

***Department of Electrical Engineering and Electronics***

# **Project Specification Form 2016-2017**

***Final Year BEng (*ELEC340*) and Year 3 MEng (*ELEC440*)***

Student Name: YUNLU WEN Module: ELEC340 / ELEC440 (delete one)

Supervisor: Roberta Piroddi Student ID No:201138656

Project Title: Hidden topic discovery from articles using machine learning

### Project Specification

**A. Project Description and Methodology:**

(Overall view of the project with proposed route to realization i.e. what are the project aims and objectives and how you are going to do it?)

The main aim of this project is finding hidden topics for given articles using some machine learning techniques like Latent Semantic Analysis and k-means clustering, but not limited to these ones. To achieve this, a mathematic model which is able to represent the relationship of terms, documents and topics will be developed and trained with bunch of data. In addition, the performance of techniques under different circumstances will be analyzed. If enough time is available, some improvements on data acquisition, model structure and model training will be made. All the algorithms developed will be implemented in MATLAB (maybe some C++ and shell script as well).

**B. Project Tasks and Milestones:** (indicate the tasks and milestones that should be achieved and their expected dates e.g. understanding of theory, designs of circuits, construction of circuits, software specifications, working demonstrations etc.)

**Tasks:** (a task is a package of work that should be completed during a particular time period)

1 Understand the background

2 Data acquisition and process

3 Implement LSA and k-means clustering

4 Learn probabilistic version of LSA and EM algorithm

5 Try to test the new model using testing data set

6 Try to realize supervised clustering

7 Compare the performance between two techniques

8 Analysis the limitation of LSA, pLSA and try to improve them

(expand text box if necessary)

**Milestones:** (an objective that should be achieved by a particular date e.g. the completion of a task)

1 By the end of 4th week, data set should be processed.

2 By the end of 6th week, documents can be clustered by topic roughly

3 By the end of 9th week, training of pLSA model should be derived mathematically

4 By the end of 12th week, finish the implementation of pLSA, also test LSA and pLSA using new added data.

5 By the end of 16th week, supervised clustering solution should be provided.

6 For the rest of time, try to improve the performance of these two models (or come up with some ideas).

(expand text box if necessary)

**C. Project Deliverables:** (Indicate what should be completed at the end of the project e.g. this list should indicate what will be presented / demonstrated at the final bench inspections)

1 The topics of documents will be visualized.

2 Given an article, some similar articles can be recommended

3 The computer will be able to tell what are the topics are about, while there’s no need for human to read the articles with this topic

4 Provide improvement solutions to LSA and pLSA (at least some ideas).

(expand text box if necessary)

**D.** A sectionon Project Rationale and Industrial Relevance must be included in the preliminary report (deadline midnight Friday 14th October 2016). This should explain how and why the project was devised, e.g. it may be a project sponsored by a company or linked to a research project.

Student Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor’s Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

By signing this form, the supervisor and student are confirming that the project is of a sufficiently demanding nature that it is suitable for the individual project component of an accredited engineering degree and that a student, who is capable of producing a first class performance, will be able to demonstrate his/her capabilities in this project.