EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-535-RC1, 2022

Comment on egusphere-2022-535

Anonymous Referee #1

Referee comment on "Potential of natural language processing for metadata extraction from environmental scientific publications" by Guillaume Blanchy et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-535-RC1, 2022

Interesting study regarding the use use natural language processing methods to extract information from the growing volume of scientific literature. The authors not only illustrate the use of different algorithms but also try to evaluate them numerically. In general, a well written manuscript. However, I think there is a lack of discussion and some of their objectives/aims are weakly met. The "relationship extraction" section is interesting and well written and the authors might want to put the same effort in the rest of the sections.

Comments

- Abstract: The beginning abstract seems a bit disconnected with the rest of the manuscript. Climate change is a hot topic but the paper itself is not related to that. I would suggest re-framing the abstract to match the content of the manuscript.

- Assessing the ability of an algorithm such as regex: I find this evaluation a bit estrange. The algorithms itself is infallible in the sense that it always finds what you tell it to find if it is present in the text. The algorithm is only restricted by the capacity of the user to generate valid regular expressions.

- Topic modelling: There is no discussion.

- How did you achieve your second aim (to illustrate the ability of topic classification to classify a new paper as relevant to a given topic)?

- You mention that topic modelling "can help identify knowledge gaps". How? Did you find any? If your aim is to present a practical workflow, perhaps you should guide the user to achieve that.

- Why did you select 6 topics instead of 9. You only mention that you are trying to maximise the coherence, which is higher for 9 topics.

- How does the number of topics might affect your workflow? Is selecting the highest coherence score infallible?

- Could you elaborate on how excluding monograms increased the coherence? From the term frequencies (Fig 7) I do not see many soil related terms, which seems strange. Perhaps they were ignored since their appeared as monograms? I do agree that bi and even trigrams are important but I have usually seen them added to a selection of monograms.