Benjamin Gutierrez Garcia bencouver@gmail.com

Sentiment analysis

Minería de datos de Twitter para evaluar "sentimiento social"

Introduction

Como proyecto de análisis de sentimiento social escogí utilizar el stream de twitter para identificar tweets positivos o negativos. De hecho ya había trabajado en esto antes y me gustaría compartirles mi codigo y lo que hice. Esta red social es considerada como una de las más ricas en información del sentir social en tiempo real en el mundo. Se ha demostrado que el "humor" de la comunicación de tweeter es un reflejo de ritmos biológicos, y puede hasta usarse para predecir los mercados bursátiles. Un estudiante de la Universidad de Washington tweets geo-codificados para crear un mapa de las ubicaciones donde la palabra "relámpago" se mencionaba en el contexto de un sistema de tormentas en el verano de 2012. Investigadores de la Universidad de Northeastern y Harvard que estudian las características de la dinámica de twitter tienen un excelente sitio sitio para aprender más acerca de twitter y como analizar "humor" o "sentimiento" a escalas nacionales.

Los tweets representan datos muy sucios y sin estructura, que por lo regular están incompletos, y que a veces ni siquiera tiene texto. A veces tiene localización geográfica, idioma, a veces están codificados en distintos lenguajes y formatos, etc.

Primer Paso: Capturar el stream de datos de Twitter

Para esto utilizamos la API (Application Programming Interface)[1] que nos permite conectar un programa en python en nuestra laptop a twitter.

Se tiene que crear una cuenta de twitter y crear una App nueva en **https://dev.twitter.com/apps**. Con esto se crear un Token de autenticación que proporciona las siguientes credenciales para nuestra aplicación de python:

```
api_key = "<Enter api key>"
api_secret = "<Enter api secret>"
access_token_key = "<Enter your access token key here>"
access_token_secret = "<Enter your access token secret here>"
```

Estas credenciales son colocadas en mi `codigo` *`twitterstream.py`* P**or seguridad removi mis credenciales del código que estoy entregándoles en el repositorio de github, así que se deben colocar otras credenciales o puedo hacer un demo de alguna forma si es requerido**.

El código está escrito en python 2.7 y se ejecuta como: *`python twitterstream.py`*

La ejecución produce el stream de tweets como salida (siguiente figura). Como se observa viene codificado en *unicode strings* y se requiere la biblioteca del API para pasarlo a JSON y sea humanamente legible.

[1] https://developer.twitter.com/en/docs

Implementación de la evaluación de Sentimiento

La estrategia a seguir es asignar un sentimiento en base a las palabras del texto del tweet. Cada palabra recibirá un score y el sentimiento es la suma de los scores de todas las palabras del tweet. Para asignar scores a palabras vamos a usar un diccionario,y si la palabra no está pues el score es cero. El diccionario que vamos a usar es:

> AFINN[2] is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. There are two versions:
>
> AFINN-111: Newest version with 2477 words and phrases.
>
> AFINN-96: 1468 unique words and phrases on 1480 lines. Note that there are 1480 lines, as some words are listed twice. The word list in not entirely in alphabetic ordering.

[2] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, Michael Etter, "*Good Friends, Bad News - Affect and Virality in Twitter*", The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011).

Vamos a utilizar el formato JSON de los datos de twitter para que lo podamos usar como un formato de diccionario.

Cada campo debe ser validado y para simplificar la tarea vamos a aplicar las siguientes reglas para asignar un sentimiento a un tweet de acuerdo al formato en que viene:

1. Nos vamos a limitar a tweets en INGLÉS. Tweets en otros idiomas reciben un sentimiento de cero. Esto es claramente porque el diccionario que tenemos es en ingles, y ademas muchos idiomas (Europa del Este, Asia, etc) tiene codificaciones binarias difíciles de limpiar.
2. Tweets sin texto reciben un sentimiento de cero.
3. Si el campo del lenguaje está vacío, sentimiento es igual a cero.

