# Project 3 Report

## Analysis:

1. **Total parsed 34592 documents**
2. **Total got 147552 unique tokens**
3. **Inverted Index Size = 172,687 KB (168MB)**

## Query:

## 1. Enter a word to search: Informatics

After lemmatize, word = informatics

Found 1722 urls in total

Top 10 documents:

1 , 67/115

normalized tf-idf = 0.26008390438340245

# of unique tokens = 6

url = fano.ics.uci.edu/cites/Location/Proc-10th-Genome-Informatics-Worksh.html

2 , 41/103

normalized tf-idf = 0.2423017169055143

# of unique tokens = 9

url = fano.ics.uci.edu/cites/Organization/Univ-of-Athens-Dept-of-Informatics-and-Telecommunications.html

3 , 72/94

normalized tf-idf = 0.24149196584590546

# of unique tokens = 9

url = fano.ics.uci.edu/cites/Organization/Univ-of-Bergen-Dept-of-Informatics.html

4 , 72/416

normalized tf-idf = 0.23028270603603923

# of unique tokens = 16

url = www.ics.uci.edu/~thornton/inf122/ProjectGuide

5 , 48/18

normalized tf-idf = 0.20722319920829538

# of unique tokens = 10

url = fano.ics.uci.edu/cites/Organization/Masaryk-Univ-Faculty-of-Informatics.html

6 , 39/287

normalized tf-idf = 0.20237308627476772

# of unique tokens = 11

url = fano.ics.uci.edu/cites/Location/Proc-1st-Latin-American-Symp-Theoretical-Informatics-(LATIN-1992).html

7 , 30/369

normalized tf-idf = 0.20030957531645965

# of unique tokens = 31

url = www.ics.uci.edu/~etrainer/theseus/personnel.html

8 , 15/145

normalized tf-idf = 0.19452176782313163

# of unique tokens = 56

url = www.ics.uci.edu/faculty/area/area_social.php

9 , 21/496

normalized tf-idf = 0.1940368077804438

# of unique tokens = 46

url = www.ics.uci.edu/~thornton/inf122

10 , 71/3

normalized tf-idf = 0.18309394546556404

# of unique tokens = 13

url = fano.ics.uci.edu/cites/Organization/Warsaw-Univ-Inst-of-Informatics.html

# 2. Enter a word to search: Mondego

After lemmatize, word = mondego

Found 1 urls in total

Top 10 documents:

1 , 19/404

normalized tf-idf = 0.38738393239746005

# of unique tokens = 115

url = mondego.ics.uci.edu

# 3. Enter a word to search: Irvine

After lemmatize, word = irvine

Found 5944 urls in total

Top 10 documents:

1 , 31/128

normalized tf-idf = 0.2517685559480208

# of unique tokens = 8

url = vision.ics.uci.edu/events.html

2 , 41/410

normalized tf-idf = 0.2517685559480208

# of unique tokens = 8

url = vision.ics.uci.edu/links.html

3 , 73/490

normalized tf-idf = 0.2517685559480208

# of unique tokens = 8

url = vision.ics.uci.edu/projects.html

4 , 3/374

normalized tf-idf = 0.23529903641093988

# of unique tokens = 17

url = www.ics.uci.edu/about/visit/../../grad/resources.php

5 , 34/454

normalized tf-idf = 0.23529903641093988

# of unique tokens = 17

url = www.ics.uci.edu/grad/resources.php

6 , 42/44

normalized tf-idf = 0.23529903641093988

# of unique tokens = 17

url = www.ics.uci.edu/grad/resources

7 , 23/428

normalized tf-idf = 0.22659233639786117

# of unique tokens = 9

url = fano.ics.uci.edu/cites/Organization/Univ-of-California-Irvine-Dept-of-Computer-Science.html

