Tasks

1. Text extraction from XML file is not working though it was working in earlier versions so I think when Farooq has implemented some more options(sometimes if we have no xml file and want to select direct text file then it also accept this file) in it then something wrong happened in the code. So you have to correct it.
2. Implementation of a similarity formula and for this formula it is needed to find path length and path depth (similarity formula file is attached) between words from Wiktionary, it is already implemented in WordNet, but it is giving some illogical error (I mean it is not finding similarity correctly, Farooq has I will change it because he has implemented and when I pointed out this problem he said yes I got it I will do it, now he has no time so maybe you say to him or in other case you will do)
3. On third form, when making compound words, system joins keyword to a very next noun, this noun may be on its (key word) right side or left side. But now we also want to make 3words compound words in which one word will be our keyword (KW) and other two nouns in any following form:

KW + NOUN+NOUN

NOUN+KW+NOUN

NOUN+NOUN+KW

1. But we will follow same rule like if before keyword there is “,” or “.” then don’t join back noun and same if after KW there is “.” Then also not join next noun with KW.
2. So after creating 3noun compound words, some rules will also consider for information gain and similarity measure, about these rules we will discuss on skype.

At the end, when we have list of words then manage categories module starts, in current version, we are taking some fixed categories(classes) with some concepts and then match our extracted list of words with these classes and put appropriate word to its respected class, but now we have to modify it.

1. It will take these classes from an OWL file (which will come from ontology) and then will perform same function which is already implemented.
2. On second form, there is a button of “add source files”, through this we transformed our word, pdf and ppt files into a text file (corpus file), it was working in earlier version but now not working but code will be there. You have to make it rich in function because like mostly online available pdf converter it merges alphabets are omitted during conversion and also have to apply ocr on it.,
3. At the end automate all this system, at this time we select one word as a keyword and then create a list of words against it but we have to automate this system it should take all words(one by one) and at the end we find many (according to key words) list of words.