Global Shark Attacks

Data Mining

Marc Solé Farré

DataSet [link]

	Case Number	Date	Year	Type	Country	Area	Location	Activity
0	2016.09.18.c	18-sept-16	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Surfing
1	2016.09.18.b	18-sept-16	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Surfing
2	2016.09.18.a	18-sept-16	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Surfing
3	2016.09.17	17-sept-16	2016	Unprovoked	AUSTRALIA	Victoria	Thirteenth Beach	Surfing
4	2016.09.15	16-sept-16	2016	Unprovoked	AUSTRALIA	Victoria	Bells Beach	Surfing
5	2016.09.15.R	15-sept-16	2016	Boat	AUSTRALIA	Western Australia	Bunbury	Fishing
6	2016.09.11	11-sept-16	2016	Unprovoked	USA	Florida	Ponte Vedra, St. Johns County	Wading
7	2016.09.07	7-sept-16	2016	Unprovoked	USA	Hawaii	Makaha, Oahu	Swimming
8	2016.09.06	6-sept-16	2016	Unprovoked	NEW CALEDONIA	North Province	Koumac	Kite surfing
9	2016.09.05.b	5-sept-16	2016	Unprovoked	USA	South Carolina	Kingston Plantation, Myrtle Beach, Horry County	Boogie boarding
10	2016.09.05.a	5-sept-16	2016	Unprovoked	AUSTRALIA	Western Australia	Injidup	Surfing
11	2016.09.04	4-sept-16	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Body boarding
12	2016.09.01	1-sept-16	2016	Unprovoked	USA	California	Refugio State Beach, Santa Barbara County	Spearfishing
13	2016.08.29.b	29-Aug-2016	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Surfing
14	2016.08.29.a	29-Aug-2016	2016	Unprovoked	USA	Florida	New Smyrna Beach, Volusia County	Surfing

DataSet link

Name	Sex	Age	Injury	Fatal (Y/N)	Time	Species	Investigator or Source
male	M	16	Minor injury to thigh	N	13h00		Orlando Sentinel, 9/19/2016
Chucky Luciano	M	36	Lacerations to hands	N	11h00		Orlando Sentinel, 9/19/2016
male	M	43	Lacerations to lower leg	N	10h43		Orlando Sentinel, 9/19/2016
Rory Angiolella	M		Struck by fin on chest & leg	N			The Age, 9/18/2016
male	M		No injury: Knocked off board by shark	N		2 m shark	The Age, 9/16/2016
Occupant: Ben Stratton			Shark rammed boat. No injury to occupant	N			West Australian, 9/15/2016
male	M	60s	Minor injury to arm	N	15h15	3' to 4' shark	News4Jax, 9/11/2016
female	F	51	Severe lacerations to shoulder & forearm	N	14h30	Tiger shark, 10?	Hawaii News Now, 9/7/2016
David Jewell	M	50	FATAL	Y	15h40		TVANouvelles, 9/6/2016
Rylie Williams	F	12	Lacerations & punctures to lower right leg	N	Late afternoon		C. Creswell, GSAF
Fraser Penman	M		No inury, board broken in half by shark	N	Late afternoon		Perth Now, 9/5/2016
Austin Moore	M	9	Foot bitten	N			Orlando Sentinel, 9/7/2016
Tyler McQuillen	M	22	Two toes broken & lacerated	N		White shark	R. Collier, GSAF
Sam Cumiskey	M	25	Lacerations to right foot	N	15h00	Bull shark, 6'	News Channel 8, 8/30/16
male	M	37	Minor injury to ankle	N	14h00		News Channel 8, 8/30/16
Laurent Chardard	M	20	Right arm severed, ankle severely bitten	N	17h00	Bull shark, 3.5 m	LaDepeche, 8/29/2016
David Cassetty	M	49	Minor injury to ankle	N	16h00		First Coast News, 7/25/2016
Johnny Stoch	M	15	Lacerations to left leg	N			ABC, 8/11/2016
Connor Baxter	M	21	No inury, shark & board collided	N		Tiger shark, 10'	SUP, 8/9/2015
Nolan Tyler	M	22	Big toe bitten	N		Blacktip shark	News 965, 8/5/2016
male	M	40	Lacerations to right hand	N	11h30	Blue shark	Informacion.es, 7/29/2016