La estructura JSON de los tweets es de esta forma:


```
{
  "tweet": {
    "created_at": "Thu Apr 06 15:24:15 +0000 2017",
    "id_str": "850006245121695744",
    "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API
platform!\nhttps:\/\/t.co\/XweGngmxlP",
    "user": {
      "id": 2244994945,
      "name": "Twitter Dev",
      "screen_name": "TwitterDev",
      "location": "Internet",
      "url": "https:\/\/dev.twitter.com\/",
      "description": "Your official source for Twitter Platform news, updates & events. Need
technical help? Visit https:\/\/twittercommunity.com\/ \u2328\ufe0f #TapIntoTwitter"
    },
    "place": {

    },
    "entities": {
      "hashtags": [

      ],
      "urls": [
        {
          "url": "https:\/\/t.co\/XweGngmxlP",
          "unwound": {
            "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
            "title": "Building the Future of the Twitter API Platform"
          }
        }
      ],
      "user_mentions": [
```

```
        ]
      }
    }
}
```

Minando Tweets

Primero vamos a tomar diez minutos de tweets (aprox 150MB) y redireccionarlos a un archivo:

```
python twitterstream.py > output.txt
```

A continuación ejecutamos el código para asignar un sentimiento, usando el diccionario más grande que tenemos:

```
python tweet_sentiment.py AFINN-111.txt output.txt > resultados.txt
```

Esto nos proporciona simplemente una línea continua de números enteros entre -5 y 5 para el sentimiento del tweet, algo así como:

-3.4.4.5.-2.3,-2,4,5,-1,0,0,0,4,0,0,0,0....

Tengo una opcion en el codigo que me permite obtener mas atributos de cada tweet en forma relativamente limpia: idioma, ubicacion, texto, fuente (IOS, Android, etc), y sentimiento. Aquí tenemos un extracto:

```
('Not English', u'in', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @lilyachty: Somebody make me into a fortnite character', 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @nganguka: Calling The Attention of all MayWard Flyers, Please Proceed to the Twitter, We Have Our Trending Party Today! \n\nLeggo, Fam!\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'pt', 'sentiment=', 0)
('Not English', u'ko', 'sentiment=', 0)
('Not English', u'es', 'sentiment=', 0)
('Not English', u'pt', 'sentiment=', 0)
('Not English', u'th', 'sentiment=', 0)
('Not English', u'th', 'sentiment=', 0)
('Not English', u'pt', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @jonnysun: whenever donald glover does
```

```
anything i am filled with an immense and deep-seated creative anxiety that does not go away for a\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'in', 'sentiment=', 0)
('Not English', u'ro', 'sentiment=', 0)
('Not English', u'es', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @NoHoesGeorge: gamer girls https://t.co/XeJ3GTUJQu', 'unknown', 'sentiment=', 0)
('Not English', u'tr', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @sttepodcast: RT &amp;amp; Follow us to #win a @SignatureEntertainment 5-film romance movie DVD bundle to celebrate the release of #MeghanMarkle\u2026', 'unknown', 'sentiment=', 5)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @Sanemavcil: LumiWatch: On-Arm Projected Graphics and Touch Input by Robert Xiao\nOriginal Video: https://t.co/y3WlnmYzkF\n\n#Technology #A\u2026', 'unknown', 'sentiment=', 0)
('Not English', u'ko', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u"RT @_ShamGod: There is a whole generation of people who don't remember when Snoop was on trial for murder. What a world. https://t.co/Fj1Lf\u2026", 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'@elonmusk Hello Elon... when You will bring tesla to India??', 'unknown', 'sentiment=', 0)
('Not English', u'ru', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'RT @Penguincallme: Yes bitch I\u2019ve been trying to reach you for days wtf https://t.co/HcUXOKGTEl', 'unknown', 'sentiment=', -8)
('Not English', u'und', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'I spoke too soon. This one wins. https://t.co/6yOWXMwPid', 'unknown', 'sentiment=', 0)
('ENGLISH', u'en', 'loc:', None, 'text:', u'corsets black chick dick porn teen white reggaeton sex videos jessica biel easy virtue sex scene tattoo  https://t.co/rLFylD4Duk', 'unknown', 'sentiment=', -3)
HAS NO TEXT,sentiment=0
HAS NO TEXT,sentiment=0
HAS NO TEXT,sentiment=0
HAS NO TEXT,sentiment=0
HAS NO TEXT,sentiment=0
('Not English', u'ko', 'sentiment=', 0)
('Not English', u'ja', 'sentiment=', 0)
('Not English', u'ja', 'sentiment=', 0)
```

Esto nos da una idea de la heterogeneidad de los datos. ES el mismo formato JSON de arriba pero “desenrollado” y con datos reales. La idea es que el potencial de minería de información es enorme, por ejemplo podría asociar ubicación geográfica con sentimiento, aislando aquellos tweets que tengan esta información:

```
('ENGLISH', u'en', 'loc:', {u'full_name': u'Manchester, CT', u'url': u'https://api.twitter.com/1.1/geo/id/f1d134c7fd204d74.json', u'country': u'United States',
```

u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[-72.583489, 41.733619], [-72.583489, 41.820226], [-72.465121, 41.820226], [-72.465121, 41.733619]]]}, u'country_code': u'US', u'attributes': {}, u'id': u'f1d134c7fd204d74', u'name': u'Manchester'}, 'text:', u'My dad truly the funniest man ever today my best friend birthday and he literally sent me pictures of when me and h\u2026 https://t.co/6iVOx11HrE', 'unknown', 'sentiment=', 3)

