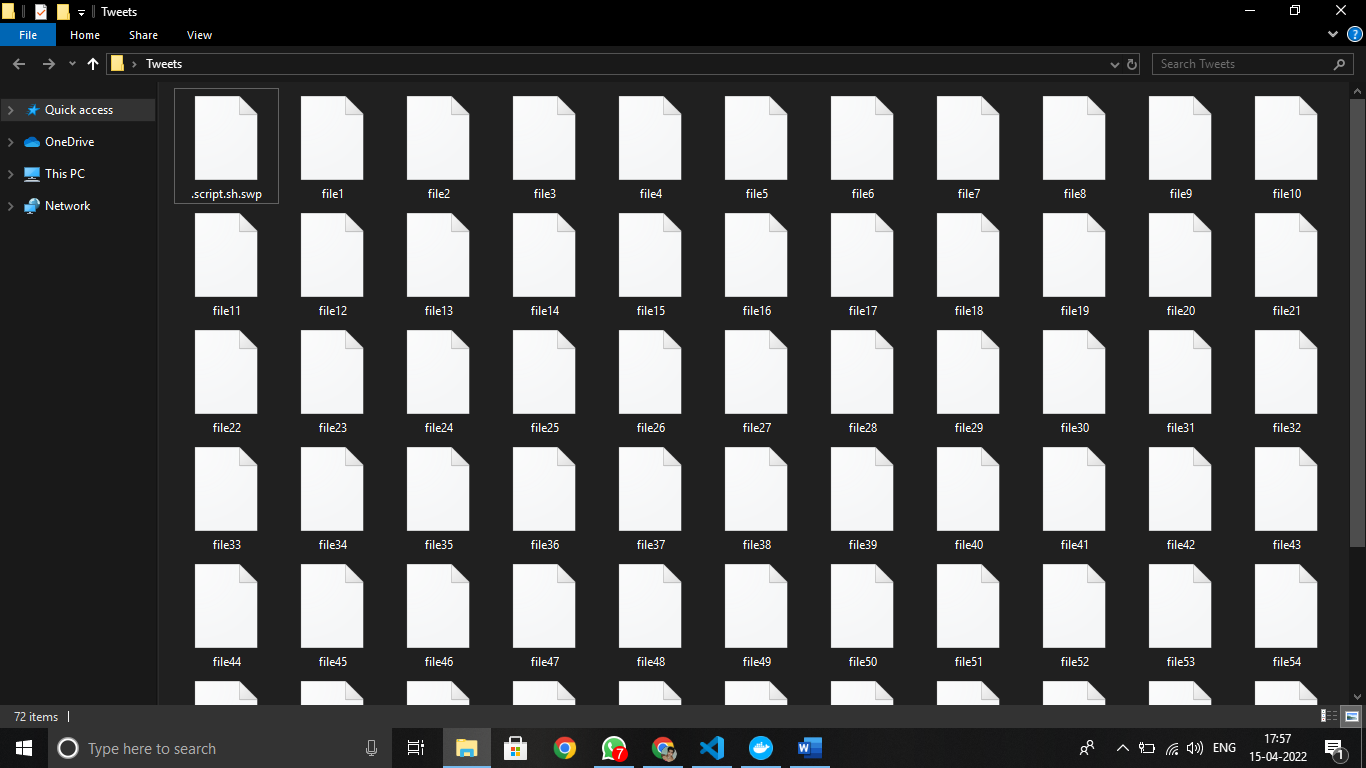
NLP Assignment 2 - Tweet Tokenizer

Problem Statement

The goal of the assignment is to write a tweet tokenizer. The input of the code will be a set of tweets and the output will be the tokens in each tweet.

Part 1 (Data Creation):

As a first step you will be asked to tokenize a small number of tweets by hand. This will allow you to understand the problem from a linguistic point of view.



