Anaphora Resolution Assignment

You may use any programming language to complete this assignment. Please submit the following documents:

1. A PDF report answering the questions mentioned below.

2. Your code and README file in a `.tar` archive.

Reference

Chapter 21 from Jurafsky and Martin (2nd edition, J&M henceforth). Available here.

Dataset

Download the **GAP (Gender-Balanced Coreference Data)** dataset from this link.

Task

Identify the anaphor of each pronoun in the dataset using a simple Naive Bayes classifier. Submit a report after performing the following operations on the dataset `gap-test.tsv`:

1. **(10 marks)** For each discourse segment in the test set, extract all the candidate noun phrases that can potentially serve as the antecedent of the given pronoun. You need to use a standard POS tagger.

2. **(5 marks)** Randomly assign an antecedent to the pronoun and calculate the average accuracy for 100 random assignment runs.

3. **(5 marks)** Select the most recent antecedent of the pronoun and compute accuracy.

4. **(10 marks)** Implement a naive Bayes classifier by extracting features from the candidate noun phrases you extracted and report precision, recall and F-score. Feature probabilities should be learned from the training data, and the evaluation needs to be done on the test set.

5. **(5 marks)** Write a report describing the errors made by your implementation and how the algorithm can be improved.