

Web Scraping y Lenguaje de Procesamiento Natural en la polaridad de opinión de usuarios en medios digitales y redes sociales

Kevin Molina-Bejar

Universidad Nacional de Ingeniería

kevin.molina@uni.pe

December 16, 2018

1 Web Scraping

- Concepto
- Maneras de evitarlo

2 Análisis de Sentimiento

- Concepto
- Proceso de entrenamiento y predicción

3 Diseño Empleado - Web Scraping

- Medio Digital
 - Data Recopilada

4 Diseño Empleado - Análisis de Sentimiento

- API Twitter
- Modelo Desarrollado
 - SVM
- Data de Entrenamiento
 - TASS
- Pre-procesamiento de Datos
- Predicción de la Polaridad

5 Conclusiones y trabajo futuro

- Conclusiones
- Trabajo futuro

Concepto

Es una técnica para simular la navegación de un usuario en la web e ir estructurando datos relevantes de manera automática.

Maneras de evitarlo

- Captchas.
- Peticiones por dirección IP.
- Fichero robot.txt.
- Uso de contenido dinámico.
- Generar etiquetas html random.

Web Scraping - Aplicaciones

ORCE

Búsqueda de Alumnos y Egresados

Nota: Para efectuar la búsqueda, sólo es necesario colocar uno o más datos solicitados. Haga click en los campos para mostrar un ejemplo.

Codigo UNI

APE: PATERNO

APE: MATERNO

NOMBRE

Buscar

Limpiar Campos

Figure: Fuente: `orce.uni.edu.pe`

Fotografia

..Datos Personales:.

Codigo UNI: 20110331F

Nombres: MOLINA-BEJAR-KEVIN DANIEL

Facultad: CIENCIAS

Especialidad: CIENCIA DE LA COMPUTACIÓN

Situación: ALUMNO REGULAR **MATRICULADO - 182**

Medida Disciplinaria: NO TIENE

Ciclo Relativo Aprobado: 1 2 3 4 5 6 7 8 9 10

Ciclo Relativo Matriculado: 1 2 3 4 5 6 7 8 9 10

Figure: Fuente: `orce.uni.edu.pe`

Recolectando datos de ORCE.

```
kevin@kevin-Inspiron-14-3467: ~/Escritorio/python
```

```
20872053H, C1, ALEGRE-MILLA-ENRIQUE GREGORIO
20872049K, M3, ALVINO-RODRIGUEZ-FRANCIS
20872021I, M6, ALVARADO-ALVA-DANIEL ANGEL
20872018H, C1, ACOSTA-LEZAMA-JOSE HUMBERTO
20872012J, L2, ANDRADE-TEEVIN-RUBEN LEONARDO
20872000B, M6, ALBITES-SANABRIA-JOSE LUIS
20872007F, C1, ACOSTA-GRANDEZ-HUGO LEONIDAS
20872004G, L2, AGAMA-MOLINA-JOSE DANIEL
20872004G, L2, AGAMA-MOLINA-JOSE DANIEL
20871402I, A1, ABSI-MEJIA-MARLO ISRAEL
20871387J, A1, AYALA-ARANGO-HERNAN MIGUEL MAC
20871355K, P3, ASCENCIOS-ALBAN-ERNESTO DAVID
20871350I, E3, APEÑA-CHILI-CHRISTIAN IVAN
20871346A, Q1, ALVARADO-BALTAZAR-ROBERT TEODORO
20871344I, N3, AYALA-PIZARRO-DAGOBERTO MANUEL
20871343B, E3, ALVA-RODRIGUEZ-EDNAR OSCAR
20871342F, Q1, AREVALO-FLORES-JORGE ARTURO
20871338I, E3, ALVARADO-VALLEJOS-ARNALDO EDUARDO
208713190, E1, AQUINO-VALLEJOS-CARLOS JUNIOR
20871318H, P3, ALAMA-OTAEGUI-ALONSO WILFREDO
20871312J, Q1, ARESTEGUI-ROMAN-NOE
20871284F, E1, ARIJATA-MANANI-JORGE LUIS
208712816, C1, ANCASTI-CANDIOTTI-FREDDY
20871280K, Q1, ANCALLE-DE LA CRUZ-ROSA MERCEDES
20871274K, G1, ALCANTARA-CHUCO-OCTAVIO FERMIN
20871262B, A1, ALVAREZ-ANGULO-MARIA VICTORIA
20871242A, S1, AURIS-CORTEZ-YESICA FIORELA
20871235E, G3, ALVAREZ-HUAYHUA-MIGUEL GRIMALDO
20871234I, Q2, ALBERTI-MAYS-MARYSABEL ELAINE
20871234I, Q2, ALBERTI-MAYS-MARYSABEL ELAINE
20871223G, C1, ALDAVE-GUILLEN-LUIS ALEJANDRO
20871217G, Q1, ALMONACID-CHURA-CARLOS ALBERTO
20871208H, M4, ACOSTA-CIRIACO-FRANK ENRIQUE
20871182I, M4, ALDANA-CORDOVA-JOSEPH JULIO
20871173J, M4, ALVAREZ-ESCOBAR-RAFAEL
20871165G, Q1, AMASIFEN-BRAVO-VIRGINIA NATALY
20871150J, G1, ARIZA-SANCHEZ-ERICK GODOFREDO
20871143C, I1, ALVITES-CARPIO-DANIEL ALEJANDRO
20871090H, L1, ACASIEDE-VEGA-KEVIN ODYNSKY
20871092J, S1, AYUQUE-MENDOZA-EDSON RHOMARIO
20871090G, I1, ARAGON-VALLADOLID-JAVIER ENRIQUE
20871088B, I1, ALVAREZ-HERRERA-RICARDO DIONICIO
20871064F, M6, ACEVEDO-PASCUAL-GLIDOVER FRANCO
20871058F, C1, ALVAREZ-JARA-KIRO IRWING
20871052H, M4, ALMIRON-FLORES-JOSEPH MIJAIL
20871049G, P2, AQUINO-HANCO-NESTOR
20871046H, M4, ALBINO-MARTINEZ-CESAR CRISTIAN
20871039A, I2, ATALLUZ-HUARI-OMAR ENMANUEL
20871034J, I2, ATOCHE-BRAVO-JUNIOR ALIN
20870435K, A1, AVILA-COLCHAO-JULIO CESAR
```

Web Scraping - Aplicaciones

ONPE

ELECCIONES REGIONALES Y MUNICIPALES 2018		ONPE
BIENVENIDO(A)	DNI: 72276988	AYACUCHO / HUAMANGA / AYACUCHO
Nombres: ADERLY KEN		Mesa de Sufragio: 008779 N° de Orden en la Mesa: 78
Apellidos: POMA CHUCHON		Local de Votación: IE MARIA PARADO DE BELLIDO
No es Miembro de Mesa.		
		Dirección: Referencia:
		FRENTE AL COLEGIO MARIA AUXILIADORA

Figure: Fuente: consultamiembrodemesa.onpe.gob.pe

Web Scraping - Aplicaciones

ONPE

Recolectando datos de ONPE.

BIENVENIDO(A)

DNI: 19998833

Nombres:
OSWALDO

Apellidos:
SUAREZ ALVARADO

DNI	AP. PATERNO	AP. MATERNO

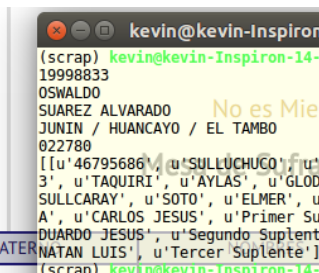


Figure: Fuente: `consultamiembrodemesa.onpe.gob.pe`

Dato curioso.

BIENVENIDO(A)		DNI: 00000317	UCAYALI / CORONEL PORTILLO / CALLERIA
Nombres: MIGUEL	Apellidos: RODRIGUEZ DEL AGUILA	No es Miembro de Mesa.	
		Mesa de Sufragio: 077363	

Figure: Fuente: `consultamiembrodemesa.onpe.gob.pe`

Concepto

Es un campo dentro del Procesamiento de Lenguaje Natural (PNL) construye sistemas que intentan identificar y extraer opiniones dentro del texto.

- Polaridad.
- Asunto.
- Titular de opinión.

Tipos

- Análisis de sentimiento de grano fino
- Detección de emociones
- Análisis de sentimiento basado en aspectos
- Análisis de intenciones

Análisis de Sentimiento

Proceso de entrenamiento y predicción

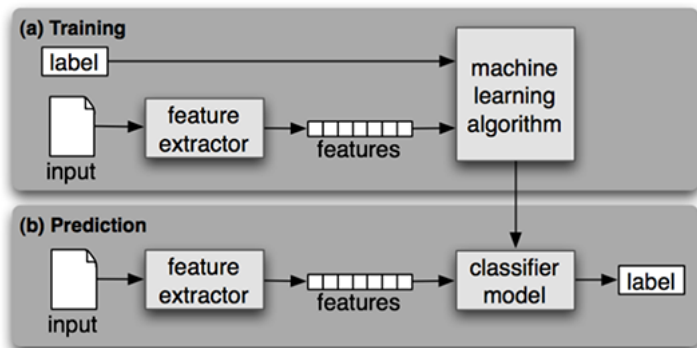


Figure: Fuente: monkeylearn.com

- Monitoreo de redes sociales.
- Monitoreo de marca.
- Servicio al cliente.

