# ¿De qué se habla en la correspondencia de Jovellanos?

## Análisis automático de textos adaptado al castellano del siglo XVIII

**Oriol Invernón Llaneza**

# 1.

# Justificación y objetivos

# Justificación del proyecto

**Información**

Transmitir la información recogida en la correspondencia de un personaje histórico como Gaspar Melchor de Jovellanos, sin que sea necesario leer las miles cartas que la componen.

**Campo de estudio**

Expande un campo multidisciplinar que resulta ampliamente beneficioso de cara a educación y cultura.

# Objetivos del proyecto

- Realizar un análisis de lenguaje natural adaptado a la correspondencia de Jovellanos.
- Identificar las temáticas presentes utilizando métodos de aprendizaje no supervisado.
- Generar visualizaciones de los resultados.
- Diseñar una web para presentar los datos.

# Republic Of Letters

Proyecto de carácter similar desarrollado por la Universidad de Stanford.

# 2.

# Planificación del proyecto

El desarrollo del proyecto se ha dividido en el análisis y la página web

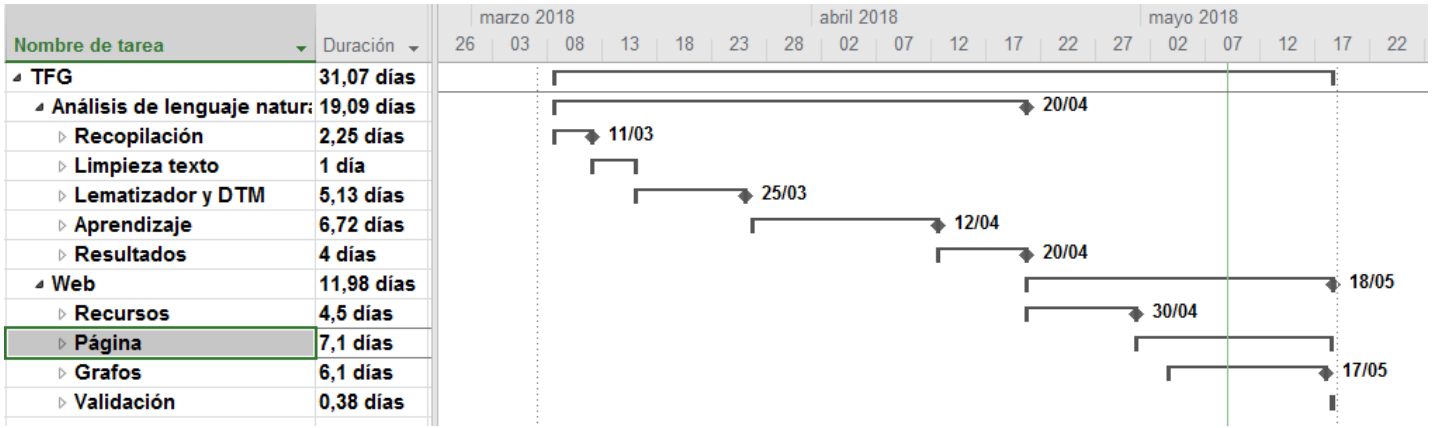
# Análisis del lenguaje

1. Recopilación de texto.
2. Lematizador.
3. Aprendizaje.
4. Resultados.

# Desarrollo de la web

1. Recursos.
2. Cuerpo de la página.
3. Grafos y funciones.
4. Validación.

# Diagrama de Gantt



# 3.

# Presupuesto del proyecto

# Suposiciones

Se han supuesto tres roles:

- Desarrollador experto en aprendizaje.
- Desarrollador web (front-end).
- Historiador.

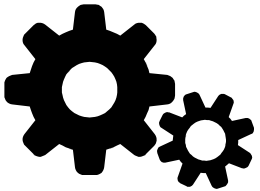
# Presupuesto cliente

- Análisis de lenguaje: 7.625,13 €
- Desarrollo web: 1.885,66 €
- Total (con impuestos): 11.508,06 €

# 4.

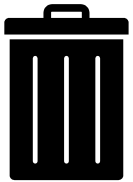
# Análisis de lenguaje


# Herramientas y lenguajes



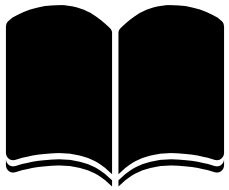
# El proceso

# Limpieza del texto



- Paso a minúsculas.
- Eliminación de números.
- Eliminación de puntuación y otros caracteres problemáticos.
- Eliminación palabras vacías.
- Eliminación espacios innecesarios.

# Lematizador



Por cada palabra:

1. ¿Es común?
2. ¿Aparece en el diccionario?
3. ¿Aparece alguna palabra con el mismo lexema en el diccionario?
4. Consulta a GRAMPAL.

# Document Term Matrix

| | Abad | Abandonar | ... |
|---|---|---|---|
| 1 | **2** | **1** | **...** |
| 2 | **0** | **1** | **...** |
| 3 | **0** | **0** | **...** |
| ... | **...** | **...** | **...** |

# 2.133 cartas

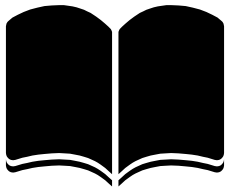
Filas

# 14.593 palabras

Columnas

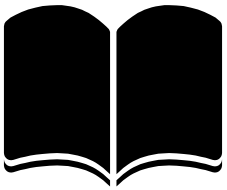
# 489 corresponsales

# Métodos aprendizaje



1. K-means.
2. Densidad (DBSCAN).
3. Jerárquicos (Agnes y Diana).
4. TopicModeling.

# K-means



**Funcionamiento**

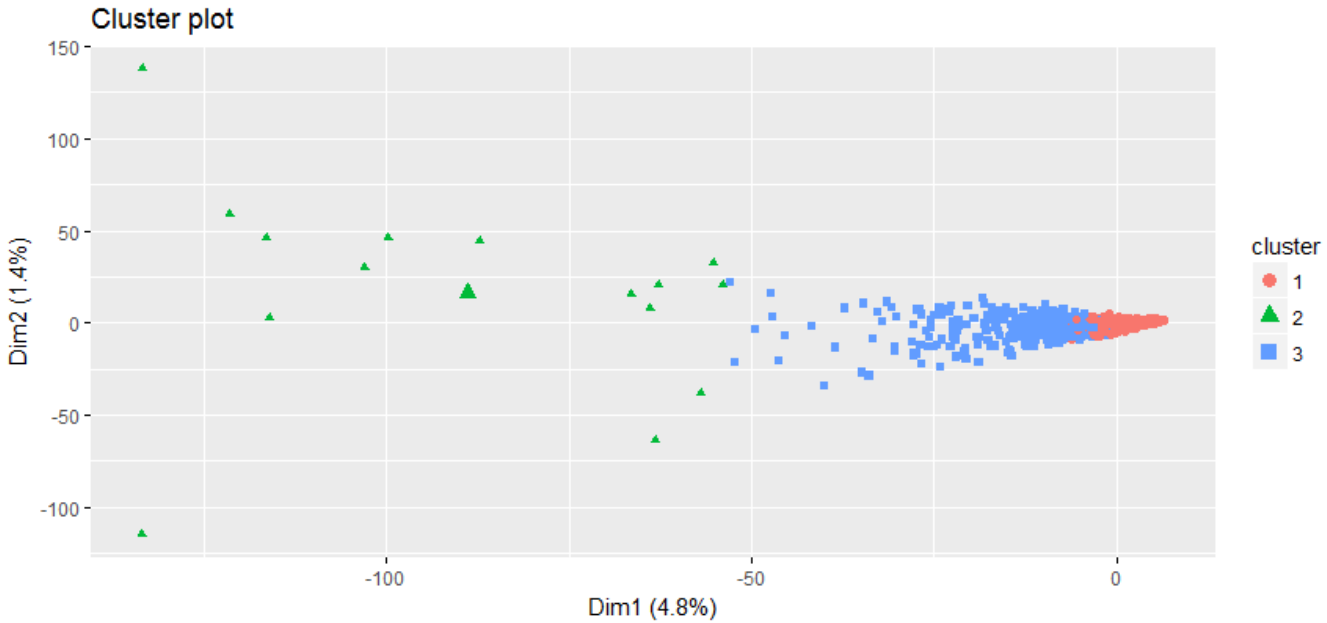
1. Estimar *K* centroides.
2. Asignar cada punto a un clúster.
3. Actualizar centroides.
4. Repetir 2 y 3 hasta criterio.

**Requisitos**

Obtener el valor óptimo de *K* previamente, mediante *elbow check* o índices de validación.

