SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation

> https://arxiv.org/pdf/2310.03991.pdf

这篇提出了一个Semantic Robust Watermark，也是基于red/green list的范式。

它的方法本质是，在生成新token的时候，把之前窗口内的sentence做一个LSH（Locality-Sensitive Hashing），目的是希望相似语义的句子Hash到相似的Signature。然后再基于这个Signature，把一个d维空间划分成2片（类似green/red list），在生成下一个token的时候只从green空间中选；

其实是一个red/green list的语义hash高维版