Assignment1

Natural Language Processing

Ziyi Hu-100981750

Dongyang Li-7996582

Part1:

a)Submit a file microblog2011\_tokenized.txt with the

tokenizer’s output for the whole corpus. Include in your

report the output for the first 20 sentences in the corpus.

1.Use the white space to tokenizer the corpus.

2.In the program, the user can choose whether corpus the

whole txt file or the first 20 sentences by entering 1 or 2

3.The result:

Save

BBC

World

Service

from

Savage

Cuts

http://www.petitionbuzz.com/petitions/savews

a

lot

of

people

always

make

fun

about

the

end

of

the

world

but

the

question

is.."ARE

U

READY

FOR

IT?..

ReThink

Group

positive

in

outlook:

Technology

staffing

specialist

the

ReThink

Group

expects

revenues

to

be

“marg...

http://bit.ly/hFjtmY

'Zombie'

fund

manager

Phoenix

appoints

new

CEO:

Phoenix

buys

up

funds

that

have

been

closed

to

new

business

and

...

http://bit.ly/dXrlH5

Latest::

Top

World

Releases

http://globalclassified.net/2011/02/top-world-releases-2/

CDT

presents

ALICE

IN

WONDERLAND

-

Catonsville

Dinner

has

posted

'CDT

presents

ALICE

IN

WONDERLAND'

to

the...

http://fb.me/GMicayT3

Territory

Manager:

Location:

Calgary,

Alberta,

CANADA

Job

Category:

bu...

http://bit.ly/e3o7mt

#jobs

I

cud

murder

sum1

today

n

not

even

flinch

I'm

tht

fukin

angry

today

BBC

News

-

Today

-

Free

school

funding

plans

'lack

transparency'

-

http://news.bbc.co.uk/today/hi/today/newsid\_9389000/9389467.stm

…

Manchester

City

Council

details

saving

cuts

plan:

http://bbc.in/fYPYPC

...Depressing.

Apparently

we’re

4th

most

deprived

&

top

5

hardest

hit

http://bit.ly/e0ujdP,

if

you

are

interested

in

professional

global

translation

services

Fitness

First

to

float

but

isn't

the

full

service

model

dead

?

http://bit.ly/evflEb

David

Cook

!

http://bit.ly/fkj2gk

has

the

mostest

beautiful

smile

in

the

world!

Piss

off.

Cnt

stand

lick

asses

BEWARE

THE

BLUE

MEANIES:

http://bit.ly/hu8iJz

#cuts

#thebluemeanies

Como

perde

os

dentes

no

World

Of

Warcraft

-

Via

Alisson

http://ow.ly/1beBPo

How

exciting!

RT

@BunchesUK:

Hello!

What's

happening

in

your

world?

We're

all

gearing

up

for

#Valentines

with

bouquets

flying

out

the

door.

I'd

very

much

appreciate

it

if

people

would

stop

broadcasting

asking

me

to

add

people

on

BBM.

@samanthaprabu

sam

i

knw

u

r

a

cricket

fan

r

u

watching

any

of

the

world

cup

matches

John

Baer:

Who

didn't

see

this

coming?:

TO

THOSE

who

know

Ed

and

Midge

Rendell

-

heck,

to

the

Philly

world

at

la...

<http://bit.ly/ii6WEO>

b)How many tokens did you find in the corpus? How many

types (unique tokens) did you have? What is the type/token

ratio for the corpus? The type/token ratio is defined as the

number of types divided by the number of tokens.

1.Use white space to split the corpus, get all the tokens

2.Use hashmap to get the number of the unique tokens(the first

time one token add to the hashmap, the unique add 1)

3.get the ratio

4.the result:



c)For each token, print the token and its frequency in a file

called Tokens.txt (from the most frequent to the least

frequent) and include the first 100 lines in your report.

1.Use hashmap to store the frequency and the token

2.Use white space to split the tokens

3.the result:

the 15778

to 12644

of 9872

in 9418

a 8580

- 7548

and 7518

for 6278

on 5513

is 5292

I 5114

The 4432

RT 4203

at 3501

with 2626

you 2615

& 2425

my 2387

that 2303

... 2098

from 2049

are 2020

be 1965

it 1878

by 1679

Egyptian 1663

have 1641

this 1600

will 1567

New 1485

A 1481

has 1430

| 1410

as 1399

not 1369

your 1327

just 1251

State 1224

was 1209

out 1202

me 1180

I'm 1177

new 1152

Super 1138

about 1110

Obama 1109

an 1079

like 1069

Egypt 1048

all 1034

via 1026

but 1013

de 987

#Egypt 981

i 976

up 954

News 948

2 944

Bowl 941

get 921

… 914

so 909

In 890

can 866

To 860

or 849

US 835

they 822

his 799

we 789

do 780

Social 769

White 762

no 744

Union 739

now 735

#Jan25 724

who 717

2011 716

more 713

people 684

President 676

if 663

what 659

You 655

World 652

For 651

our 650

love 619

: 613

Media 606

when 595

says 584

#jan25 578

their 575

My 574

its 573

some 567

This 564

– 562

one 561

d)How many tokens appeared only once in the corpus?

