Project 1
Text synthesis

The goal of the project is to provide a data text synthesis of a massive corpora of scientific articles about natural language processing.
This amount of data is not really big, just 44Mo, but is big enough to be too complicated to be processed by hand.

The way to generate the synthesis and what kind of synthesis is completely open !
You can use all tools and models seen during sessions.
For example you can focus on keywords and show a clustering of topics involved in NLP.
You can exploit time information and show evolution of topics or methods in NLP like apparition of deep learning techniques...
You can focus on abstracts or titles (less computational effort) or select a specific subset.

Your report must explain what technics/approachs you use, how you use them and the results obtained. If an approach don't work as planned you can show and explain (It will be very appreciate).
You can work in pairs of students. Your report must contain the names of students involved.
Your report must explain the logic of your approachs and results.
You can write in English or French.
Your report must contain your link to your Colab Notebook.
Your report must be deposited on DVO before Thuesday **8 december**.

Dataset (44Mo) : https://www.ortolang.fr/market/corpora/corpus-taln (download link at page bottom)

To parse XML with Python : https://docs.python.org/3/library/xml.etree.elementtree.html

Good Luck