CE807 – Assignment 2 - Final Practical Text Analytics and Report

School of Computer Science and Electronic Engineering - University of Essex

Assignment Due at 11:59:59am on 02-08-2022

Electronic Submission

URL: https://www1.essex.ac.uk/e-learning/tools/faser2/

Please also see your student handbook for rules regarding the late submission of assignments

On Plagiarism

The work you submit must be your own. Any material you use, whether it is from textbooks, classmates, the web or any other source must be acknowledged in your work.

All submissions are fairly and transparently checked for plagiarism. Please make sure that you provide frequent citations. But also make sure that each sentence written is originally yours, i.e. the material is read, understood and the report is written using your own words and own language only. Do not copy and paste and rephrase copied text.

There are many different forms of what is considered plagiarism. For example, based on the feedback from the SAO officer, many students were not aware that, e.g. copying entire paragraphs without clearly identifying them as quote etc. is a form of plagiarism etc. Thus, please check back with your scientific writing module, before you submit!

Further note that also plainly reusing software code or merely slightly adapting existing software code and submitting as one's own fulfils the matter of plagiarism. Cite any code that you reuse, too.

In 2019, 20% of the submitted reports where plagiarised. There were also multiple cases of software code plagiarism. This number is too high and shall be 0% in 2020!

MOTIVATION: The task of eXtreme Multi-Label Classification (XMLC) deals with the problem of assigning *multiple* labels to a data object. The specific challenge is that the classification of a data object is conducted in *k* few labels given a pool of hundreds or thousands of labels, i.e., $k << n$. XMLC is of tremendous practical importance and industries and organisations in various fields such as retailing, web-content recommender, scientific libraries, new providers and others are dealing with it.

OBJECTIVE: After Assignment 1 focused on the theoretical aspects of text classification, the objective of this assignment is to get practical experience in designing, implementing, running and scientifically evaluating your own XML classifier. The dataset will be a collection of two very large-scale resources of scientific papers in economics and medicine. Each `document' in this dataset represents a record from a large document corpus. You can also use further datasets of your choice.

SUBMISSION, ASSESSMENT AND RULES

- This assignment counts towards 75% of the overall mark for CE807.
- The assignment is to be done individually.
- Be sure to put your registration number as a comment at the top of all code and other files. Furthermore, the assessment is blind, i.e. **do not put your name on any document or provide personally identifiable information**.
- The assignment must be submitted in a single zipped archive containing the following exactly three subfolders:

CE807/Assignment2/	All files
CE807/Assignment2/Task1	The report produced for Task 1. It is mandatory to include the source files (Word/LaTeX) and a PDF.
CE807/Assignment2/Task2	The commented code written to perform the classification Task 2 (with comments describing the features and further information external to the original datasets you used, e.g. some entity graph or linguistic resource).
CE807/Assignment2/Task3	The scientific report produced for Task 3. It is mandatory to include the source files (Word/LaTeX) and a PDF.

Important Note: You are free to use any software you like for this assignment. Your software should run first on your laptop or in one of the CSEE labs. You start with using a small portion of the dataset. Subsequently, you scale the experiments as much as possible by using the HPC of the university. For details see: https://hpc.essex.ac.uk/

Furthermore, you are encouraged to explore and exploit additional datasets for XMLC. You can also use any additional data sources such as thesauri and lexical resources.

Task Description: eXtreme Multi-Label Classification of Scientific Papers

Multi-label classification is one of the standard tasks in text analytics. The objective of the assignment is to perform an eXtreme multi-label classification, short XMLC. In an XMLC setting, there are *k* many labels from a large pool of *n* labels to be assigned to the data objects. The classification task is extreme in two senses: First, the number of *n* labels is very large with hundreds or thousands of labels. Second, there are only very few *k* labels to assign, i. e. it holds $k << n$. Thus, it is likely to have false positives.

Data description
We compiled two English datasets from two digital libraries, EconBiz and PubMed.

EconBiz

The EconBiz dataset was compiled from a meta-data export provided by ZBW - Leibniz Information Centre for Economics from July 2017. We only retained those publications that were flagged as being in English and that were annotated with STW labels. Afterwards, we removed duplicates by checking for same title and labels. In total, approximately 1,064k publications remain. The annotations were selected by human annotators from the Standard Thesaurus Wirtschaft (STW), which contains approximately 6,000 labels.

PubMed

The PubMed dataset was compiled from the training set of the 5th BioASQ challenge on large-scale semantic subject indexing of biomedical articles, which were all in English. Again, we removed duplicates by checking for same title and labels. In total, approximately 12.8 million publications remain. The labels are so called MeSH terms. In our data, approximately 28k of them are used.

Fields Both datasets share the same set of fields:

- **id**: An identifier used to refer to the publication in the respective digital library.
- **title**: The title of the publication
- **labels**: A string that represents a list of labels, separated by TAB.
- **fold**: For reproducibility of the results in our study: Number of the fold a sample belongs to as used in our study. 0 to 9 correspond to the samples that have a full-text, fold 10 to all other samples.