**Validación**

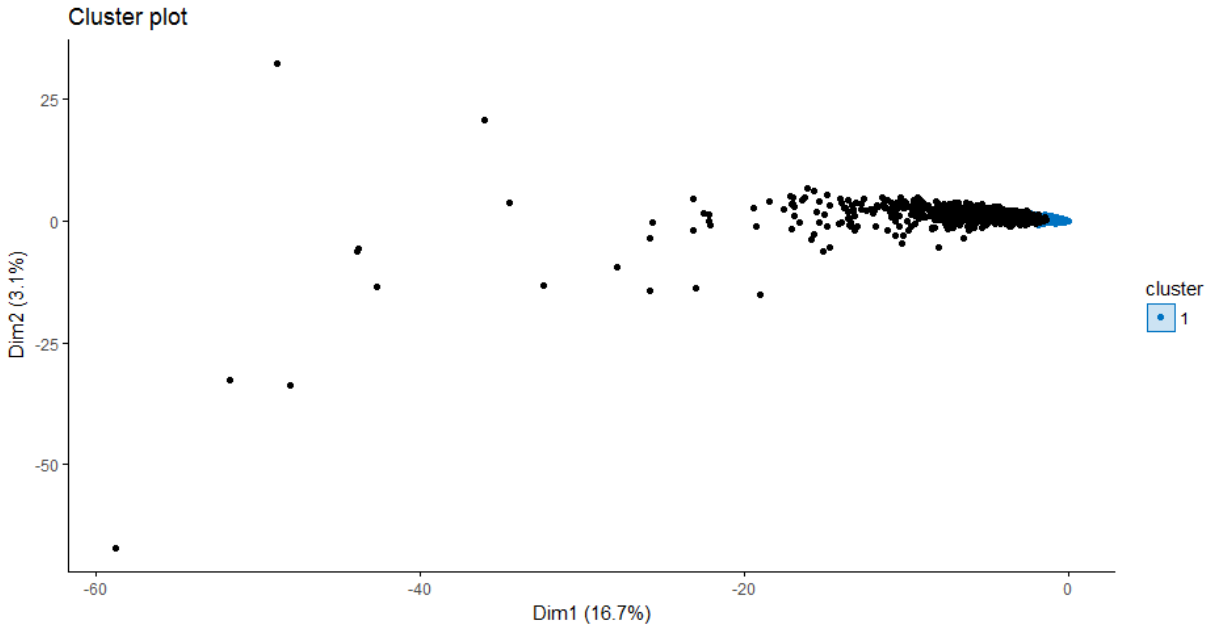
Validación de clústeres realizada mediante *silhouette*.




**Resultado K-means**

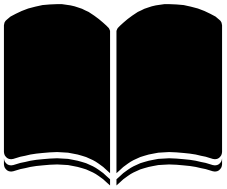
# Validación K-means

| | Agrupamiento 1 | Agrupamiento 2 | Agrupamiento 3 |
|---|---|---|---|
| Nº observaciones | **1.656** | **15** | **462** |
| Valor *silhouette* | **-4,808** | **4,455** | **-3,322** |




# Resultado DBSCAN

# Agnes



**Funcionamiento**

Algoritmo jerárquico aglomerativo.

1. Asigna dato a clúster.
2. Calcula distancias entre clústeres.
3. Junta los más similares.

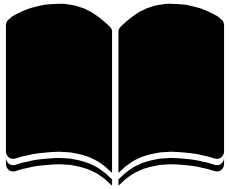
**Requisitos**

1. Seleccionar el método para calcular la distancia: singular, completo, medio, centroide o Ward.
2. Cortar el árbol para el número óptimo de clústeres.

**Validación**

Validación de clústeres realizada mediante *silhoutte*.

