**CS372 Natural Language Processing with Python**

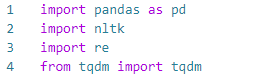
**Homework 3**

Python version : 3.8.10

Nltk version : 3.5

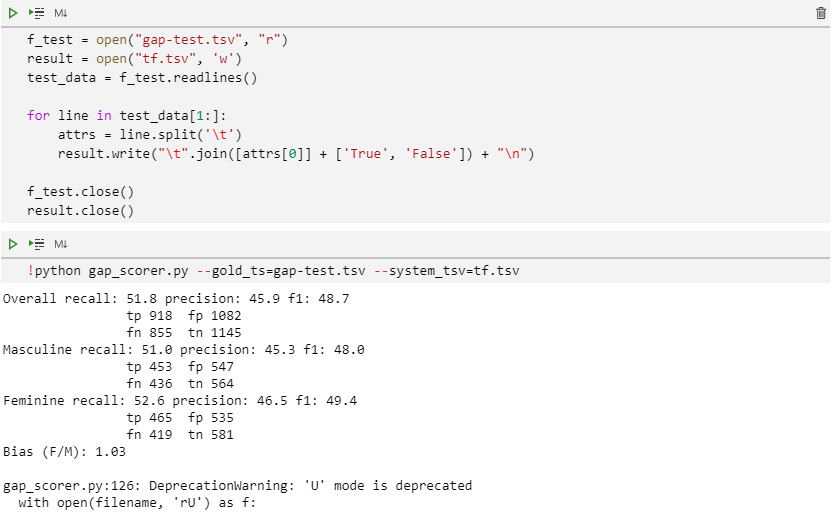
This homework is kind of open-ended problem. So instead of writing the final result of my program, I will show how I improved my model, step by step. There are some ideas that turned out to be a failure, and some of successful ideas remained to the next level.

For external modules, I imported pandas, to handle tsv file easily, tqdm, to estimate the remaining running time of the program, and nltk.



**Design 0 – tf**

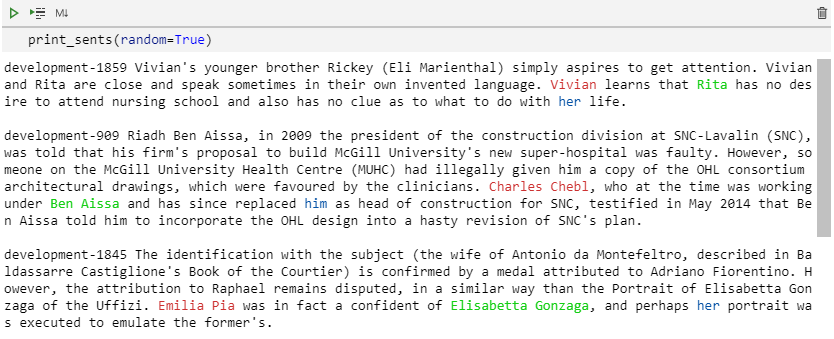
For two given words, I will always predict first word as True, and second word as False.



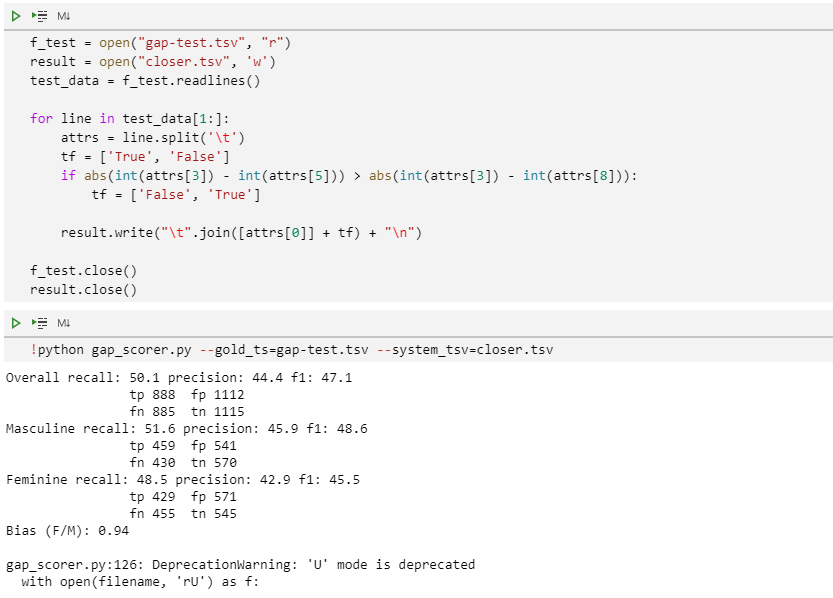
Overall f1-score is 48.7, which means almost nothing for such a binary classification problem.