Análisis de Sentimiento - Aplicaciones

Monitoreo de redes sociales

Beneficios

- 1 Priorizar la acción.
- 2 Seguimiento de tendencias.
- 3 Competencia.



Figure: Ejemplo Saga Falabella. Fuente: elcomercio.pe

Seminario de Tesis

Resumen

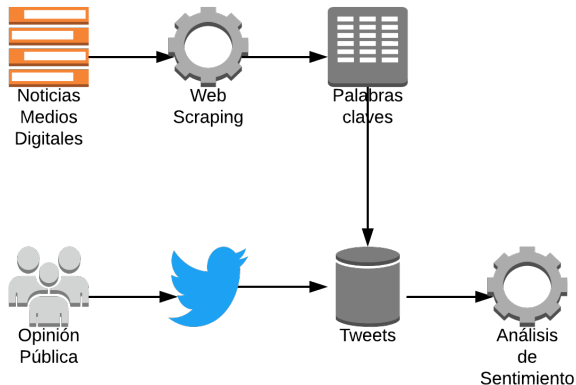


Figure: Fuente: elaboración propia

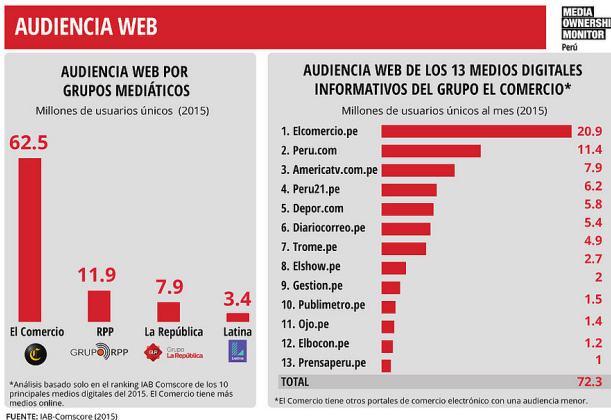


Figure: Fuente: IAB-Comscore(2015)

Diseño Empleado

Web scraping

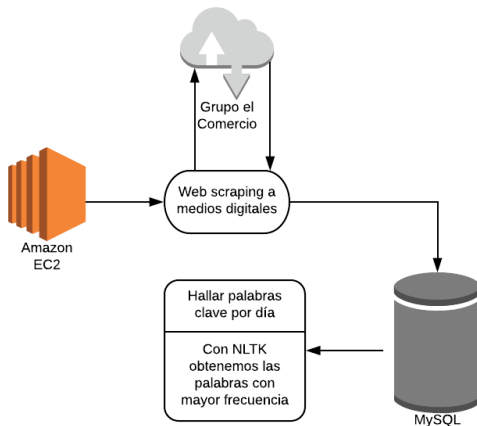


Figure: Fuente: elaboración propia

Data Recopilada

Elcomercio.pe, peru21.com, trome.pe, gestion.pe, peru.com

<https://elcomercio.pe/rss/>



El Comercio

CANALES RSS

Ya no necesitas buscar las noticias: ahora ellas te encuentran. Anímate a usar la tecnología RSS (**Real Simple Syndication**), un sistema que te alertará automáticamente de las novedades de elcomercio.pe y de otros sitios de su interés.


Para ello necesitas de un **agregador RSS**, un programa que te permitirá visualizar las últimas noticias cuando lo desees. Algunos pueden ser descargados de Internet en tu computadora, mientras que otros funcionan en un sitio web, sin necesidad de instalar ningún programa.

PORTADA

 Política

 Perú

 Lima

 Tecnología y ciencias

 Mundo

 Economía

Figure: RSS. Fuente: elcomercio.pe

Data Recopilada

Palabras más frecuentes

```
freq[freq.fecha==4][['palabra','frecuencia','fecha']]
```

Out[20]:

	palabra	frecuencia	fecha
16	Elecciones	35	4
17	Fujimori	31	4
18	candidatos	23	4
19	Alberto	17	4
20	indulto	16	4
21	Congreso	15	4

Figure: 04-10-2018. Fuente: elaboración propia

Diseño

Análisis de sentimiento

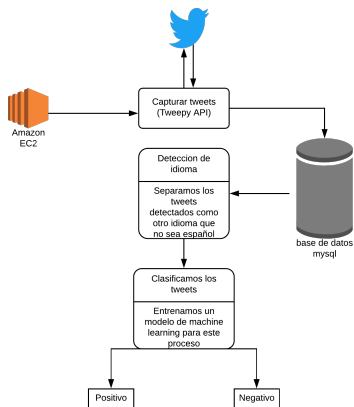


Figure: Fuente: elaboración propia

API Twitter

Beneficios premium

Tweets	Standard (free)	Premium	Enterprise
Publish and engage	✓		
Search Tweets: 7-days	✓		
Search Tweets: 30-days		✓	✓
Search Tweets: Full-archive		✓	✓
Filter Tweets	✓		✓
Sample Tweets	✓		✓
Batch Tweets			✓

Figure: Fuente: developer.twitter.com

Concepto

Las máquinas de vectores de soporte (SVM) son un conjunto de métodos de aprendizaje supervisado que se utilizan para la clasificación o la regresión.