# Selección de método

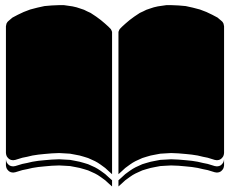


| | silhouette | frey | ball | Pseudot2 | cindex | Media |
|---|---|---|---|---|---|---|
| Singular | **3** | **-** | **3** | **3** | **2** | **3** |
| Completo | **2** | **8** | **3** | **2** | **2** | **4** |
| Medio | **3** | **-** | **3** | **2** | **2** | **3** |
| Centroide | **2** | **-** | **3** | **2** | **2** | **2** |
| Ward | **2** | **-** | **3** | **9** | **9** | **6** |

# Validación Agnes

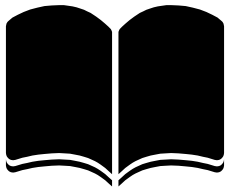
| | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| Nº observaciones | **1.402** | **560** | **162** | **7** | **1** | **1** |
| Valor *silhouette* | **0,4976** | **-0,2182** | **-0,240** | **-0,1707** | **0** | **0** |

# TopicModeling



- Tipo de minería de texto.
- Identifica temas (*topics*) en colecciones de documentos.
- Un tema es un patrón recurrente de palabras concurrentes.
- Algoritmo utilizado → LDA.

# TopicModeling



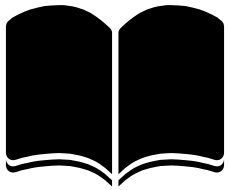
**Requisitos**

Número de temas que aparecen (igual que el valor de *K* en K-means).

**Validación**

Comprobación “manual” de la información obtenida.

# *Latent Dirichlet Allocation*



**Consideraciones básicas**

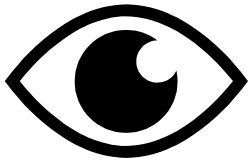
1. Cada documento es una mezcla de temas presentes en toda la colección.
2. Cada palabra es asignable a, como mínimo, uno de esos temas.

**Creación de un documento según LDA**

1. Se decide el número de palabras que tendrá el documento.
2. Se elige una mezcla de temas cada uno con una probabilidad.
3. Para cada palabra:
    1. Se escoge un tema.
    2. Se genera la palabra en base a la distribución de palabras del tema.


1
ver
año
real
servir
consejo
instituto
establecer
orden
guardar
informar
0.00
0.01
0.02
0.03
2
decir
hacer
poder
querer
conocer
deber
coser
pasar
pensar
tratar
0.000
0.005
0.010
0.015
0.020
0.025
3
perder
fechar
diario
citar
carta
enviar
referir
extractar
fuerte
papel
0.00
0.02
0.04
0.06
4
carta
ver
deseo
amar
querer
hermano
dios
esperar
venir
volver
0.000
0.005
0.010
0.015
5
nombrar
copiar
obrar
noticiar
hallar
san
libro
historiar
tomar
santo
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
6
ver
obrar
idear
escribir
gusto
tomar
verso
materia
sentir
amor
0.000
0.005
0.010
7
juntar
cortar
gobierno
pueblo
hombre
escribir
españa
ejército
partir
enemigo
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
term
beta

# Temas identificados



**Jovellanos, el político**

Cartas de carácter institucional y formal. Tratan con frecuencia sobre reales decretos y órdenes, además del Instituto Asturiano.

**Círculo cercano, tono formal**

Tema dominado por conversaciones de un cierto carácter formal con su círculo cercano.

**Cartas perdidas**

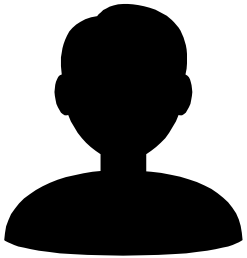
Correspondencia cuya existencia es conocida pero carecemos de su contenido.

*Excelentísimo señor:*

*Cumpliendo con la Real Orden que V.E. se ha servido comunicarme con fecha de 7 de setiembre del año anterior, paso a sus manos el adjunto Informe a S.M. sobre la Representación que en 30 de abril del mismo había dirigido a S. R. P. el Director General de Minas, don Francisco de Angulo. El deseo de tomar una plena instrucción del objeto que trata, me ha hecho suspenderle hasta ahora, que con esta misma fecha dirijo a S.M. las resultas de mi principal comisión por la vía reservada de Marina, de quien dimana.*

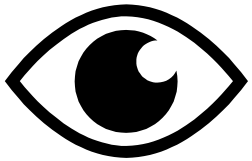
*Nuestro Señor guarde a V.E. muchos años.*

# Corresponsales tema 2



| Personaje | Cartas |
|---|---|
| González de Posada | **30** |
| Juan Meléndez Valdés | **10** |
| Lord Holland | **8** |
| Francisco de Paula Jovellanos | **7** |
| Gertrudis del Busto | **6** |
| Josefa Jovellanos | **4** |

# Temas identificados



**Círculo cercano, tono íntimo**

Conversaciones en tono cariñoso e íntimo con, principalmente, sus mejores amigos y miembros de su familia.

