



### Step 1.2: Remove Mentions and Hashtags (5 Marks)


// Notice: the reason I include a "( )?" at the end is that each word  
// contains at most one space at the end. If we can replace the space in  
// the first place, it will be even cleaner.


Find: @[a-zA-Z\_0-9]\* ( )? 

Replace:  nothing in here

Match Case: Yes

Mode: Regular Expression


Find: #[a-zA-Z\_0-9]\* ( )? 

Replace:  nothing in here

Match Case: Yes

Mode: Regular Expression

### Step 1.3 Remove any -, = or Space at the Start of a Tweet (5 Marks)

Find: text: [-= ]+ 

Replace: text:

Match Case: Yes

Mode: Regular Expression

### Step 1.4 Cleaning up the Spaces (3 Marks)

Find: ( ){2,} 

Replace:

Match Case: Yes

Mode: Regular Expression

### Step 1.5 Anonymize the Screen Names (8 Marks)

// Notice: In NotePad++, we can use either \$1 or \1 for referencing group.

Find: (screen\_name: [0-9A-Za-z\_])[0-9A-Za-z\_]\*([0-9A-Za-z\_])

Replace: \$1\*\*\*\*\$2

Match Case: Yes

Mode: Regular Expression

### Step 1.6 Remove the Data Value Names (6 Marks)

Find: (\t|^)([0-9A-Za-z\_]+:\s)

Replace: \$1

Match Case: Yes

Mode: Regular Expression