1. 使用的Java库

在本次作文文本特征提取的程序中，采用了四种开源工具包，分别是OpenNLP、Stanford NLP、lingppipe和Nulooper。其中Nulooper是可自定义实现的NLP工具包，用来对文本处理的粒度进行定制和控制，可以根据不同的想自定义实现或者综合其他三种自然语言处理工具来满足特定的需求。

1. 在词性标注方面

初步采用OpenNLP的词性标注方案。OpenNLP的词性标注方案采用的词性标注集是Penn Treebank tag set。在该标注集中，对分词进行如下的叙述：

Our tokenization is fairly simple:

* most punctuation is split from adjoining words
* double quotes (") are changed to doubled single forward- and backward- quotes (`` and '')
* verb contractions and the Anglo-Saxon genitive of nouns are split into their component morphemes, and each morpheme is tagged separately.
  + Examples
    - children's --> children 's
    - parents' --> parents '
    - won't --> wo n't
    - gonna --> gon na
    - I'm --> I 'm

This tokenization allows us to analyze each component separately, so (for example) "I" can be in the subject Noun Phrase while "'m" is the head of the main verb phrase.

* There are some subtleties for hyphens vs. dashes, elipsis dots (...) and so on, but these often depend on the particular corpus or application of the tagged data.
* In parsed corpora, bracket-like characters are converted to special 3-letter sequences, to avoid confusion with parse brackets. Some POS taggers, such as [Adwait Ratnaparkhi](http://www.cis.upenn.edu/~adwait)'s [MXPOST](http://www.cis.upenn.edu/~adwait/statnlp.html), require this form for their input.   
  In other words, these tokens in POS files: ( ) [ ] { }   
  become, in parsed files: -LRB- -RRB- -RSB- -RSB- -LCB- -RCB-   
  (The acronyms stand for (Left|Right) (Round|Square|Curly) Bracket.)

[Here](http://www.cis.upenn.edu/~treebank/tokenizer.sed) is a simple sed script that does a decent enough job on most corpora, once the corpus has been formatted into one-sentence-per-line.