

Currently the number of scientific journals is on the order of thousands, which means that in a given field there may be dozens of journals. Therefore, a dilemma that researchers face is deciding which journal to submit a scientific paper to. In addition to quality factors, an element to take into account is that the works published in said journal have a similar theme since it increases interest and visibility among other researchers in the field.

Therefore, the objective set out in this work is to carry out a Data Science project with the aim of creating an intelligent system that recommends a scientific journal to submit an article for publication. In this way, by giving the researcher the title of the work, the abstract and the keywords, the intelligent system will recommend the journal to which the article should be sent. The recommender will have learned this knowledge from examples of works previously sent to each of the candidate journals.

To do this, the techniques seen during the course related to preprocessing, obtaining models, evaluating results, and more specifically natural language processing, as well as others that the student considers necessary, will be used.

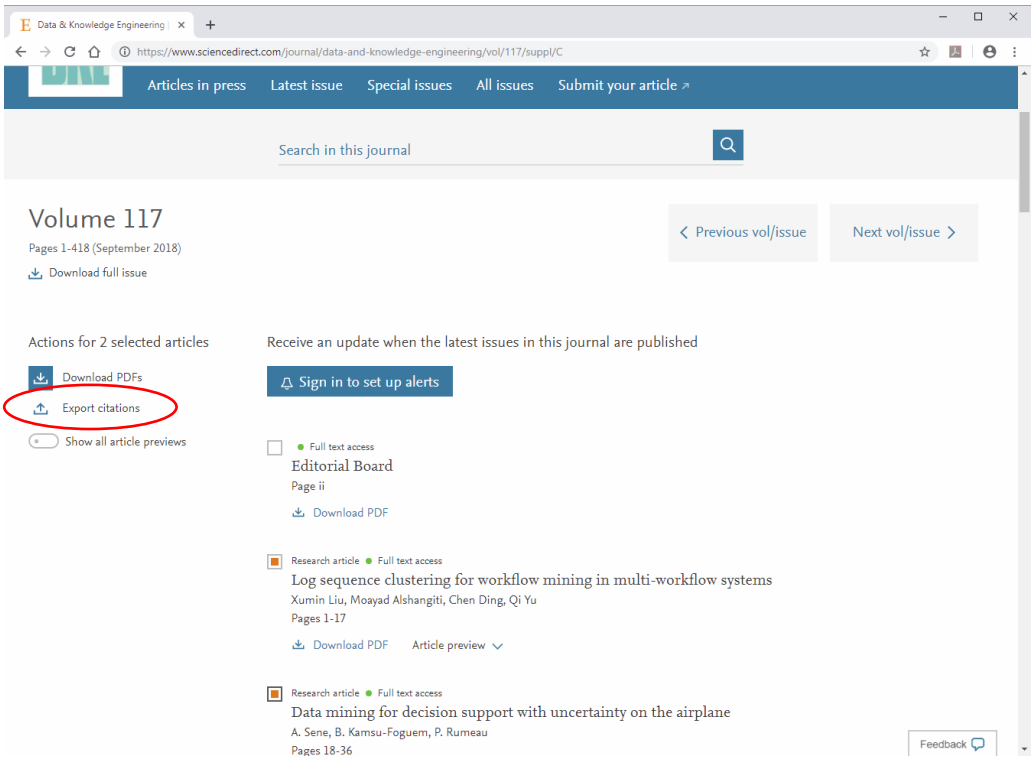
As a result of the course work, a report must be written explaining the process followed and justifying the design decisions made, and an analysis and discussion of the results obtained. The programs made in Python for the development of the work and the data sets used must also be delivered, as well as the presentation that will serve as a guide for the defense of the work.

Ambit

For the implementation of the intelligent system, a set of journals from the Elsevier publishing house will be used. The corpus or set of documents with which the work will be carried out will be the articles published in 2018, 2019, 2020, 2021 and 2022 in the indicated journals, using the following data: title, abstract (*abstract*) and keywords.

These data are obtained through the web portal of the publisher ScienceDirect (https://www.sciencedirect.com/) which can be accessed through the university library.

In order to download the basic data mentioned above, it can be done by exporting all the references of each issue to a RIS or Bibtex format, incorporating the abstract.



Development

As stated above, coursework will be completed using the Python programming language. For the text preprocessing phase, apart from the feature_extraction.text module of Sklearn, the package may also be considered *Natural Language Toolkit* (NLTK) available as free software, or similar. To obtain the models, the Sklearn library will also be used. For the optional part it will be done using the Keras or Pytorch development framework. However, the above does not exclude the use of other packages that the student considers necessary to carry out the work.

The development of the work consists of a **mandatory part** and one **optional part**.

- The mandatory part will be based on the classic approach that, from the data obtained from the journals, the term-document matrix will be obtained. Once the matrix is obtained, a document classification model(s) will be trained and validated. (80% of the grade)
- The optional part will be based on the application of connectionist techniques to obtain the recommender, with the student being the one to propose the solution that they consider most appropriate under this paradigm. As in the mandatory part, the model(s) for document classification must be trained and validated. (20% of the grade)

Delivery

The delivery of the work will be done through the virtual campus of the subject. The following must be delivered in a compressed file:

- Descriptive memory.
- PowerPoint presentation of the work done.
- Python code generated during the execution of the work.
- Data sets used in carrying out the work.

Defending

The work must be defended in a tutoring session whose date will be agreed between the student and the teacher.

Important dates

- **Delivery**through the virtual campus (deadline):**December 30, 2023**
- **Defending**(deadline)**: January 30, 2024**

Journals

1. Applied Ergonomics
2. Data & Knowledge Engineering
3. Expert Systems with Applications
4. Journal of Visual Communication and Image Representation
5. Pattern Recognition
6. Robotics and Autonomous Systems