TFG: Aplicación web para clasificación de documentos de texto

Carlos Sanabria Miranda

Índice

1.
ALCANCE Y OBJETIVOS

Alcance y objetivos

- **Aplicación** que permita la **clasificación** de documentos de texto
  - Esta clasificación se realizará en base al contenido de los documentos, sin que exista una organización previa

Alcance y objetivos

- Para ello, procesa una colección de documentos, para:
  - **Identificar las temáticas (topics)** presentes en la colección
  - **Identificar la relación** de cada documento con cada topic
  - **Asignar cada documento** de la colección a un topic
- Para identificar los topics se utiliza **aprendizaje no supervisado**
  - Los datos no tienen por qué estar etiquetados
  - El número de topics óptimo se determina en base a los documentos de la colección

Alcance y objetivos

- Funcionalidades adicionales:
  - Identificar los **documentos más representativos** de cada topic
  - Identificar los **documentos** de la colección **más similares** a un nuevo documento
  - Generar un **resumen extractivo** de un documento de texto
- Implementar una **API REST** que encapsule esta funcionalidad
- Implementar un **frontend** que consuma la API REST

Alcance y objetivos

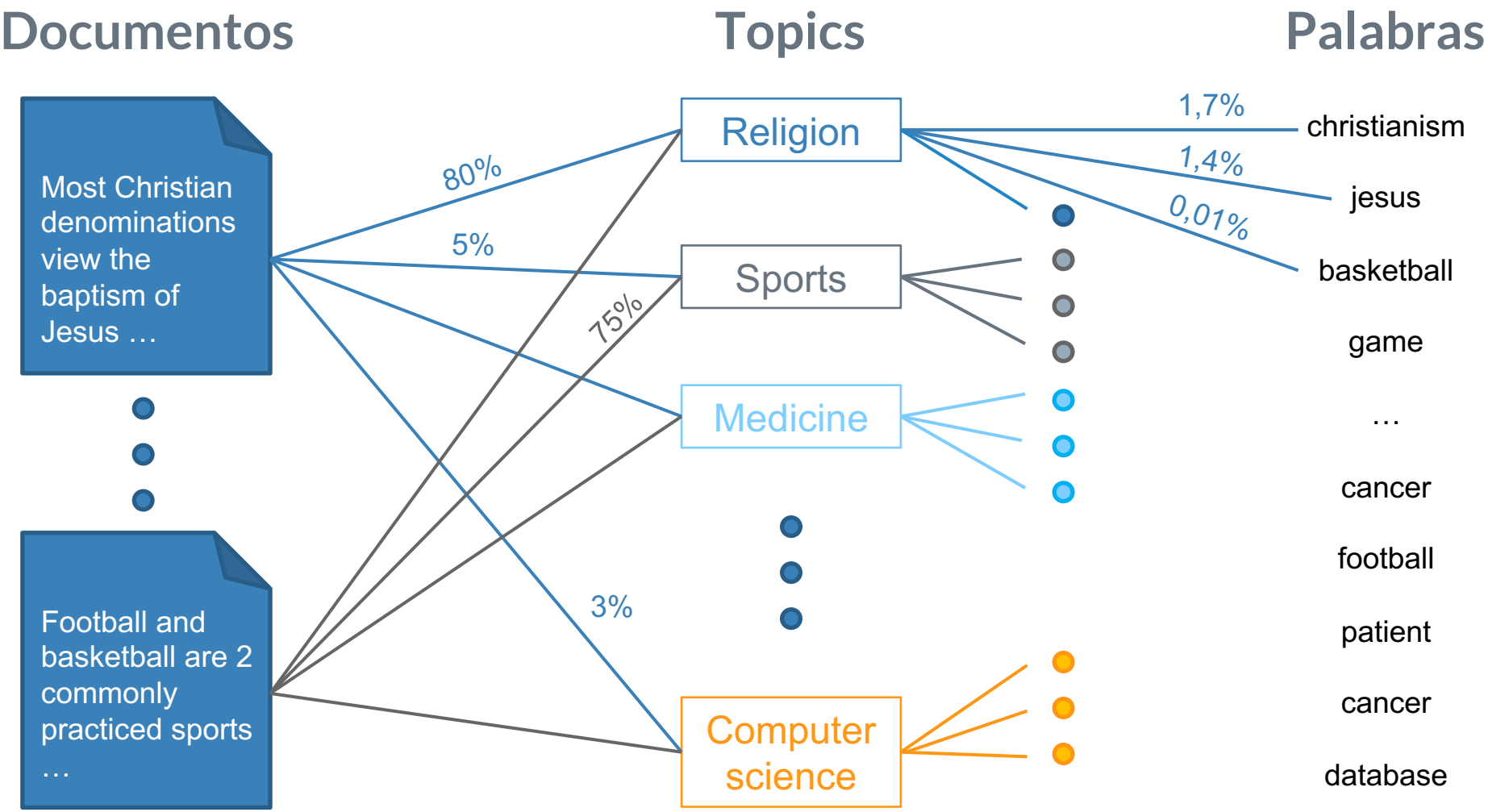
- Sus **aplicaciones** son muy **diversas**
- Ejemplos:
  - Organizar en grupos un conjunto de documentos desordenados (noticias, medicina, …)
  - Clasificar un correo electrónico en una determinada carpeta
- En general, cualquier caso de uso que involucre:
  - Documentos de texto
  - Su clasificación

