uOttawa

Faculty of Engineering

Course

Natural Language Processing

Assignment 1

Professor

Prof. Diana Inkpen

Students

Diana Lucaci

Mozhgan Nasr Azadani

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

Table of Content

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

Part 1

a.

We have used the twitter tokenizer from nltk python's library. In addition, we have used a list of contractions (from wikipedia and other web resources).
For the words in this list, we have used the nltk's word tokenizer in order to obtain a better accuracy.
We have conducted a comparison between the nltk's word tokenizer and the nltk's twitter tokenizer which led us to the conclusion that the first one is more accurate for contracted forms such as (we're, I'm, can't), while the second one is more suited for social media data, where the used language is closer to the user, thus often containing unconventional words and phrases (abbreviations, links, emojis, shortening of the words, jargon, etc.).
Output of the first 20 messages in the corpus:
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 sum 1 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 is n'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 did n't see this coming ? : to those who know ed and midge rendell - heck , to the philly world at la ... http://bit.ly/ii6weo

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

b.

Total tokens: 331860

Unique tokens: 49454

Type/token ratio: 0.1490206713674441

c. Token frequency

:	**8629**
.	**7755**
the	**7738**
,	**6903**
...	**5099**
to	**5040**
!	**4390**
a	**3986**
in	**3947**
-	**3896**
of	**3856**
and	**3056**
i	**2714**
for	**2567**
on	**2319**
"	**2146**
is	**2138**
?	**2131**
rt	**1786**
(	**1753**
)	**1538**
you	**1530**
'	**1493**
my	**1338**
it	**1245**
with	**1197**
at	**1184**
new	**1106**
news	**1083**
that	**949**
this	**942**

&	**925**
be	**894**
release	**877**
from	**870**
are	**830**
world	**791**
by	**787**
me	**769**
just	**735**
security	**734**
not	**732**
have	**722**
will	**715**
u	**707**
/	**669**
as	**649**
has	**641**
now	**639**
your	**637**
white	**634**
s	**629**
no	**626**
phone	**625**
rite	**603**
all	**596**
..	**590**
out	**587**
return	**586**
was	**576**
egyptian	**571**
so	**555**
|	**542**
we	**531**
up	**530**
i'm	**521**
but	**515**
like	**514**
if	**508**
crash	**501**
toyota	**498**
#jan25	**496**
an	**493**
bbc	**490**
#egypt	**481**
2	**474**
…	**472**

us	**470**
what	**470**
about	**466**
de	**464**
get	**459**
’	**436**
do	**434**
can	**432**
via	**426**
egypt	**424**
*****	**424**
they	**414**
love	**408**
2011	**398**
one	**387**
more	**387**
people	**380**
time	**376**
day	**368**
how	**367**
$	**359**
or	**356**
peace	**356**

d. **Number of tokens that appear only once: 34593**

e. **Number of words: 246314;**

Type/token ratio: 0.11752884529502992

The words were determined based on a regular expression that checks whether the token is only formed from letters and digits, having at least one letter.
We have also considered a version of the list of words that were found in WordNet, but this was significantly smaller due to the misspellings, the contraction forms such as “don t” or “dont”, proper nouns, and internet slang such as “lol”, words that are not present in WordNet.

the	**7739**
to	**5040**
a	**3986**
in	**3957**
of	**3858**
and	**3056**

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

i	**2714**
for	**2567**
on	**2319**
is	**2138**
rt	**1791**
you	**1530**
my	**1338**
it	**1248**
news	**1240**
with	**1197**
at	**1184**
new	**1107**
that	**949**
this	**942**
egypt	**905**
be	**894**
release	**879**
from	**870**
are	**830**
world	**803**
by	**787**
me	**769**
security	**760**
just	**735**
not	**732**
have	**722**
will	**715**
u	**707**
as	**649**
now	**644**
has	**641**
white	**640**
your	**637**
phone	**630**
s	**629**
no	**626**
rite	**604**
all	**596**
return	**589**
out	**588**
egyptian	**587**
was	**576**
so	**555**
toyota	**537**
we	**532**
up	**530**

