FELIPE@IFT6285 A2019

QUELQUES OBSERVATIONS
CONCERNANT LE

TP2 - IFT6285

REPORT IS IMPORTANT !!!

- The presence of code is almost useless (unless clarification needs to be made).
- The abstract should describe what has been done (not just paraphrasing the subject)
- **Reporting is one thing, analyzing the limitations is better**

DIRTY HANDS

What has been done

NORMALISATION CONDUCTED

- PUNCT … => DOTS
- EMOJI :-) => EMOJI
- DEFORMED cooooolllllll => COOL
- TIME / DATE / MONTH / DAY/DECADES
  - 80s / 80's / '80s => DECADE
- RIRES ahahah -> RIRE
- EMAILS, URL
- ORTHOGRAPHE
- VARIANTES
  - PHONETIQUES 2d => today
  - BRITISH/AMERICAN neighbour => neighbour
  - CONTRACTIONS didn't => did not
- GIBBERISH dlkfjhglskdjfhg => GIBBERISH

- NUMBER
  - SMALL/LARGE-NUMBER
  - ORDNUM st / nd / thrd / th => ORDNUM
  - MEASUREMENTS
  - MONEY
  - PERCENTAGE
  - BINNUM 00010100101
  - PHONE NUMBER
- EMAILS
- BULLES 1. Topic => <bulle> Topic
- GROGNEMENTS grrrr => ROARING
- KISSES xox,xo => KISSES/HUGS
- PHOTO (136 blogs contain *flickr*)
- FILENAME

NORMALISATION (CONT)

ARROWS —-> RIGHT-ARROW

POSSESSIVES ('s might interfere with other steps)

INTERROBANG. !?!?!??!?!

REFLEXION hmmmm, mmmmm => REFLECTION

AWE awhhhh -> AWE

NOTUNDERSTAND huh

I.E. / E.G. i.e / ie. / i.e. /ie => IE

SOME « FINDINGS »

- Unclear whether the text was or not lowercased
  - Several reported to have conducted a conversion, because some material was not lowercased
  - *grep '[A-Z]' train_posts.csv* revealed nothing
- Malaysian slang words (merci Leila)
  - https://www.ordinaryreviews.com/2017/01/25/51-slangs-malaysians-love-to-use/
- Non English sequences
  - Filtering by anti-list of ngrams not possible in English
    - *jq, qg, qk, qy, qz, wq, wz* n'existent pas en anglais (merci Hubert) http://norvig.com/mayzner.html
  - langid.py

JUNK: PGP ENCODED MATERIAL (MERCI HUBERT)

gpg --keyserver ""hkp://pgp.mit.edu"" --recv-keys 6db089d5 -----**begin pgp** public key block----- version: gnupg v1.2.4 (gnu/linux) mqghbedqv+arbadc4tesywsyvh4+l4sol3zwtjguo4tchgotq9bxjc99juywezdb
j7bwsyptt6gn//m0fi+ow0+q/saqq5o5xm9uqjjljmbxm79fbwkfesryv5nh4z54
fzjvxdq4unwkycq+v76hdzoynmwktcj0n6wzmwvtjwcxegyhg5ydqpt4jwcg+7zn
2761xp43e+lztdfat8zdzjud+pnykigmm1rvabmqmgeg+9oiapslqmzwm2eolge/ hiauesapk89u6xqbk7sgj/
9jcc54gfxdfi0nbvf2xdpcawh89hzyper9v3qhnsvk
rxxjoniiiuxj8bo96fybcdfidsuairpgliyk0xffzupysrdftkbmbj4hky8re7ze /fgd/0uj5ikcijwqpgqcuaajoni/
dyoglqa5qbsslm9p7ekqkvhr42zd98v0lq4z
cbqu2eynbdtr0qcbvhmy52po1uk0url+woqf4o7afays3rx7kq+81khqviauxquw
puuon4tzcsmuoyd8rztnpogj6mxfmbw+gt9b53y1eifjn4attehmdwlziefszm9u
c28gvmvnysbhyxjjw2lhichhbhzlz2encybrzxkgzm9yigfsbckgpgvsdmvnyw14
qhlhag9vlmnvbs5ted6ixgqteqiahgucqnc/4aibawylcqghawidfqidaxycaqie
aqixgaakcrcdp4kbdx1gejhqakdaob7au2qxre5m6sxvbma45i12oacft46nv25c h1mat6rnzfzkbczimzi5aq0eq
4xaeanc6/tuqs82krzvmus54xav/eo+ewxy6 d2r/sq20zu/
2lejbfvbjiaf01tu6hmlaitilrvgfc7frrp9zbsontkjxvaxg7ydr +op5kzqtcxk2qnhx1tedhn47yrdvdaxul4rsssvodvhu
at+uq5+zroozf9s2hqn u6vlhv/dttfnaaqla/9xnr44e2e57meld9h28lsqvslvwqv9ai/tgwhpp/vhww3h
wecplfl2tiiubuz2e9rlurf2eytgtgtvcc6fcvltpyibzh+2g6dcpohd6hm/pvf1
hjn4lz83coax6ljp0jyxx8wcqsgvfnjtkjxj3xwumo4v99sk3/eqgk3sdijxtihj bbgragajbqja0l/jahsmaaojeim/
irt1fuysjkkaol7s85vltdn0uxvwt2cd3o4w 3ktgaj9ww9bvdz3yaaimt8mn3t080/xk6a== =u/nj -----**end pgp** pu
key block----- "