2.
ASPECTOS TEÓRICOS

Aspectos teóricos: **LDA**

- Latent Diriletch Allocation
- **Algoritmo** utilizado para la **identificación** de los **topics** y la **clasificación** de los documentos
- Ideas principales:
  - **Agrupa** las **palabras** presentes en los documentos **en topics**
  - Cada **topic** se describe mediante una **distribución** de **palabras** (keywords)
  - Cada **documento** se describe mediante una **distribución** de **topics**
  - Las palabras de cada documento definen pertenencia a los topics
  - No tiene en cuenta el orden de las palabras

Aspectos teóricos: **LDA**



Aspectos teóricos: **TextRank**

- **Algoritmo** utilizado para la **generación de resúmenes**
- Basado en el algoritmo PageRank
  - Asigna puntuación numérica a las páginas web / documentos en función de su importancia

Aspectos teóricos: **TextRank**

Usando word embedding

Dividir texto en frases → Preprocesar cada frase → Transformar cada frase en un vector

I. Se transforma cada palabra de la frase en un vector
II. Se calcula la media de los vectores de las palabras de la frase

Obtener la similitud entre las frases → Transformar la matriz en un grafo → Aplicar PageRank al grafo

I. Matriz triangular
- Filas y columnas son las frases

II. Se rellena con los valores de similitud entre frases
- Similitud del coseno

- **Nodos** = Frases
- **Arcos** = Valores de similitud

Hasta convergencia

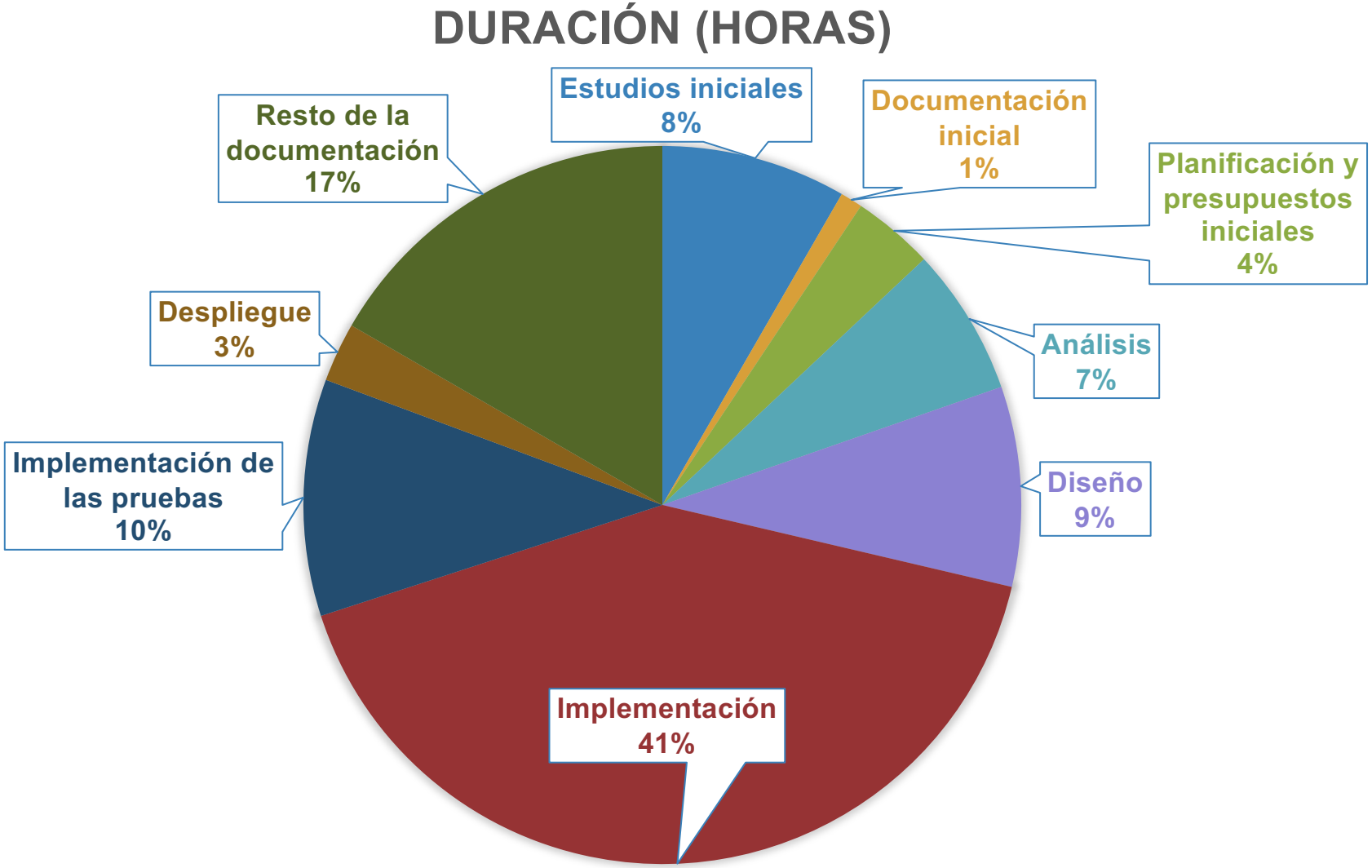
Ordenar las frases por PageRank

3.
PLANIFICACIÓN Y PRESUPUESTO INICIALES

Planificación inicial

- **01 febrero 2019 – 11 junio 2019**
- **9 etapas:**
  - Estudios iniciales
  - Documentación inicial
  - Planificación y presupuestos iniciales
  - Análisis
  - Diseño
  - Implementación
  - Implementación de las pruebas
  - Despliegue
  - Resto de la documentación
- Duración total: **300h**