but	**515**
like	**514**
bbc	**511**
crash	**510**
if	**508**
jan25	**496**
an	**493**
us	**476**
what	**470**
about	**467**
de	**464**
get	**459**
do	**434**
can	**432**
via	**426**
love	**416**
they	**414**
one	**387**
more	**387**
peace	**386**
pakistan	**386**
mexico	**381**
people	**380**
time	**377**
fifa	**377**
haiti	**377**
day	**368**
how	**367**
police	**357**
or	**356**
over	**352**
soccer	**344**
good	**338**
when	**338**
who	**336**
service	**335**
his	**334**
he	**326**
computer	**321**
after	**314**
go	**314**
lol	**312**
british	**311**
video	**307**
protesters	**304**
its	**302**

today 301
date 286

f. **Top most frequent words that are not stopwords:**

rt	**1791**
news	**1240**
egypt	**905**
release	**879**
world	**803**
security	**760**
white	**640**
phone	**630**
rite	**604**
return	**589**
egyptian	**587**
toyota	**537**
bbc	**511**
crash	**510**
jan25	**496**
love	**416**
peace	**386**
pakistan	**386**
mexico	**381**
people	**380**
fifa	**377**
haiti	**377**
police	**357**
soccer	**344**
service	**335**
computer	**321**
lol	**312**
british	**311**
video	**307**
protesters	**304**
today	**301**
date	**286**
drug	**284**
museum	**280**
assange	**276**
war	**271**
says	**264**
man	**259**
murder	**257**
back	**250**
wikileaks	**248**

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

stripes 239
car 237
cup 230
cairo 226
top 216
know 213
live 210
cuts 209
national 205
protests 204
released 204
jobs 203
staff 203
press 201
free 199
think 198
state 195
mubarak 194
business 191
online 187
ap 186
clinton 183
twitter 177
iphone 176
cut 173
work 173
home 173
kate 171
big 169
watch 169
president 169
cloud 166
game 162
right 162
uk 162
recall 162
movie 160
post 160
life 157
media 155
help 154
social 152
check 152
hacking 152
blog 149
black 149

job	**144**
hit	**144**
known	**144**
attack	**143**
reuters	**143**
facebook	**142**
government	**141**
nobel	**141**
julian	**139**
oprah	**139**
internet	**138**
tv	**138**
london	**137**

g. **Type/toke ration no stopwords: 0.19280711794201003**

white stripes	**199**
world cup	**198**
bbc news	**141**
press release	**134**
rt rt	**130**
julian assange	**128**
egypt jan25	**107**
jan25 egypt	**103**
release date	**103**
world service	**92**
prime minister	**86**
hillary clinton	**81**
bbc world	**79**
world news	**68**
phone hacking	**65**
social media	**64**
egyptian protesters	**60**
anthony hopkins	**60**
kate middleton	**60**
fifa soccer	**58**
fifa world	**55**
shorty award	**53**
egyptian museum	**51**
super bowl	**50**
tahrir square	**50**
cell phone	**49**
toyota recalls	**49**
strings attached	**48**
white house	**46**
global war	**46**

customer service	**45**
security forces	**45**
state hillary	**45**
car crash	**43**
windows phone	**43**
nobel peace	**42**
box office	**41**
egypt museum	**40**
wikileaks founder	**38**
egyptian protests	**38**
breaking news	**37**
oprah winfrey	**37**
family secret	**37**
airport security	**37**
egypt protests	**36**
plane crash	**36**
egyptian police	**35**
youtube video	**34**
united states	**34**
cloud computing	**34**
peace prize	**34**
kim clijsters	**34**
lol rt	**33**
mexico city	**33**
olympic stadium	**32**
ap ap	**31**
blog post	**30**
ca wait	**29**
egyptians form	**29**
justin bieber	**28**
iphone ipod	**28**
national security	**27**
national museum	**27**
egyptian security	**26**
drug war	**26**
squad iphone	**26**
ipod ipad	**26**
house arrest	**26**
australian open	**26**
andy gray	**25**
latest news	**25**
war online	**25**
iphone click	**25**
social security	**25**
tear gas	**25**
prize winner	**25**

mobile phone	**24**
raymond davis	**24**
toyota motor	**24**
looting jan25	**24**
nobel laureate	**24**
egyptian army	**23**
global security	**23**
british people	**23**
wikileaks assange	**23**
weight loss	**22**
tax return	**22**
egyptian government	**22**
cell phones	**22**
immediate release	**22**
nobel prize	**22**
peace rt	**22**
south africa	**21**
ipad global	**21**
war gwo	**21**
egypt army	**21**
hosni mubarak	**21**
toyota corolla	**21**
egyptian embassy	**21**
egyptian president	**21**
special olympics	**21**
prince william	**21**

h.

