Nokia Interview Presentation

Yayang Tian

University of Pennsylvania

yaytian@cis.upenn.edu

August 20, 2013

Education

University of Pennlsyvania

- M.S.E. in Computer and Information Science, Big Data

Shanghai Jiao Tong University

- B.E. in Information Science, Honor Class

Shenzhen Middle School

- Physics, Honor Class, CPO 1st Prize

Experience

Research Assitant - NLP, University of Pennlsyvania

- Emotion classification with unsupervised learning.
- Implicit expressions of emotion in text. Python.

Software Developer - CS Dept, University of Pennlsyvania

- Automatic methods to measure quality in text.
- Crowdsourcing websites for data mining. Php/MySQL.

Software Engineer Intern - Cisco Systems, Shanghai

- J2ME applications for digital TV on set-top box terminals.
- Video on demand and interactive services. Java.

Projects Overview

Distributed Systems

- Distributed Web Search Engine: MiniGoogle
- DHT-based Search Engine: PennSearch
- Distributed Caching and Web Services
- Web Crawler with Multithreaded Server

Data Mining

- Amazon Reviews Data Mining
- Tweets Emotion Prediction
- New York Times Summarization
- Wall Street Journal Crowdsouring Website

Others

- Mobile Applications
- Database Website on GAE

MiniGoogle

Techniques

- Java, Hadoop, MapReduce, AWS, FreePastry, Berkeley DB.

Components

- Crawler, indexer, PageRank, web UI.

Contribution

- MapReduce: TF-IDF information retrieval for indexing;
- Ranking/Search Relevance: scoring, weighting, ranking; presented new ranking algorithm including 10 features;
- Website: designed and implemented the website with Twitter Bootstrap and HTML/CSS;
- Web Services: REST APIs, recommendation engine integration.

MiniGoogle


Globus@Penn
Globus
Search engine makes global
food
Advanced
Yahoo
Amazon
YouTube
Safe
Copyright © 2013 - Michael, Yayang, Angela, Krishna

MiniGoogle

Globus | Search... | Advanced Search | About Team

Web | Flickr | YouTube | eBay | Maps

Results Integration

Powered by Michael, Yayang, Angela, Krishna

Home / Web

Globus Results

Including results for food? Feedback changed the results. Click to Reload.

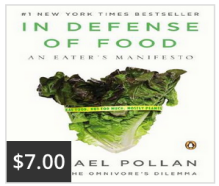
Sort by Relevence Authority Filter by Nearby All Time

201 results
2.81 seconds

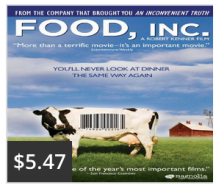
Top amazon

Food (4 customer reviews)

Happy Baby Organic .. (180 customer reviews)



In Defense of Food:.. (558 customer reviews)



Food, Inc. (1,268 customer reviews)

Recipes & Cooking Tips - Yahoo!7 Food:
http://www.boston.com/lifestyle/food/
Delicious recipes with photos and easy to follow instructions, categorised by cuisine and

yelp Recommendation

Denise's Soul Food Restaurant
Soul Food 203.7 M
Was back in Philly last weekend. Worry not - Denise's was ba

Tang's
Chinese 191.3 M
This is my absolute favorite food truck hands down. It's dir

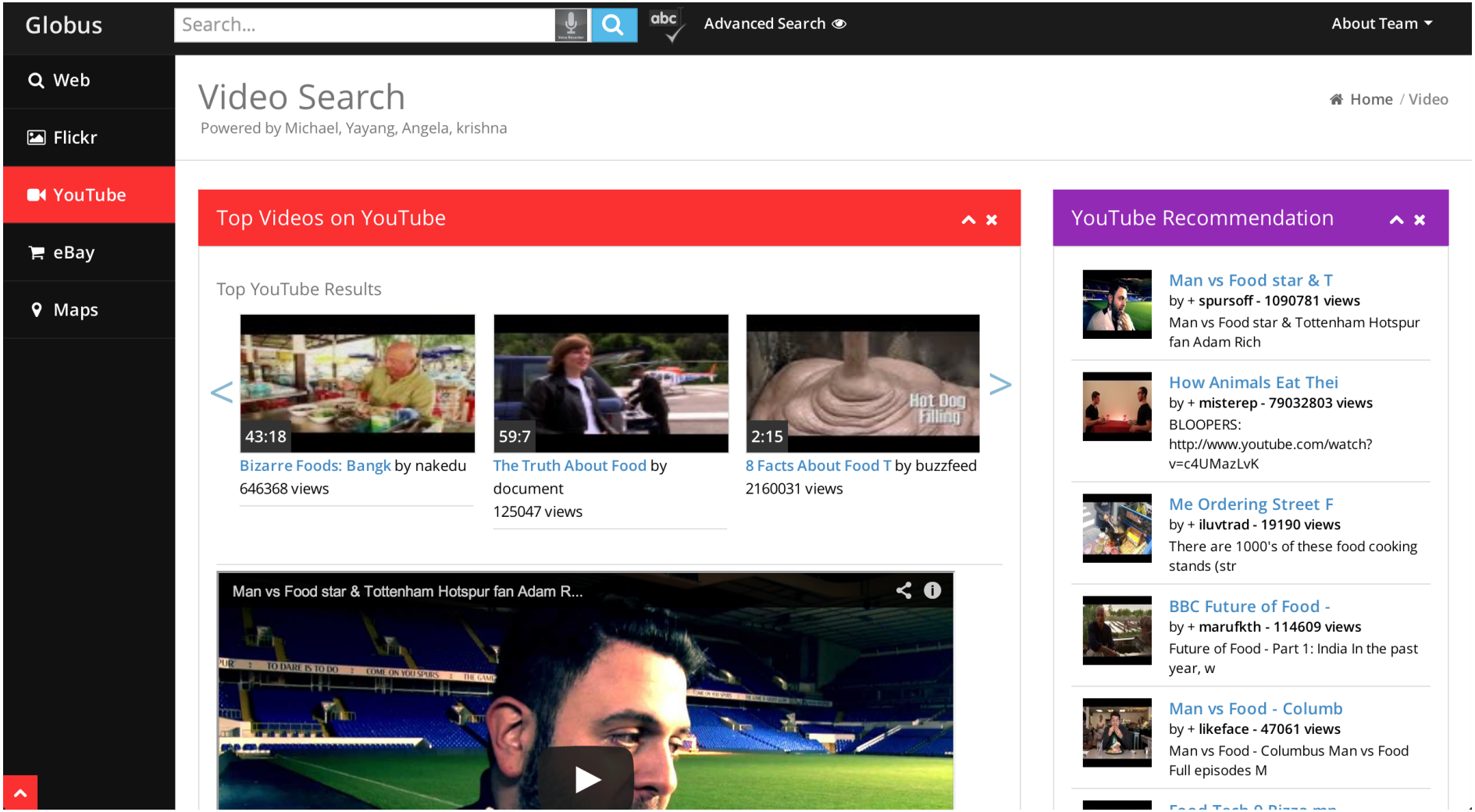
Han Dynasty
Chinese 340.7 M
Our experience was amazing! I've been to the Old City Han Dy