Getting started

You will first have to register with Kaggle (https://www.kaggle.com/) and download the files econbiz.csv and pubmed.csv from the dataset titled "Title-Based Semantic Subject Indexing". The direct link to the dataset is: https://www.kaggle.com/hsrobo/titlebased-semantic-subject-indexing.

Read the paper by Mai et al., 2018. Download the paper's source code and get used to both, the large-scale datasets and Python code.

Tasks
Your tasks will be as follows:

- In Task 1, you will – based on the basic review of text classification approaches in Assignment 1 – add a **review of literature on the state of the art in multi-label classification. Particular attention shall be given on XMLC approaches.** We expect this to be between two to four pages using ACM style plus references, which do not count to the page limit.

- In Task 2, you will **devise and train your own XML classifier using a suitable set of features** extracted from the given dataset. You can use any tools you like for extracting the features, e.g. any of the tools you have come across in the labs such as NLTK, ScikitLearn etc. You can also use the implementation given with the paper by Mai et al., 2018. **But please note that simply rerunning the code is not sufficient. You need to design your own new classifier, implement them and run the classifier and evaluate their performance on XMLC datasets.**
- In Task 3, you will **write a scientific report** explaining what you did in Task 2 together with some motivation (why you did it) and reflection (which alternatives did you consider, lessons learned etc). You will compare and contrast your results with what you found in Task 1, i. e. discuss and reflect on results in a scientifically sound manner. We expect this to be between eight to ten pages using ACM style plus references, which do not count to the page limit. **Please note: You can integrate the report of Task 1 in Task 3. In this case, indicate this clearly in the report of Task 3.**

Organisation and content of the report shall follow a scientific paper

As template for your report, please use the ACM style for conferences, preferably using LaTeX. There is even an Overleaf version of it available.

You can find the template here: https://www.acm.org/publications/proceedings-template

Important Note: It is mandatory to use the ACM style for formatting the results for reasons of comparability of the different reports being submitted.

_____ TASKS _____

TASK 1: XMLC literature review (report of 3-4 pages in ACM style plus references)

Whenever you develop a new classifier or other text analytics software, you have to show that your system outperforms the state of the art. The goal of this part of the assignment is to explore the landscape of XMLC approaches. This review should highlight what multi-label classification algorithms are out there and which approaches and features have led to the best classification accuracy.
In Assignment 1, you have already identified simple approaches that can easily be implemented and which serve as **simple baseline systems**. An example of a baseline system could be one that simply uses tokens as features. We would expect that your own XML classifier will outperform a simple baseline system but may also obtain or even exceed results of state-of-the-art approaches.

You can start by looking at the papers referenced in this assignment.

TASK 2: Devising and training your own classifier

This task will only be marked if you have completed Task 1.

For this task, you will develop your own XML classifiers using the features of your choice and the classification algorithms of your choice. The pre-processing steps and classification tools you have come across in the labs should get you started. Doing this will involve:

- Identifying the approaches and features you want to extract;
- Developing code to extract these features from text (and weigh them if you want to use more than simply binary features);
- Train a classifier that uses these features;
- Evaluate the performance of the classifier using a scientifically sound methodology.
- Initially start with a subset of the large datasets. Subsequently, scale the classifier to include more and more data until you are using the entire datasets or very large parts of if. You can request access to the HPC CERES with your student account, see: https://hpc.essex.ac.uk/

TASK 3: Report (8-10 pages in ACM style plus references)

This task will only be marked if you have completed Tasks 1 and 2.

Finally, you will write a report documenting what you did in Task 2 and comparing and contrasting your approach with what you discussed in Task 1. You should explain why you decided on the algorithms and features you used and how this compares to the state of the art. You should discuss the performance of your approach and reflect on what you have learned.

Bibliography

L. Galke, F. Mai, A. Schelten, D. Brunsch, and A. Scherp: *Using Titles vs. Full-text as Source for Automated Semantic Document Annotation*, Knowledge Capture (KCAP); Austin, TX, USA, 2017. Open Access URL: https://arxiv.org/abs/1705.05311

F. Mai, L. Galke, and A. Scherp: *Using Deep Learning For Title-Based Semantic Subject Indexing To Reach Competitive Performance to Full-Text*, Joint Conference on Digital Libraries (JCDL); Fort Worth, TX, USA, 2018. Open Access URL: https://arxiv.org/abs/1801.06717

MARKING BREAKDOWN (out of 100%)

Task 1. XMLC literature review (20%)

- Appropriate coverage and contextualisation: up to 10%
- Critical discussion: up to 10%

Task 2. Design and implementation of a XML classifier (40%) - Task 1 required

- Training of XML classifier and features using a framework: up to 10%
- High-quality code including comments up to 10%
- Achieving state-of-the-art performance: up to 10%
- Scale classifier to very large datasets: up to 10%

Task 3. Report (40%) – Task 2 required

- Discussion of work carried out: up to 10%
- Comparison with the state of the art: up to 10%
- Lessons learned: up to 10%
- Material submitted as requested in ACM style etc: up to 10%