# NLP方向总结-新词发现

- 定义
  - 更准确的叫法应该是“无监督构建词库”，因为原则上它能完整地构建一个词库出来，而不仅仅是“新词”。你可以将它跟常用词库进行对比，删掉常见词，就可以得到新词了。
- 各种旧词
  - https://github.com/johnson7788/funNLP
- 参考工具：SmoothNLP，HanLP，jieba，Hellonlp
  - Hellonlp优于SmoothNLP，提高的点在于对于英文的处理和自由度的计算有些许区别
    - https://zhuanlan.zhihu.com/p/210584733
    - https://zhuanlan.zhihu.com/p/80385615
  - https://www.spaces.ac.cn/archives/3491
- JioNLP
  - https://github.com/dongrixinyu/JioNLP
    - 还在更新
  - 新词发现方法依然采用统计词汇的内聚度(PMI点间互信息熵)、以及与边界字的分离度(词汇左右信息熵) 进行统计
- HarvestText
  - https://github.com/blmoistawinde/HarvestText
- Jiagu
  - https://github.com/hellonlp/Jiagu
- 苏建林
  - https://github.com/bojone/word-discovery
- 商业领域
  - https://fenci.weiciyun.com/
- 属于无监督任务范畴
- 生成候选词
  - 将文本切词后的结果拼接为候选词
    - 弊端： 很大程度上依赖于分词的效果。有的分词工具将“楚乔传播出之前”切成了“楚乔/传播/出/之/前”，这样一来，就算之后的工作做得再好，想要抽出“楚乔传”一词也是无力回天。
    - HanLP使用的此方法
  - 直接将文本按字符分割后拼接为候选词
    - 例如：提取每句话的2元组,3元组,...,k元组作为候选词(一般k<=5就够用了)。比如“专注于自然语言处理技术”这句话产生的2元候选词有：[“专注”、“注于”、“于自”、“自然”、“然语”、“语言”、“言处”、“处理”、“理技”、“技术”]。
- 问题
  - Word Tokenization问题， 对于阿拉伯数字和英文字母，选择的最大词窗口为4或者5。但是几个数字或者几个英文字母随便组合一下，他们的程度就大于5
- Trie
  - https://en.wikipedia.org/wiki/Trie
  - https://github.com/google/pygtrie