Anaphora Resolution Assignment

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

1. A PDF report answering the questions mentioned below
2. Your code and README file in a tarball archive

Deadline: 5:30pm on Nov 18

Reference: Chapter 21 from Jurafsky and Martin 2nd edition (J&M henceforth):

https://courses.iiit.ac.in/draftfile.php/159447/user/draft/977663181/jurafsky-2ed-chap21-comp-disc.pdf

Download the Gap (Gender-Balanced Coreference Data) dataset from the following link:

https://www.kaggle.com/datasets/thedevastator/gap-unlocking-gender-balanced-coreference-data-f

Task: Identify the anaphor of each pronoun in the dataset using the centering theory algorithm in J&M Chapter 21. Then submit a report after performing the following operations on the dataset gap-test.tsv you downloaded:

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 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 centering theory assumptions described in J&M Section 21.6.2 titled "A Centering Algorithm for Anaphora Resolution" and test in on the following discourse segment in J&M:

John saw a beautiful 1961 Ford Falcon at the used car dealership. (U_1)
He showed it to Bob. (U_2)
He bought it. (U_3)

6. (10 marks) Run the centering algorithm on the entire Gap dataset you downloaded and report accuracy. Write a report describing the errors made by your implementation and how the algorithm can be improved.