Beneficios LinearSVC

- 1 Versátil.
- 2 Eficiente en memoria
- 3 Validación cruzada

TASS

El TASS se ha llevado a cabo desde 2012, en el marco de la Conferencia Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN)

Tokenizar

- 1 Función `nltk.word_tokenize`.
- 2 Signos de puntuación.

Lematizar

- 1 Función `nltk.stem`.
- 2 Luego de tokenizar.

Stopwords

- 1 Stopwords de la librería `nlTK.corpus`.
- 2 Palabras sin valor semántico.

Cambiar a minúsculas y **CountVectorizer**

- 1 **CountVectorizer** de la librería `sklearn.feature_extraction.text`
- 2 Parámetro `lowercase`.

Ejemplo

```
7 Más de mañana en Gaceta. UPyD contará casi seguro con grupo gracias al Foro Asturias. Eso se dice
so
7 Las remuneraciones económicas son lo que todos esperan, pero un "me encanta lo que has hecho", lo
l cliente es primero.
Name: content, dtype: object
['asturi', 'casi', 'client', 'congres', 'cont', 'dic', 'econom', 'encant', 'esper', 'for', 'gacet', 'gr',
p', 'hech', 'mas', 'mañan', 'per', 'primer', 'remuner', 'segur', 'tod', 'upyd', 'val']
[[1 1 0 1 1 1 0 0 0 1 1 1 1 0 1 1 0 0 0 1 0 1 0]
 [0 0 1 0 0 0 1 1 1 0 0 0 0 1 0 0 1 1 1 0 2 0 1]]
```

Figure: Fuente: elaboración propia

Predicción de la Polaridad

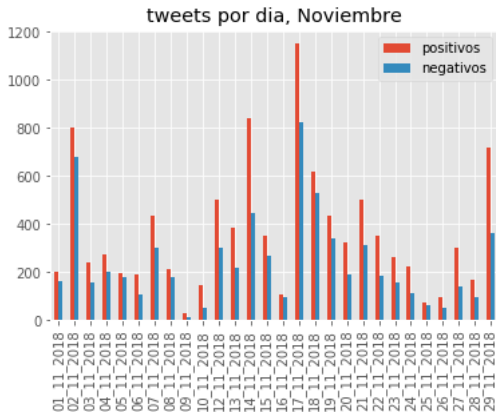


Figure: Fuente: elaboración propia

Predicción de la Polaridad

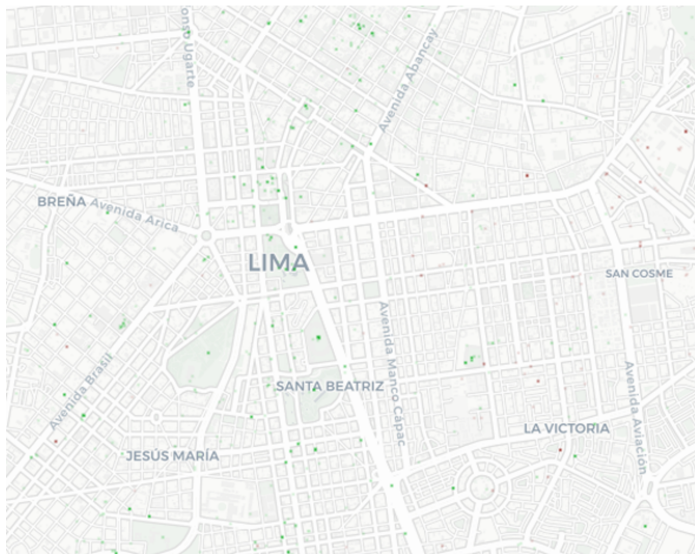


Figure: Fuente: elaboración propia

Conclusiones y Trabajo Futuro

Conclusiones

- Twitter gran fuente de datos no estructurado.
- Palabras más usadas, fútbol.
- Corpus TASS como muestra de entrenamiento.
- Objetivo inconcluso.
- Faltas ortográficas y jergas.

Conclusiones y Trabajo Futuro

Trabajo futuro

- Clasificador multiclase.
- Incluir otra rama de análisis de sentimiento.
- Focused crawler.
- Grupo interdisciplinario.

Gracias