Data Extraction Using Semantic Similarity

Kavita S. Oza
Dept. Computer Science
Shivaji University, Kolhapur
kso_csd@unishivaji.ac.in

Swati S. Patil
Dept. Computer Science
Shivaji University, Kolhapur
swati.483@rediffmail.com

Abstract - Nowadays the Syllabus of most of the Courses in some extent is common. Means if there is an Engineering Student and other doing Msc(maths) then they both have the common subject named Fundamentals of Maths.In that Situation Instead of teaching that subject by more than one teacher,one teacher can teach them without any difficulty.This may save the time and improve the efficiency of the Teacher.This paper includes the dataset of syllabus of computer networks of more than one courses.The language used is python for processing the dataset and the result contains the syllabus which is common in the subjects.

Keywords - extraction,semantic,similarity

I. INTRODUCTION

Data mining refers to the process of extracting useful, non-trivial knowledge from data .The extracted knowledge is typically used in business applications, for example fraud detection in financial businesses or analysis of purchasing behavior in retail scenarios. In recent years data mining has found its way into many scientific and engineering disciplines. As a result the complexity of data mining applications has grown extensively. To address the arising computational requirements distributed and grid computing has been investigated and the notion of a data mining grid has emerged. Semantic Similarity is a metric defined over a set of documents or terms,where the idea of distance between them is based on the likeliness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation.Python is high-level, interpreted, interactive and object-oriented scripting language.It is designed for to be highly readable.Python is derived from many other languages including ABC,Modula-3, C, C++, Algol-68, SmallTalk and Unix shell and other scripting languages.It provides very high-level dynamic data types and supports dynamic type checking.

II. LITERATURE REVIEW

In current the Ontologies used with semantic similarity measures.Wordnet,SENSUS,Cyc KB are the general purpose ontologies.The Semantic Measure have the Categories named Structure-based measures,Information Content Measures,Feature based Measures,Hybrid Measures.Semantic Similarity is useful in many Web-related tasks.It is used in web related applications such as automatic annotation of Web pages,community mining and keyword extraction for inter-entity relation representation.Different approaches have been followed in the past decade to measure the semantic similarity and they are broadly categorized into two ways: Distance based approach & Corpus based approach.Snippets are a brief window of text extracted by a search engine around the query term in a document. They give the useful information regarding the local context of the query term. Semantic similarity measures defined over snippets have been used in query expansion, personal name disambiguation and community mining.

III. DATASET PREPARATION

Data set consist of syllabus of Computer Network of six courses-
MCA(Sci.),MCA(RI),B.E(Electronics),M.E(Electrical),B.E(Computer),BCAIII.There are some points of syllabus which are common in between them.That is some points are common in two courses,three courses.First the syllabus of all the subjects are copied in notepad file with .txt extension,one file for each course.Then by using the language python each course file is opened in read mode by using the file read technique.Then for each course-subject a set object is created. By using for loop in each part of the courses all the files were read one after another and added that file contents in a different variable for different courses each word at a time by removing the leading and trailing whitespaces and doing all the letters in each words as lowercase.Because the words in one file may be upper case and in another may be in a lower case.

After that those variables values are added one by one in a set object.One set object for each course subject.Finally the six sets were created which has the words as values listed.Then by using the intersection operation of set each set object is compared with different set object. After that files were closed.And if the common points were found then that points are displayed on the screen of python IDLE.The Python IDLE prompt runs this program and displays the output in the IDLE window.

Semantic Similarity:

fig a.

The above fig is showing the syllabus of BCAIII course under Shivaji University

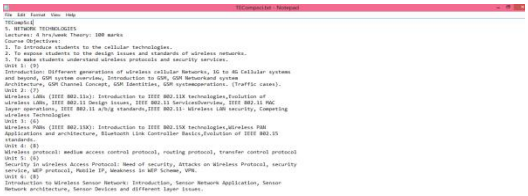

fig b

The above fig b shows the syllabus of T.E.Comp.Sci course under the Shivaji University course.

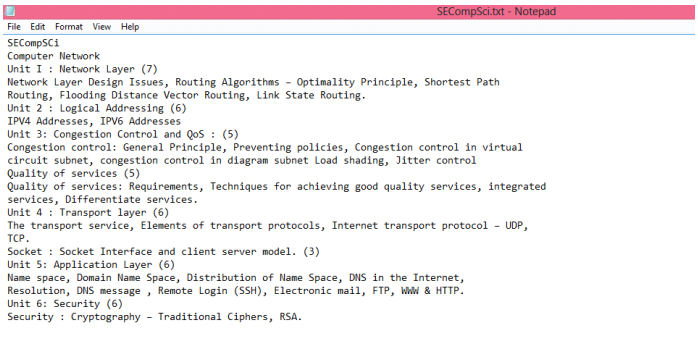

fig c

The above figc shows the syllabus of SECompSci course under the Shivaji University Kolhapur.

fig d

The above fig d shows the syllabus of MCASCi in Shivaji university kolhapur.