YouTube Recommendation

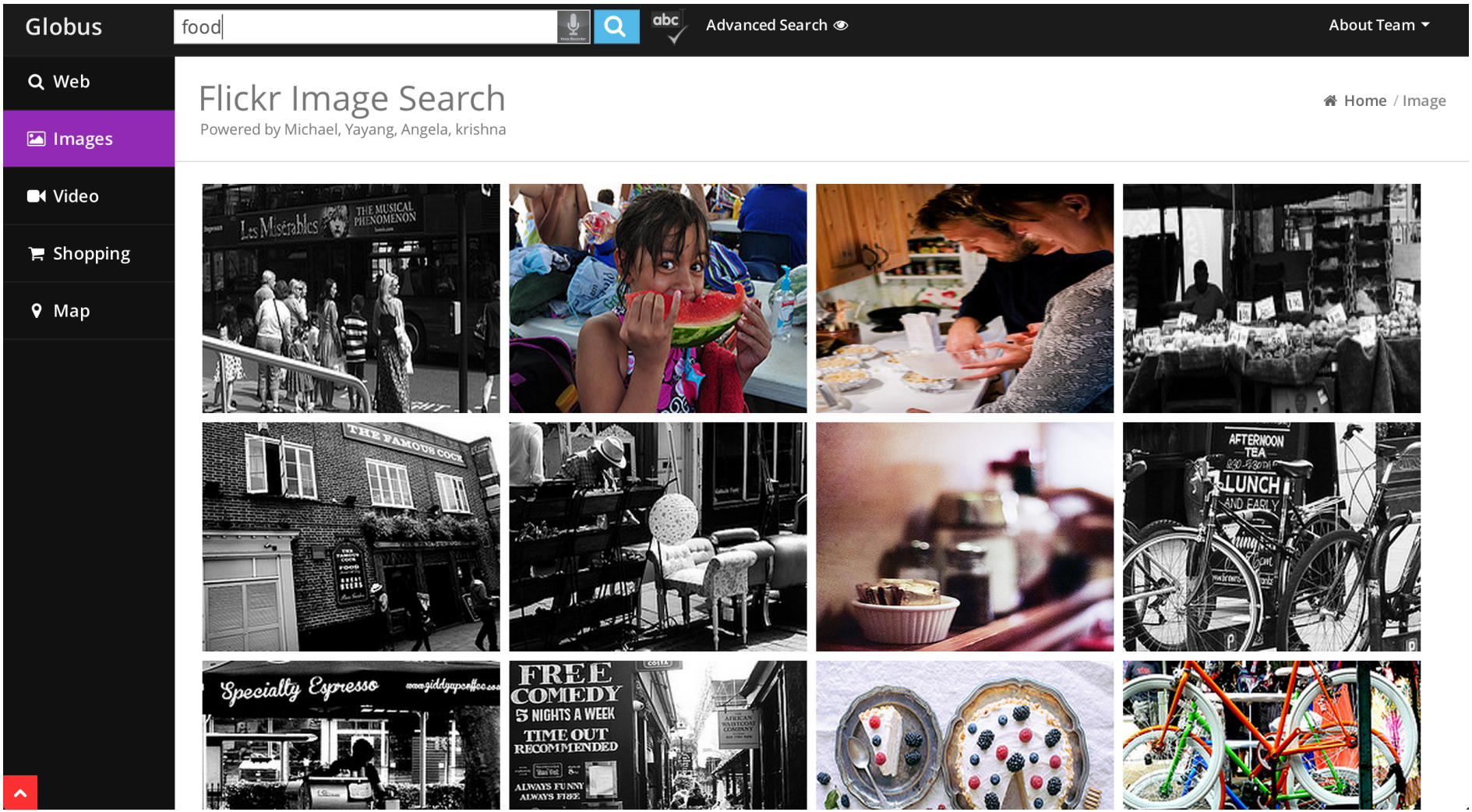
Man vs Food star & T
by + spursoff - 1090781 views
Man vs Food star & Tottenham Hotspur fan Adam Rich

