Unit 5

Arjun Chandrasekaran
Jouella Fabe
Jonathan Tran
Saurav Sharma
Peter Sulucz

Our investigation of nltk.ne_chunk

We applied the ne_chunk tagger to the ClassEvent and found that from the tagged set of Named Entities, only the top 15 were relevant to our collection :

Type of NE	NE	Frequency
GPE	New York	9
GPE	Islip	9
GPE	Suffolk	7
PERSON	Islip	7
ORGANIZATION	Islip	5
GPE	Detroit	5
GPE	Baltimore	5
ORGANIZATION	University	4
GPE	New York City	4

One interesting thing that we noticed is that Islip, which is a location near New York is tagged thrice! Once as a Person, once as a GPE (correctly) and once as an Organization. All other locations are tagged correctly.

We also tagged the Named Entities from our small collection on the Cloudera Virtual Machine. We experimented with different NamedEntity types and have listed our most relevant Named entities along with frequency of occurrence sorted by Type of entity:

Type of NE	NE	Frequency
GSP	Connecticut	4015
GPE	Newtown	3545
PERSON	Sandy Hook Elementary School	2507
PERSON	Sandy Hook Shooting Victims	110
PERSON	School Shooting	1281
PERSON	Adam Lanza	125
ORGANIZATION	CNN	2733

We found that when we sorted based on just frequency, there were We notice that Connecticut and Newtown are perfectly tagged as Locations, however, Sandy Hook Elementary School is tagged as a Person. Adam Lanza, is correctly identified as a person.
Another interesting result is the term School Shooting. This phrase is a topic of the event, and not a named entity - it does not fit either of the fields of GPE, Person or Organization.

Problem 2.3

StanfordNER default model
Hadoop information: application_1412653258045_0314
START: 2014-10-22 16:54:36
END: 2014-10-22 17:17:02
Total Duration: 22 minutes and 26 seconds for about 4000 documents = 1346 seconds

StanfordNER custom model
Hadoop information: application_1412653258045_0386
START: 2014-10-23 15:01:12
END: 2014-10-23 15:04:01
Total Duration: 2 minutes and 49 seconds for about 4000 documents = 169 seconds

Using the stanford NER tagger with the default english model, we were able to tag our small collection, with decent accuracy. Although, there is plenty of room for improvement.

- Location -- The first location tagged is Connecticut, which is correctly tagged as the State. This is correctly followed by the town: Newton. Realistically here, we would also be looking for “Sandy Hook” as the location, but both “Sandy” and “Hook” were tagged as Organizations. Interestingly, the tagger tagged Connecticut and Newton as both both locations and organizations.
- Organization -- The first organization which is tagged is google… Which is correctly tagged as an organization, but is irrelevant to our data. “School” is tagged as an organization too, but it should be a Location.

- Person: The first person tagged is “Lanza”, who was in fact the shooter involved in this event. We also have “adam” and “obama” tagged.

All in all, the tagger tagged some important features of our collection, but seemed to give them the wrong type of tag in a lot of the cases.

Using a stanford NER tagger with a custom training set, we were able to get a decent set. It did a good job tagging words that were important to our collection. Also it was a small training set, so it ran quite fast.

- Time -- This returned 3 items, “friday”, “saturday”, “dec”
- Person -- This tag seemed to function much better with our custom tagger, than with the Stanford tagger. All of the items that were tagged as Person were names. With the top two being ‘lanza’ and ‘adam’.
- Organization -- this tag was kind of neglected in our custom set, and it only tagged cnn and patch.com. We could have improved by better tagging organizations, but most of our training data was clean of them.
- Location performed the best of the tags. The first couple exactly describe the location of the shooting. “school”, “connecticut”, “newton”, “conn.”, “hook”, “sandy”. We can even see that Connecticut was very commonly abbreviated to “conn.”. Another interesting thing to note, is that “Newtown” was misspelled as “Newton” 8 times.

Discuss your investigation of Stanford NER

The Stanford NER is a Named Entity Recognizer that tags entities based on their grammatical structure. In comparison to the NLTK NER, we found that Stanford NER’s performs better after training it for the dataset. We see that both Stanford NER and nltk.ne_chunk have mistakenly tagged Sandy Hook Elementary School. However, after training, the Stanford NER tagged this correctly, along with other named entities.

Discuss the value of different types of named entities in summarizing your collection

ClassEvent

These are the top named entities in class event are GPE, Person, Organization and Location that are related to our corpus using the nltk chunk. Class event was best described by named entities having to do with location. Because most of the relevant data was location based. As you can see, most of the Organization data is unrelated.