Input File – file23:

Free Demos @APPCRAFT Marines Cut Off From Social Networks: The U.S. Marine Corps has banned .. http://bit.ly/1vsk5

@tackyspoons sorry! the next one won't help either...

Confira a minha nova imagem: http://bit.ly/2ngr9D

Integrity commissioner from THIS council? It's like asking the rapist to be gentle and to wear a condom. #tostrike

@Kirstiealley, I am with @teeco71- Ur not sloppy seconds. You brought the pizazz to Cheers post SL....

@LucasBlack The older version 406 is a great car, but the Passat will run forever. I do 450m a week in my Golf TDI. Cruise 80mph at 60mpg! x

" RT @Furyofsolace Joss Whedon talks ""Dr. Horrible"" sequel: http://bit.ly/9wq5Y"

RT @mortgageporter: Top Twitter etiquette tips http://tinyurl.com/q7ev8n

RT @RealTalibKweli 2. I will never say something on twitter I cant say to your face. Thats for the e goons

Might go 2 movies n the park,showin The Goonies! lol How ironic,was just tlkn bout it other day&showin our dorky side that we CAN quote it

At the@AlexAlbrecht Diggnation show I'm certain I could have given birth to everyone here

Just because I can: Will install MAMP on a Flash drive tonight.

@jasoncastro Gosh, I hope you guys aren't doing this while the car is in motion... distracted drivers #1 cause of accidents... headlines...

I love the mall lol. hung out with sam and allan and traded bracelets. and then katie thought my tennis case was a banjo.

" @cnn Petition for ""All Congressional Seats & President on 2010 Ballot"" http://bit.ly/7tY7b"

about to go to the store..weeee

" Just found a piece of paper... ""people buy value and values"". Words of wisdom from someone who shared with me + I share with you."

@savvybaby MY FAM ON MY DADS SIDE FROM MIDWEST EAST COAST AND SHIT

#uknowusprung wen u argue on the phone hang up and text a paragraph rite as soon as u hang up

@HisFriendlyWay @AndyGroenink @Gagrin @mariahdawn @seekzpeace @fish\_lover Thank you for the RTs and fun :)

@DotsRach hmmm...how long. i gots lots to do tomorrow night.

vanessa hudgens is someome nice piece of a\*\*. shoot me!! but she is. like i said disney knows how to breed these skanks pretty good lol

h130084370：ネットビジネスアーカイブ(NBA)＋ブランディング大百科＋PPC大百科.. http://twit.ac/TDMe

@miriamq i'll be a good example to my subordinates so that they'll be good as well! like Tita Cory :)

@miayunn and why are you going BKK this sunday?

going into town, need to get a bag and stationary. (:

Finally Nordic Nosebleed has started their own twitter blog. Players from the team will be discussing the EASHL and match results.

@nayes1982 I like Adele's Make YOu Feel My LOve - Please play it @algoafm !!

Tender: London: transport systems consultancy services http://u.mavrev.com/z3bk

The guidelines for tweet:

1. All punctuations are individual tokens. This includes double-quotes and single quotes also. No need to normalize quotation marks further.

Code:

﻿

Text

Description automatically generated

Output:

Text

Description automatically generated

1. Clitics are separated into component words, with the clitic normalized into the original word.

Example: I’m I am; she’d she would; can’t can not; etc. Possessive apostrophes are

Individual tokens. Example: “NLP’s assignment” NLP ‘s assignment. If a word-ending

apostrophe is possessive, convert it into apostrophe s. Example: “Jones’ golf clubs” Jones ‘s

golf clubs.

Code:

Text

Description automatically generated

Output:

All The clitics detected in the given input file:

A picture containing text

Description automatically generated

1. Each hashtag is an individual token. Each user reference is an individual token.

Code:

Text

Description automatically generated

Output:

All The Hashtags and user references detected in the given input file:



1. Each smiley is a separate token.

Code:



Output:

All the smileys detected in the given input file:

