Homework 2 : Information Extraction

<u>Task 1 : Preliminary Work</u>

1. Manually selected sentences as per given instructions from five sources.

Name of Source	Quality of Source
The Guardian	High
CNN	High
USNews	High
Twitter	Low
Facebook	Low

2. Selected the following labels to be extracted, keeping in mind the domain of the task – recent disasters :

Label	Definition	Examples
Location	Any Place, State, Country name - could give us important information about where disaster took place	Italy, LA, CA, Mt. Fuji
Time	Any day, month, year or combination of those - giving us information about when disaster took place	August 21th, Jan., 2014
Disaster	Type of disaster - agreed that this is not a closed set, but ideally we'd want CRFs to capture syntactic information and predict doers of destruction as disasters	Earthquake, quake, flood, wildfire
Number	Specifically number of people - articles usually talk about casualties and damages - in this case, I considered only number of people affected and not buildings, properties etc.	**<u>2500</u>** people, more than **<u>500</u>** deaths
Irrelevant	Fallback default category for all other non-labelled tokens.	An, wiped, heavy,erupt

<u>Task 2 : Training the CRF</u>

1. I used NLTK for tokenizing words, I found that it is a reasonably accurate way to get tokens than manually splitting by spaces. It was also beneficial for using NLTK's support for part of speech tagging.

2. I selected 9 different features (description provided later in Task 3) and appended to the file. Also used a few automated ways to assign labels, thus saving manual effort wherever possible. Code is in the source folder.

3. Final training.txt in the given format can be found in the training folder for 104 sentences. Also used the same code for test files.

Task 3 : Questions

1. A brief description about the sources – (Total number of annotated units – 104)

Name of Source	URL	Quality of Source	Number of units	Explanation of Quality of Source
The Guardian	https://www.theguardian.com/	High	20	Trusted & 2nd Most Popular UK Online Newspaper. Won Best Newspaper awards multiple times.
CNN	http://www.cnn.com/	High	23	The online news website reports incidents instantly along with latest tweets by officials, photos and Live TV.
USNews	http://www.usnews.com/	High	20	Provides latest information in concise manner, less trustworthy than two above, more than below two ~ High Quality.
Twitter	https://twitter.com/	Low	20	Famous for capturing recent trends first but tweets can be unreliable, biased, unstructured and without citations.
Facebook	https://www.facebook.com	Low	21	Has larger audience base than twitter. Facebook hashtags have same drawbacks as Twitter or any identical social network.

2. Description about Features selected – (Total number of features – 17)

Feature Name	Reasoning	Range	Examples
PartOfSpeech	Gives meaningful information about sentence constructions and syntactic similarity	examples include NNP, VP, DT, etc. Full set can be obtained by using nltk.help.upenn_tagset() in terminal.	A-DT, ancient- JJ, people-NNS
IsCapital	More probable to be Proper nouns - like Location, Organization or disaster Name	[1,0] 1 if first letter is capitalized else 0	China - 1, an - 0
IsNumber	Suggests strong indication to year, value of money, number of people etc.	[1,0] 1 if text is all numbers else 0	123 - 1, a123 - 0, keep - 0

IsIndicator	Give weightage to presence of words like "by", "at", "in" suggesting doer of action, location, time etc.	[1,0] 1 if token∈{'by','at','in',''s','over','about','of'} else 0	floods **at** Bay, China**'s** quakes wiped away, Hurricane **in** OK,CA
IsPunctuation	Check if punctuation marks like ".", ",", "?", "!", etc.	[1,0] 1 if token is a punctuation else 0	: - 1, abc - 0
IsLink	Suggests presence of hyperlinks - largely found in tweets or reference links	[1,0] 1 if link is present else 0 , used regex to identify regex	www.firstpost.com - 1, hello world - 0
IsMonth	Closed set of 12 -> give weightage to direct correlation between "Time" and Month	[1,0] if token ∈[January-December] else 0	March -1 , Wednesday - 0
FirstLetter	To capture the appearance of the word	[A-Z a-z 0-9 punctuation]	China - C
LastLetter	To capture the appearance of the word	[A-Z a-z 0-9 punctuation]	China - a
Prefix3	Get first three letters of token	Any combination of 3 alphabets	China - Chi
Prefix4	Get first four letters of token	Any combination of 4 alphabets	China - Chin
Suffix3	Get last three letters of token	Any combination of 3 alphabets	China – ina
Suffix4	Get last four letters of token	Any combination of 4 alphabets	China - hina
Len1	Check if length is 1	[1,0] if len==1 else 0	China – 0, C -1
Len2	Check if length is 2	[1,0] if len==2 else 0	China – 0, Ch - 2
Len35	Check if length>=3 and <=5	[1,0] if len>=3 and <=5 else 0	China – 1, Ch - 0
Len6	Check if length>=6	[1,0] if len>=6 else 0	Function – 1

3. Test data is collected as testing_low, testing_high, testing_combined. Available in testing folder. Total number of entries – 20 + 20 = 40.

   Results are as below :

Precision	**Recall**	**F1**	**Category**	**Dataset**
0.9041	1	0.9497	Irrelevant	Low Quality
1	0.475	0.6441	Disaster	Low Quality
0.9592	0.8103	0.8785	Location	Low Quality
1	0.5	0.6667	Time	Low Quality
1	1	1	Number	Low Quality
Macro-average				
Precision	**Recall**	**F1**		
0.972666	0.757069	0.82778		

Precision	Recall	F1	Category	Dataset
0.9695	0.9922	0.9807	Irrelevant	High Quality
1	0.6667	0.8	Disaster	High Quality
0.9667	0.9062	0.9355	Location	High Quality
0.9167	1	0.9565	Time	High Quality
0.8889	0.8889	0.8889	Number	High Quality
Macro-average				
Precision	**Recall**	**F1**		
0.948353	0.890803	0.912328		

Precision	Recall	F1	Category	Dataset
0.9372	0.9959	0.9657	Irrelevant	Combined
1	0.5469	0.7071	Disaster	Combined
0.962	0.8444	0.8994	Location	Combined
0.9474	0.72	0.8182	Time	Combined
0.9	0.9	0.9	Number	Combined
Macro-average				
Precision	**Recall**	**F1**		
0.949315	0.801446	0.858063		

Yes, personally I think quality of datasets matters because of two major reasons –

1. A) High quality data was more structured and syntactically uniform. Proper nouns were capitalized and all grammar rules were followed.
   B) While on the other hand, low quality data had just words thrown because of how people casually write on blogs or social networking websites.
   It is also apparent from the F1 scores stated above high quality data gives better results.

2. A) I tried removing punctuations from the list of features. The scores decreased, just for low quality dataset.
   B) Upon further exploration – low quality data was from Twitter and Facebook which will tend to have more punctuations like people exclaiming (!!!!), or extra colons (…) etc. While on high quality dataset, the scores varied depending on data. Because we don't know the quality of dataset beforehand, we'll naturally include all the features and hence scores might change.