LOCATION_islip 18
ORGANIZATION_weather 14
LOCATION_long 14
LOCATION_island 14
LOCATION_york 12
LOCATION_new 12
ORGANIZATION_noaa 10
ORGANIZATION_news 9
ORGANIZATION_national 9
ORGANIZATION_service 8
ORGANIZATION_sports 7
ORGANIZATION_entertainment 7
ORGANIZATION_health 6
ORGANIZATION_traffic 5

ORGANIZATION_tech 5
ORGANIZATION_world 4
ORGANIZATION_newsday 4
ORGANIZATION_new 4
ORGANIZATION_business 4
ORGANIZATION_york 3
ORGANIZATION_social 3
ORGANIZATION_contests 3
ORGANIZATION_all 3
ORGANIZATION_alerts 3
ORGANIZATION_watercooler 2
ORGANIZATION_u.s. 2

CollectionSmall

Our small collection responds best to being classified by locations and names, since those are what is most important in the event. Locations and names did the best job of being relevant, while organizations seemed largely irrelevant to our data.

Comparison of Taggers by Type of Entity

LOCATION

CUSTOM STANFORD NER		STANFORD NER	
school	2577	connecticut	1362
connecticut	1638	newtown	757
newtown	869	u.s.	599
conn.	329	conn.	275
hook	168	washington	91
sandy	166	ohio	75
control	79	east	74
washington	45	city	71
detroit	45	new	63
huffington	31	america	56
congressman	29	mississippi	54
elementary	22	united	48
company	18	hollywood	48
congress	10	san	45
newton	8	states	44
capitol	8	middle	44
chardon	7	york	43
connect	6	michigan	43
concussion	5	county	43
columbine	5	south	42
china	5	africa	42
schools	4	turnpike	38

nelson	4	europe	38
conn	4	asia	38
cohen	4	australia	35
city	4	oklahoma	34
christmas	4	latin	34
chinese	4	chicago	34
deadliest	3	carolina	33
county	3	florida	31
confronting	3	japan	28
canton	3	maryland	27
saturday	2	tennessee	25
pigeon	2	detroit	25
nation	2	twp	24
nancy	2	pittsburgh	24
monroe	2	rosanapaconcordoaklandlivermorehaywardsan	23
european	2	franciscosanta	23
deadly	2	bronx	23
darkens	2	knoxville	22
daily	2	boise	21
council	2	cleveland	19
could	2	calif.	19
controversial	2	usa	18
contributor	2	fla.	18
contract	2	china	18
contact	2	reddit	17
conneticut	2	francisco	16
comforting	2	benghazi	16
code	2	valley	15

PERSON

CUSTOM STANFORD NER		**STANFORD NER**	
lanza	217	lanza	339
adam	197	obama	264
obama	33	adam	246
ryan	16	sandy	140
nancy	9	hook	129
vance	4	john	102
paul	4	edwards	97
summers	3	kasich	84
puja	3	paul	82
parti	3	jennifer	69
nather	3	christina	67
lanzas	3	scarlett	63
kenen	3	apatow	63
jean	3	mccartney	62
bill	3	johansson	62

anna	3	aguilera	62
marsha	2	digg	58
wikipedia	1	david	55
stulhman	1	taylor	54
nick	1	james	54
martin	1	hamaguchi	54
marbella	1	lawrence	53
from	1	swift	51
dana	1	soto	50
cannon	1	samantha	49
		barack	49
		polly	46
		robert	43
		victoria	41
		clinton	40
		feinstein	35
		bieber	34
		dan	33
		huffington	32
		jr.	31
		nancy	30
		susan	29
		rice	29
		justin	29
		dick	29
		krull	28
		masafumi	27
		mary	27
		griffin	27
		george	26
		andrew	26
		shootadam	25
		ryan	25
		ron	25

Describe surprises and confirming results arising from evaluating your answers with the help of Solr

ClassEvent:

Based on the results of query for the random 5 words - Long Island New York Islip Suffolk Detroit, we found similar events related to our corpus - Islip13. However, we also found other events such as Shooting and Hurricane that is not related to the corpus.

SmallEvent:

Based on the query for the random 5 words - Sandy Hook Elementary School, Sandy Hook Shooting Victims School Shooting, Adam Lanza, CNN. We found that based on all of the event_types, we only found one discrepancy that is not related to our corpus, which was “Fire”.

We were surprised by the results we had gotten from Solr on the ClassEvent corpus since we had gotten 20% accuracy. In contrast, to the results that we had gotten from SmallEvent where have 90% accuracy on similar content.