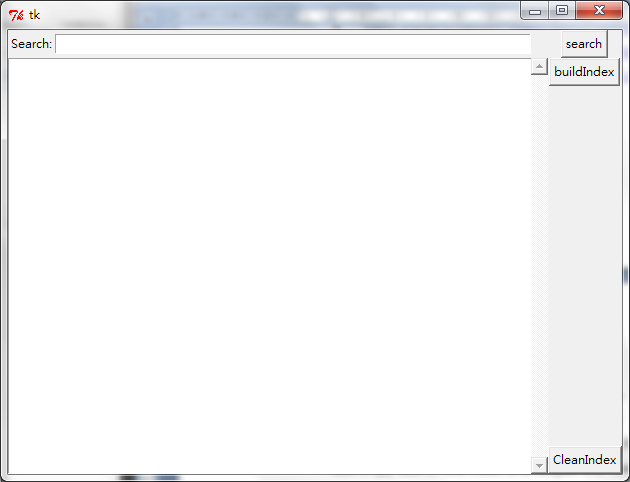
**Informaiton Retrieval and Applications**

**Homework01: Vector Space Retrieval**

1. Major features
   * GUI介面搜尋與建立Index、並可清空Index
   * Index約是原本檔案大小的一半
   * 提供Boolean Search方法
   * 自由選擇檔案建立Index
2. Instructions for compilation/execution environments
   * Runtime:Python 2.7.6
   * Open Source Library
     1. Warc-ClueWeb: <https://github.com/cdegroc/warc-clueweb>
     2. Whoosh 2.5.7 : <https://pypi.python.org/pypi/Whoosh/>
3. Major difficulties encountered
   * Warc File讀檔問題：  
      雖然說網路上有許多Open Source可以使用，但是還需要看老師提供的範例檔案能夠正常地讓程式讀取，像是Open Source:Warc，這就沒有辦法解讀ClueWeb09的WARC/0.18的檔案，並且會發生錯誤，最後我們找到warc-ClueWeb Python是基於Warc製作的，因此API與Warc相似，但是能夠讀取老師所提供的範例檔案
   * HTML Parsing 效能問題：  
      我們製作Index，需要將HTML Tag移除，因此在HTML Parsing會發現效能很差，因此我們改用正規表示法來解決Parsing緩慢的問題。
4. How to build the run-time environment in the OS of Windows.
   * Please install the **python-2.7.6.msi** in the folder of python library at first.
   * The next step is unzip the pip.zip, and then run the **get-pip.py** with python.
   * After we have the pip tool, we can install the libraries of Warc and Whoosh.
   * Unzip **warc-clueweb-clueweb09.zip** and **Whoosh-2.5.7.zip**, and then run the setup.py with python for each other.
   * After we build the environment in success, we can run the **gui.py** to start our homework program.
5. Reference:
   * Python Documentation: <https://docs.python.org/2/>
   * Warc Documentation: <http://warc.readthedocs.org/en/latest/>
   * Whoosh Documentation: [http://pythonhosted.org//Whoosh/](http://pythonhosted.org/Whoosh/)
   * TkInter GUI Documentation: <https://wiki.python.org/moin/TkInter>
6. Team members list:  
   99310317 陳冠廷  
   99590330 呂學昱  
   99590337 徐振峰
7. Detail

gui.py : the gui groups the need functions.



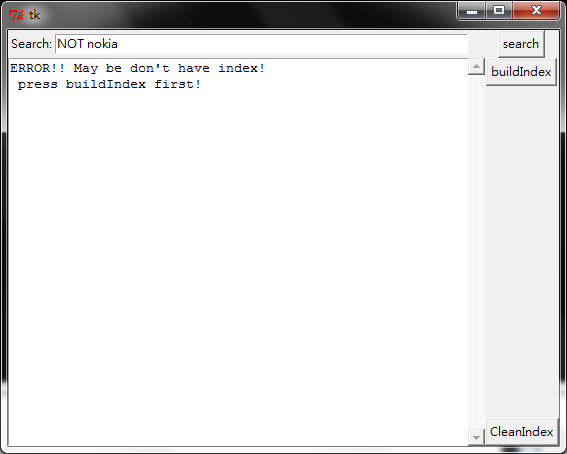
Search-button: after you input the textbox that near the search-button

