Annexure 10

TEXT AND SOCIAL MEDIA ANALYTICS

Course Code: CSD 501 **Credit Units: 03**
Total Hours: 30

Course Objective:

The objective of the course is to provide the understanding of the fundamental graphical operations and the implementation on computer, the mathematics behind computer graphics, including the use of spline curves and surfaces. It gives the glimpse of recent advances in computer graphics, user interface issues that make the computer easy, for the novice to use.

Course Contents:

Module I: Introduction: (5 Hours)

Introduction of social media and natural language processing research. Collecting and Extracting Social Media Data using API'S.

Module II: Language Identification and Naïve Bayes: (7 Hours)

Domain/Genre Difference Language Identification Supervised Learning and Classification Naïve Bayes Algorithm + feature selection (Information Gain) Tokenization, Emoticons, Noisy Text Normalization

Module III: Overview of paraphrase research: (6 Hours)

WordNet, DIRT, MRPC (Microsoft Research Paraphrase Corpus), PPDB (Paraphrase Database), etc Linear Regression Cost Function, Gradient Descent Logistic Regression, Decision Boundary.

Module IV: Vector Semantics: (5 Hours)

Unsupervised Learning Class-based Clustering: Brown Clusters Soft Clustering: Singular Value Decomposition (SVD) Neural Word Embeddings: Word2vec (CBOW and Skip-gram)

Module V: Deep Learning for NLP: (7 Hours)

Neural Network Basics: Neuron, Activation Function, Non-linearity, Learning Recurrent Neural Network Long Short-Term Memory Networks Neural Machine Translation Neural Conversation Generation Sentiment Analysis, Convolutional Neural Networks and Attention Sentiment Analysis Attention Model Convolutional Neural Network

Course Outcomes:

After taking this course, you will be able to:

- Utilize various Application Programming Interface (API) services to collect data from different social media sources such as YouTube, Twitter, and Flickr.
- Process the collected data - primarily structured - using methods involving correlation, regression, and classification to derive insights about the sources and people who generated that data.
- Analyze unstructured data - primarily textual comments - for sentiments expressed in them.
- Use different tools for collecting, analyzing, and exploring social media data for research and development purposes.

Examination Scheme:

Components	**A**	**CT**	**S/V/Q/HA**	**ESE**
Weightage (%)	5	15	10	70

CT: Class Test, HA: Home Assignment, S/V/Q: Seminar/Viva/Quiz, ESE: End Semester Examination;A: Attendance

Text & References*:*

Text:

- Mining Text Data. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.
- Speech & Language Processing. Dan Jurafsky and James H Martin, Pearson Education India, 2000.

References:

- Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

Annexure 10

TEXT AND SOCIAL MEDIA ANALYTICS LAB

Course Code: CSD 521 **Credit Unit: 01**

Total Hours: 20

Course Objective :

The objective of the course is to provide the understanding of the fundamental graphical operations and the implementation on computer, the mathematics behind computer graphics, including the use of spline curves and surfaces. It gives the glimpse of recent advances in computer graphics, user interface issues that make the computer easy, for the novice to use.

SOFTWARE REQUIREMENTS: Python 3.6, Anaconda IDE with Spider

List of experiments/demonstrations:

1. Write python code to flatten and evaluate a deep tree in NLP
2. Create Shallow Tree in NLP and print its height
3. Download wine quality data set from the UCI Machine Learning Repository which is available for free. Then print data of five rows of red and white wines. Check for NULL Values in red wine. Create a histogram to show distribution of alcohol and finally split the data for training and validation.
4. Avengers Endgame and Deep learning. Write python code to implement Image Caption Generation using the Avengers End Games Characters
5. Create a Neural network using Python (you can use NumPy to implement this)
6. Implement Word Embedding using Word2Vec
7. Collocations are two or more words that tend to appear frequently together, for example – United States. Implement this using Python.
8. WordNet is the lexical database i.e. dictionary for the English language, specifically designed for natural language processing. Synset is a special kind of a simple interface that is present in NLTK to look up words in WordNet. Synset instances are the groupings of synonymous words that express the same concept Show working of these using Python
9. Implement Naïve Baye’s Classifier using python.
10. Twitter Sentiment Analysis using Python. Fetch tweets from twitter using Python and implement it.

Examination Scheme:

IA			EE			
A	PR	Practical Based Test	Major Experiment	Minor Experiment	LR	Viva
5	10	15	35	15	10	10

Note: IA –Internal Assessment, EE- External Exam, A- Attendance PR- Performance, LR – Lab Record, V – Viva.

Course Outcomes:

After taking this course, you will be able to:

- Utilize various Application Programming Interface (API) services to collect data from different social media sources such as YouTube, Twitter, and Flickr.
- Process the collected data - primarily structured - using methods involving correlation, regression, and classification to derive insights about the sources and people who generated that data.
- Analyze unstructured data - primarily textual comments - for sentiments expressed in them.
- Use different tools for collecting, analyzing, and exploring social media data for research and development purposes.

Text & References*:*

Text:

- Mining Text Data. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.
- Speech & Language Processing. Dan Jurafsky and James H Martin, Pearson Education India, 2000.

References:

- Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.