Suggested structure for the report

**Part 1 – Search engine**

**1. Crawler:**

1.1 Number of staff whose Google Scholar profiles are crawled (approximately).

1.2. Which parts of a Google Scholar profile is retrieved (e.g. the title of the publications only or any additional part)

1.3. Which pre-processing is performed before passing data to Indexer/Elastic Search (preferably, screenshot of data before passing to Indexer).

1.4. When the crawler operates, e.g. scheduled or run manually

1.5. Explanation of how it works step by step based on the written code (one or two pages should be enough)

**2. Indexer**

2.1. Whether you implemented the index or used Elastic Search

2.2. If you implemented it, which data structure is used (for example, incidence matrix or inverted index)

2.3. If you implemented it, whether it is incremental, i.e. it grows and gets updated over the time, or it is constructed from scratch every time your crawler is run

2.4. If you implemented it, some screenshots of its content

2.5. How it works step by step based on the written code (a few pages should be enough)

**3. Query processor**

3.1. Do you only support Boolean queries (using AND, OR, NOT, etc.) or accept keywords like Google does (without any need for AND, OR, NOT etc.)

3.2. If Elastic Search is used, how you convert a user query to an appropriate query for Elastic Search

3.3. If Elastic Search is NOT used, whether or not you perform ranked retrieval; if yes, the method used to calculate the ranks

3.4. Screenshots of the search results found for various queries

3.5. How it works step by step based on the written code and screenshots, where applicable (a few pages should be enough)

**4. (Optional)**

Any other important point you may want to mention, including any restriction, extras, issues.

**Part 2 – Subject classification (text classification)**

1. How training data are gathered, and how many records are there

2. Which machine learning method (e.g. Multinomial NB) has been used and how its performance is measured

3. How it works step by step based on the written code and screenshots of its answers for various inputs (a few pages should be enough).

4. (Optional) any other important point you may want to mention

**Appendix**

1. Complete source code

2. Training dataset used in Part 2

**Note:** This is only a suggestion and you may change it as you wish, as long as you provide enough evidence for meeting the requirements specified in the assignment brief. Also, please be generous and provide enough screenshots where applicable.