Manual classification of documents for the purpose of training a classifier is a tedious and expensive job. Motivated by this field, which is represented by a letter in the classification coding system. For example, a module assigned the classification scheme with three levels of classification: broad field, narrow field, and detailed field. Accordingly, the scheme uses a 3-digit code in a hierarchical fashion for classifying fields of education and training, such that the first digit represents 'broad field', the second digit represents the 'narrow field' and third digit represents the 'detailed field' of a given document. There are 9 broad fields, 35 narrow fields and about 80 detailed fields. We have extended the by adding a forth level of classification, subject field, which is represented by a letter in the classification coding system. For example a module assigned the classification code 'GSCB' would indicate that the module belongs to the broad field of 'Science, Mathematics and Computing', the narrow field of 'Computing', the detailed field of 'Information Systems' and the subject field of 'Databases', where the broad fields, narrow fields and detailed fields represent the branches of the upper three levels of the classification hierarchical tree, from top to bottom respectively, and the subject fields represent the leaves of the tree.

**SYLLABUS CLASSIFIER**

- The task of the syllabus classifier is to automatically assign a classification code to each individual course/module based on a pre-defined education classification scheme.
- The Higher Education Authority (HEA) and Higher educational institutions in Ireland use the International Standard Classification of Education (ISCED) to provide a framework for describing statistical and administrative data on educational activities and attainment in Ireland.
- The need for a more detailed national education classification standard than that provided by the ISCED has already been recognised by educational authorities within other jurisdictions. This has led some other countries to develop their own national classifications of education standards such as ACS in US and AQF in Australia.
- In order to standardise the classification of modules among Irish higher education institutions, HEA is currently considering the development of Irish Standard Classification of Education.
- The current version of the classifier classifies the syllabus documents based on a draft extended version of ISCED which will be replaced by the Irish Standard Classification of Education in future.
- The syllabus classifier component is based on the widely used Naïve Bayes algorithm. We are also experimenting with the application of a search engine in automatic classification of a training set or creating a fully unsupervised document classification system.

**TRAINING THE CLASSIFIER**

- A major difficulty of supervised-approaches for text classification is that they require a great number of training instances in order to construct an accurate classifier.
- Manual classification of documents for the purpose of training a classifier is a tedious and expensive job. Motivated by this position, the semi-supervised and unsupervised training methods are being researched to train a classifier with a limited number of training documents and no training documents, respectively.
- In this work, we experimented with a semi-supervised web-based approach to train a Naïve Bayes classifier used for classifying syllabus documents based on a hierarchical education classification scheme.
- The classification scheme used here is an extended version of ISCED represented in XML. The ISCED is a hierarchical scheme with three levels of classification: broad field, narrow field, and detailed field. Accordingly, the scheme uses a 2-digit code in a hierarchical fashion for classifying fields of education and training, such that the first digit represents 'broad field' and the second digit represents the 'narrow field' of a given document. There are 9 broad fields, 35 narrow fields and about 80 detailed fields. We have extended the by adding a forth level of classification, subject field, which is represented by a letter in the classification coding system. For example a syllabus code 'GSCB' would indicate that the syllabus module belongs to the broad field of 'Science, Mathematics and Computing', the narrow field of 'Computing', the detailed field of 'Information Systems' and the subject field of 'Databases', where the broad fields, narrow fields and detailed fields represent the branches of the upper three levels of the classification hierarchical tree, from top to bottom respectively, and the subject fields represent the leaves of the tree.

**RESULTS**

The performance of the classifier was measured using a held-out undergraduate and a held-out postgraduate syllabus documents. The micro-averaged precision achieved for undergraduate syllabi was 0.75 and it decreased to 0.90 for postgraduate syllabi. In general, the low micro-averaged precision values indicate that some syllabus documents may contain components belonging to different fields of study. For example, a syllabus document could be describing a course which contains both databases and web design components. Classifying such documents which belong to more than one class is more error-prone and requires the classifier to recognize the core component of the course. Since the number multi-component courses is substantially higher among the group of postgraduate courses, therefore the classification accuracy achieved for this group of syllabus documents is about 15% lower than undergraduate syllabi. Also it should be noted that this level of accuracy is achieved without using any manually classified training document to train the classifier.