Planificación inicial

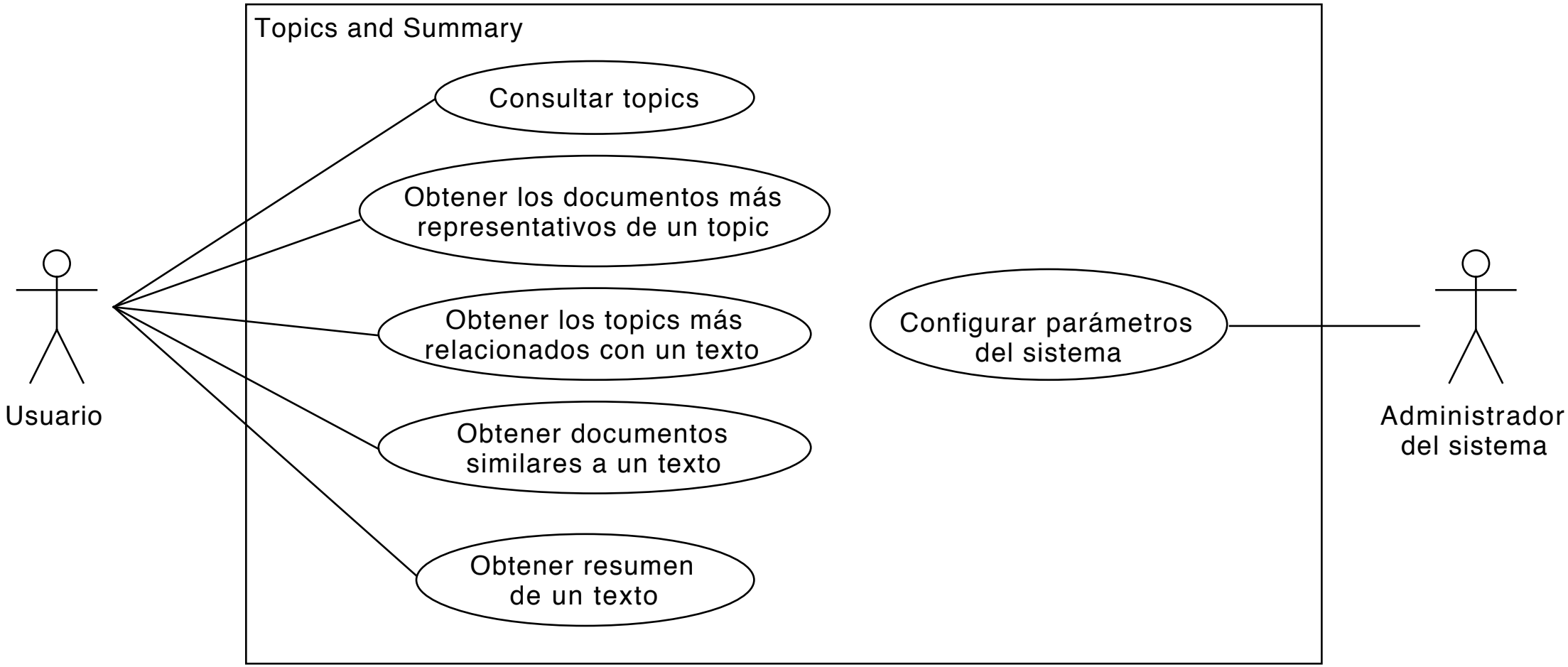


Presupuesto inicial

Presupuesto de costes inicial		
Cód.	***Partida***	***Total***
01	Estudios iniciales y planificación inicial	1.350,00 €
02	Análisis y diseño del sistema	1.762,50 €
03	Implementación y despliegue del sistema	6.248,40 €
04	Documentación	1.987,50 €
05	Otros costes	1.134,60 €
	Total Coste	**12.483,00 €**

4.
CASOS DE USO

Casos de uso


Topics and Summary
Consultar topics
Obtener los documentos más representativos de un topic
Obtener los topics más relacionados con un texto
Obtener documentos similares a un texto
Obtener resumen de un texto
Configurar parámetros del sistema
Usuario
Administrador del sistema

5.
SUBSISTEMAS, COMPONENTES Y DESPLIEGUE

Subsistemas

Topics and Summary

- Generación, análisis, comparación y utilización de **modelos** de:
  - Identificación de **topics**
  - Generación de **resúmenes**

