**BLEU(Bilingual evaluation understudy)**

取将翻译结果分为多个连续的元组，依次观察每个元组在参考翻译中是否出现，计算n-gram精度

其中c是机器翻译长度，r是参考翻译长度

BLEU分数取值范围[0,1]，分数越接近1，说明翻译的质量越高

**nltk.translate.bleu\_score.corpus\_bleu(references, canditate)**

**参数解释**

**references**：[ref1,ref2,..,refbatchsize ]=[[ref1\_1,ref1\_2,ref1\_3], [ref2\_1,ref2\_2]…[refbatchsize\_1,refbatchsize\_2,…]]

references为一个列表；ref1,ref2,..,refbatchsize为各句子的不定数量的参考翻译版本组成的列表；ref1\_1,ref1\_2,ref1\_3为第一个句子参考翻译的各版本，该版本对应的词字符串或词索引组成的列表

**hypotheses**：[hyp1,hyp2,…hypbatchsize]

hyp1,hyp2,..,hypbatchsize为各句子的机器翻译，实际为各个翻译位置上机器翻译得到的概率最大的词对应的词字符串或词索引组成的列表

**corpus\_bleu计算过程**

得到全部翻译句子和其对应的参考句子的各版本的组合(hyp1,ref1\_1),(hyp1,ref1\_2), (hyp1,ref1\_3), (hyp2,ref2\_1), (hyp2,ref2\_2),

…(hypi,refi\_1), (hypi,refi\_2),…, (hypi,refi\_5),

…,(hypbatch\_size,refbatch\_size\_1), (hypbatch\_size,refbatch\_size\_2)

令n=1,2,3,N=4，计算出相应的Max\_Ref\_Count和Count，其中Max\_Ref\_Count是翻译句子hypi中的连续n元组总数，Count是在refi\_j中出现的n元组数量，将全部翻译句子与其参考句子各版本组合的Max\_Ref\_Count和Count累加，并以此计算

根据每个翻译句子hyp1,hyp2,...,hypi,...,hypbatch\_size的长len\_hypi，找到最接近的参考句子长度len\_refi，将len\_hypi和len\_refi分别对i累加，作为的翻译句子和参考句子的长度，并以此计算BP

from nltk.translate.bleu\_score import corpus\_bleu

references = [[['this', 'is', 'a', 'test'], ['this', 'is' 'test']],[['I','Will','Eat','hamburgers','tomorrow']]]

hypotheses = [['this', 'is', 'a', 'test'],['I','Will','Eat','hamburgers','today']]

score = corpus\_bleu(references, hypotheses)

print(score)

>>>

0.7984079523098931