Minutes of Meeting

Project: Lab 8 - Representing Document Concepts with Embeddings
Course: DSCI-560: Data Science Practicum
Instructor: Young H. Cho
Team Members: Jaival, Mayank, Pratham
Meeting Dates: 4th March 2025 – 7th March 2025

Meeting 1: 4th March 2025

Agenda:

- Understanding Lab 8 requirements and deliverables.
- Assigning responsibilities to team members.
- Planning workflow and setting milestones.

Discussion Points:

- Reviewed the **Lab 8 assignment document** and discussed the tasks.
- Agreed to experiment with **three different Doc2Vec configurations**.
- Decided to explore **Word2Vec and Bag-of-Words embeddings** for comparison.
- Planned to **use cosine similarity for clustering**.
- Distributed tasks:
    - Doc2Vec experimentation and initial clustering.
    - Word2Vec and BOW embeddings.
    - Documentation, result interpretation, and GitHub updates.

Meeting 2: 5th March 2025

Agenda:

- Discuss progress on Doc2Vec and Word2Vec embeddings.
- Identify challenges and refine strategies.

Discussion Points:

- Successfully generated embeddings with different vector sizes. Encountered a minor issue with **memory allocation** when training large models.

- Implemented Word2Vec and BOW embeddings but faced difficulty in defining the optimal bin size.
- Structured the **report outline** and ensured the code repository was set up on GitHub.

Meeting 3: 6th March 2025

Agenda:

- Analyze clustering results and compare methods.
- Start drafting the report.

Discussion Points:

- **Doc2Vec results:** Identified **strong clustering performance** with a vector size of 200.
- **Word2Vec results:** Different bin sizes **affected accuracy significantly**.
- **Comparison:** Word2Vec worked better for short posts, while Doc2Vec was **more effective for longer documents**.
- **Challenges:**
    - Computational cost was **higher for larger embeddings**.
    - Choosing the best **clustering metric** for Word2Vec was tricky.

Meeting 4: 7th March 2025

Agenda:

- Finalizing the report, README, and submission files.
- Reviewing GitHub commits and verifying code.

Discussion Points:

- **Final results:** Concluded that **Doc2Vec (vector size: 200) performed best overall**.
- **Report writing:**
    - Clearly stated methodology, results, and conclusion.
    - Included qualitative and quantitative comparisons.
- **GitHub verification:** Ensured all **code updates** and history were properly logged.
- **Submission:** Prepared all required files:
    - Code files
    - Report with results
    - README for execution steps
    - Meeting notes

Final Notes:

- The team successfully **completed the assignment** as per the requirements.
- **Challenges faced**: Memory usage in Doc2Vec, binning issues in Word2Vec, computational cost.
- **Key Takeaway**: Doc2Vec was **better for longer documents**, while Word2Vec worked well with **shorter texts**.