

LAB MANUAL

**Text Analytics(22CSE652)**

Semester V

B. Tech (All Specializations) 2021-2025

|  |  |
| --- | --- |
| **Faculty Name** | Prof. Rajan Kumar |
| **Affiliation** | School of Technology |
| **Email ID** | [rajankumar.pandit@woxsen.edu.in](mailto:rajankumar.pandit@woxsen.edu.in) |

# Brief Description and Relevance of the Course

## Text Analytics is an interdisciplinary field that focuses on enabling computers to understand, interpret, and generate human language. In this course, students will explore various techniques and tools for processing and analyzing text data. The syllabus covers fundamental concepts, practical applications, and recent trends in Text Analytics. It plays a crucial role in building efficient search engines. Techniques like tokenization, stemming, and semantic analysis enhance search results by understanding user queries and matching them with relevant documents. It powers machine translation systems (e.g., Google Translate). Understanding syntax, semantics, and context helps translate text between languages. Techniques like Latent Dirichlet Allocation (LDA) help discover latent topics within a corpus. Researchers and analysts use topic modeling to explore large text collections.

**Course Objectives**

1. Understand the varieties of text preprocessing and exploratory methods and discuss the various ways in which text can be analyzed.
2. Adapt considerations to both long and short texts of appropriate text classification and clustering algorithms.
3. Apply techniques for collecting and preparing text data for computational analysis, Use open-source text analytic tools.

# Intended Learning Outcomes

On completion of this course, the students will be able to:

1. Understand the basics of natural language processing, including tokenization, stemming, and part-of-speech tagging.
2. Apply regular expressions for text manipulation and normalization.
3. Explore semantic relationships between words.
4. Collect and preprocess text data from various sources (corpus creation, web scraping).
5. Analyze word frequency and co-occurrence patterns.
6. Use Naïve Bayes for sentiment classification.

# Suggested Reference Books for solving the problems.

1. Jurafsky, D. and J. H. Martin, “Speech and language processing: An Introduction to Natural Language

Processing, Computational Linguistics, and Speech Recognition”

# Software Requirements for the Course

## The programs may be executed using any available Open Source/ Freely available IDE. Some of the Tools available are:

* 1. Integrated Development and Learning Environment (IDLE): Download the software from <https://docs.python.org/3/library/idle.html>
  2. Jupyter: Download the software from <https://jupyter.org/install.html>
  3. Google Colaboratory: It is a cloud-based web IDE for Python <https://colab.research.google.com>[/](https://colab.research.google.com/)
  4. PyCharm : Download the software from <https://www.jetbrains.com/pycharm/>

**\*Note:** This list is not exhaustive and is NOT in any order of preference.

# List of Experiments:

**Experiment 1: Regular Expression Basics**

Demonstrate Noise Removal for any textual data and remove regular expression patterns such as hashtag from textual data.

**Experiment 2: Text Normalization Techniques**

Perform lemmatization and stemming using the Python library NLTK.

**Experiment 3: Part-of-Speech Tagging**

Perform parts of speech tagging on any textual data. Use NLTK or spaCy library.

**Experiment 4: N-gram Language Models**

Build an N-gram language model and evaluate its performance on a given corpus.

**Experiment 5: Web Scraping and Corpus Creation**

Collect data from websites and create a corpus for analysis.

**Experiment 6: Word Embeddings and Vector Semantics.**

Implement word2vec or GloVe to create word embeddings and explore their semantic relationships**.**

**Experiment 7: Naïve Bayes Classifier.**

Implement Text Classification using naÃ¯ve Bayes classifier and text blob library.

**Experiment 8: Latent Dirichlet Allocation(LDA) in Python.**

Implement topic modeling using Latent Dirichlet Allocation(LDA) in Python.

**Experiment 9: Clustering and Categorization.**

Apply clustering algorithms(e.g., k-means) to group similar documents.

**Experiment 10: Case study 1:- Identify the sentiment of tweets.**

In this problem, you are provided with tweet data to predict sentiment on electronic products of netizens.

**Experiment 11: Case study 2:- Detect hate speech in tweets.**

The objective of this risk is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has racist or sexist sentiment associated with it. So, the task is to classify racist or sexiest tweets from other tweets.

**Programming for Problem-Solving Course Index**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No.** | **Week** | **Experiment** | **Page No/s** |
| 1. |  |  |  |

**Performance Evaluation Components for the Course**

|  |  |  |
| --- | --- | --- |
| **Week No.** | **Marks** | **Evaluation Form** |
| **Continuous Evaluation** | | |
| 1-10 | 100 | 10 Marks Per Lab |
| **End Term Examination** | | |
|  |  |  |

**Attendance & Punctuality**

Learning is an interactive process. Woxsen students are admitted partly based on the experience they bring to the school and what they can add to class discussions. Therefore, attendance is an important aspect of studying here. Students are expected to be present in all the classes. Absence is only appropriate in cases of extreme personal illness, injury, or close family bereavement. Voluntary activities are never valid reasons for missing any class. The faculty, with the assistance of the Faculty Associates shall keep track of students’ attendance and decide on the nature and extent of penalty for any absence from the class. Penalty may include reduction in grade or repetition of the course.

Late arrival is disruptive to the learning environment; students are expected to be in class before the scheduled commencement time. Students arriving for class after the scheduled commencement time should be turned away unless they have a valid reason to be permitted to attend.

Faculty should not consider attendance of sessions as a component of performance evaluation. The grading system at Woxsen accounts for this.

# External Websites Disclaimer

Neither the instructor nor Woxsen University is responsible for the content of external websites discussed in the classroom and/or linked to via online course materials, e-mail messages, message boards, or other means. Referred websites are for illustrative purposes only and are neither warranted nor endorsed by the faculty or Woxsen School of Business. Web pages change frequently, as do ownership of domain names. While every effort is made to ensure proper referencing, it is possible that students may, on occasion, find materials to be objectionable for reasons beyond our control.

# Copyright

The content provided by the faculty in the class is copy-righted. Students are instructed not to distribute or share content used during courses with external entities without the explicit written consent of the author and/or faculty.

# Student Code of Ethics

Each student enrolled in this course accepts personal responsibility to uphold and defend academic integrity and to promote an atmosphere in which all individuals may flourish. The Students’ Code of Ethics strives to set a standard of honest behaviour that reflects well on students and the school. All students enrolled in these courses are expected to follow the Students’ Code of Ethics, which they have been given at the time of enrolling for the program and pledged to adhere to. Unethical and unfair practices adopted by students may lead to penalties such as having to repeat the course or having the student’s enrolment cancelled.

7 | P a g e