Objective

Predicting the probability a shark attack with the following variables

- Month
- Hour
- Country
- Age
- Sex
- Activity

Data Cleaning

Time

1. Get the most repeated words in the “Time” column for obtaining the String types and define types and mapped to time field like afternoon -> 17h00.
2. Get the first time in the cases of two times

```
[u'13h00', u'11h00', u'10h43', nan, u'15h15', u'14h30', u'15h40',
 u'Late afternoon', u'15h00', u'14h00', u'17h00', u'16h00', u'11h30',
 u'12h00', u'19h05', u'After noon', u'1300', u'10h00',
 u'14h30 / 15h30', u'22h00', u'16h20', u'14h34', u'15h25', u'14h55',
 u'06h00', u'17h30', u'08h30', u'Morning', u'Afternoon', u'15h45',
 u'18h00', u'17h46', u'Morning ', u'13h20', u'15h49', u'07h00',
 u'08h00', u'10h45', u'19h00', u'13h30', u'Midnight', u'12h30',
```

Data Cleaning

Date

1. Clean the year of the “case number” for get the date

Reported 03-
2-mar-16
22-feb-16
19-Feb-2016
12-feb-16
Reported 10-
10-feb-16
5-feb-16
4-feb-16
29-Jan-2016
28-Jan-2016
25-Jan-2016
24-Jan-2016
24-Jan-2016
42392

Case Number
2016.09.18.c
2016.09.18.b
2016.09.18.a
2016.09.17
2016.09.15
2016.09.15.R
2016.09.11
2016.09.07
2016.09.06

Data Cleaning

Date

2. Some rows are B.C. and some others don't has the data and the year. For this rows I remove the data because it's not completed and has ambiguous data

5866	0077.00.00	77 A.D.	77
5867	0005.00.00	Ca. 5 A.D.	5

3 Join the year, month, and day cleaned and the Time values mapped to time.

- For the month of lost data, we put the average of the month of the others years.
- For the day of lost data, we put the average of the days of it's month and year.
- For the hour of lost data, we put the average of the hour of it's year, month and year.

Data Cleaning

Activity:

1. Get the most frequents words in activities and mapping to an identifier, like this.

```
Surfing
Swimming
Fishing
Spearfishing
Bathing
Wading
Diving
Standing
Snorkeling
Scuba diving
Body boarding
Body surfing
Swimming
Treading water
Pearl diving
Boogie boarding
Kayaking
Free diving
```

```
Swimming, using bundles of sticks as raft
Fishing, holding fish
Wading with surfboard
Fisherman
Swimming along side N.E.L. vessel Saluda
Spearfishing / freediving
The schooner Tahitienne foundered in a hurricane
Free diving with seals
Sinking of the M/V Mindoro during a typhoon
Abandoning burning ship Captain George in raging seas
```

```
if "SURFING" in activity:
    return 0
elif "SWIMMING" in activity:
    return 1
elif "FISHING" in activity:
    return 2
elif "SPEARFISHING" in activity:
    return 3
elif "BATHING" in activity:
    return 4
elif "WADING" in activity:
    return 5
elif "DIVING" in activity:
    return 6
elif "STANDING" in activity:
    return 7
```

Data Cleaning

Age:

1. Clean the concatenate chars and mapping the youngest or oldest words to a numerical age.
2. For the empty values I put the average of the age.

```
u'adult', u'9 & 12', u'? & 19', u'9 months', u'25 to 35',
u'23 & 26', u'1', u'(adult)', u'33 & 37', u'25 or 28',
u'37, 67, 35, 27,  ? & 27', u'21, 34,24 & 35', u'30 & 32',
u'50 & 30', u'17 & 35', u'X', u'"middle-age"', u'13 or 18',
u'34 & 19', u'33 & 26', u'2 to 3 months', u'MAKE LINE GREEN',
u' 43', u'81', u'"young"', u'7 or 8', u'78', u'17 & 16', u'F'
```

```
if "teen" in age.lower():
    clean_age = "10"

if "adult" in age.lower():
    clean_age = "18"

if "young" in age.lower():
    clean_age = "14"

if "middle-age" in age.lower():
    clean_age = "35"
```

