NLP 期末项目 实验报告

上 - 编译与运行

一. 概述：

这是实验报告的上篇，重点在于一些重要的注意事项，以及怎么搭环境，怎么把程序跑起来。为了方便阅读，报告语言用中文。

二. 重要(IMPORTANT)：

1. 项目用 python 完成。关于不同文件编码，经过本人测试，python 对 ANSI 编码的支持相当好，对 UTF-8 也还行，但是对 UNICODE 非常差！

所以，请 TA 高抬贵手，将测试数据用记事本打开，另存为，选择 ANSI 编码，保存。重新打开之后，某些标点可能会捣乱（比如缩写's 的'），麻烦 TA 替换一下，谢谢！^_^

2. 关于长句运行时间过长：由于 parser 性能所限，单个句子的运行时间，和单词数量的三次方大概成正比，本人的经验公式是：

```
t = 1/1000 * wordNum^3，单位为秒
```

也就是说，30 个单词的句子，就要 parsing 半分钟，50 个单词就要两分多钟，可能会造成运行者的不耐烦。

为了缓解情绪，消除烦躁，我在 config.py 配置了 ignore_long_sentence 选项，初始值设为：

```
ignore_long_sentence    = 1
max_word_num            = 25
```

请根据需要，随意更改！

（本人电脑为 i3，假如壕电脑，或许统统秒杀也说不定？）

三. 运行环境：

win8.1 专业版

这个换一个也没关系

python 2.7.3, 32 位版本

3.x 版本不支持 pyStatParser，对 nltk 支持性也不好，所以用 2.x

64 位版本不支持 nltk，所以用 32 位版本

pyStatParser

安装方法：在 cmd 中 python setup.py install 即可

nltk 3.0.0

这个是可选的，没有它少了语法树可视化功能，但不影响主程序

以上 4 项中，2,3 两项已经放入我的百度云，方便下载：

http://pan.baidu.com/s/1bnCqEeF

nltk 3.0.0 需要语料库支持，比较大，而且和主线没关系，所以就不放了。

（nltk 名气这么大，TA 应该已经装了，else 请联系我）

四. 压缩包文件：

main.py

主程序

config.py

用来配置参数，比如是否打印 parsing 时间，是否打印 NN-VB 依赖关系，是否无视太长的句子等等。大部分是方便 coding 和 debug 所设置的接口，以及一些 unit test。可以随便玩一玩，看看效果。

show_nlkt_tree 为语法树可视化，画树形图，需要 nltk 的支持。假如没装的话，请不要设置成 1。

五. 如何运行：

python main.py test_data output_data

如有任何问题，feel free to contact me: vendisky@163.com