1.the token first appear, add 1 to the number

2.the token appear the second time, minus 1 to the number

3.the token appear more than two times, do nothing

4.the result:



e)From the list of tokens, extract only words, by excluding

punctuation and other symbols. How many words did you

find? List the top 100 most frequent words in your report,

with their frequencies. What is the type/token ratio when

you use only word tokens?

1.replace all the other symbols and punctuation to null

2.using string.replaceAll(“[^a-zA-Z]”,””);

3.the top 100 most frequent words and their frequencies:

the 21031

to 13844

in 10716

of 10554

a 10397

and 8313

for 7091

i 6323

on 6241

is 6074

rt 4448

at 3909

you 3737

with 3136

my 3103

egypt 3086

it 2981

new 2954

that 2696

news 2546

from 2466

this 2422

are 2363

be 2260

us 2145

by 2038

will 1924

egyptian 1879

have 1863

not 1846

your 1814

me 1774

state 1762

just 1732

jan 1717

as 1686

out 1601

im 1568

has 1560

its 1523

no 1494

all 1490

obama 1456

we 1427

now 1412

super 1390

so 1387

an 1374

up 1357

was 1333

via 1320

social 1312

media 1308

like 1292

get 1280

white 1277

about 1275

world 1267

what 1263

but 1233

if 1209

can 1137

more 1110

do 1104

how 1097

de 1092

union 1043

they 1026

people 1009

security 1007

airport 991

love 983

or 978

u 973

day 960

release 948

his 915

one 912

who 903

time 897

dont 890

today 880

good 879

video 869

house 860

jobs 859

over 847

show 837

service 817

our 814

he 791

go 782

mubarak 770

cairo 769

4.total words:713637, unique words:64779, the result:



f)From the list of words, exclude stopwords. List the top 100

most frequent words and their frequencies. You can use [this](http://www.site.uottawa.ca/~diana/csi5180/StopWords)

[list](http://www.site.uottawa.ca/~diana/csi5180/StopWords) of stopwords (or any other that you consider adequate).

1.If the word is in the list, delete it, do not add that word to the

hashmap

2.the result:

rt 4448

egypt 3086

news 2546

egyptian 1879

state 1762

jan 1717

obama 1456

super 1390

social 1312

media 1308

bowl 1280

white 1277

world 1267

union 1043

people 1009

security 1007

airport 991

love 983

release 948

president 915

dont 890

today 880

video 869

house 860

jobs 859

protests 837

service 817

mubarak 770

cairo 769

job 768

lol 743

energy 735

police 718

global 707

phone 701

free 679

dog 673

taco 664

back 660

bbc 641

protesters 635

return 630

live 623

bell 616

rite 613

special 611

toyota 597

know 594

here 557

think 556

court 539

crash 523

health 522

twitter 521

tv 519

cuts 512

watch 507

budget 505

man 497

weather 496

pm 493

top 490

home 488

business 485

cant 472

online 470

post 469

th 464

food 463

tcot 457

organic 452

blog 448

right 447

attack 446

car 444

peace 434

help 418

big 415

protest 408

pakistan 402

haiti 395

fifa 394

government 384

reuters 383

life 382

mexico 381

work 380

recovery 376

date 373

jordan 372

g)Compute all the pairs of two consecutive words (excluding

stopwords and punctuation). List the most frequent 100

pairs and their frequencies in your report.

1.Use a hashmap that contains another hashmap

2.The first hashmap contains the first word and the second

hashmap, the second hashmap contains the second word and

the frequency

3.Get the high frequency of the two consecutive words

4.the result:

