# PROGRAMMING ASSIGNMENT 3

## Search Engine Design

### Assignment Summary

Nirav Patel, Vaibhav Kamble and Yogesh Wattamwar

# Table of Contents

## Team Information

1. Team Name - **Hell Raisers**
2. Members
   a. Nirav Patel
   b. Vaibhav Kamble
   c. Yogesh Wattamwar

## Wikipedia Dump

1. enwiki-latest-pages-articles.xml2 (Size 45.97 GB)

## Implementation and observation

Phase wise development details.

### 1. Phase 1 (Page title and Page id)

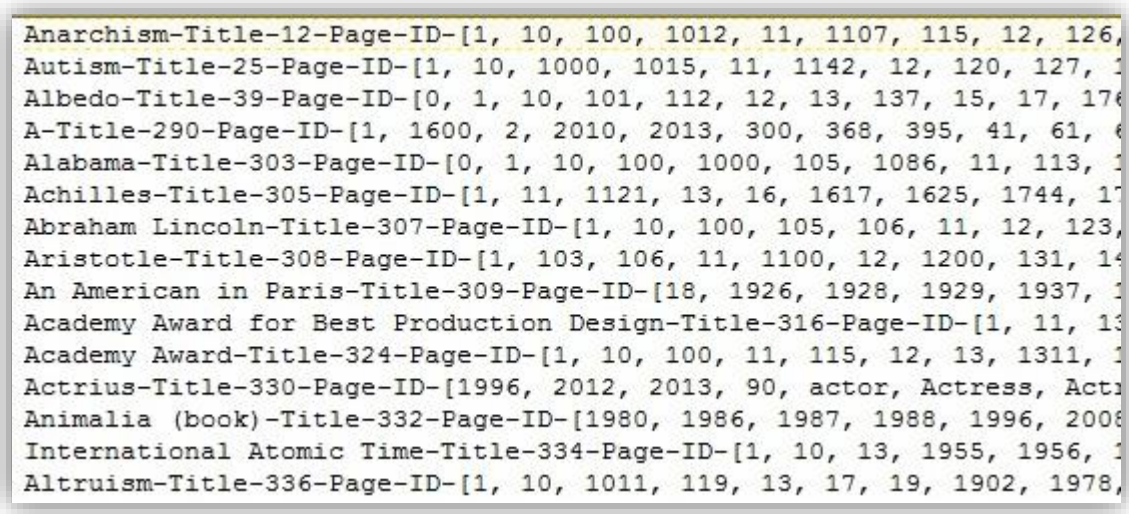
- Parsed whole wiki dump and created separate arrays for page title and page ids.
- One to one relation is maintained between page title array and page id array
- So, element at index 'i' in page title array contains title of some page in wiki and its corresponding page id is maintained at index 'i' of page id array.
- Both the array serialized, as an intermediate output.
- Result
  a. Time required - 754.626 seconds
  b. Valid page count - 5956707
  c. Bogus page count - 8172270
  d. Total page count in wiki dump – 14128977
- Serialized output
  a. TitleArray_45G_1.ser of 235 MB
  b. PageIDArray_45G_1.ser of 103 MB

### 2. Phase 2 (In-link and Out-link calculation)

- Calculated page wise in-link count and out-link count.
- Created arrays of in-link count and out-link count.
- One to one relation is maintained in page title array, page id array, in-link array ad out-link array. Such that, element in index 'i' in all array represent title, page id, in-link count, out-link count information of particular page.
- Both the array serialized, as an intermediate output.
- Result
  a. Time required for in-link calculation - 2672.057 seconds
  b. Time required for out-link calculation - 2600.797 seconds

