**Ginko workplan**

**Overview – the chronological timeline of ginko**

* Technical changes to the package
* Methodological changes to the package
* Testing
* Publishing R package + Tutorial

**Work Packages**

|  |  |  |  |
| --- | --- | --- | --- |
| **function** | **what** | **Who wants it? Who is doing it?** | **How long?** |
| searchScopus | Save Metamatrix from searchScopus also in Wd | Lisa/Jian | 30min |
| calculateModels | While loop always until 12 clusters and then what number of words was the highest | Henrik/Lisa | 1h |
| - | Understand try catch | ../Julius | 1h |
| - | implement try catch in entire script (occurred in beginning of textMatrix, 135) | ../Julius |  |
| processMetaDataMatrix | Use tf.idf instead of the first column without frequencies – right now, we only save 1 or 0 if a word occurs in a document | Julius & Lisa/ | 2h |
| calculateModels | Extract the 5% of the words (that make it into the analysis) as a file, give the user the option to alter them and read them in again (Henrik and the other changed quite a lot by hand..) | Henrik |  |
| searchScopus | Why error messages when retrieving from searchScopus (if more than 200 results) | Lisa |  |
| calculateModels | own Warning messages for dir create instead of unreadable Default | Lisa | 1h |
|  | Convert the binary wordmatrix from a matrix to a documentTermMatrix (package tm) as it is much better at saving sparse matrices. I see this as a very challenging task as we’d have to change the entire code (processMetaMatrix, calculateModels) | Lisa |  |
| ordinationCluster | Get the wrapper function to work without API key | Lisa/Lisa |  |
| createOrdinationPlot | Make plots work even if no metadata | Lisa/Lisa |  |
| getScopusMetaData | “If catching the metadata worked…” not only exemplary, try multiple ones as maybe after the third one we’ll get an NA? | Henrik/Lisa |  |
| - | All RDS files we save in the working directory: maybe save them in their own log folder and automatically delete older files? |  |  |
| (long term task) | change cluster to LDA or any other word/NLP model (LDA). We have a bag of words representation, there is other representations |  |  |
| calculateNetwork | Whenever a package is attached, it seems to prompt with a message, let’s catch these |  |  |
| (long term task) | Get rid of all commented code |  |  |

Some additional tasks I am not sure about:

* If the data comes from scopus, don’t catch the metadata
* Drop duplicate row names modeled data from search scopus
* At least 5 significant Indicator values per cluster
* Work package 1h: First Metadaten, then process meta…, also change in help files & ginkgo explained

**We have some questions by Hanna Weber:**

I don’t think all of them are necessary, additionally I am not sure if I understood all of them correctly.

- Is there a ranking of the most significant indicator words per cluster? How can I access that?

- How can I view the "agglomerative coefficient" of the cluster analysis?

- Could I see the *journals* of the papers as a table including their cluster? (Lisa: I think this is already what GinkoAnalysis$MetaMatrix does)

- I would like to access the exact number of publications per year that are also shown in the plot