**Design 1 – closer word**

In order to find any characteristics of the sentences, I had to take a closer look at them. But I found it too difficult to read them with just ‘offset’. So I made a program that prints the sentences with colors so that I could identify which one is the target word/pronoun.



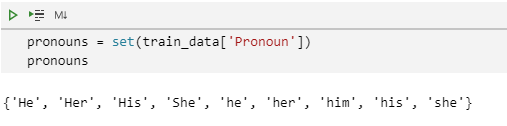
What I realized from doing this was that the word that appears closer with the pronoun are more likely to be true. So I used the idea to my model.



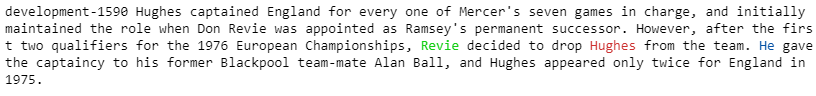
But unfortunately, f1-score was even lower.

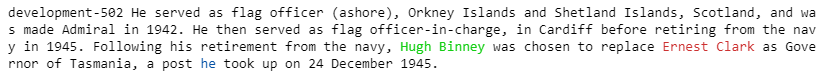
**Design 2 – Subject pronoun**

I learned that there are three different kinds of pronoun : subject pronoun(he, she), object pronoun(him, her), possessive pronoun(his, her). And they showed different characteristics of different kinds.



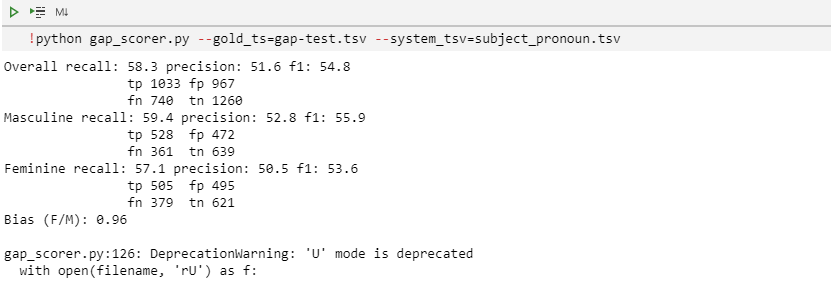
For example, unlike object and possessive pronouns, subject pronouns are more likely to refer farther word.





So I divided pronouns into two groups : subject pronouns & others. My second model predicts farther word as true, and closer word as false when the pronoun is subject pronoun, and do the opposite when the pronoun is object or possessive pronoun.

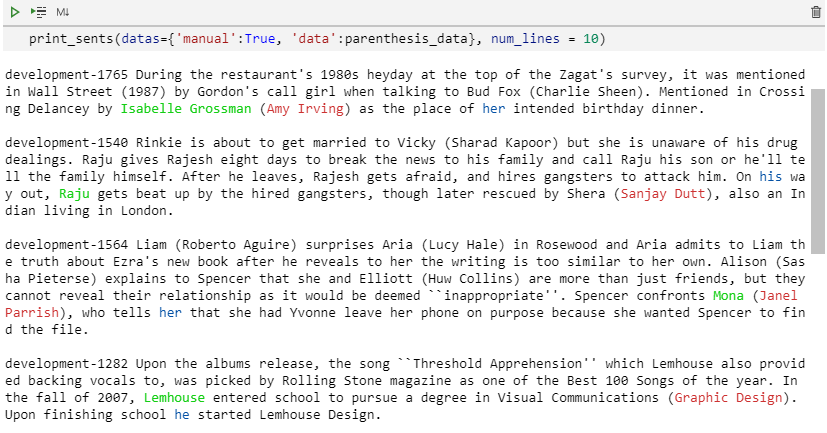




Now the f1-score is higher than Design0 and 1.

**Design 3.1 – parenthesis**

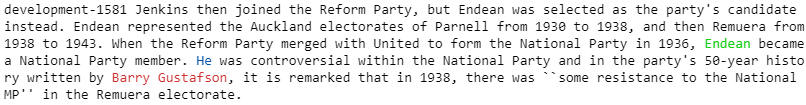
I found some words are inside the parenthesis, whose coreference in most of the cases were false.

