1. **Evaluation of Clustering and Categorization**

Our method of large content clustering and categorization needs to go through a thoughtful validation process taking into account appropriate means of qualitative and quantitative assessment. We have a need for validation of three types of results. First, we automatically decomposed texts into highly relevant keywords, which we call “concept tokens.” These keywords should be relevant, the error rate low, and the set of keywords produced for each text should sufficiently capture the main ideas of the text. Second, we match texts by their shared keywords. These matches should be of high quality with few errors. Finally, we validate the results of clustering which we believe produces a reasonable categorization scheme for the Map of Sciences. We perform this analysis in both a top-down and bottom-up fashion. This is done in two steps: 1) quantitative assessment of our clustering method based on measures of cohesiveness for shared concepts and minimized errors, and 2) qualitative assessment of our automated categorization by evaluating against some standard pre-existing categorization scheme.

**4.1. Clustering Validation**

<To be filled by AY from **Thesis section 4.3, Appendix B.2.** and <https://gist.github.com/andrewdyates/40c29902b4217123036f> >

1. Cluster coherence, and 2. Cluster stability.

What are stable clusters from a general statistical point of view? These clusters can be confirmed and re- produced to a high degree. To define stability with respect to the individual clusters, measures of correspondence between a cluster E and a cluster F like [1]

τ(E,F)= |E∩F| , γ(E,F)= |E∩F| , η(E,F)= |E∩F| (1) |E ∪ F| |E| |E| + |F|

**4.2. Categorization Validation**

We expect that our clustering should produce categories at a higher level that resemble (but not necessarily match) existing methods of categorization, and that these categories should be clearly informative and may produce new insights about how to best organize the domain of topics in the corpus categorized. Since this validation is performed at a much higher-level it is primarily qualitative in nature and requires human evaluation of whether certain dominant categories appear at the right level within the concept hierarchy and if these are being broken down into appropriate sub-categories thereafter or not. We believe that the role of statistical tests of significance to rule out the chance presence of certain subject areas at a higher-level is not as crucial at this level of analysis, as we do expect the clusters to be very diverse at the top levels. However, we do expect to see a pattern in the way the mix of subject areas at a top-level cluster, are broken down into their respective sub-categories at the next level. These, we believe should match within a certain margin of error, and an acceptable level of confidence, with the categorization performed by human experts.

We want to validate our hypothesis that our clustering method, in addition to producing a nice shared-concept network for large content, produces a categorization scheme that is comparable to that of human experts. To do this, we need to be able to satisfactorily answer the question: “as we walk down levels in our hierarchy, for documents belonging to a certain subject area or category at a higher level cluster, do we see an acceptable breakdown into its appropriate respective sub-categories at the corresponding lower-level clusters, that is consistent with how humans would categorize these documents into their respective sub-categories?” We also want to be able to draw some interesting statistics on the overall OA corpus such as: 1) What are the proportions of high-level subject areas (e.g. Biology v/s Math) in all Open Access Science, e.g. what percentage of all OA is Physics? Or, 2) Confirm that our top-level clusters (e.g. the first 8-15, roughly until level 4 or 5) correspond to high-level topics identified by experts.

Thus what is needed most is a believable categorization for comparison to ours, such that "blind categorizing" a set of samples from our corpus using some well-known categorization scheme such as the Dewey Decimal Classification (DDC) or Library of Congress Classification (LCC) gives us a set of hierarchical categories that we can use as "a truth" for comparison and validation with categories produced by our clustering method. Conversely, a set of documents that are categorized by our method also need to be presented to experts to see if they are in agreement with our labeling, in order to complete the cycle of validation for our scheme. The following sections describe our experimental setup for such validation.

