Project Proposal

Ruhui Shen A20410233

Name of our project: Hierarchical Dataless Categorization.
Goal: Extend the dataless classifier approach to extract hierarchical categories from Wikipedia, construct summary representations for the higher levels of categories and analyze if they can be used for hierarchical categorization.
Data: Wikipedia dump, A list of all available database dumps is available here: https://dumps.wikimedia.org/backup-index-bydb.html. and also, there are also SQL version online available.
I have already download both XML and SQL one. For English (enwiki), the download size is 13 GB at the time of writing, for Dutch (nlwiki) it is 1.3 GB.
Approaches:
I plan to use the Wikiextractor to extracting plain text from Wikipedia dumps and gather page id for each page we collect. Or Use categorylinks table like [[Category:Title|sortkey]] to extract link files per topic. Or the "category", "pages" table to get the whole point of the dataset. These methods should work due to the original structure of Wikipedia is regular. So, it is easy to form an overall organization, and there are some websites like https://en.wikipedia.org/wiki/Portal:Contents/Overviews could prove our results.
By using the Probability-proportional-to-size sampling method, I sample several thousands of pages and decide which categories to use, then use the SQL data to gather the actual pages. And if there's certain documents belong to multiples categories, I would use the page id to solve it, which means keep track of the page id and not add it to categorylinks table if we have already processed it.
Then we build up a tree-like hierarchical data structure with the data I collect beginning from the broad category and traverse through its sub category to get deeper.
Apply Explicit Semantic Analysis, a way to derive meaning based on Wikipedia, to build up word vectors for each concept and use both the pages directly under the category and pages under its sub category. The text is transformed into a vector of Wikipedia articles. The vectors of two different texts can then be compared to assess the semantic similarity (e.g., cosine) of those texts. According to some paper, this method does work. Compared with the previous state of the art, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgments: from r = 0.56 to 0.75 for individual words and from r = 0.60 to 0.72 for texts. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.
Finally, we could analyze if summary representations for the higher levels of categories can be used for hierarchical categorization.
PS: I got my idea after discussing with my teammates. Therefore, even if I used "I" in the individual proposal, it should be each member in the group (Of course they may have different idea but we have achieved basic agreement so far). Thank you!