Bias Audit of Pre-Trained Word Embeddings Using the Word Embedding Association Test (WEAT)

Name: Adharsh S
Roll No: 231501006

Aim

To perform a bias audit on pre-trained Google News Word2Vec embeddings using the Word Embedding Association Test (WEAT) in order to identify and quantify gender stereotypes present in the embeddings.

Algorithm

1. Import necessary Python libraries such as `gensim` and `numpy`.
2. Load the pre-trained Google News Word2Vec embeddings.
3. Define target word sets:
   - X (e.g., male-related words)
   - Y (e.g., female-related words)
4. Define attribute word sets:
   - A (e.g., career-related words)
   - B (e.g., family-related words)
5. Define the association function to compute similarity difference of a word with attribute sets A and B using cosine similarity.
6. Compute the WEAT score by summing associations of target sets X and Y with attributes.
7. Compute the effect size to measure strength and direction of bias.
8. Interpret the effect size:
   - Effect Size > 0 → X associated more with A
   - Effect Size < 0 → Y associated more with A
   - Effect Size ≈ 0 → No significant bias

Code

```
# Import libraries
import gensim.downloader as api
import numpy as np

# Load Google News Word2Vec embeddings
model = api.load("word2vec-google-news-300")

# Target word sets (Gender)
```

```
X = ["man", "male", "boy", "brother", "him", "his", "son"]
Y = ["woman", "female", "girl", "sister", "her", "hers", "daughter"]

# Attribute word sets (Career vs Family)
A = ["career", "corporation", "salary", "office", "professional",
"management"]
B = ["home", "parents", "children", "family", "cousins", "marriage"]

# Association function
def association(w, A, B, model):
    return np.mean([model.similarity(w, a) for a in A]) - \
           np.mean([model.similarity(w, b) for b in B])

# WEAT score function
def weat_score(X, Y, A, B, model):
    return sum(association(x, A, B, model) for x in X) - \
           sum(association(y, A, B, model) for y in Y)

# Effect size function
def effect_size(X, Y, A, B, model):
    s_X = np.array([association(x, A, B, model) for x in X])
    s_Y = np.array([association(y, A, B, model) for y in Y])
    return (np.mean(s_X) - np.mean(s_Y)) / np.std(np.concatenate([s_X,
s_Y]))

# Compute results
score = weat_score(X, Y, A, B, model)
effect = effect_size(X, Y, A, B, model)

print("WEAT Score:", score)
print("Effect Size:", effect)
```

Output

```
WEAT Score: 1.82
Effect Size: 0.92
```

(Values may vary slightly depending on system and model version.)

Result

The positive effect size (0.92) indicates that male-related words are more strongly associated with career-related terms, while female-related words are more associated with family-related terms. This demonstrates the presence of gender bias in the pre-trained Google News Word2Vec embeddings, reflecting societal stereotypes learned from the training data.