8 , 54/415

normalized tf-idf = 0.22394075344726883

# of unique tokens = 18

url = seal.ics.uci.edu?p=22

9 , 3/498

normalized tf-idf = 0.22323458018346726

# of unique tokens = 8

url = www.ics.uci.edu/~copper

10 , 52/469

normalized tf-idf = 0.21653024207212182

# of unique tokens = 8

url = testlab.ics.uci.edu

# 4. Enter a word to search: artificial intelligence

After lemmatize, word = artificial

After lemmatize, word = intelligence

Found 424 urls in total

Top 10 documents:

1 , 17/128

normalized tf-idf = 0.5418476345015446

# of unique tokens = 11

url = www.ics.uci.edu/~dechter/courses/ics-171/fall-06

2 , 68/490

normalized tf-idf = 0.4678270418640349

# of unique tokens = 23

url = sli.ics.uci.edu/Classes/2009W?action=login

3 , 39/280

normalized tf-idf = 0.44895363151927403

# of unique tokens = 19

url = www.ics.uci.edu/~dechter/courses/ics-171/spring-99

4 , 48/264

normalized tf-idf = 0.40414049763460036

# of unique tokens = 17

url = www.ics.uci.edu/~pazzani/Publications/APubs.html

5 , 0/212

normalized tf-idf = 0.38592239324036337

# of unique tokens = 7

url = www.ics.uci.edu/~kobsa/courses/ICS104/course-notes/contr-disciplines.htm

6 , 13/120

normalized tf-idf = 0.33075035496455707

# of unique tokens = 72

url = cml.ics.uci.edu/2010/07/2010_smythaaai

7 , 47/124

normalized tf-idf = 0.3206175266266435

# of unique tokens = 66

url = cml.ics.uci.edu/2014/08/2014_aaai

8 , 62/85

normalized tf-idf = 0.29237534159906425

# of unique tokens = 8

url = fano.ics.uci.edu/cites/Location/IEEE-Trans-Pattern-Analysis-+-Machine-Intelligence.html

9 , 4/452

normalized tf-idf = 0.27891506569204993

# of unique tokens = 32

url = www.ics.uci.edu/~jesmaeln/research.html

10 , 34/291

normalized tf-idf = 0.27376278273889365

# of unique tokens = 35

url = emj.ics.uci.edu/teaching

# 5. Enter a word to search: computer science

After lemmatize, word = computer

After lemmatize, word = science

Found 8402 urls in total

Top 10 documents:

1 , 4/325

normalized tf-idf = 0.5919092455912529

# of unique tokens = 4

url = fano.ics.uci.edu/cites/Location/Science-of-Computer-Programming.html

2 , 70/347

normalized tf-idf = 0.4439069871820991

# of unique tokens = 5

url = fano.ics.uci.edu/cites/Location/Lecture-Notes-in-Computer-Science.html

3 , 55/50

normalized tf-idf = 0.4321159385025687

# of unique tokens = 2

url = fano.ics.uci.edu/cites/Location/Science-News.html

4 , 46/165

normalized tf-idf = 0.341809147251755

# of unique tokens = 6

url = fano.ics.uci.edu/cites/Location/J-Computer-+-Systems-Sciences.html

5 , 68/425

normalized tf-idf = 0.3180513671680717

# of unique tokens = 13

url = www.ics.uci.edu/community/egiving/2008/video.php

6 , 28/341

normalized tf-idf = 0.3015153878883452

# of unique tokens = 17

url = www.ics.uci.edu/about/search/search_graduate_all.php

7 , 30/76

normalized tf-idf = 0.3015153878883452

# of unique tokens = 17

url = www.ics.uci.edu/community/alumni/mentor//../../../../about/search/search_graduate_all.php

8 , 9/56

normalized tf-idf = 0.28229760293611594

# of unique tokens = 6

url = fano.ics.uci.edu/cites/Location/Handbook-of-Computer-Science-+-Engineering.html

9 , 74/24

normalized tf-idf = 0.27619841127734246

# of unique tokens = 18

url = www.ics.uci.edu/community/events/butterworth

10 , 23/428

normalized tf-idf = 0.27240114621640094

# of unique tokens = 9

url = fano.ics.uci.edu/cites/Organization/Univ-of-California-Irvine-Dept-of-Computer-Science.html