JUNK: FLICKR MARKUP

".flickr-photo {border: solid 1px #000000;}.flickr-yourcomment {}.flickr-frame {float: left;width: 150px;text-align: center;padding: 3px;margin-right: 10px;}.flickr-caption {font: 75%;/*color: #666666; */margin-top: 0px;}.flickr-buddyicon {margin-right:5px; vertical-align:middle;border: solid 1px;}.flickr-postedby {font: 75%;} urllink urllink week 2 - side view , originally uploaded by urllink bigbelly. ",1

".flickr-photo {border: solid 1px #000000;}.flickr-yourcomment {}.flickr-frame {float: left;width: 150px;text-align: center;padding: 3px;margin-right: 10px;}.flickr-caption {font: 75%;/*color: #666666; */margin-top: 0px;}.flickr-buddyicon {margin-right:5px; vertical-align:middle;border: solid 1px;}.flickr-postedby {font: 75%;} urllink urllink week 2 - front view , originally uploaded by urllink bigbelly. ",1

".flickr-photo {border: solid 1px #000000;}.flickr-yourcomment {}.flickr-frame {float: left;width: 150px;text-align: center;padding: 3px;margin-right: 10px;}.flickr-caption {font: 75%;/*color: #666666; */margin-top: 0px;}.flickr-buddyicon {margin-right:5px; vertical-align:middle;border: solid 1px;}.flickr-postedby {font: 75%;} urllink urllink week 1 - side view , originally uploaded by urllink bigbelly. ",1

".flickr-photo {border: solid 1px #000000;}.flickr-yourcomment {}.flickr-frame {float: left;width: 150px;text-align: center;padding: 3px;margin-right: 10px;}.flickr-caption {font: 75%;/*color: #666666; */margin-top: 0px;}.flickr-buddyicon {margin-right:5px; vertical-align:middle;border: solid 1px;}.flickr-postedby {font: 75%;} urllink urllink week 1- front view , originally uploaded by urllink bigbelly. ",1

".flickr-photo {border: solid 1px #000000;}.flickr-yourcomment {}.flickr-frame {float: left;width: 150px;text-align: center;padding: 3px;margin-right: 10px;}.flickr-caption {font: 75%;/*color: #666666; */margin-top: 0px;}.flickr-buddyicon {margin-right:5px; vertical-align:middle;border: solid 1px;}.flickr-postedby {font: 75%;} urllink urllink week 3 - side view , originally uploaded by urllink bigbelly. ",1

COOL TOOLS USED

- Spacy
  - spacy_hunspell
- Nltk
  - tweettokenizer, wordnet
- emot
- SymSpell, autocorrect, PySpellChecker
- dateutil
- Monoise — some benchmarks on text normalization
- Senticnet

```
import emot
text = "I love python 👨 :-)"
emot.emoji(text)
[{'value': '👨', 'mean': ':man:', 'location': [14, 14], 'flag': True}]
emot.emoticons(text)
{'value': [':-)'], 'location': [[16, 19]], 'mean': ['Happy face smiley'],
'flag': True}
```

- How long is my normalisation process ?