In order to do multi-word expressions acquisition, we have considered a few options:

1. using the LocalMax algorithm, which extracts MWEs by generating all possible n-grams from a sentence and then further filtering them based on the local maxima of a customisable Association Measure's distribution (Silva and Lopes 1999)[1].

One of the methods that can be used to detect MWE is by determining whether the expression that has one of the words replaced with a synonym makes sense or not. This can be approximated using a probability function and a large corpus. If the probability of the new expression is very small, than we can conclude that the initial expression is a MWE expression because replacing one of its words with a synonym leads to an expression that does not make sense.

Thus, for the LocalMax score function, one could use to sum up the probabilities of the expressions that are formed by replacing each word of the initial expression with a synonym and take either the reverse sign of this expression (since the algorithm looks for the maximum value) or one could use to look for the minimum value of this probabilities.

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

2. using word for word translation and the idea that a multiword expression, translated in another language would most probably lead to a non-sense expression.
This method would require a translation tool and a corpus of the destination language that would tell us the probability of the translated expression.

3. using the skip-gram model available through the Gensim python library, which detects phrases based on collocation counts. Potential phrases are scored according to the formula presented in **Mikolov, et. al: "Distributed Representations of Words and Phrases and their Compositionality"**[2] (pag. 6). The formula is based on the unigram and bigram counts from the training corpus.
For the model, we have used a twitter corpus with 5980324 tweets, gathered from 2 sources: a SemEval task since 2013 and Kaggle dataset of customer support twitter data.
For this method, the top 100 most frequent bigrams that were identified are the following:

press_release	**134**
release_date	**98**
has_been	**96**
new_york	**85**
thousands_of	**75**
social_media	**63**
more_than	**56**
strings_attached	**48**
right_now	**47**
cell_phone	**46**
at_least	**45**
customer_service	**44**
car_crash	**42**
so_much	**39**
new_album	**35**
youtube_video	**34**
united_states	**34**
cloud_computing	**34**
last_night	**33**
mexico_city	**33**
part_of	**30**
blog_post	**29**
looks_like	**27**
social_security	**25**
lot_of	**24**
so_far	**24**
first_time	**23**
de_la	**23**
i_dont	**23**

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

people_who 22
cell_phones 22
south_africa 21
great_news 21
middle_east 19
lots_of 19
last_year 18
new_zealand 18
bad_news 17
as_well 17
cut_off17
los_angeles 17
tell_me 17
too_much 16
of_duty 16
how_many 16
those_who 15
rest_of15
wake_up 15
shut_down 15
would_like 15
each_other 15
good_news 15
kind_of 14
how_much 14
i_cant 14
naman_ako 14
last_week 13
hundreds_of 13
instead_of 13
black_ops 13
woke_up 13
more_details 13
email_address 13
said_he 12
must_be 12
tired_of 12
police_officer 12
social_networking 12
brand_new 12
breast_cancer12
star_wars 12

1000s_of 12
recently_added 12
inspired_by 12
whole_world 11
member_of 11
lil_wayne 11
every_time 11
rock_band 11
soccer_game 11
better_than 11
never_heard 11
he_said 11
white_girl 11
u_r 10
next_week 10
his_own 10
talking_about10
no_longer 10
en_el 10
west_ham 10
what_happens 10
looking_forward 10
make_sure 10
en_mexico 10
search_engine 10
security_guard 10
cyber_security 10
less_than 10
no_matter 10

Part 2

a.

We have used the CMU Twitter POS tagger[1] in order to extract the part-of-speech of the tokens. To do so, there were some steps to take, which are explained in the following:
1. First, the tagset style of the mentioned tagger was not exactly the same as the Penn TreeBank style and the problem was that the tagger used its own style. As a result, we had to use other models for the POS tagger that matched the output format of the tags as the PTB

[1] http://www.cs.cmu.edu/~ark/TweetNLP/

Diana Lucaci - SN: 300098053 - dluca058@uottawa.ca responsible for part 1 of the assignment
Mozhgan Nasr Azadani - SN:300085577- mozhgan.nasr91@gmail.com responsible for part 2

style. In order to do that, we have employed two of the already existing models in their website, called model.ritter_ptb_alldata_fixed.20130723 and model.irc.20121211, to train the POS tagger. The reason why we have used two models is to improve the accuracy.
2. Second, the output format of the POS tagger needed to be changed so that we could do the comparison. Regarding the format change, we programmed in Java to get the favorite output format, which has been attached.
3. The complete output of the tagger has been submitted as POS_results.txt.