the program will search the index and get rank list in text-view

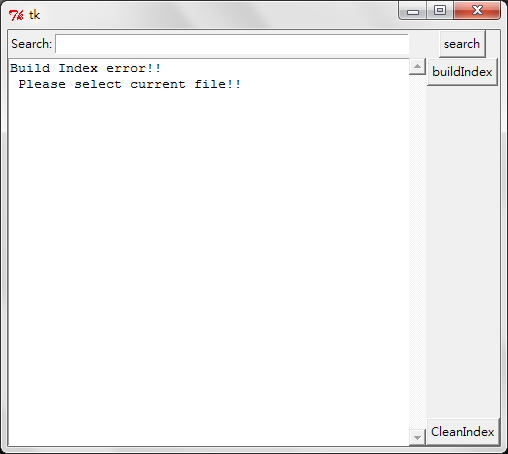
BuildIndex-button: build indexs by user input file.

If choose the uncurrent fomat file , the console log print the "error infor".

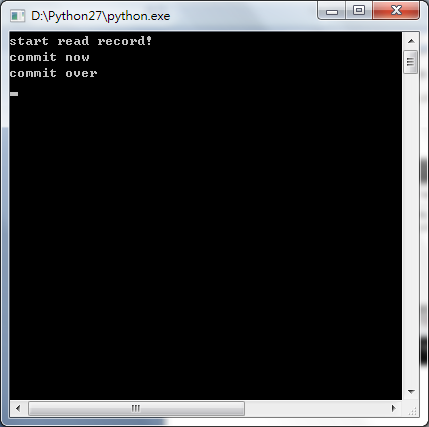
First time using, indexs hasn't created, searching will get the "error infor".



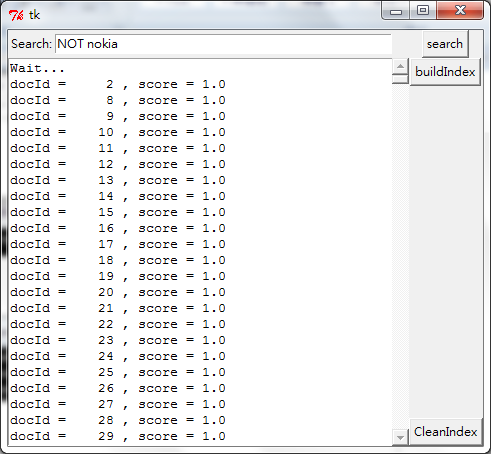
indexs hasn't created



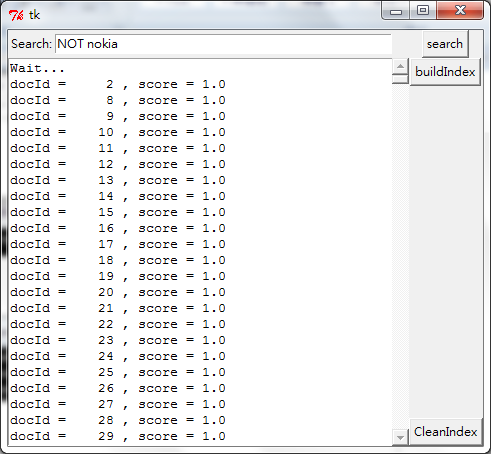
uncurrent fomat file



current fomat file and succeed



after create index, user can searching



CleanIndex-button: clean the Index. If you want change the index,before creating, cleaning Index is suggestion.

test.py: read the warc file and build indexs

the detail please look the source code and comments.

createIndex.py: create index folder and initial file

the detail please look the source code and comments.

search.py: search the index and output the rank list

the detail please look the source code and comments.