How Animals Eat Thei

MiniGoogle


Globus
Search...
Advanced Search
About Team
Web
Flickr
YouTube
eBay
Maps
Video Search
Powered by Michael, Yayang, Angela, krishna
Home / Video
Top Videos on YouTube
Top YouTube Results
43:18
Bizarre Foods: Bangk by nakedu
646368 views
59:7
The Truth About Food by document
125047 views
2:15
Hot Dog Filling
8 Facts About Food T by buzzfeed
2160031 views
Man vs Food star & Tottenham Hotspur fan Adam R...
YouTube Recommendation
Man vs Food star & T
by + spursoff - 1090781 views
Man vs Food star & Tottenham Hotspur fan Adam Rich
How Animals Eat Thei
by + misterep - 79032803 views
BLOOPERS: http://www.youtube.com/watch?v=c4UMazLvK
Me Ordering Street F
by + iluvtrad - 19190 views
There are 1000's of these food cooking stands (str
BBC Future of Food -
by + marufkth - 114609 views
Future of Food - Part 1: India In the past year, w
Man vs Food - Columb
by + likeface - 47061 views
Man vs Food - Columbus Man vs Food Full episodes M

MiniGoogle


Globus
food
Advanced Search
About Team
Web
Images
Video
Shopping
Map
Flickr Image Search
Powered by Michael, Yayang, Angela, krishna
Home / Image
Les Misérables
THE MUSICAL PHENOMENON
THE FAMOUS COCK
AFTERNOON TEA
LUNCH
AND EARLY
Specialty Espresso
FREE COMEDY
5 NIGHTS A WEEK
TIME OUT RECOMMENDED
ALWAYS FUNNY
ALWAYS FREE

Crawler


Topic Specific Web Crawler
cis555-hw2ms2 | Yayang Tian
Login
Register
XPath Channels
Logout

Crawler

```
GET /target/about.txt HTTP/1.1
Host: localhost

HTTP/1.1 200 OK
Content-Type: text/plain
Date: Tue Feb 08 16:39:46 EST 2011
content-length:46
Server: HTTP Server

This directory can contain HTML or JSP pages.
Connection closed by foreign host.
```

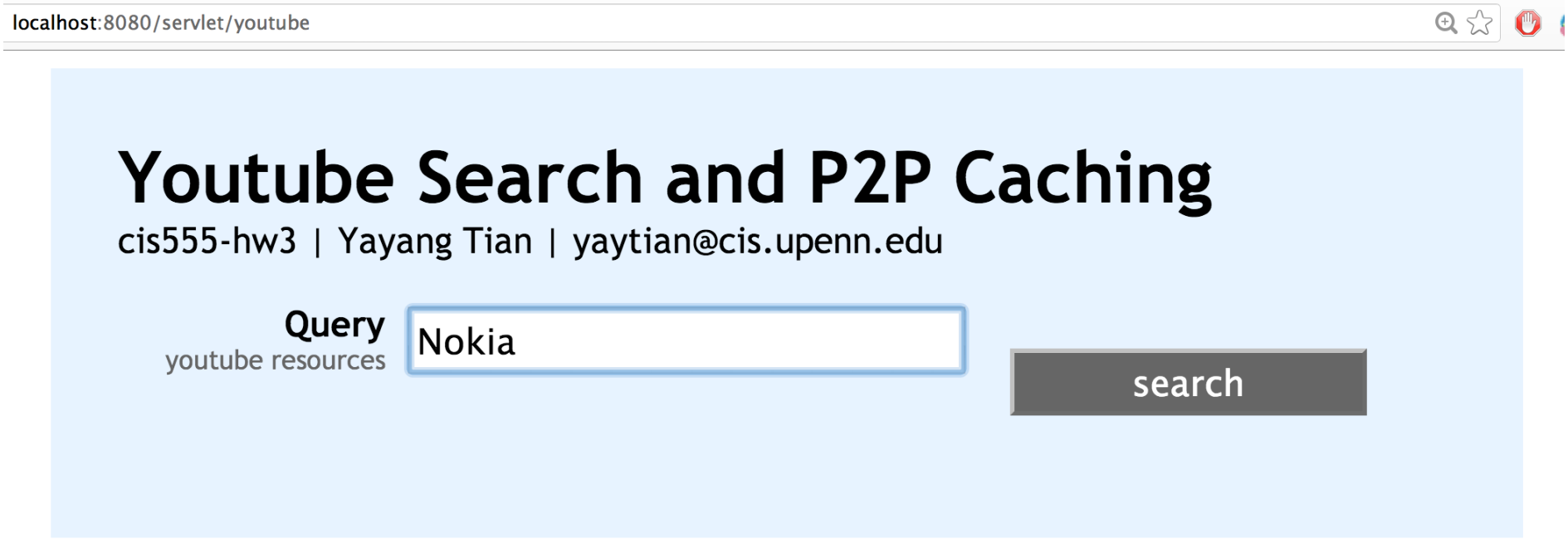
Crawler

```
[status] Database started at /Users/Alantyy/Desktop/BerkeleyDB

startURL = http://crawltest.cis.upenn.edu/
doRoot = /Users/Alantyy/Desktop/BerkeleyDB
maxSize = 100
maxNum = 100
Start Crawling.



------------------------------------------------------------
http://crawltest.cis.upenn.edu/
1. Politeness:           -> [Downloading /robots.txt]  -> Wait 5 s... -> Polite Now.
3. Get file:              -> Send HEAD ->Got cached file.
3. Process HTML:         -> Links Parsed -> to Frontier.
------------------------------------------------------------
http://crawltest.cis.upenn.edu/nytimes/
1. Politeness:           -> Wait 4 s... -> Polite Now.
3. Get file:              -> Send HEAD ->Got cached file.
3. Process HTML:         -> Links Parsed -> to Frontier.
------------------------------------------------------------
http://crawltest.cis.upenn.edu/bbc/
1. Politeness:           -> Wait 4 s... -> Polite Now.
3. Get file:              -> Send HEAD ->Got cached file.
3. Process HTML:         -> Links Parsed -> to Frontier.
------------------------------------------------------------
http://crawltest.cis.upenn.edu/cnn/
```

