看论文《Generating\_Topical\_Poetry-FSA》

想法：

1. 构建词典

* 淘汰模糊的词
* 淘汰频率太高的词
* 淘汰没有韵律前20，000中的词语

1. 如何找到主题相关的词

* WordNet分数评判相似分数
* PMI指数：
* word2vec：训练词袋(windows:8/40 dimension:200)，

1. 选择韵律词

不知道**Strict Rhyme，Slant Rhyme，Non-Topical Rhyming Words**的押韵方式

生成韵律对方式是：hash，碰撞

使用韵律对的方式是：加入到头尾中。

相应的问题

针对一、

* 模糊的定义是什么？可有现成的词典？

* 频率高的阈值是什么？
* 那么是否会导致结果出现大量的生僻词
* 这样淘汰是否会影响语义通顺
* 是否应该也淘汰频率很低的词语？
* 韵律前20，000是什么词表？怎么定义实现？
* PMI指数倾向于出现频率较低的词语

* Word2Vec词袋是基于词语的？还是句子？

后续找到代码组的小伙伴讨论一下具体细节。

其他模型的学习：

N-GRAM

FSA模型