The first twenty sentences are tagged as follows:

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_NNP
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
The_DT fact_NN that_IN @Brittney_9_USR and_CC @brynnmariecee_USR gain_VB up_RP on_IN me_PRP in_IN child_NN development_NN <<<#realjerks_NN
Dreaming_VBG about_IN you_PRP ._.

b.

In order to calculate the accuracy, we have programmed in java and calculate the accuracy as the number of the correctly tagged tokens divided by the total number of tokens in the corpus. The accuracy is 95.02%. The java source code to calculate the accuracy has been attached as accuracy.zip. In the following, we have shown some incorrect tags for some random sentences:

results:Remember_VB money_NN cant_MD buy_VB us_PRP true_JJ happiness_NN
expected:Remember_VB money_NN cant_MD buy_VB us_PRP true_JJ happiness_NNS

results:Olifs_NNS RT_RT @TheDakari_USR :_: #PussyTasteLike_HT Vanilla_NNP ,_, tasty_JJ hoe_NN ._.
expected:Olifs_NNP RT_RT @TheDakari_USR :_: #PussyTasteLike_HT Vanilla_NN ,_, tasty_JJ hoe_NN ._.

results:#IFOLLOWBACK_HT :_:)_-RRB-
expected:#IFOLLOWBACK_HT :_.)_)

results:RT_RT @_mpeterrrs_USR :_: It_PRP hurts_VBZ to_TO know_VB you_PRP like_IN someone_NN else_RB ._.
expected:RT_RT @_mpeterrrs_USR :_: It_PRP hurts_VBZ to_TO know_VB you_PRP like_VBP someone_NN else_RB ._.

results:@aleonard4_USR I_PRP miss_VBP this💔_DT
expected:@aleonard4_USR I_PRP miss_VBP this💔_PRP

results:#WaysToGetShot_HT mess_NN or_CC flirt_VBP with_IN MY_PRP$ boyfriend_NN @cornhole696_USR 😘😉_UH
expected:#WaysToGetShot_HT mess_NN or_CC flirt_VBN with_IN MY_PRP$ boyfriend_NN @cornhole696_USR 😘😉_UH

results:Haha_UH going_VBG to_TO the_DT store_NN real_JJ quick_JJ #lilbro_HT #smashin_HT http://t.co/N1Rheg1S_URL
expected:Haha_UH going_VBG to_TO the_DT store_NN real_RB quick_JJ #lilbro_HT #smashin_HT http://t.co/N1Rheg1S_URL

The obtained results show that first, POS tagger has more difficulty finding the correct format of the verbs and nouns in the context of the sentences. Second, it has difficulty finding the correct tag for some words having more than one part-of-speech such as like or real. It should be noted that for two of the tags such as Close parenthesis and Open parenthesis, the POS tagger used the style (-RRB- and -LRB-). However, these tokens are tagged differently in the expected_output file. As a result, although the tagging is correct, i.e. the tagger finds them open and close parenthesis, we consider these tags incorrect since the notation is different in comparison with the expected ones.

c.

The source code to calculate the frequency of each tag has been attached as frequency.zip. The frequency of each POS tag in the corpus is as follows:

NN	9910
RB	5101
JJ	3668
VBZ	1812
PRP	9340
VBP	4711
:	4049
UH	3145
.	6817
CD	716
NNP	3090
MD	1303
DT	4758
IN	5768
NNS	2485
VBG	1853
,	1381
VBD	1809
VB	4701
WRB	753
USR	6271
PRP$	1879
HT	2480
POS	131
RBS	6
VBN	491
CC	1361
RP	436
WP	551
TO	1786
URL	1094
RT	2415
JJS	233
-RRB-	161
"	460
-LRB-	66

RBR	62
EX	18
JJR	114
WDT	6

Bibliography

1. Carlos Ramisch, "Multiword Expressions Acquisition: A Generic and Open Framework", Theory and Applications of Natural Language Processing series, XIV, Springer, ISBN 978-3-319-09206-5, 230,2015.
2. Mikolov, et. al: "Distributed Representations of Words and Phrases and their Compositionality"