FIT5196-S2-2019 assessment 1

This is a group assessment and worth 25% of your total mark for FIT5196.

Due date: 11:59 PM, Sunday, August 25, 2019
Text documents, such as crawled web data, are usually comprised of topically coherent text data, which within each topically coherent data, one would expect that the word usage demonstrates more consistent lexical distributions than that across data-set. A linear partition of texts into topic segments can be used for text analysis tasks, such as passage retrieval in IR (information retrieval), document summarization, recommender systems, and learning-to-rank methods.

Parsing Raw Text Files

This assessment touches the very first step of analyzing textual data, i.e., extracting data from semi-structured text files. Each group is provided with a data-set that contains information about grants given for IP patent claims (please find your group file on the GDrive, i.e., **<your_group_number>.txt**). Each data-set contains information about several patent grants, e.g., patent title, patent ID, citation network, abstract etc. (**see sample_input.txt**). Your task is to extract the data and transform the data into the **CSV** and **JSON** format with the following elements:

1. **grant_id:** a unique ID for a patent grant consisting of alphanumeric characters.
2. **patent_kind:** a category to which the patent grant belongs.
3. **patent_title:** a title given by the inventor to the patent claim.
4. **number_of_claims**: an integer denoting the number of claims for a given grant.
5. **citations_examiner_count:** an integer denoting the number of citations made by the examiner for a given patent grant (0 if None)
6. **citations_applicant_count:** an integer denoting the number of citations made by the applicant for a given patent grant (0 if None)
7. **inventors:** a list of the patent inventors' names ([NA] if the value is Null).
8. **claims_text:** a list of claim texts for the different patent claims ([NA] if the value is Null).
9. **abstract:** the patent abstract text ('NA' if the value is Null)

The output, methodology, and documentation will be marked separately for this task, and each carries its own mark as follows:

Output (60%)
Carefully examine **sample_output.csv** and **sample_output.json** for detailed information about the output structure. Note that the sample outputs are your only ground truth, your output must exactly follow the sample outputs' structure and any deviation from this structure (e.g. wrong key names which can be caused by different spelling, different upper/lower case, ... or wrong hierarchy which can be caused by the wrong usage of '[' in the json files) will result in zero marks for the respective output. So please be careful.
Please note that for this task, **re** and **pandas** packages in **Python** are the only packages that you are allowed to use and the following must be performed to complete the assessment.

- Designing efficient regular expressions in order to extract the data from your dataset **<your_group_number>.txt**
- Storing and submitting the extracted data into a CSV file, **<your_group_number>.csv** following the format of **sample_output.csv** (you are allowed to use pandas for this)
- Storing and submitting the extracted data into a JSON file **<your_group_number>.json** following the format of **sample_output.json** (you are not allowed to use pandas for this)
- Explaining your code and your methodology in **<your_group_number>.ipynb**
- Documenting contribution percentage of each member of the group using the provided assignment-cover-group template. The submitted file should be named **assignment-cover-<your_group_number>.pdf** and must be signed by each member of the group.
- Documenting the details of the individual contributions of each member of the group using the provided reflective-diary template. The submitted file should be named **reflective-diary-<your_group_number>.pdf**
- All submitted files must be zipped into a single file **<your_group_number_ass1>.zip**

Note: Please specify how you divided different tasks between group members and summarise your management of completing the task in a group thoroughly. Failure to do so will severely impact your final mark.

Methodology (20%)

The report should demonstrate the methodology (including all steps) to achieve the correct results.

Documentation (20%)

The solution to get the output must be explained in a well-formatted report (with appropriate sections and subsections). Please remember that the report must explain both the obtained results and the approach to produce those results. You need to explain both the designed regular expression and the approach that you have taken in order to design such an expression.

Note: all submissions will be put through a plagiarism detection software which automatically checks for their similarity with respect to other submissions. Any plagiarism found will trigger the Faculty's relevant procedures and may result in severe penalties, up to and including exclusion from the university.