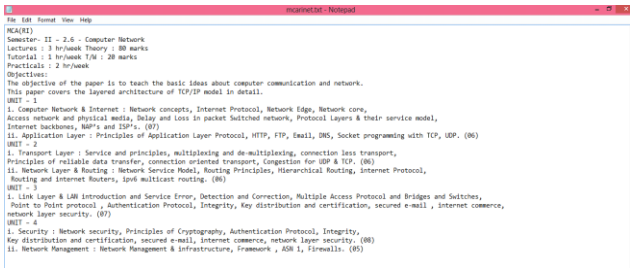

Fig. e

The above fig e shows the syllabus of MCA(RI) in shivaji University Kolhapur.

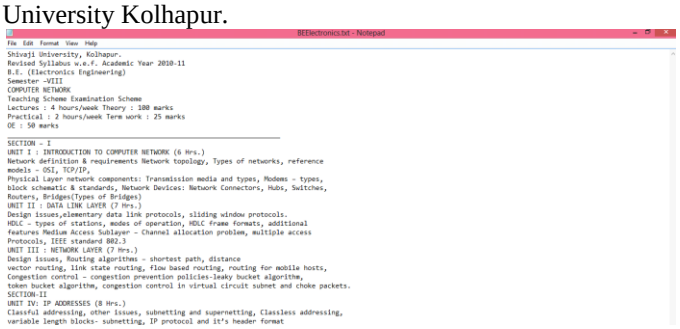

Fig f

The above fig f shows the syllabus of B.E(Electical Engineering) in Shivaji University Kolhapur.

Results :

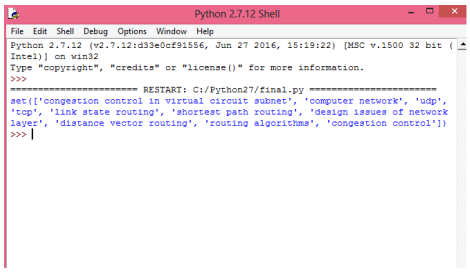

fig 1

The above fig1 shows the common syllabus in between the B.E(Electical) and SECompSci.

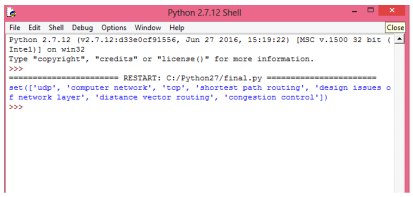

fig 2

The above fig2 shows the common syllabus in between the courses BE(Electrical),SECompSci and BCAIII.

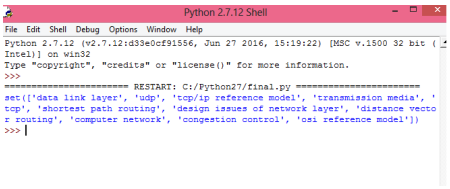

fig 3

The above fig3 shows the common syllabus in between the B.E and BCAIII courses.

fig4

The above fig 4 shows the common syllabus in between SEComp and BCAIII courses

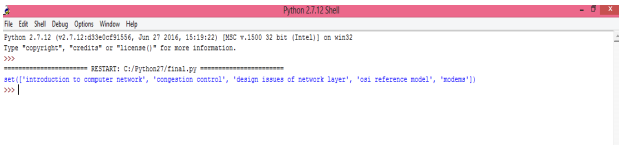

fig5

The above fig5 shows the common syllabus in between MCA(RI) and TECompSciin Shivaji University Kolhapur

IV. CONCLUSION

In many courses the syllabus of the subjects are common in some extent.So, If to sort out that common syllabus manually,instead we can use this method for listing the common syllabus.It will help the teachers to teach the same topic two times,only in once he can teach that points by combining the classes.Another is it will assist the universities or institutes that the one teacher can teach that common points by taking the lecture in his college and by taking the guest lecture in another college.It will save time also the workload of teachers may be decreased in some extent.

REFERENCES

1. http://www.tutorialspoint.com
2. https://www.python.org
3. Data Mining Techniques in Grid Computing Environments Editor Werner Dubitzky University of Ulster, UK
4. Description and Evaluation of Semantic similarity Measures Approaches by ThabetSlimani(Computer Science Department,Taif University & LARODEC Lab.
5. Measuring Semantic Similarity between Words Using Web Search Engines by DanuskaBollegala,Yutaka Matsuo,Mitsuru Ishizuka.
6. Semantic Similarity Measurement Between Words using Swd&Snippets byR.MenahaG.Anupriya.(IT and CSE department Dr.MCET, Pollachi, Coimbatore)