('ENGLISH', u'en', 'loc:', {u'full_name': u'Mandaluyong City, National Capital Region', u'url': u'https://api.twitter.com/1.1/geo/id/005de1fe214f002d.json', u'country': u'Republic of the Philippines', u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[121.016761, 14.567448], [121.016761, 14.602063], [121.06176, 14.602063], [121.06176, 14.567448]]]}, u'country_code': u'PH', u'attributes': {}, u'id': u'005de1fe214f002d', u'name': u'Mandaluyong City'}, 'text:', u'Atm\U0001f60a swimming time\U0001f3ca\U0001f3fb\u200d\u2640\ufe0f @ Tivoli Garden Residences https://t.co/EiRKuSzMMs', 'unknown', 'sentiment=', 0)

('ENGLISH', u'en', 'loc:', {u'full_name': u'Kuching, Sarawak', u'url': u'https://api.twitter.com/1.1/geo/id/c9250e46aa4bac0b.json', u'country': u'Malaysia', u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[110.128315, 1.052488], [110.128315, 1.805971], [110.537741, 1.805971], [110.537741, 1.052488]]]}, u'country_code': u'MY', u'attributes': {}, u'id': u'c9250e46aa4bac0b', u'name': u'Kuching'}, 'text:', u'Secrets Ive held in my heart are harder to hide than thought.', 'unknown', 'sentiment=', -1)

('ENGLISH', u'en', 'loc:', {u'full_name': u'Ohio, USA', u'url': u'https://api.twitter.com/1.1/geo/id/de599025180e2ee7.json', u'country': u'United States', u'place_type': u'admin', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[-84.820309, 38.403186], [-84.820309, 42.327133], [-80.518626, 42.327133], [-80.518626, 38.403186]]]}, u'country_code': u'US', u'attributes': {}, u'id': u'de599025180e2ee7', u'name': u'Ohio'}, 'text:', u'Let\u2019s play a game on how many tweets I can post of lyrics from 3:15! We already have two from within the last hour\u2026 https://t.co/X7qBOckRwQ', 'unknown', 'sentiment=', 0)

('ENGLISH', u'en', 'loc:', {u'full_name': u'Queens, NY', u'url': u'https://api.twitter.com/1.1/geo/id/00c39537733fa112.json', u'country': u'United States', u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-73.699793, 40.541722]]]}, u'country_code': u'US', u'attributes': {}, u'id': u'00c39537733fa112', u'name': u'Queens'}, 'text:', u'@ZarekValentin Hard fought win tonight!!! Thanks for leaving it all on the pitch!!! \U0001f49a\U0001f49b\U0001f49a\U0001f49b #PTFC #RCTID #lovemyteam @MLS @TimbersFC', 'unknown', 'sentiment=', 4)

('ENGLISH', u'en', 'loc:', {u'full_name': u'Minneapolis, MN', u'url': u'https://api.twitter.com/1.1/geo/id/8e9665cec9370f0f.json', u'country': u'United States', u'place_type': u'city', u'bounding_box': {u'type': u'Polygon', u'coordinates': [[[-93.329515, 44.889964], [-93.329515, 45.051257], [-93.194578, 45.051257], [-93.194578, 44.889964]]]}, u'country_code': u'US', u'attributes': {}, u'id': u'8e9665cec9370f0f', u'name':

u'Minneapolis'}, 'text:', u'Holy Fuck. https://t.co/ApipRmOjx3', 'unknown', 'sentiment=', 0)

Un proyecto futuro es asociar sentimiento a un mapa.

Sentimiento de tweets en inglés por un cierto periodo de tiempo

Un histograma que puedo construir es cuántos tweets negativos y positivos existen en un periodo de tiempo, en ingles. Con los datos de diez minutos tenemos:

Removi aquellos son sentimiento cero ya que de otra forma dominan el histograma y no se podrían apreciar las proporciones. Parece que los sentimientos positivos dominan en el periodo de tiempo muestreado. Después de un incidente o catástrofe quizá las tendencias sean otras, con más palabras de estrés. Se incluye un script en bash para obtener los datos para el histograma y uno en python para graficarlo:

```
sh hist.sh > sentiment_data.csv
```

```
python plot_histogram.py
```

Limitaciones

Es claro que la limitación principal es el diccionario y estar restringidos a inglés.