Armenian Word2Vec Model Development Task

1. Project Overview

Create a Word2Vec model for the Armenian language using news article data to generate high-quality word embeddings that capture semantic relationships in Armenian text.

2. Dataset Structure

Training Dataset

- Categories:
  - accidents/
  - culture/
  - economy/
  - politics/
  - society/
  - sport/
  - weather/

Each category contains multiple text files (text-64.txt, text-144.txt, etc.) with Armenian news articles.

Data Format

```
File format: UTF-8 encoded .txt files
Size: 2-3 KB per file
Content: Natural Armenian text with:
- Full sentences
- Punctuation
- Numbers
- Dates
- Special Characters
```

3. Task Requirements

3.1 Data Preprocessing

- Clean and normalize Armenian text:
  - Remove numbers
  - Handle punctuation
  - Normalize Armenian characters

    - Handle dates and times
    - Split into sentences
    - Tokenize words
    - Remove rare words (frequency < 5)

3.2 Model Development

- Implement Word2Vec model using either:
    - TensorFlow
    - Gensim
    - Custom implementation

3.3 Model Parameters

```
Required parameters:
- vector_size: 300
- window_size: 5
- negative_samples: 4
- epochs: 20
```

3.4 Model Features

- Skip-gram architecture
- Negative sampling
- Support for Armenian Unicode characters