# Installation

**Programming language:** Python

**Source code location:** /home/sbillah/nlp2/

**Documentation:**

sbillah@turing:~/nlp2$ ./NLPEngine.py -h

usage: NLPEngine.py [-h] [-o DST\_DIR] [-s SRC\_DIR] [-nlp COMMANDS]

This is Syed's NLP program for HW02

optional arguments:

-h, --help show this help message and exit

-o DST\_DIR, --output DST\_DIR

The directory where output goes

-s SRC\_DIR, --source SRC\_DIR

The directory where raw files reside

-nlp COMMANDS, --nlp COMMANDS

index of all bigrams in src directory, or compute

colocaiton. e.g, -nlp bigrams, -nlp colocation

**Here is a complete command line example:**

sbillah@turing:~/nlp2$ ./NLPEngine.py -s /home/sgauch/public\_html/5013IR/files/ -o parsed/ -nlp colocations

Done!

# Algorithm

I use an in-memory Hash-table to count all bigrams. The pseudo code is given bellow:

*initialize hash-table* ***ht<(tuple), int>***

*foreach file* ***f*** *in input\_directory:*

***pain\_text*** *= html\_parser(****f****.read())*

***tokens*** *= tokenizer(****plain\_text****)*

***i = 0***

*foreach token* ***t*** *in* ***tokens****:*

*if i>0:*

***ht[(t[i-1],t[i]***)***]*** *+= 1*

***i++***

***sort ht***

***write ht to file***

**Time Complexity:**

N = num\_files

M= avg num\_of\_words\_per\_file

Bigram generation complexity: O(N\*M)

Hash-table sorting complexity: O(N\*M\*log(N\*M))

Total complexity: O(N\*M) + O(N\*M\*log(N\*M)) = **O(N\*M\*log(N\*M))**

# Parser Configuration

Here is the configuration of my html parser and tokenizer:

|  |  |
| --- | --- |
| str\_src\_dir | /home/sgauch/public\_html/5013IR/files/ |
| str\_dst\_dir | parsed/ |
| str\_doc\_id\_file\_name | bigram.txt |
|  |  |
| min\_token\_freq | 3 |
| max\_token\_freq | 1000 |
| min\_token\_len | 3 |
| max\_token\_len | 12 |
| str\_stop\_list | Stoplist from this link:  http://www.csce.uark.edu/~sgauch/5013IR/S12/index.html |

# Runtime & Memory usage

|  |  |  |  |
| --- | --- | --- | --- |
| Input size (# files) | Run time (sec) | Memory size (MB) | Total bigrams |
| 100 | 5.60 | 110 | 55,013 |
| 200 | 19.21 | 210 | 116,697 |
| 300 | 30.31 | 277 | 166,179 |
| 505 | 54.64 | 440 | 272,646 |

# Top 50 bigrams

|  |  |  |
| --- | --- | --- |
| risks | jul | 607 |
| net | alter | 344 |
| alter | dynip | 340 |
| health | care | 215 |
| paper | title | 208 |
| com | interramp | 204 |
| net | sunbelt | 189 |
| edu | psu | 177 |
| net | mci | 152 |
| edu | nodak | 144 |
| edu | uiuc | 142 |
| critical | analysis | 142 |
| mass | media | 141 |
| edu | umich | 137 |
| net | idt | 133 |
| edu | umn | 132 |
| mil | navy | 130 |
| political | science | 129 |
| rights | reserved | 127 |
| edu | arizona | 120 |
| human | rights | 117 |
| edu | indiana | 115 |
| social | security | 111 |
| hogy | nem | 111 |
| edu | utexas | 111 |
| nemzet | magyar | 110 |
| los | angeles | 109 |
| horn | gyula | 109 |
| send | comments | 108 |
| mci | campus | 108 |
| black | studies | 107 |
| world | war | 106 |
| home | page | 106 |
| book | report | 106 |
| term | papers | 105 |
| http | www | 105 |
| written | price | 104 |
| urban | studies | 104 |
| termpaper | com | 104 |
| term | paper | 104 |
| subject | index | 104 |
| sports | recreation | 104 |
| specific | paper | 104 |
| paper | written | 104 |
| paper | click | 104 |
| description | paper | 104 |
| copyright | asm | 104 |
| comments | termpaper | 104 |
| comments | comments | 104 |
| cold | surges | 104 |

# Bottom 20 bigrams

|  |  |  |
| --- | --- | --- |
| abacs | kiskun | 1 |
| ababa | response | 1 |
| aau | zoo | 1 |
| aau | psy | 1 |
| aau | hum | 1 |
| aas | nearly | 1 |
| aarp | national | 1 |
| aarp | american | 1 |
| aaron | word | 1 |
| aaron | netland | 1 |
| aaron | moshiashwili | 1 |
| aaron | jon | 1 |
| aaron | happened | 1 |
| aaron | comparison | 1 |
| aalen | image | 1 |
| aaemassago | sem | 1 |
| aaeliberalis | demokraciato | 1 |
| aachen | rad | 1 |
| aachen | oph | 1 |
| aaa | passed | 1 |