f at : 152

a good : 155

a new : 309

i will : 142

i can : 133

i just : 321

i need : 151

i think : 206

i am : 259

i dont : 256

i love : 261

i was : 162

i want : 156

i have : 242

go to : 137

do you : 173

by the : 173

be a : 145

at least : 196

at the : 534

as a : 145

as the : 128

al jazeera : 220

of a : 264

of my : 130

of the : 2420

on a : 206

on my : 194

on the : 973

if you : 325

is a : 486

is not : 151

is the : 431

it was : 136

it is : 195

im at : 157

im not : 121

in a : 369

in egypt : 598

in cairo : 263

in my : 176

in the : 1473

we are : 161

to a : 145

to make : 167

to get : 311

to see : 171

to go : 186

to do : 139

to be : 558

to the : 817

president barack : 132

president obama : 199

return to : 193

keith olbermann : 239

about the : 170

right now : 132

rite now : 144

more than : 155

rahm emanuel : 175

press release : 143

bowl xlv : 209

want to : 231

barack obama : 162

super bowl : 1186

winds are : 147

with a : 224

with the : 295

will be : 413

when i : 139

thousands of : 178

check out : 162

protests in : 159

need to : 242

budget cuts : 142

world cup : 206

hosni mubarak : 123

white house : 358

white stripes : 198

supreme court : 121

state of : 1063

tahrir square : 121

egypt jan : 327

egypt protests : 136

union address : 361

united states : 174

taco bell : 532

this is : 265

that the : 154

customer service : 138

time to : 121

julian assange : 135

moscow airport : 192

release of : 163

bbc news : 183

but i : 133

and a : 166

and i : 169

and the : 400

all the : 190

are you : 135

from the : 384

jan egypt : 321

has a : 127

has been : 214

how to : 311

for a : 455

for the : 730

going to : 276

due to : 128

have a : 274

have to : 172

new post : 120

new york : 341

out of : 265

global warming : 308

one of : 188

you can : 165

you are : 156

way to : 141

via addthis : 163

the super : 248

the world : 369

the white : 276

the best : 223

the most : 140

the us : 191

the new : 243

the state : 379

the same : 132

the rite : 171

the union : 894

the egyptian : 367

the first : 141

the last : 131

health care : 157

like a : 168

social media : 974

Part2:

a) Submit a file POS\_results.txt with the tagger’s output for the whole corpus. Include in your report the POS tagger’s output for the first 20 sentences in the corpus.

We used GATE Twitter POS tagger

The result for the first 20 sentences:

DREAM\_NN

Too\_RB much\_JJ hw\_NN

high\_JJ school\_NN is\_VBZ weird\_JJ

I\_PRP feel\_VBP ..\_: Blah\_UH .\_.

I\_PRP Love\_VBP One\_CD Direction\_NN

Can\_MD I\_PRP make\_VBP a\_DT pie\_NN with\_IN potatoes\_NNS ?\_.

After\_IN so\_RB many\_JJ days\_NNS of\_IN just\_RB trying\_VBG ,\_, finally\_RB

made\_VBD it\_PRP of\_IN bed\_NN for\_IN a\_DT run\_NN at\_IN 6\_CD .\_. Hah\_UH

I\_PRP ca\_MD n't\_RB express\_VB how\_WRB I\_PRP feel\_VBP in\_IN a\_DT text\_NN !\_.

Finally\_RB

@smosh\_USR awesome\_JJ about\_IN food\_NN battle\_NN 2012\_CD

I\_PRP should\_MD probably\_RB finish\_VB my\_PRP$ homework\_NN

I\_PRP 'm\_VBP so\_RB sleepy\_JJ right\_RB now\_RB !\_. !\_. #earlybedtime\_HT

Life\_NN 's\_POS most\_RBS important\_JJ promises\_NNS might\_MD never\_RB be\_VB

spoken\_VBN .\_.

@JCSweetGirl\_USR Hi\_UH !\_.

@nessamaders\_USR aaaawn\_UH \*-\*\_UH

@djherrold\_USR just\_RB ask\_VB if\_IN you\_PRP can\_MD get\_VB a\_DT picture\_NN

with\_IN him\_PRP .\_. I\_PRP 'm\_VBP sure\_JJ it\_PRP 'll\_MD make\_VB his\_PRP$

day\_NN .\_.

Me\_PRP beating\_VBG this\_DT trend\_NN bad\_JJ tonight\_NN #ThugLife\_HT

@ALAXASS\_USR #idontevenknowyournamebro\_HT

b) What is the POS tagging accuracy for the whole corpus?

1.Compare every two POS tagging tokens from the two txt file(expected.txt, POS\_results.txt)

2.Get the number of the same tokens and the number of the different tokens

3.Caclulate the accuracy:



c)  Include in your report the frequency of each POS tag in the corpus.

NN=9564

NNPS=13

NNP=2812

NNS=2499

MD=13365

POS=141

PDT=1

PRP$=1874

PRP=9389

RBR=75

RBS=24

DT=4748

JJR=209

JJS=232

HT=2524

FW=2

USR=6320

URL=1091

RT=2419

UH=3128

TO=1874

VBD=1760

SYM=5

VBP=5042

VBN=638

VBZ=1813

WDT=10

VBG=1855

WP=539

RP=403

RB=5022

IN=5678

CC=1360

JJ=3810

Work:

Ziyi Hu

Part 1: a, b, c

Part 2: a, c

Dongyang Li

Part 1: d, e, f

Part 2: b