**4.2.1. Validation methodology rationale**

Following steps outline our methodology to validate our primary hypothesis of *does the categorical breakdown provided by our process have a high degree of overlap with how experts break down higher-level subject areas into their respective sub-categories for a given set of documents*: 1) Have a subset of documents annotated by human experts into their respective hierarchical categorizations, using a pre-existing scheme such as LCC or DDC. 2) Align these set of hierarchical labels against the right “levels” in the tree produced by our clustering, corresponding to the right sub-clusters 3) Do frequency counts of the top 3 expert labels at those same levels for those sub-clusters 4) Compare the top 3 expert labels at a given sub-cluster, against the top 3 “concept tokens” generated by our method, at that same level, for that same sub-cluster, and see if there is a high degree of meaningful overlap between concepts – it is important to note here, that we may not have an exact overlap in the “terms” from the expert labels and our concept tokens for a sub-cluster, since the expert labels are more generic subject area labels and our concept tokens are mined from terms in the title and abstract, but what we would expect to see would be a high degree of “conceptual overlap” determined by *subjective evaluation* of the expert labels versus our tokens. For example, if the top 3 tokens for cluster 0.1.1.2 coming from the experts are {‘Life Sciences’, ‘Biology’, ‘Nucleotides’} and our top 3 concept tokens for the same cluster are {‘Nucleotides’, ‘Amino acids’, ‘RNA’}, and we see this type of conceptual match repeatedly for many different sub-clusters on different paths of the tree then we would conclude there is a high-degree of overlap between our categorization and that of the experts. 5) Conversely we repeat this process in the exact reverse fashion, where we pick a small subset of articles from our corpus, label them in a hierarchical way per the ranked concept tokens we have for those articles and have the experts evaluate “our” labels free of any prior notion or knowledge of any pre-existing categorization scheme and provide us a subjective evaluation or score between 1-10 for the level of their general agreement with our labels.

* + 1. **Expert Evaluation**

For the process outlined in section 4.2.1, we involved a group of human experts, namely Library Sciences specialists from three broad subject areas, viz. Medicine, Biology and the general non-Life Sciences domain that include areas such as Physics, Mathematics & Engineering, all of which we know to be representative of the articles in our corpus. We suggested the use of DDC and LCC for hierarchical labeling of the articles, but finally left it up to the experts to decide the best system to use for categorization, as we wanted the process to be truly unbiased.

When we approached the subject area experts for the annotation task they advised us that LCC and DDC were primarily “cataloging” systems that did not directly apply to OA articles such as ours, i.e. journal articles are not typically "catalogued" but rather are "indexed" with the subject headings. E.g. a journal title, like the “Journal of the American Medical Association” could be cataloged, but individual articles typically would not. The experts suggested that the Library of Congress system would be much better for classifying items, as most large academic libraries did not use Dewey Decimal Classification or DDC anymore, and that many librarians would likely not be anywhere near as familiar as with Library of Congress Classification or LCC either. For Medicine, we were told the largest used hierarchical classification for journal articles is MeSH or Medical Subject Headings. Thus, we were advised that classification systems that categorized articles by subject area such as LCSH or Library of Congress Subject Headings and MeSH or Medical Subject Headings, were much more appropriate for this task, hence we decided to proceed with primarily LCSH and MeSH for annotation of our articles. The important point again to be noted here is that neither of these systems have a “fixed” number of levels in a hierarchical label, hence this top-down analysis is highly qualitative in nature.

We perform two rounds of expert evaluation. For the first, we select a random set of 500 articles, from the same set that was used for token and edge validation for cluster evaluation in section 4.1. These articles are from a diverse set of subject areas. For the second round of evaluation we select a specific branch in our concept tree and choose a narrow set of Subject areas and qualitative assess the performance of our categorization.

4.2.2.1.The acid test - black box evaluation of a random sample

Method:

How sample created

How labelled by experts (what schemes used, why)

Include a summary of what Stephanie wrote about hierarchies

Breaking hierarchy into tags

Analysis

Preprocessing

Tree creation “algorithm”

Example - show snippet of tree

Variation of levels

Results and challenges

4.2.2.2. A weaker test - proof of “viability” (i.e. not correctness)

Method:

Validate a promising branch

Example

What could this show and why (you have to explain this since you picked this method)?

Analysis (just explain whatever is different from 2.1, if anything)

Results

Later on:

<Table 1. for hierarchical concept comparison goes here>: below is a fake e.g.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bin (Cluster) | Top 3 Labels by count | | | Top 3 concept tokens by count | | |
| 0.1.1.2 | Life Sciences=42 | Biology=  30 | Nucleotides=28 | Nucleotides=36 | Amino acids=30 | RNA=20 |
| 0.2.1.0 | … |  |  |  |  |  |

Table 2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bins (Cluster) | Our hierarchical labels | | | Expert score | | | | |
| Expert 1 | Expert 2 | | Expert 3 | |
| 0.0.0.6 | Life Sciences=42 | Biology=  30 | Nucleotides=28 | 9 | | 7 | | 8 |
|  |  |  |  |  | |  | |  |

Table 3.

<Do scoring here and include score table to measure level of agreement between experts>

Results and Future Work

Summary of validation