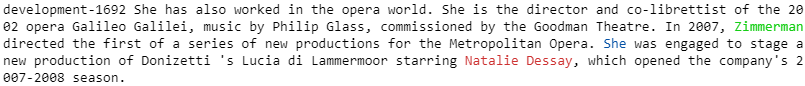


When I predict the word inside the parenthesis as false, 22 out of 31 sentences were predicted correct, which means accuracy is 0.71.(pretty good)

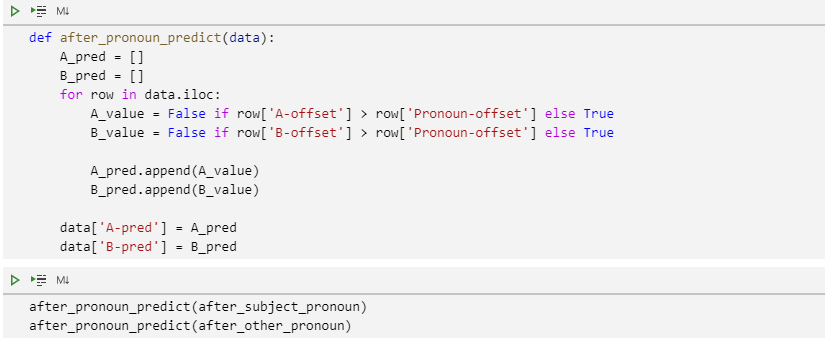
**Design 3.2 – word coming after pronoun**

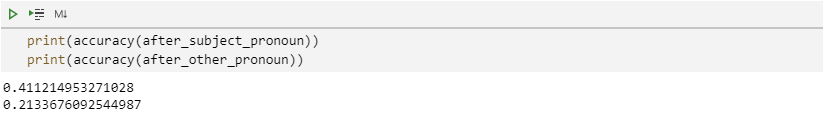
Sometimes one of the words comes after the pronoun.





And the word coming after the pronoun is more likely to be false. I tested this idea in two divided datas: when the pronoun is subject pronoun & others.

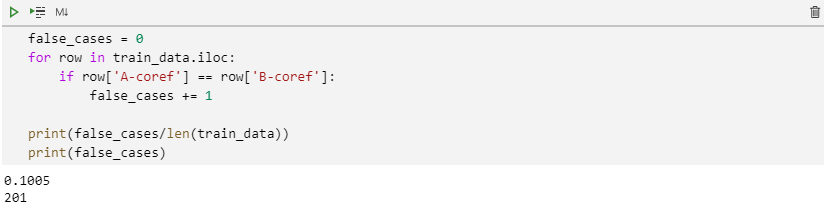




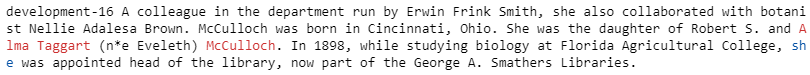
But the accuracy wasn’t so good. I think the hypothesis I made was too general.

**Design 4 – both false cases**

I realized that there are 200 cases, which is 10% of total cases that the both coreference of words are false.



While watching ‘both false cases’, I found some characteristics they have in common. First, another pronoun already appears before the words.

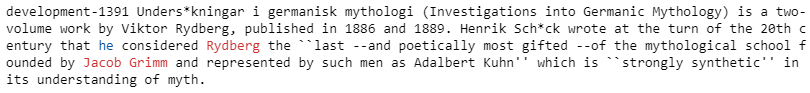


These pronouns are likely to refer to the same words as each other, while most of the words that are referenced appear much earlier. So when another pronoun in same gender appears before the words, I predicted them as both false.

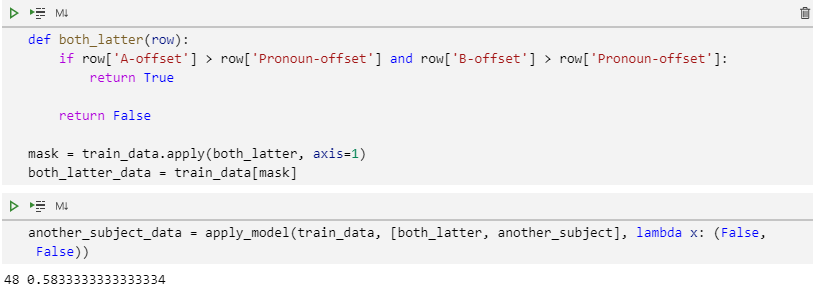


However, the performance was really bad.

Second idea is that when the words both come after the pronoun, and there is another subject exist in the text, predict the both coreference of the words false. (Another subject means that proper noun that comes in the beginning of the sentence)







**\*apply\_model() : apply the conditions with condition functions to given data, and predict with prediction function(Appendix)**

The accuracy is not bad, but size of total cases is too small(48 out of 2000 cases).

**Design 5 Subject word**

I defined subject word like this : Subject word is a word that comes right before verbs, modals, or adverbs.

Design 5-1 Only subject word