## 3. Phase 3 (Compression of original wiki dump to text file)

- Original xml wiki dump of 45.97 GB is parsed and compressed in text file.
- Compressed text file is 7.6GB, which is used in further phases as a replacement to original wiki dump
- Each document is maintained as a single row in text file in following format
  **{Page title}-Title-{Page Id}-Page-ID-[{Token1}, {Token2},..,{Token10}]-E-O-C-**
  *Anarchism-Title-12-Page-ID-[1, 10, Anchor, bad, money, power]-E-O-C-*
- Stemming
  a. Porter stemmer is applied to tokens
- Only word with length between 3 and 16 are allowed.
- Regular expression is applied to avoid invalid words.
- Result
  a. Time required - 7992.499 seconds
  b. Size of compressed text file - 7.6GB



```
Anarchism-Title-12-Page-ID-[1, 10, 100, 1012, 11, 1107, 115, 12, 126,
Autism-Title-25-Page-ID-[1, 10, 1000, 1015, 11, 1142, 12, 120, 127, 1
Albedo-Title-39-Page-ID-[0, 1, 10, 101, 112, 12, 13, 137, 15, 17, 176
A-Title-290-Page-ID-[1, 1600, 2, 2010, 2013, 300, 368, 395, 41, 61, 6
Alabama-Title-303-Page-ID-[0, 1, 10, 100, 1000, 105, 1086, 11, 113, 1
Achilles-Title-305-Page-ID-[1, 11, 1121, 13, 16, 1617, 1625, 1744, 17
Abraham Lincoln-Title-307-Page-ID-[1, 10, 100, 105, 106, 11, 12, 123,
Aristotle-Title-308-Page-ID-[1, 103, 106, 11, 1100, 12, 1200, 131, 14
An American in Paris-Title-309-Page-ID-[18, 1926, 1928, 1929, 1937, 1
Academy Award for Best Production Design-Title-316-Page-ID-[1, 11, 13
Academy Award-Title-324-Page-ID-[1, 10, 100, 11, 115, 12, 13, 1311, 1
Actrius-Title-330-Page-ID-[1996, 2012, 2013, 90, actor, Actress, Acti
Animalia (book)-Title-332-Page-ID-[1980, 1986, 1987, 1988, 1996, 2008
International Atomic Time-Title-334-Page-ID-[1, 10, 13, 1955, 1956, 1
Altruism-Title-336-Page-ID-[1, 10, 1011, 119, 13, 17, 19, 1902, 1978,
```

*Figure 1 Compressed text version of wiki dump*

## 4. Phase 4 (Dictionary creation)

- Parsed compressed text file and created dictionary of unique tokens.
- Same token acceptance rules of phase 3 are applied in dictionary.
- Size of dictionary - 92.4 MB

## 5. Phase 5 (Posting list creation)

- Posting lists are created for small chunks of words from dictionary.
- As dictionary is sorted, we have created posting list for 100000 words. So we have maintained 86 posting lists for whole dictionary.

| Word Range | Posting list file to refer |
|---|---|
| **AAB** to **Afulai** | Postinglist1.txt |
| **arrivedthei** to **Avradh** | Postinglist4.txt |
| **Hikoma** to **hubsplainlist** | Postinglist30.txt |
| **Yannarilyi** to **Zerekli** | Postinglist81.txt |

*Figure 2 Word rang- Posting list file mapping*

- We have tried variable-byte-encoding, but it was not providing expected compression.
- Time required to calculate posting list for 100000 tokens – 35 Minutes (Average)

```
Alpenu=,157446,7715007,8143115,11554049,14492473,26080392,31330397,31331
alonethi=,1306,34237314
altaral=,2902552
alphaxtmath=,45305
Alphcat=,37576558
Altamontso=,309682,25736770
Altamaps=,1509898,5735149,5736637,8804403,12734176
Alsenesi=,6719431
ALPILIGNUM=,8233552
Alpenz=,2042648
Altargan=,156461,1666422
Altaramisch=,24566,5540419,9700083
alsopurifi=,4520753
alphabetsvg=,69874,316936,40660329
Alpinism=,20341,220861,256310,286864,289703,410279,577872,598371,653411,
alphapolymorph=,5636766
Altamasi=,19437594
Altamash=,224331,462318,1017015,1299098,2071928,2142556,2977706,2977818,
alonetim=,2266626,9431806,11801068,16390576,35384040
```

*Figure 3 Sample posting list*

## 6. Phase 6 (Search service)

- Finally, search service is implemented which accept query string as an input and returns search results.
- For search operation, we are loading dictionary, page title array, page id array, in-link count array and out-link array in memory.
- Page rank is decided on the basis of in-link count, out-link count and zone scoring.

## Languages/API Used

1. Java
2. SAXParser
3. Porter Stemmer (For stemming only)
4. Google collection (For posting list intersection only at the time query processing)
5. Commons Lang API (For alphanumeric check)

## Statistics

1. Document count: 5956707
2. Tokens Count: 8152344
3. Time required to initiate search service (To load all required data) - 436 seconds

## Development Infrastructure

1. 3 PC's with following configuration
   a. Windows 7 operating system
   b. Intel Core I5 processor
   c. 4 GB RAM
2. Eclipse IDE

