Group Members:

1) Aanshula Shetty (45)
2) Harshvardhan Thakur (55)
3) Nidhi Worah (60)

PBL Assignment 4

Problem Statement: Text Summarization

Software Requirements:

Pycharm/VS Code for python, python libraries like spaCy etc

Hardware Requirements:

Laptop for coding

The problem:

Textual data on the internet is decreasing gradually and people are not much interested in textual data anymore. People these days are usually busy and hence summaries come in as a great help. Nobody has to time to read out entire articles and would rather prefer reading small portions of the text to get a basic gist of the entire article. For this purpose the technique text summarization is used which helps us create summary of any text material and it removes data which does not change the overall meaning of the context. This technique has a greater scope in the field of research as there are research papers containing thousands of pages having important documentation. Text Summarization would ehlp scientists in focusing only on the key phrases from all the data.

Flowchart/DFD:

```
          ┌─────────────┐
          │    START    │
          └──────┬──────┘
                 │
                 ▼
          ┌─────────────┐
          │    Text     │
          └──────┬──────┘
                 │
                 ▼
  ┌──────────────────────────────────────┐
  │ Sentence segmentation, word segmentation │
  └──────────────────┬───────────────────┘
                     │
                     ▼
  ┌──────────────────────────────────────┐
  │ Iterative calculation based on word vector, │
  │ word-sentence relationship and graph model  │
  └──────────────────┬───────────────────┘
                     │
                     ▼
      ┌──────────────────────────┐
      │ Summarization generation │
      └────────────┬─────────────┘
                   │
                   ▼
          ┌─────────────┐
          │     END     │
          └─────────────┘
```

```
┌─────────────┐    ┌──────────────────┐    ┌──────────────────┐    ┌────────────────┐    ┌─────────────┐
│   Input     │───▶│ Words Extracted  │───▶│  Pre-processing  │───▶│     Method     │◀───│  Algorithms │
│  Documents  │    │   from each      │    │ (Stop-word       │    │ Identification │    └─────────────┘
└─────────────┘    │    document      │    │  removal,        │    └────────┬───────┘
                   └──────────────────┘    │   Stemming)      │             │
                                           └──────────────────┘             │
                                                                            ▼
┌─────────────┐    ┌──────────────────┐    ┌──────────────────┐    ┌────────────────┐
│   Summary   │◀───│     Summary      │◀───│ Content Selection│◀───│    Analysis    │
└─────────────┘    │   Generation     │    │ (Sentence Scoring,│   │   (Parsing)    │
                   └──────────────────┘    │  Classification) │    └────────────────┘
                                           └──────────────────┘
```