Part 1

1- Write a program to compute word n-grams (up to 4-grams) of a given input sentence. For example:

Input: "cat sat on the mat"
Output:

> 1-grams: {cat, sat, on, the, mat}
> 2-grams: {cat sat, sat on, on the, the mat}
> 3-grams: {cat sat on, sat on the, on the mat}
> 4-grams: {cat sat on the, sat on the mat}

2- Write a program to evaluate a translation output (sentence) against a reference translation (sentence) in BLEU score. The input to your program should be two sentences, the first representing the translation to be scored and the second representing the gold standard reference translation (i.e. your target language test file). The output to your program should be a score (float).

Hint:

> (1) Interpolation of n-grams extractor (using Q1) could be a good idea;
> (2) the BLEU score is calculated as shown in lecture slides (week 8):

$$BLEU = min(1, \frac{output\ length}{reference\ length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

Input:

> Translation Output: "The gunman was shot dead by police ."
> Reference: "The gunman was shot dead by the police ."

Example output:

> Score = 0.680183007338

Part 2

3- Try to adapt your program from Q2 in order to evaluate the translation output comparing to multiple references.

Input:

> Translation Output: "The gunman was shot dead by police ."
> Reference 1: "The gunman was shot to death by the police . "
> Reference 2: "The gunman was shot to death by the police . "
> Reference 3: "Police killed the gunman ."
> Reference 4: "The gunman was shot dead by the police ."

Output:

> Score =0.680183007338

4- Compare your score in Q2 to BLEU Toolkit.

Hint: Put a translation output in a file named trans.txt and put a reference in another file named ref.txt. For example:

trans.txt: "The gunman was shot dead by police ."
ref.txt: "The gunman was shot dead by the police ."

You can use the following BLEU Toolkit
https://pypi.org/project/bleu/

then you can compare this score (from the bleu library) with your previous score (from your Q2 program).

    a) How good/meaningful is your proposed metric?
    b) How correlated is your metric to human evaluation and BLEU?

Part 3

5 - (**Optional)** Compare with other metrics
- About TER: https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf
Compare its results to BLEU. *(Note: TER is referred to as 'translation edit rate in the above link)*
TER link: http://www.cs.umd.edu/~snover/tercom/
Please download the Java code of Version 0.7.25.

$ java -jar tercom.7.25.jar -r ref.txt -h trans.txt -n TER_output > TER_result

Note: When you try TER tool, please carefully read its README file and prepare the reference and translation files following their format.