ProMine Functioning

1. "ProMine” is basically a text mining tool. it takes an xml file as an input file and extract text from a tag "activity" that contain some text in paragraphs,

2. After extracting it makes a text file against each activity and place them a folder.

3. When user select this folder on second form, it shows on left side these text files

4. When click on any file, it performs different preprocessing tasks on these files

5. These preprocessing tasks are

1. Tokenization
2. Stopped words removal
3. Lemmatization/stemming
4. Frequency Count

6. At the end we find keywords of all files

7. When click on any word it extracts synonyms from WordNet and Wiktionary and merge them

8. Then we click on corpus file/glossary file and it performs all preprocess steps on that file and then comes to next form

9. Here system takes one by one all words from merged list (keyword and synonyms of keyword from WordNet and Wiktionary) and match in corpus file and make compound words (keyword +noun)

10. When this list of words come with frequency word then we apply information gain formula (I will provide you a paper to check either its formula is accurately implemented or not)

11. In this way rank this list of words

12. On this list apply our proposed similarity measure, its function will explain you on skype

13 after this we ranked this list of words according to our similarity measure and apply threshold on it and save it.

Tasks

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.

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)

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

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.

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.

It will take these classes from an OWL file (which will come from ontology) and then will perform same function which is already implemented.

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.,

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.