SET DATA MINING ENVIRONMENT IN MAC

Preprocess准备工作

- 采用 homebrew 安装 python 以及必要的库。
- 将 xcode 升级到最新版。
- 用 homebrew 安装所需的工具，gfortran 等。
- 如果没有配置 Homebrew，用下面的命令进行配置：
- `ruby <(curl -fsSkL raw.github.com/mxcl/homebrew/go)`
- Homebrew 安装的软件都在 usr/local 下需要把该路径加到 PATH 变量中 (~/.profile or ~/.bashrc)，=而且要放在 /usr/bin 之前，确保使用homebrew 安装的python，而不是系统自带的（/usr/bin/python）。
- 不试用命令窗运行的话，可以省略

- `export PATH=/usr/local/bin:$PATH`
  修改 PATH 后最好重新打开终端（shell）以更新路径。
- 注意: Homebrew 的安装目录默认是用户可写的，因此不需要 sudo 。

- 安装 numpy、scipy 等需要 gfortran、swig、umfpack：
  `brew install gfortran`
  `brew install swig`
  `brew install umfpack`

用 Homebrew 安装 Python

```
brew install python --framework --universal
```

- 运行 python，看是否已经是最新版本了。（可以省略）

- 此后的 Python 包都尽量用 pip 进行安装和管理。安装好的 Python 自带了一个 pip，其版本是 1.2，如果想用更新一点的，用 easy_install 新安装一个 1.21 版(pip 无法安装自己)：
  `easy_install pip`

安装Numpy 和 scipy

- 先用 pip 安装 numpy：
  `pip install numpy`

```
pip install scipy
```

- 查看 numpy 和 scipy 的版本和路径 (可以省略)：

```
import numpy
print numpy.__version__
print numpy.__file__
import scipy
print scipy.__version__
print scipy.__file__
1.8.0
/usr/local/Cellar/python/2.7.6/Frameworks/Python.framework/Version
s/2.7/lib/python2.7/site-packages/numpy/__init__.pyc
0.13.2
/usr/local/Cellar/python/2.7.6/Frameworks/Python.framework/Version
s/2.7/lib/python2.7/site-packages/scipy/__init__.pyc
```

matplotlib

- 仍然是 pip

```
pip install matplotlib
```

Install Library 安装其他库

- 安装 NLTK 用里面的Stopword库来去除停词

```
sudo pip install nltk
```

- 安装 pyenchant 基于enchant，内置英语、法语字典，可以自定义字典。建议只用Wiktionary作为词库，缺点：文件太大，运行慢

```
sudo pip install pyenchant
```

- 安装 pattern 利用其自带的lemma来恢复动词形态。

```
sudo pip install pattern
```

- 安装 sklearn 包含了大量Data mining库

```
sudo pip install sklearn
```

- 例如：
-
```
from sklearn import feature_extraction
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer
```

```
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.decomposition import PCA
from sklearn.decomposition import NMF
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import KMeans
```

- 安装 BeautifulSoul 用于抓取网页中的内容
  `sudo pip install BeautifulSoul`

添加字典和定义个人字典

- 字典文件包含 .dic 和 .aff文件
- 目录在
- 调用语句—程序中，非安装中
- ```
  def correctword(words):
  pwl = enchant.request_pwl_dict("/Users/lxy/PycharmProjects/data mining/enwiktionary.txt”)
  d_gb = enchant.Dict("en_GB”)
  d_g = enchant.DictWithPWL("grc_GR", "/Users/lxy/PycharmProjects/data mining/enwiktionary.txt")
  return [word for word in words if d_gb.check(word) or d_g.check(word)]
  ```

安装Pycharm

- Mac-普通安装
- Linux-安装后调用python打开
- Window-安装后双击打开

配置库文件路径

- Mac——performance——project interpreter
- 选择系统自带的python路径
  ```