Walchand College of Engineering,Sangli

DEPARTMENT OF INFORMATION TECHNOLOGY

PROJECT PRESENTATION ON:

Watch Docs

Categorizing & Searching documents using NLP and Search Engine with efficient retrieval

Presented by:

Abhinav Shirur	2014BIT004
Sujit Ghatage	2014BIT007
Akash More	2014BIT010

Guided by:

Dr. D. B. Kulkarni

Objectives of the Project

- To categorize documents based on NLP
- To provide a search engine with efficient retrieval mechanism
- To compare the performance of LSA, LDA and word2vec in categorization task

LSA's performance for extracting topics

Number of documents	Time taken (in seconds)
1	4
100	100
1000	500
10000	3000
17000	5000

Time taken by word2vec for training the model

Number of documents	Time taken (in seconds) Window size=5, dimensions=7, worker=1	Time taken (in seconds) Window size=5, dimensions=10, workers=4
100	30	5
500	120	20
1000	300	30
10000	1500	200
17000	2000	300

Accuracy of relevant words according to the trained model

Number of documents for which model is trained	Correct relevant words out of 10 (window size=3)	Correct relevant words out of 10 (window size=5)
100	3	5
500	3	5
1000	4	5
10000	5	7
17000	5	7

Algorithms

- LSA (Latent Semantic Analysis)
- word2vec

Database

- simplewiki-20171020 (AA)

Languages & Tools

- Flask
- MongoDB
- Python

Thank You !