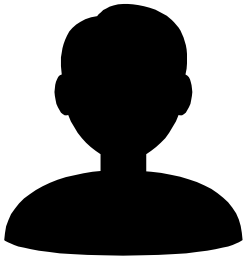
**Jovellanos, el recopilador**

Cartas intercambiadas durante la década que pasó visitando monasterios e iglesias del norte de España y copiando escrituras antiguas que encontraba.

**Jovellanos, el poeta**

Correspondencia en la que trata sobre poesía y otras obras literarias. Incluye consejos y revisiones del trabajo de sus conocidos, recomendaciones de obras, etc...

# Corresponsales tema 4



| Personaje | Cartas |
|---|---|
| Lord Holland | **>35** |
| Josefa Jovellanos | **42** |
| González de Posada | **39** |
| Baltasar González de Cienfuegos | **16** |
| Catalina de Sena | **15** |

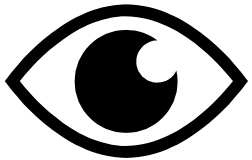
*Muy señor mío y mi más estimado dueño:*

*Quedará esta tarde efectuado el andamio que V.S. se sirve encargarme, y mañana haré copiar la consabida inscripción, [...] de otras inscripciones de esta nación. Esta misma tarde paso a copiar la inscripción que dije a V.S. de Corao, pues la otra está bien cerca de Santa Cruz, en una casería, y es regular quiera verla V.S. más en su original que en copia, bien que elegirá lo que guste, pues la tendrá también. Ambas son sepulcrales, y ésta es muy parecida a otra que halló Sandoval junto a Burgos, [...] A la hora que V.S. me dice procuraré hallarme en Santa Cruz, [...] su más seguro servidor Josef Antonio Ru.*

*Muy señor mío y mi estimado paisano: Doy a usted muy finas y sinceras gracias por el romance [...] el entusiasmo poético arrebataron su imaginación de usted y colocaron sus héroes entre los signos del Zodíaco; [...] atribuir a los colores de la poesía, ya sabe usted que la poesía didáctica no concede tantas licencias. Pero si considero el romance como poeta, hallo en él mil gracias: muchos pensamientos sublimes y brillantes, muchos versos correctos y armoniosos, algunas ideas originales, y sobre todo un estilo fácil, noble y de bastante majestad. Seguramente usted podrí] hacer grandes cosas en poesía, [...].*

*Gaspar de Jovellanos.*

# Temas identificados



**Jovellanos durante la guerra:**

Contiene la correspondencia intercambiada durante la Guerra de Independencia, la mayor parte con Lord Holland. Destaca el número de menciones a las Cortes de Cádiz y la Junta Suprema Central.

*Mi muy amado Lord:*

*Por fin usted llegó a Lisboa, como dije, antes que mis cartas a su mano, y así me lo confirma la del 5, escrita de Badajoz, que me entregó el calesero. [...] ataque, y tal vez a esta hora estarán empeñados en él nuestros ejércitos. Hemos entrevisto a medias el plan, y aunque no entiendo la materia, me gusta poco. No me parece que hay bastante unión en los tres cuerpos, que deben obrar a mucha distancia contra un enemigo reunido. [...]. Es el día de buen hado; hoy hemos celebrado en la capilla de San Fernando la batalla de Bailén. Asistieron el nuncio y los ministros de Inglaterra, Austria, Portugal y Provincias Unidas. Capmany está ya libre de la Gaceta y agregado a los trabajos de Cortes. Pero nos ocupan demasiado los negocios de la guerra y el temor de sus resultas; si malas, al pueblo, si buenas, al general victorioso. Amable Milady: me llama la hora de la Junta nocturna. [...]*

# 5.

# Desarrollo web





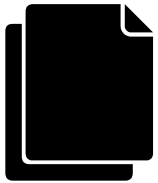
# Herramientas y lenguajes

# Alcance

El alcance del desarrollo comprende:

- Creación de varios grafos dinámicos e interactivos con la información de la correspondencia y el análisis de la misma.
- Generación de una galería con gráficas sobre el resultado del análisis.

# Archivos



**BBDD / Datos entrada**

Archivos JSON con la información de la correspondencia (corresponsales, nº cartas, temas, ...). Leídos dinámicamente por D3.js para generar los grafos.

