# Procesamiento de Lenguaje Natural como herramienta para el estudio de la conflictividad

Algunas ideas

Germán Rosati (CONICET-UNSAM, PIMSA)

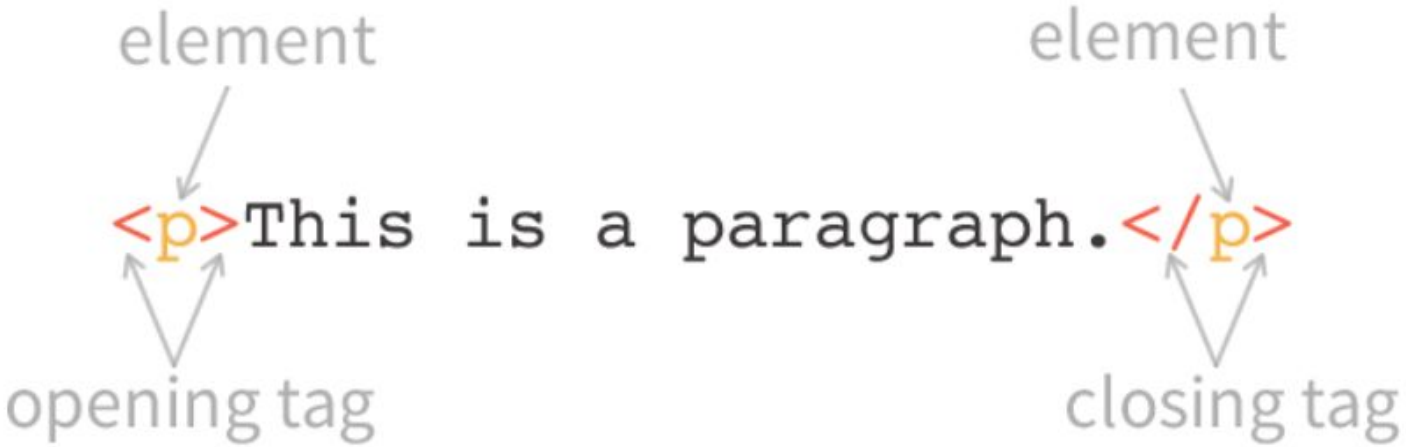
# Producción de información sobre conflictos

- Uso de fuentes hemerográficas
- Problemas de validez
  - Teóricos - Operacionalización
  - Sesgos en la fuente (omisión, completitud, etc.)
    - **Sesgo geográfico => diarios nacionales => subregistro de hechos en unidades subnacionales**
- Problemas de confiabilidad
  - **Problemas de codificación**

- **Resultado => diferencias notables**

# Webscraping

- Literalmente: ‘rascar de la web’
- Técnicas para extraer del código html de un sitio información
- Caso clásico: webscraping de noticias
- Código html => Elementos, tags, documentos
- Fácil escribir un programa que pueda extraer texto



div.title 375.2 × 476.4

Quinta huelga contra Mauricio Macri

# Sin transporte, el paro mostró una alta adhesión y para el Gobierno tuvo fines electorales

Fueron clave la falta de colectivos, trenes y subte. La CGT calificó la medida de contundente. Dante Sica acusó a los sindicalistas de pensar en la carrera electoral.



Elements Console Sources Network Performance Memory Application Security Audits

```
<!doctype html>
<html class="no-js" lang="es" xmlns="https://www.w3.org/1999/xhtml">
▶ <head>…</head>
▼ <body data-adspath="clarin/politica/nota" data-adskv="{"section":["politica"],"kw_referer":["https://www.clarin.com/tema/paro-general.html"],"kw_ck":"-2","kw_pw":["0"],"kw_temperatura":[19],"kw_pw_counter":[0],"kw_time":[1573341563228],"kw_choque":["0"],"kw_width":[1536],"kw_height":[864],"kw_keywords":["transporte","paro","mostro","alta","adhesion","gobierno","fines","electorales"],"kw_tags":["paro-general","cgt","hector-daer","fotogalerias"],"kw_login_edad":[],"kw_login_sexo":[]}"> == $0
  ▶ <div id="fb-root" class=" fb_reset">…</div>
  ▶ <noscript>…</noscript>
  ▼ <div class="main-menu off-canvas-wrap" data-offcanvas>
    ▼ <div class="inner-wrap">
      ▼ <section class="main-section">
          <div class="pase_login header-logged-back" style="display: none;"></div>
        ▼ <div data-role="content" class="mainPage ">
          ▶ <header id="header-interior" style="padding-top: 45px;">…</header>
          ▶ <div style="text-align: -webkit-center;">…</div>
          ▶ <div id="flotante3" class="ad-slot" data-adtype="flotante3" style="height: 0px;">…</div>
          ▶ <div id="horizontal1" class="ad-slot" data-adtype="horizontal1" style>…</div>
            <link href="https://static.clarin.com/contents/news/css/news.css?05c7a42" type="text/css" rel="stylesheet">
            <div id="mostSignificantTag" data-value="CGT" data-value-parse="CGT"></div>
          ▶ <div class="container-fluid headTop">…</div>
            <link href="https://static.clarin.com/contents/news/css/news.normal.css?05c7a42" type="text/css" rel="stylesheet">
          ▼ <div class="news container newsNormal no-p stickyBar politica nota-unica" data-content-id="hFFQKjrM1">
              ::before
            ▼ <div class="contentFlex">
              ▼ <div class="entry-title col-lg-6 col-md-12 col-sm-12 col-xs-12">
                  <p class="volanta">Quinta huelga contra Mauricio Macri</p>
                ▶ <h1 id="title">…</h1>
                  <meta itemprop="headline" content="Sin transporte, el paro mostró una alta adhesión y para el Gobierno tuvo fines electorales">
                ▼ <div itemprop="description" class="bajada">
                  ▼ <h2>
                      "Fueron clave la falta de colectivos, trenes y subte. La CGT calificó la medida de contundente. Dante Sica acusó a los sindicalistas de pensar en la carrera electoral."
                      <p></p>
                    </h2>
                  </div>
                </div>
              </div>
            ▶ <div id="galeria-trigger" class="entry-media no-p main-image pointer col-lg-12 col-md-12 ">…</div>
            ▶ <script type="text/javascript">…</script>
            </div>
          ▶ <div class="interior-1 page">…</div>
          ▼ <div class="interior-1 page bottom-page">
            ▶ <script type="text/javascript">…</script>
            ▶ <div id="taboola-below-article-thumbnails" class=" trc_related_container trc_spotlight_widget trc_elastic trc_elastic_thumbnails-b " data-placement-name="Below Article Thumbnails" observeid="tbl-observe-0">…</div>
            ▶ <script type="text/javascript">…</script>
```

