2 - CRAWLING THE DATA

Q2.1- Which Wikipedia category is crawled in this script?

- The category of Wikipedia which is crawled is **Biology category**

Q2.2- What does this script output?

- This script, after extracting webpage from Wikipedia biology category, keep their title into wiki.lst file.

Q2.2- When running the script like python3 crawl.py, what should the file wiki.lst contain?

- The file wiki.lst contains all the webpage gotten

3 - DOWLOADING THE DATA

Q3.1- How many pages per batch is downloaded?

- 3000 pages per batch are downloaded

Q3.2- What API of wikipedia is used to download a set of pages?

- Wikidata API

Q3.3- How does the crawling work here?

•

Q3.4- By going to the API page in your browser, and reading the documentation paragraph, can you tell in what format the pages will be encoded?

- XML format

4 - PARSING THE DATA

Q4.1 - From the code, how are encoded the two matrices (i.e. what type of Python object)? What is the name of this encoding?

- The two matrices are encodes as a dictionary
- It's the sparse encoding

Q4.2 -Take a look at the database of Wikipedia documents in the dws folder, for example using the command vi or less. How are the links encoded in the wiki language?

- In the wiki language, links are encoded as internal link, external link and tooltips.