**JavaScript**

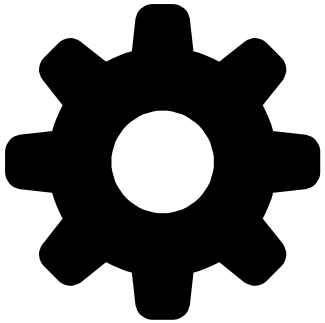
Documentos JS con las diferentes funciones de los grafos. Incluye: búsqueda de personajes, consultas a Wikipedia, filtros aplicables al grafo, etc...

**HTML y CSS**

Estilo realizado con Bootstrap y librerías auxiliares.

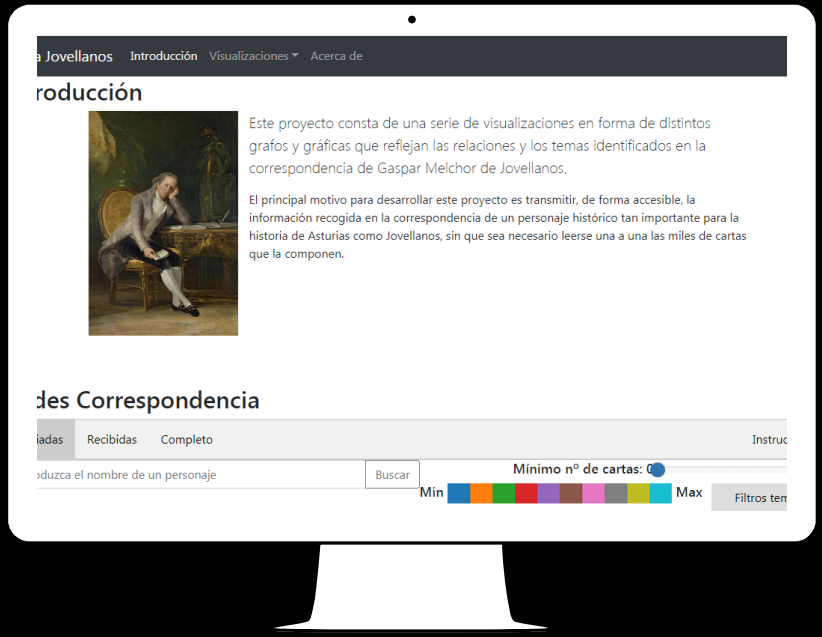
Se ha conseguido cumplir con el nivel 2.0 (AA) de WCAG.

# Funcionalidades Grafo



| |
|---|
| Selección de nodos (múltiple y sencila) |
| Movimiento de nodos. |
| Consulta a Wikipedia sobre el personaje seleccionado |
| Búsqueda de personajes |
| Filtrar por cantidad de cartas intercambiadas |
| Filtrar por sexo |
| Filtrar por los temas identificados durante el análisis |
| Destacar nodos |
| Zoom |

# Demo


Jovellanos
Introducción
Visualizaciones
Acerca de
roducción
Este proyecto consta de una serie de visualizaciones en forma de distintos grafos y gráficas que reflejan las relaciones y los temas identificados en la correspondencia de Gaspar Melchor de Jovellanos.
El principal motivo para desarrollar este proyecto es transmitir, de forma accesible, la información recogida en la correspondencia de un personaje histórico tan importante para la historia de Asturias como Jovellanos, sin que sea necesario leerse una a una las miles de cartas que la componen.
des Correspondencia
iadas
Recibidas
Completo
Instruc
oduzca el nombre de un personaje
Buscar
Mínimo nº de cartas:
Min
Max
Filtros ten

# 6.

# Conclusiones y Ampliaciones

# Problemas encontrados

- Herramientas de minería de texto poco desarrolladas o inexistentes en castellano.
- Restricciones del equipamiento.
- Trabajo con herramientas y lenguajes sin conocimiento previo.

# Conclusiones

Tras solventar los distintos problemas y realizar las pruebas, se ha considerado que los objetivos han sido cumplidos y el resultado ha sido el esperado.

# Ampliaciones

- Mejorar el lematizador.
- Actualizar librerías.
- Construir una red social completa.
- Desarrollar una herramienta.
- Traducir la página.


“Bien están los buenos pensamientos,
pero resultan tan livianos como burbuja de jabón,
si no los sigue el esfuerzo para concretarlos en acción.”