20160426 作业

作业要求：

1、 利用新浪网新闻历史信息，收集相关汉语词汇信息。新浪网的历史新闻地址 http://news.sina.com.cn/hotnews/20040705.shtml，其中 2004 是年份，07 是月份，05 是日期。

2、 语料收集程序。语料需要达到一定量级，不少于 1GB 等，网址不能重复，即去重；不进行深度搜索。

3、 nGram 程序。统计出：单字频、双字组合频率，三字组合频率，四字组合频率，五字组合频率。以三字组合为例，以“语料收集程序，语料需要达到”为例，三字组合为“语料收、料收集、收集程、集程序”(逗号后没有考虑)。为了考虑专业词汇，可以将中文标点符合视为停词以及英文的双引号、单引号、叹号等。本部分功能可以独立成一个程序，能从命令行输入参数从语料文本中生成相关文件 nGram 文件。

4、 上述功能完成优秀者，计 100 分。其后为附加练习，总分不超过 120 分。

5、 通过 nGram 数据及其相关算法，发现双字词，三字词、四字词、五字词。

6、 个别同学告知老师后，可将此作为期末大作业，继续完善。

7、 感兴趣的同学可以将 2006 的数据作为基准，分析 2007 年、2008 年增频词汇。