## Memory Consumption Statistics

1. Memory consumption < 3GB

## Stop Words Used

1. a,able,about,above,abst,accordance,according,accordingly,across,act,actually,added,adj,affected,affecting,affects,after,afterwards,again,against,ah,all,almost,alone,along,already,also,although,always,am,among,amongst,an,and,announce,another,any,anybody,anyhow,anymore,anyone,anything,anyway,anyways,anywhere,apparently,approximately,are,aren,arent,arise,around,as,aside,ask,asking,at,auth,available,away,awfully,b,back,be,became,because,become,becomes,becoming,been,before,beforehand,begin,beginning,beginnings,begins,behind,being,believe,below,beside,besides,between,beyond,biol,both,brief,briefly,but,by,c,ca,came,can,cannot,cant,cause,causes,certain,certainly,co,com,come,comes,contain,containing,contains,could,couldnt,d,date,did,didnt,different,do,does,doesnt,doing,done,dont,down,downwards,due,during,e,each,ed,edu,effect,eg,eight,eighty,either,else,elsewhere,end,ending,enough,especially,et,et-al,etc,even,ever,every,everybody,everyone,everything,everywhere,ex,except,f,far,few,ff,fifth,first,five,fix,followed,following,follows,for,former,formerly,forth,found,four,from,further,furthermore,g,gave,get,gets,getting,give,given,gives,giving,go,goes,gone,got,gotten,h,had,happens,hardly,has,hasnt,have,havent,having,he,hed,hence,her,here,hereafter,hereby,herein,heres,hereupon,hers,herself,hes,hi,hid,him,himself,his,hither,home,how,howbeit,however,hundred,i,id,ie,if,ill,im,immediate,immediately,importance,important,in,inc,indeed,index,information,instead,into,invention,inward,is,isnt,it,itd,itll,its,itself,ive,j,just,k,keep,keeps,kept,kg,km,know,known,knows,l,largely,last,lately,later,latter,latterly,least,less,lest,let,lets,like,liked,likely,line,little,ll,look,looking,looks,ltd,m,made,mainly,make,makes,many,may,maybe,me,mean,means,meantime,meanwhile

,merely,mg,might,million,miss,ml,more,moreover,most,mostly,mr,mrs,much,mug,must,my,myself,n,na,name,namely,nay,nd,near,nearly,necessarily,necessary,need,needs,neither,never,nevertheless,new,next,nine,ninety,no,nobody,non,none,nonetheless,noone,nor,normally,nos,not,noted,nothing,now,nowhere,o,obtain,obtained,obviously,of,off,often,oh,ok,okay,old,omitted,on,once,one,ones,only,onto,or,ord,other,others,otherwise,ought,our,ours,ourselves,out,outside,over,overall,owing,own,p,page,pages,part,particular,particularly,past,per,perhaps,placed,please,plus,poorly,possible,possibly,potentially,pp,predominantly,present,previously,primarily,probably,promptly,proud,provides,put,q,que,quickly,quite,qv,r,ran,rather,rd,re,readily,really,recent,recently,ref,refs,regarding,regardless,regards,related,relatively,research,respectively,resulted,resulting,results,right,run,s,said,same,saw,say,saying,says,sec,section,see,seeing,seem,seemed,seeming,seems,seen,self,selves,sent,seven,several,shall,she,shed,shell,shes,should,shouldnt,show,showed,shown,showns,shows,significant,significantly,similar,similarly,since,six,slightly,so,some,somebody,somehow,someone,somethan,something,sometime,sometimes,somewhat,somewhere,soon,sorry,specifically,specified,specify,specifying,still,stop,strongly,sub,substantially,successfully,such,sufficiently,suggest,sup,sure,textxwiki,Wikipedia,wikitext,I,a,about,an,are,as,at,be,by,com,for,from,how,in,is,it,of,on,or,that,the,this,to,was,what,when,where,who,will,with,the,a,about,above,after,again,against,all,am,an,and,any,are,arent,as,at,be,because,been,before,being,below,between,both,but,by,cant,cannot,could,couldnt,did,didnt,do,does,doesnt,doing,dont,down,during,each,few,for,from,further,had,hadnt,has,hasnt,have,havent,having,he,hed,hell,hes,her,here,heres,hers,herself,him,himself,his,how,hows,i,id,ill,im,ive,if,in,into,is,isnt,it,its,its,itself,lets,me,more,most,mustnt,my,myself,no,nor,not,of,off,on,once,only,or,other,ought,our,ours,ourselves,out,over,own,same,shant,she,shed,shell,shes,should,shouldnt,so,some,such,than,that,thats,the,their,theirs,them,themselves,then,there,theres,these,they,theyd,theyll,theyre,theyve,this,those,through,to,too,under,until,up,very,was,wasnt,we,wed,well,were,weve,were,werent,what,whats,when,whens,where,wheres,which,while,who,whos,whom,why,whys,with,wont,would,wouldnt,you,youd,youll,youre,youve,your,yours,yourself,yourselves,www,div,class,dates,date,See,Notes,References,reflist,sites,Thumb.

## External references


1. http://www.sfs.uni-tuebingen.de/~parmenti/code/VariableByte.java
2. http://en.wikipedia.org/wiki/Stemming