DHTCaching

```
[bindAddr] http://158.130.213.1:9010  [daemonAddr] http://158.130.213.1:10010
  [db] /Users/Alantyy/Desktop/database
Daemon thread waiting for WSDL query from servlet ...
Received PING to <0x77C6B4..>from node <0x4CD59D..>[port: 9008] Returning PONG.
Sending PING to <0xB05E72..>
Received PONG from node <0xB33F6D..>[port: 9002]

Received PING to <0x6C641D..>from node <0x42B4A8..>[port: 9001] Returning PONG.
Sending PING to <0x6563FC..>
Received PONG from node <0x4CD59D..>[port: 9008]

Sending PING to <0x67E3B6..>
Received PING to <0x67E3B6..>from node <0x8101D2..>[port: 9010] Returning PONG.
Received PONG from node <0x8101D2..>[port: 9010]

Received PING to <0x83F0D5..>from node <0xA5E5DC..>[port: 9007] Returning PONG.
Received PING to <0x778B4E..>from node <0xB33F6D..>[port: 9002] Returning PONG.
Sending PING to <0x39B22B..>
Received PONG from node <0x381C0E..>[port: 9005]

Received PING to <0x72C072..>from node <0xA5E5DC..>[port: 9007] Returning PONG.
Sending PING to <0x557CBE..>
Received PONG from node <0x4CD59D..>[port: 9008]

Received PING to <0x766AF5..>from node <0x4AB445..>[port: 9004] Returning PONG.
```

DHTCaching


localhost:8080/servlet/youtube
Youtube Search and P2P Caching
cis555-hw3 | Yayang Tian | yaytian@cis.upenn.edu
Query
youtube resources
Nokia
search

TwitterAffect

TwitterAffect

CIS630 Project 2 - Twitter Affect

Tao Feng, Yayang Tian, Chun Chen

Tweets: Latest Stream Crawled	Emotion Prediction
RT @JoeBieke: Two more weeks of this "school" malarchy, and I'll be lovin life. Just gotta finish strong. #selfmotivation	happy
@TheeRealJayy_ ohh alright fasho killa, see ya when I get out school lmao ✊✌	sad
RT @RandomPuber: Op elke school word er wel een docent 'pedo' genoemd..	angry
RT @TheFactsBook: Didaskaleinophobia is actually the fear of going to school.	afraid
Debating If I Wanna Go Up To My Old School Today Since I Dont Have A Fucking Umbrella😒	angry
Look! Maggie's On the Cover of the Kansas State High School Activities Journal!: http://t.co/IiusRL5e5Y via	afraid

NLP-Crowdsourcing

Logout

Read the following text and answer the questions below

Rater: 101 Remaining snippets to rate: 122

Instruction*

(FileName: snippetFiles/2000_01_25_1171222.xml-28)

In general, Dr. Weinberg said, he believes that ''half-baked philosophy has sometimes gotten in the way of doing science.''

And then there are his pronouncements on religion and deism, including his much-quoted aphorism, ''The more the universe seems comprehensible, the more it also seems pointless.''

But in the seldom-cited passages that follow, Dr. Weinberg professes belief in his own kind of conviction, the idea that the scientific effort to uncover a complete theory of the universe is one of the things that can in itself add dignity and meaning to human existence.

As for conventional religion, though, his views are uncompromising: it is not only silly but damaging to human civilization. ''The whole history of the last thousands of years has been a history of religious persecutions and wars, pogroms, jihads, crusades,'' he said. ''I find it all very regrettable, to say the least.''

Actually, Dr. Weinberg does occasionally entertain the possibility that there might be a God. While sitting in his study, with its striking view of Lake Austin, he imagined himself in the role of the biblical Abraham, whose faith God tested by commanding that he sacrifice his own son.

Rate the sentiment value of the above text.

○1(no presence of aspect)	○2	○3	○4	○5	○6	○7	○8	○9	○10(very high degree)

Please enter comments if any*

Submit

NLP-Crowdsourcing

Read the following pair of texts and answer the questions below. Rater: 201 Remaining pairs to rate: 47

A:

Health Agency Tightens Rules Governing Federal Scientists

After accusations that some government scientists used their official positions for private gain, the National Institutes of Health announced rules on Thursday that ban its scientists from consulting for drug companies.

"Our research should be based on scientific evidence that is not influenced by any other factors," Dr. Elias A. Zerhouni, director of the health institutes, said at a news conference.

The rules are being issued after disclosures that scientists at the institutes leveraged their positions to land lucrative consulting contracts that seemed to conflict with their official duties or at least overlap with them. Those contracts caused some critics to worry that research by the agency could be tainted.

An investigation by the agency concluded that 44 of its 1,200 senior scientists appeared to have violated rules governing consulting and that 9 might have violated criminal laws.

The conflicts were first reported by The Los Angeles Times.

B:

Agency Scientists Divided Over Ethics Ban on Consulting

New rules prohibiting outside consulting arrangements by researchers at the National Institutes of Health have been welcomed by some scientists, with the agency's director saying that group has "said that we needed this." But others, said the director, Dr. Elias Zerhouni, have threatened "to walk across the street" to work for organizations that do not have such a ban.

The rules, formally announced at a news conference on Tuesday, ban consulting arrangements by scientists at the agency and pharmaceutical and biotech companies.

"There's no doubt that among the majority of the 5,000 scientists who never did anything wrong and never broke any rules, they see this as being taken into a tsunami of regulations," Dr. Zerhouni said.

Dr. Sheldon Krimsky, a researcher at Tufts University who specializes in science policy and ethics, said the changes were welcome, albeit overdue.

1. Rate how similar the topic of the two articles are :

○ 1(not similar at all)	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10(very simiar)

2. Which article did you find more interesting?

Logout

Manager Page:

You can either see the annotation results, or clear all the results to re-rate.

