Trabajo Práctico 1: NLP + Clarín

Maximiliano Lombardia - Legajo: 56276

Motivación detrás del corpus

Motivación detrás del corpus

- Disparadores:
  - ¿Qué información consumimos?
  - ¿Es adecuada a corto y largo plazo?
  - ¿Cuál será su grado de relevancia a lo largo del tiempo?

- Preguntas formuladas:
  - ¿Será posible conseguir “La tapa de la época”?
  - ¿Existirán patrones en la forma en la que se desenvuelve la historia?
  - ¿Habrá eventos (o personalidades) que marcarán una época?

Obtención del corpus

Obtención del corpus

- Primer paso: Obtención de imágenes
  - Script propio con librerías de Python
- Segundo paso: Proceso de OCR
  - Tesseract OCR + Pytesseract
  - Dividido en 4 lotes de ~20 años c/u
  - Para los últimos 20 años, 20 archivos extra de 1 año c/u
- Tercer paso: Análisis sobre lo obtenido
  - Cloud of Words
  - Referencia cruzada con el dataset

OCR y Tesseract

OCR y Tesseract

- OCR:
  - Conversión de imágenes de texto a texto codificado por la máquina
  - Existen técnicas de pre y post-procesamiento
  - Para reconocimiento de caracteres usa técnicas de reconocimiento de patrones (pattern matching) y extracción de parámetros (feature extraction)

- Tesseract:
  - Software de Google de libre acceso, actualmente en la versión 5
  - Últimas versiones cuentan con implementación de redes neuronales
  - Dos parámetros importantes
    - oem: Engine Mode, sirve para especificar qué motor de Tesseract usar. Recomendable usar 1
    - psm: Page Segmentation Mode, sirve para especificar qué info se le está enviando al motor y de qué forma debería procesar esa data.

Análisis del corpus

Cloud of Words

28/8/1945 - 28/8/1962

28/8/1962 - 28/8/1982

Cloud of Words

28/8/1982 - 28/8/2002

28/8/2002 - 28/8/2022

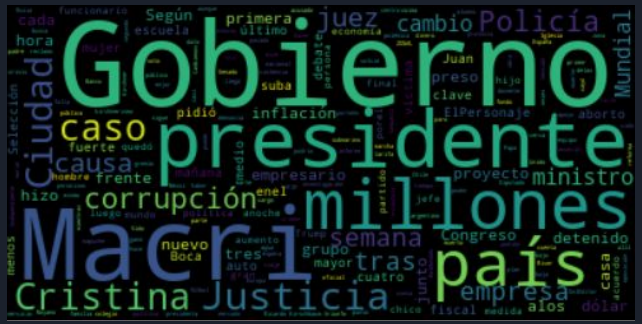
Cloud of Words - “Anatomía de un período”

Clarín en época de Mundial…

2006

2010

2014



2018

Cloud of Words - “Anatomía de un período”

2019-2020

2020-2021

Un evento que modificó tantos aspectos de nuestras vidas…

Algunas conclusiones finales

Algunas conclusiones finales

- Respecto al proceso:
  - Importante tener buena calidad en las imágenes
  - Experimentar con las opciones del procesador (en caso de que existan)
  - Prepararse para sumar muchas stopwords

- Respecto al corpus:
  - Notable diferencia entre impacto del momento vs. impacto a largo plazo
  - No resulta idóneo para cuestiones como “La tapa de la época”
  - Ideal para cuestiones que se pueda contemplar la información como un todo

¡Muchas gracias!