CSCI 548 HW2 FALL 2016

Name : Arpit Nigam

USC ID : 1933743387

Part1

Source: Blog Website – Globaldisasterwatch
URL: http://globaldisasterwatch.blogspot.com/
Quality: Low
Number of annotated units: first 25 in input file
Explanation of quality of source: Since this source is a blog we can't we sure about the accuracy of information provided in the text.

Source: Blog Website – Risk-macc
URL: http://risk-macc.blogspot.com/
Quality: Low
Number of annotated units: next 15 in input file
Explanation of quality of source: Since this source is a blog we can't we sure about the accuracy of information provided in the text.

Source: News Website – The Guardian
URL: https://www.theguardian.com/world/natural-disasters
Quality: High
Number of annotated units: next 20 in input file
Explanation of quality of source: Since this source is a news website most of the information which is available is accurate to the best possible value.

Source: News Website – Fox News
URL: http://www.foxnews.com/category/us/disasters.html
Quality: High
Number of annotated units: next 20 in input file
Explanation of quality of source: Since this source is a news website most of the information which is available is accurate to the best possible value.

Source: Government Website – FEMA
URL: http://www.fema.gov/disasters
Quality: High
Number of annotated units: next 20 in input file
Explanation of quality of source: Since this source is a US government website all information which is available thre are the most accurate ones.

Part 2

Name of feature: POS
Representative example: California NNP
Range of Values:
CC Coordinating conjunction
CD Cardinal number
DT Determiner
EX Existential there
FW Foreign word

IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS List item marker
MD Modal
NN Noun, singular or mass
NNS Noun, plural
NNP Proper noun, singular
NNPS Proper noun, plural
PDT Predeterminer
POS Possessive ending
PRP Personal pronoun
PRP$ Possessive pronoun
RB Adverb
RBR Adverb, comparative
RBS Adverb, superlative
RP Particle
SYM Symbol
TO to
UH Interjection
VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non3rd person singular present
VBZ Verb, 3rd person singular present
WDT Whdeterminer
WP Whpronoun
WP$ Possessive whpronoun
WRB Whadverb

Name of feature: Is alphanumeric?
Representative example: California is alphanumeric while Low-intensity is not
Range of Values: [0,1]

Name of feature: Is digit?
Representative example: 1215 is digit while Gaston is not
Range of Values: [0,1]

Name of feature: Is float?
Representative example: 7.2 is float while 273 is not
Range of Values: [0,1]

Name of feature: Starts with capital letter?
Representative example: California starts with capital letter while islands doesn't
Range of Values: [0,1]

Name of feature: Contains punctuations?
Representative example: west-north contains punctuation while word California doesn't
Range of Values: [0,1]

Name of feature: word in [week, month, year] **Representative example**: words which are equal to any of the three [week, month, year] Range of Values: [‘week’, ‘month’, ‘year’]
Name of feature: word is disaster type **Representative example**: Valcano is a disaster type while heavy is not **Range of Values**: ['quake','earthquake','typoon','volcano','eruption','storm','hurricane','flood','fire','cyclone','wildfire','landslide']
Name of feature: causalities **Representative example**: words which are from any of these three [‘dead’,’injured’,’rescued’] **Range of Values**: [‘dead’,’injured’,’rescued’]
Name of feature: counting words **Representative example**: words which are from any of these [‘tens’, ‘hundred’, ‘thousands’, ‘million’, ‘billion’] **Range of Values**: [‘tens’, ‘hundred’, ‘thousands’, ‘million’, ‘billion’]
Name of feature: Direction **Representative example**: words from any of these ['east','west','north','south'] **Range of Values**: ['east','west','north','south']
Name of feature: Week days **Representative example**: words from any of these ['sunday','monday','tuesday','wednesday','thursday','friday','saturday'] **Range of Values**: ['sunday','monday','tuesday','wednesday','thursday','friday','saturday']
Name of feature: Month names **Representative example**: words from any of these ['january','february','march','april','june','july','august','september','october','november','december'] **Range of Values**: ['january','february','march','april','june','july','august','september','october','november','december']

Part 3

Precision	Recall	F1	Category	Dataset
0.224621	0.323986	0.220324	10 low quality	First 10 from input file
0.221111	0.204762	0.154839	10 high quality	Last 10 from input file
0.229674	0.264236	0.206244	All 20 sentences	First and Last 10 from input file

Quality matters for the results:

Not really, if there is not much difference between precision and recall value then our accuracy is going to be somewhere in between mid of these two. While if in case one of either precision or recall value then our accuracy tends to be more on the lower side. Moreover, in this whole assignment the quality of output or in short accuracy of our model depend a lot on the features which we are generating. If attributes which we are choosing are good features, then definitely the end result will give us high accuracies. A lot of model performance also depends on the label which we are assigning and relation of those labels with our selected features.

Overall for above assignment in testing phase for low I got ~54% accuracy, for high I got ~51% accuracy and for both combined I got ~ 52%.