Result of Task2:
[rating_similar] 1~10 means 'no similar at all' to 'extremely simiar';
[rating_interesting] A++: 'A is way more interesting', A+: 'A is more interesting', A=B: no preference, and etc.

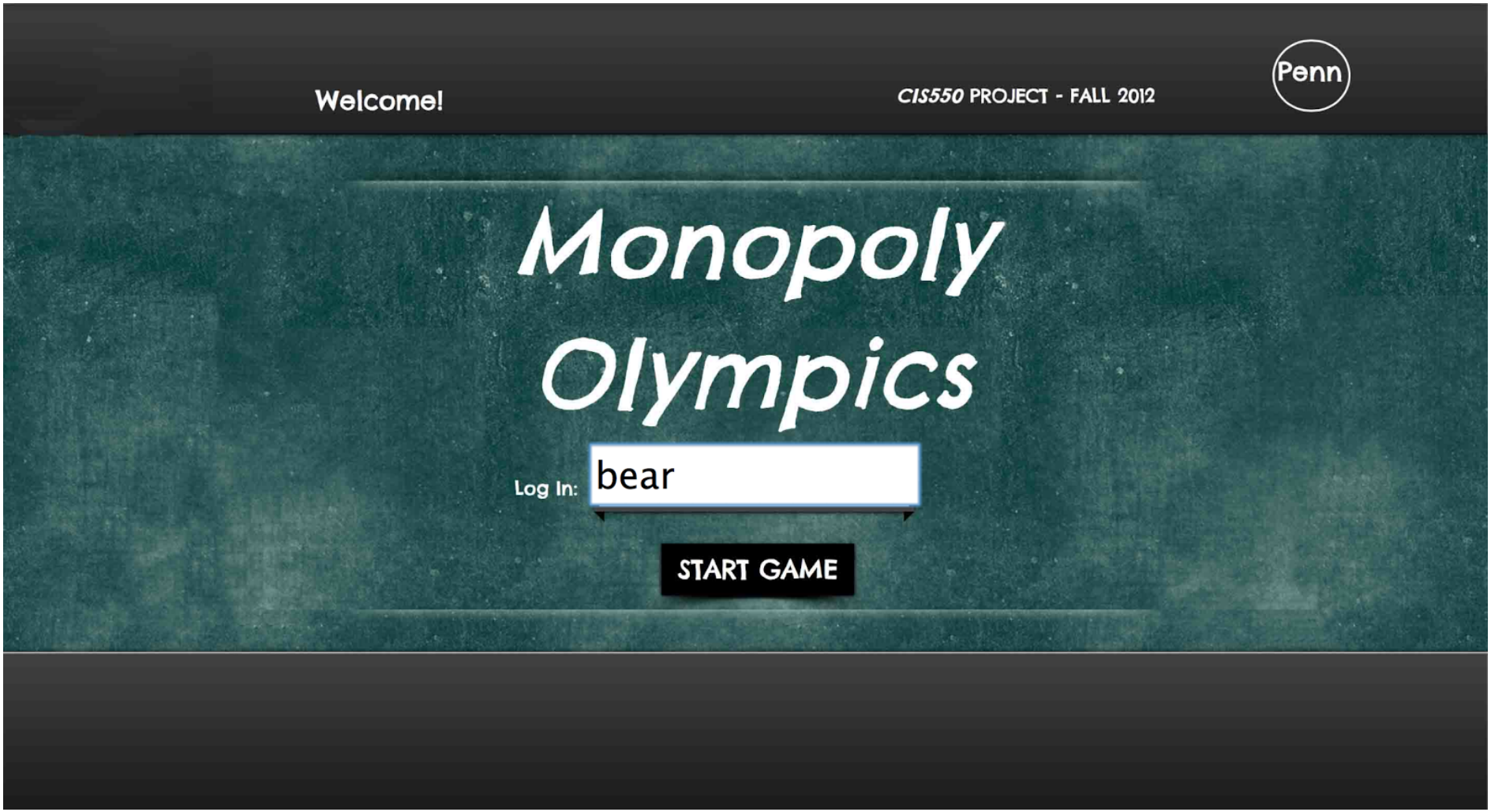
user_id	pair_id	pair_filename1	pair_filename2	rating_similar
202	1	articleFiles/2005_07_05_1685151.xml	articleFiles/2005_08_28_1697943.xml	8
202	2	articleFiles/2007_03_04_1830281.xml	articleFiles/2005_04_24_1667418.xml	9
201	1	articleFiles/2005_07_05_1685151.xml	articleFiles/2005_08_28_1697943.xml	9
203	1	articleFiles/2005_07_05_1685151.xml	articleFiles/2005_08_28_1697943.xml	5
201	2	articleFiles/2007_03_04_1830281.xml	articleFiles/2005_04_24_1667418.xml	2