# Procesamiento de Lenguaje Natural

- Sub campo de la inteligencia artificial
- Técnicas computacionales para procesar y analizar grandes cantidades de texto
- Texto como dato
- Problemas abordables
  - Traducciones automáticas
  - Detección y modelado automático de tópicos
  - Clasificación de textos
  - Sentiment Analysis,
  - Etc.

# Etapas fundamentales del procesamiento

1. Vectorización de texto
2. Eliminación de stopwords
3. Normalización
4. Ponderación de la matriz de términos

| TANGO | Bolivia | contra | de | defensa | en | Evo | huelga | Macri | nueva | protestas |
|---|---|---|---|---|---|---|---|---|---|---|
| Nueva huelga en contra de Macri | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| Protestas en Bolivia en defensa de Evo | 1 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 1 |

# Clasificación - Aplicaciones

**Automatización de procesos para la construcción de bases de datos de protestas**

[Hannah, 2017]

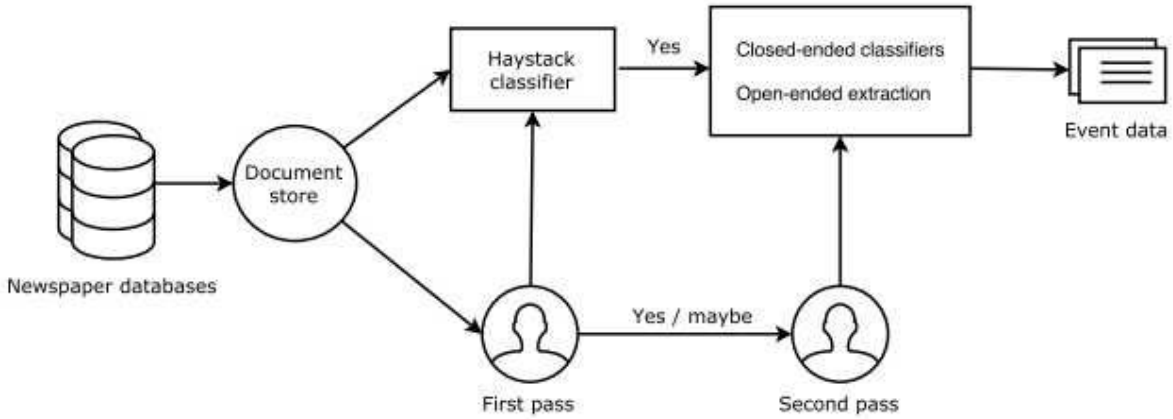



Figure 1: MPEDS pipeline with training.

# Herramientas Python y R

Beautiful Soup

scikit learn

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

gensim

topic modelling for humans

tidyverse

rvest

TIDYTEXT

StanfordNLP

pypi v0.2.0 | python 3.6 | 3.7

Overview

Usage

Pipeline

Resources

StanfordNLP 0.2.0 - Python NLP Library for Many Human Languages

Table of Contents

- About
- Get Started
- License
- Citing StanfordNLP in papers
- Links

# Comentarios finales

- Problemas de validez
  - Aumento de la cobertura de medios nacionales y locales

Web scraping

- Problemas de confiabilidad
  - Automatización de selección de noticias sobre conflictos
  - Automatización de procesos de codificación

NLP, topic modeling, clasificación, etc.

- Problemas de caracterización de fuentes

# ¿Preguntas?

german.rosati@gmail.com

https://gefero.github.io/