Tokenizer	Time (s)
NLTK	1821.54
Spacy	26205.13
Space	47.52

(Merci Yutao)

Time for tokenizing the full dataset

- How much vocabulary do I reduce ?
- Normalizing the text can also mean encoding the pertinent information
  - It is not because you normalize coooooooolllllllll to cool that you can not keep the information that an accented form has been used !

```
find tp2 -name "normalise.out" -exec wc -l {} \;
```

```
1000 tp2//Ireti Eriola Gloria Loko_13988690_assignsubmission_file_/normalise.out
1000 tp2//Zachary Barillaro_13988703_assignsubmission_file_/normalise.out
1000 tp2//Khalil Slimi_13988693_assignsubmission_file_/normalise.out
1000 tp2//Martin Weyssow_13988688_assignsubmission_file_/normalise.out
1000 tp2//Olivier Salaün_13988710_assignsubmission_file_/normalise.out
1000 tp2//Lucas Pages_13988729_assignsubmission_file_/normalise.out
1000 tp2//David Ferland_13988730_assignsubmission_file_/normalise.out
1000 tp2//Mahmoud Abou Nassif_13988717_assignsubmission_file_/normalise.out
1001 tp2//Nithin Anchuri_13988697_assignsubmission_file_/normalise.out
 999 tp2//Lu Yuchen_13988732_assignsubmission_file_/normalise.out
1000 tp2//Philippe Gagné_13988705_assignsubmission_file_/normalise.out
1000 tp2//Michel Ma_13988720_assignsubmission_file_/normalise.out
1000 tp2//Fanny Salvail-Bérard_13988709_assignsubmission_file_/normalise.out
 969 tp2//Suzy Edith Moukala Both_13988696_assignsubmission_file_/normalise.out
1000 tp2//Yan Zeng_13988723_assignsubmission_file_/normalise.out
 999 tp2//Emmanuel Eytan_13988702_assignsubmission_file_/normalise.out
 121 tp2//Marie-Ève Malette-Campeau_13988691_assignsubmission_file_/normalise.out
```

769:"HI! WELCOME TO MY FIRST BLOG. CREATED ON FRIDAY THE 13. (OPPS, NOT TO FORGET -- AUGUST 2004) HA HA!! SINAGPORE TIME = 2:45PM "

ireti hi! welcome to my first blog. created on friday the 13. (opps, not to forget -- august 2004) ha ha STRONG? ! sinagpore time = 2:TIME

khalil hi ! welcome to my first blog . created on friday the <bullet> (opps , not to forget - august <num>) ha ha ! sinagpore time = <time>

olivier hi ! welcome to my first blog . created on friday the DIGITS . (opps , not to forget FACE august DIGITS FACE DIGITS) ha ha SERIESEXCLAMATION sinagpore time = DIGITS : TIME

lucas hi ! welcome to my first blog . created on friday the NUMBER . (opps , not to forget -- august NUMBER) ha ha INTENSE sinagpore time = 2:45pm

davidf hi ! welcome to my first blog . created on friday the 13 . (opps , not to forget -- august 2004) LAUGHTERALL EXCLAMMARK sinagpore time = TIMEEE

philippe hi ! welcome to my first blog . created on friday the 13 . (opps , not to forget - august 2004) ha ha ! sinagpore time = 2:45pm

michel hi ! welcome to my first blog . created on friday the 13 . (ops , not to forget - - august 2004) ha ha STRONG-EXCL sinagpore time = 2:45 pm

yan hi ! welcome to my first blog . create on friday the 13 . (opus , not to forget -- august 2004) ha ha ! ! singapore time = 2:45pm