Task1 Task2 Task3 Clear T1 Clear T2 Clear T3

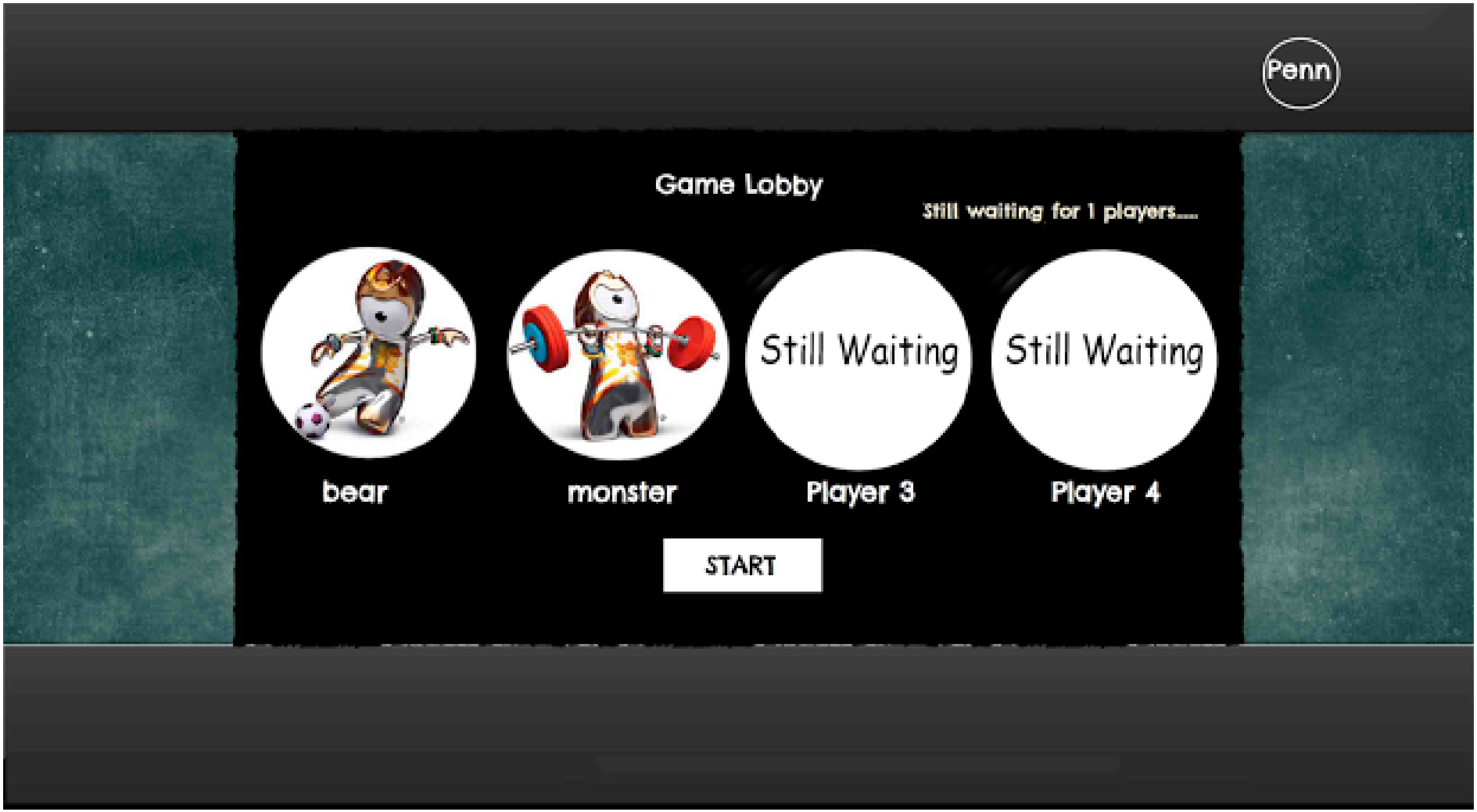
Summarization

```
#Step1. Convert sentences into math -- vector
def create_feature_space(sentence_list):
    #This creates a mapping between each word and vector.
    joint_sentences=' '.join(sentence_list)
    split_words=joint_sentences.split()
    word_type=list(set(split_words))
    dict_map=[(word_type[i],i) for i in range(len(word_type))]
    return dict(dict_map)

def vectorize(feature_space, sentence):
        #This creates a vector space
        k=feature_space.keys()
        s_list=sentence.split()
        return [val in s_list for val in k]


#Step2. Ranking Sentences by centrality
def rank_sentences(sentence_list, sim_func):
    #feature_space
    fs = create_feature_space(sentence_list)
    #vector_list
    vl = [(sent,vectorize(fs,sent)) for sent in sentence_list]
    centrality=[(vl[m][0],sum([sim_func(vl[m][1],vl[n][1]) for n
    return sorted(centrality,key=itemgetter(1),reverse=True)

#Step3. Derive a summary taken from the top ranked sentences
def summarize(sentence_list):
    ranked_sents=rank_sentences(sentence_list, sim_func)
    summary = []
```