Data Cleaning

Sex:

- Mapping M and F to 0 and 1 and the others are mapped to 0 because is the most frequent

```
data["Sex "].value_counts()
```

```
M      4835
F       585
M         2
N         1
.         1
lli       1
```

Data Cleaning

Country:

- Some Name of countries are completed with the region in the same cell and removes it, next I mapped the name of countries with an identifier

Data Cleaning

Some rows of the csv are broken or empty. In this case I remove the row

2014.10.11	11-oct-14	2014	Boat	AUSTRALIA	
,M,,Kayak deflate	no injury to	N	17h20		
2014.10.05.b	5-oct-14	2014	Unprovoked	USA	

Partial clean

Clean Date	Clean Time	Hour	Sex Fixed	DateTime	Month	Id Type Attack	Clean Age	Country Id	Activity Id	Clean Fatal
18/9/16	13:00:00	13	0	18/9/16 13:00	9	1	16	0	0	1
18/9/16	11:00:00	11	0	18/9/16 11:00	9	1	36	0	0	1
18/9/16	10:43:00	10	0	18/9/16 10:43	9	1	43	0	0	1
17/9/16	13:00:00	13	0	17/9/16 13:00	9	1	14	1	0	1
15/9/16	13:00:00	13	0	15/9/16 13:00	9	1	14	1	0	1
15/9/16	13:00:00	13	0	15/9/16 13:00	9	2	14	1	2	1
11/9/16	15:15:00	15	0	11/9/16 15:15	9	1	60	0	5	1
7/9/16	14:30:00	14	1	7/9/16 14:30	9	1	51	0	1	1
6/9/16	15:40:00	15	0	6/9/16 15:40	9	1	50	2	0	2
5/9/16	20:00:00	20	1	5/9/16 20:00	9	1	12	0	13	1
5/9/16	20:00:00	20	0	5/9/16 20:00	9	1	14	1	0	1
4/9/16	13:00:00	13	0	4/9/16 13:00	9	1	9	0	10	1
1/9/16	13:00:00	13	0	1/9/16 13:00	9	1	22	0	2	1
29/8/16	15:00:00	15	0	29/8/16 15:00	8	1	25	0	0	1
29/8/16	14:00:00	14	0	29/8/16 14:00	8	1	37	0	0	1
27/8/16	17:00:00	17	0	27/8/16 17:00	8	1	20	3	0	1
25/8/16	16:00:00	16	0	25/8/16 16:00	8	1	49	0	5	1
7/8/16	13:00:00	13	0	7/8/16 13:00	8	1	15	4	8	1
6/8/16	13:00:00	13	0	6/8/16 13:00	8	1	21	0	18	1

Data mining

Now we had the data cleaned and the next step is apply the Naive Bayes algorithm, for do this we generate an array of labeled points in order to predict if the attack of a shark can be fatal or not, according to the next variables:

- Country
- Age
- Month
- Type activity
- Hour
- Sex

Data mining

When we have a model trained, we give a simple interface for input the variables and get the prediction of the fatality of the attack

Activity

- 0 SURFING
- 1 SWIMMING
- 2 FISHING
- 3 SPEARFISHIN
- 4 BATHING
- 5 WADING
- 6 DIVING
- 7 STANDING
- 8 SNORKELING
- 9 SCUBA DIVING
- 10 BODY BOARDING
- 11 TREADING WATER1
- 12 PEARL DIVING
- 13 BOOGIE BOARDING
- 14 KAYAKING
- 15 FREE DIVING
- 16 FELL OVERBOARD
- 17 WINFSURFING
- 18 OTHER

```
text = widgets.Text()
display(text)
```