Subject area classification via Text analysis at Linköping University
http://www.ep.liu.se/hsv_categories/index.en.asp

In Sweden, research is to be classified using SCB's classification system for subject areas (http://www.scb.se/sv_/Dokumentation/Klassifikationer-och-standarder/Standard-for-svensk-indelning-av-forskningsamnen-2011/). The system has three levels, a top level, a middle level and a detail level (the latter two levels sometimes referred to as 3 and 5-digit levels, respectively, corresponding to the length of the numerical codes for the levels).

Classification of research publications to the middle or detail level is becoming a necessity and experience has shown that manual approaches involving authors or librarians are unreliable or ineffective. Automatic or partially automatic techniques based on affiliation of the authors or the journal in which a publication is published are also unreliable (in both cases, because either affiliations or journals are too broad to indicate a unique detailed-level classification. The ideal is publication-based classification.

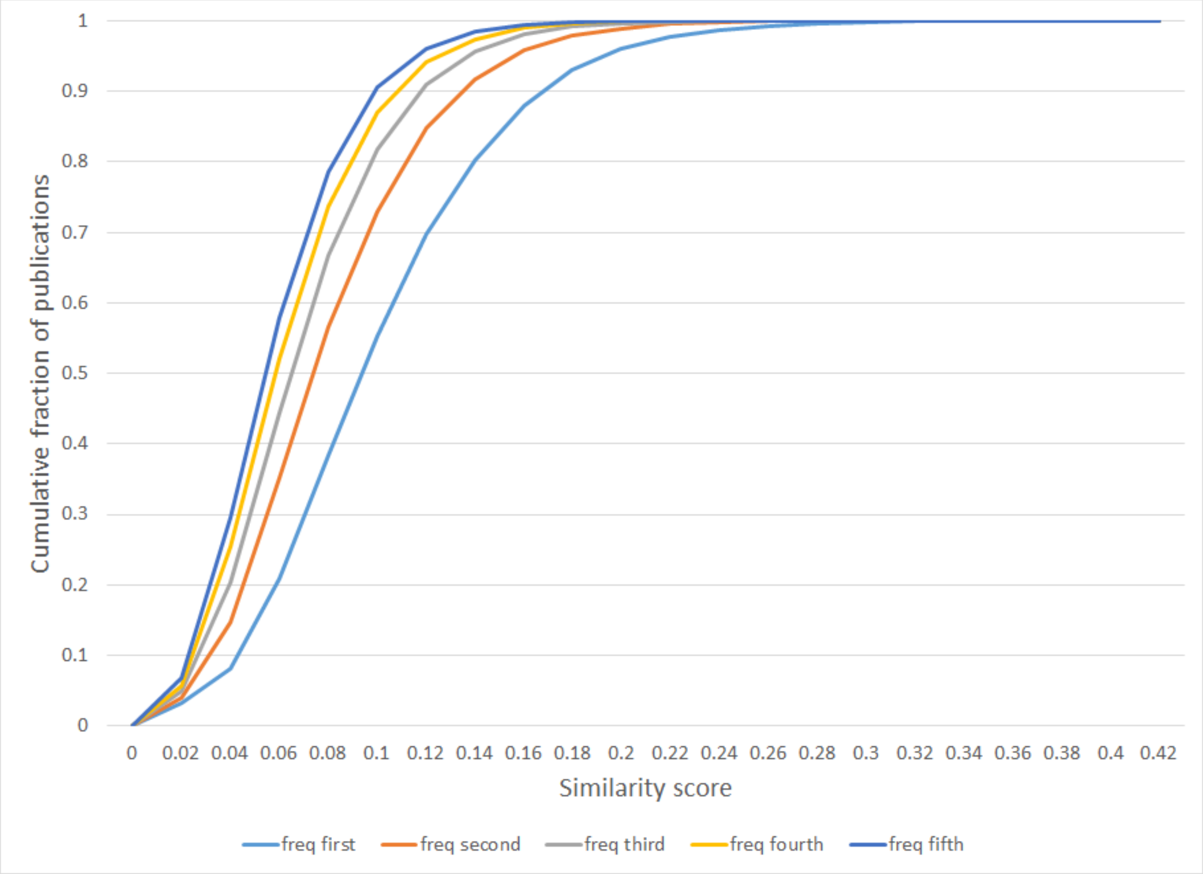
In this light text analysis is used to suggest detailed-level subject areas for publications based on a publication's abstract. Text analysis works by first creating a model, a definition of each of the detailed-level subject areas (some 255 in the SCB system), comprising a list of important keywords and their relative frequency for the subject area. To create such a model one needs a source of data comprising abstracts for publications together with the detailed-level classification for that publication. SwePub was used as the source of data. In the first model-building exercise, some 170,000 posts in SwePub had a combination of abstract and detailed-level subject area.

To create models from the SwePub data, one begins by collecting all the abstracts for each of the subject areas together; i.e. each subject area is connected with a large mass of text for subsequent analysis. For each subject area, insignificant or uninteresting words are removed (stop words) using a library of some 550 words. A porter stemming algorithm (http://tartarus.org/martin/PorterStemmer/) was then used to reduce words to their roots (e.g. ran, run, runs, running all come from the same root). The frequency of unique words is determined and a vector ordered from most frequent to least created comprising unique keywords and their frequency. The most frequent 150 words and their frequencies are retained and become the model or definition of a given subject area.

An abstract for a publication that is to be classified is analyzed similarly: stop words are removed and words are reduced to their roots using the same data and algorithms as above. Since abstracts have relatively little text, the model for an abstract rarely has 150 unique keywords. Experience has shown that having at least 50 is best for a reliable result.

The model for an abstract is then compared against the models for each of the subject areas by calculating the cosine similarity (https://en.wikipedia.org/wiki/Cosine_similarity) between normalized vectors for the abstract and a given subject area. Experience shows that the raw cosine similarity had a tendency to over predict in cases where there was relatively little overlap between the abstract and subject area models but where the overlap that did occur was for the most important keywords. To correct for this the cosine similarity is multiplied by a normalized count of the number of keywords that were the same for the abstract and subject area (normalizing is against the number of unique keywords in the abstract). This combined similarity measure varies between 0 and 1. Results normally fall in the range 0 to 0.25. The higher the number the better the agreement between abstract and subject area.

An important question is what result indicates that a subject area is a good description of an abstract. Ideally, one would collect data of the best similarity result and a “correct” subject area classification for many abstracts. This data has not been available. However, we do have the top five similarity results for some 35,000 matchings between abstracts and subject areas. Experience shows that the 5th best result only very occasionally gives a “good” subject area suggestion (i.e. a score for the 5th best result should usually not be high enough to be considered “good”) while the first is very often ok. Second, third and fourth best lie in a graded scale between the two extremes. Given this and plotting the cumulative distribution of the fraction of publications as a function of similarity score, gives the five curves in the figure below (e.g. the orange curve is for all of the 2nd best similarity scores). In some senses a cut off is arbitrary, but in a practical sense, a conservative cut off similarity score can be taken as 0.1, i.e. if a similarity score is 0.1 or higher, then a subject area suggestion would be accepted automatically. Here some 50% of 1st scores would be accepted while 90% of 5th-best scores would be rejected, in line with the experience above. If one took a cut-off of 0.08, then 60% of 1st scores would be accepted while 80% of 5th best scores would be rejected.


Cumulative fraction of publications
Similarity score
freq first
freq second
freq third
freq fourth
freq fifth