MonopolyOlympics

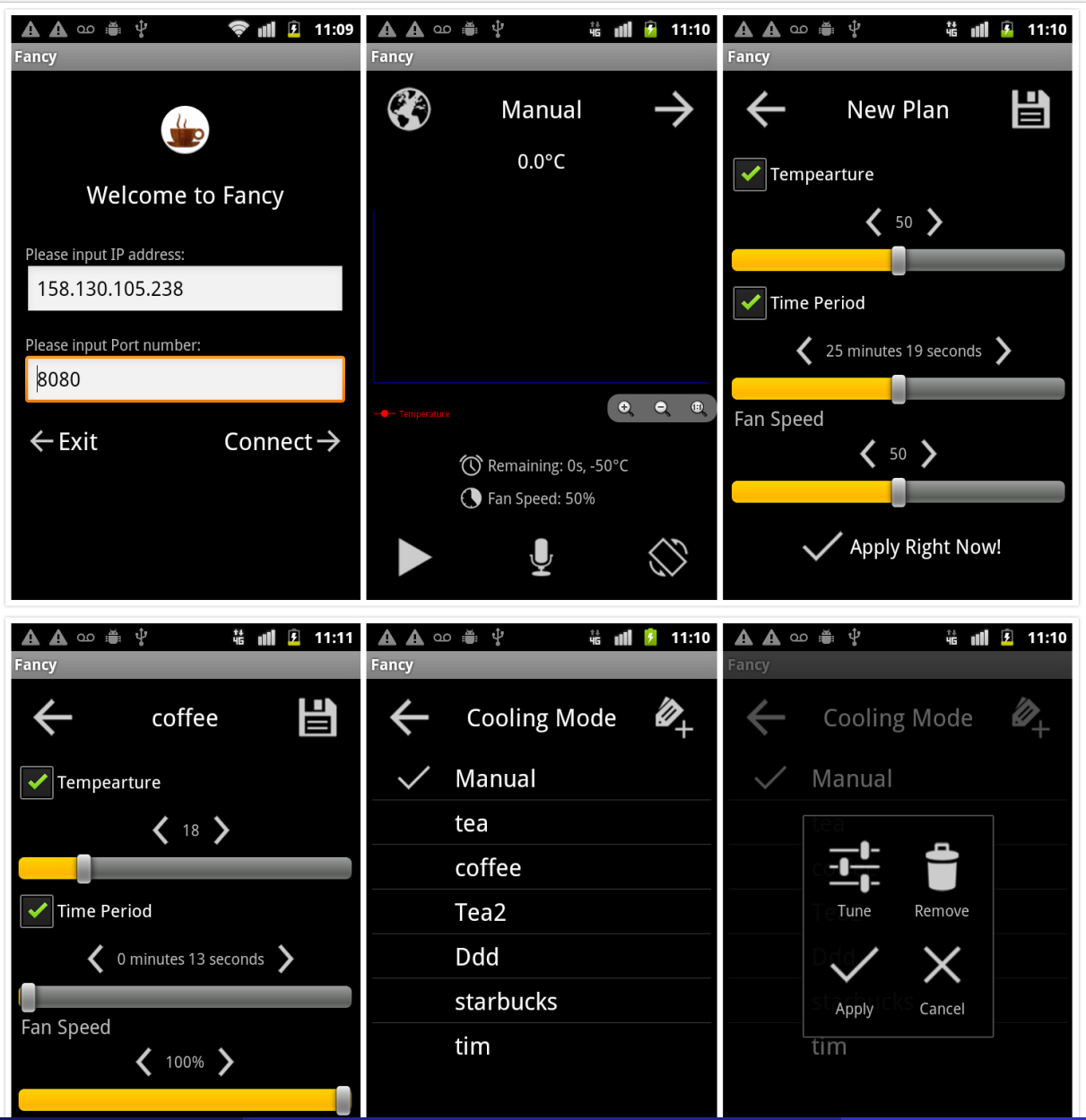


MonopolyOlympics

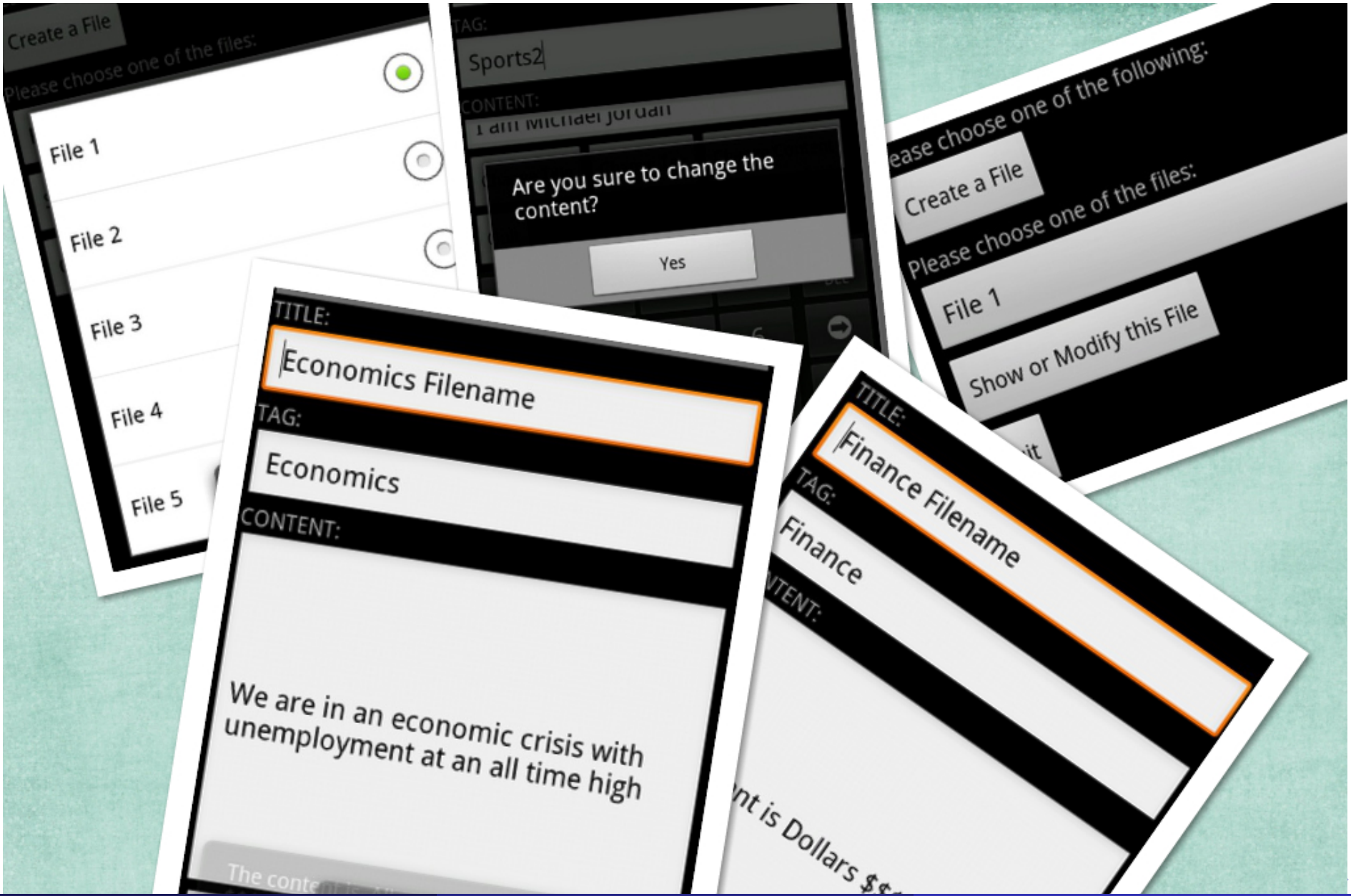


MonopolyOlympics

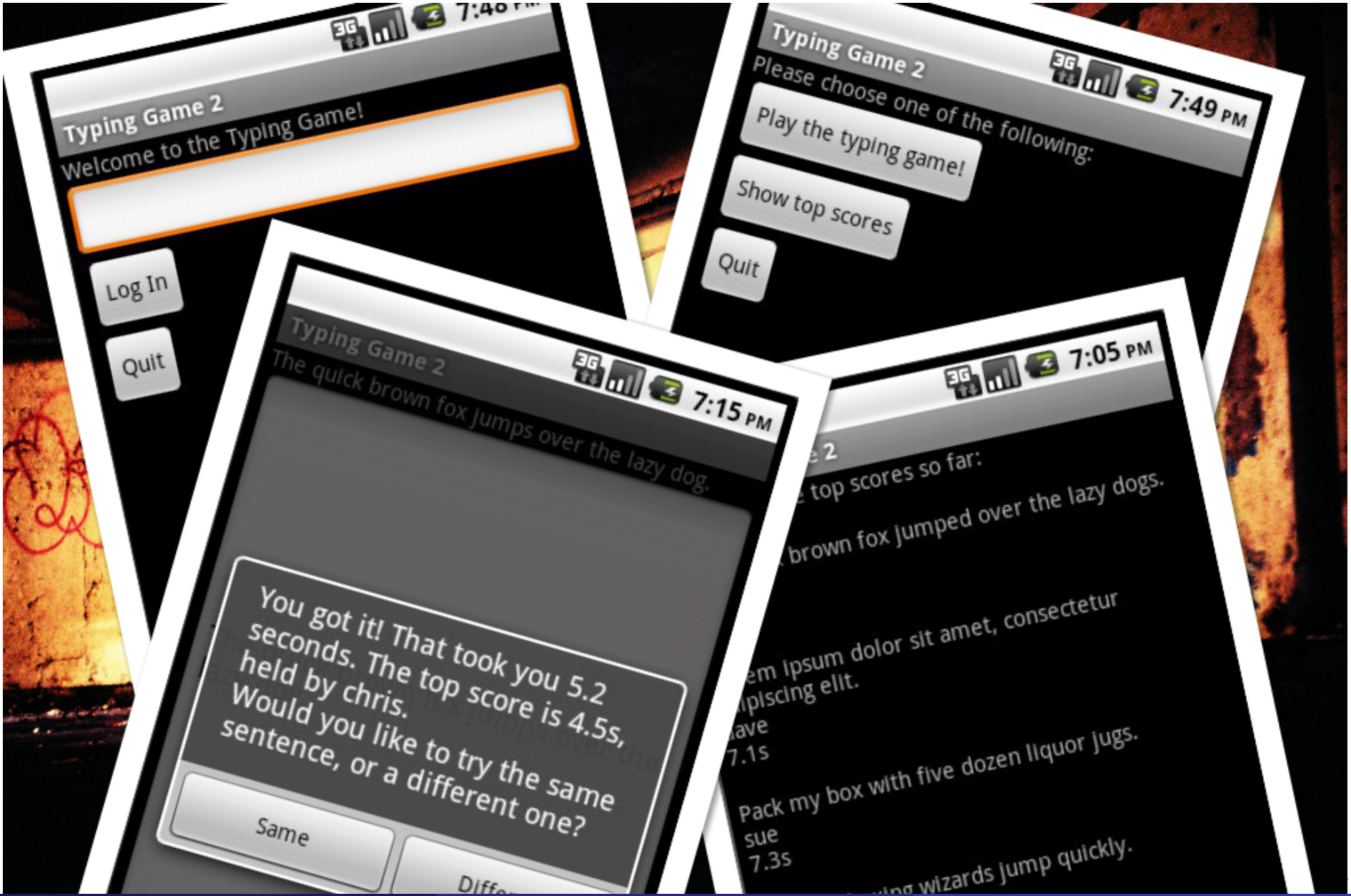
Mobile - FancyCooler



Mobile


Create a File
Please choose one of the files:
File 1
File 2
File 3
File 4
File 5
TAG:
Sports2
CONTENT:
Are you sure to change the content?
Yes
ease choose one of the following:
Create a File
Please choose one of the files:
File 1
Show or Modify this File
TITLE:
Economics Filename
TAG:
Economics
CONTENT:
We are in an economic crisis with unemployment at an all time high
TITLE:
Finance Filename
TAG:
Finance
CONTENT:

Mobile


Typing Game 2
Welcome to the Typing Game!
Log In
Quit
Typing Game 2
Please choose one of the following:
Play the typing game!
Show top scores
Quit
7:49 PM
Typing Game 2
The quick brown fox jumps over the lazy dog.
7:15 PM
You got it! That took you 5.2 seconds. The top score is 4.5s, held by chris.
Would you like to try the same sentence, or a different one?
Same
7:05 PM
top scores so far:
brown fox jumped over the lazy dogs.
em ipsum dolor sit amet, consectetur
ipiscing elit.
ave
7.1s
Pack my box with five dozen liquor jugs.
sue
7.3s
wizards jump quickly.

Projects Summary

Distributed Systems

- Distributed Web Search Engine: MiniGoogle
- DHT-based Search Engine: PennSearch
- Distributed Caching and Web Services
- Web Crawler with Multithreaded Server

Data Mining

- Amazon Reviews Data Mining
- Tweets Emotion Prediction
- New York Times Summarization
- Wall Street Journal Crowdsouring Website

Others

- Mobile Applications
- Database Website on GAE

The End