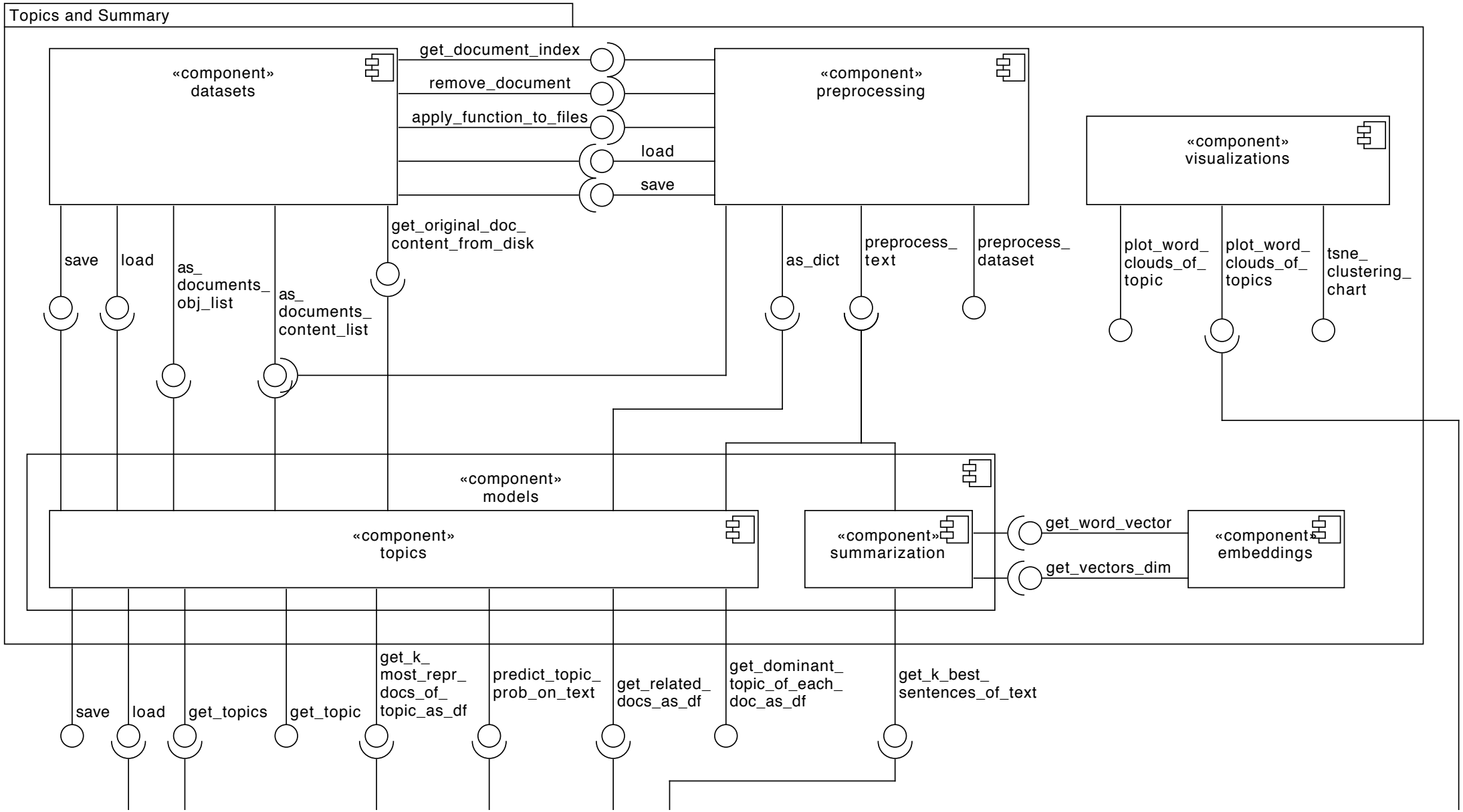
Web backend

- Implementado en **Flask**
- Expone una **API REST**

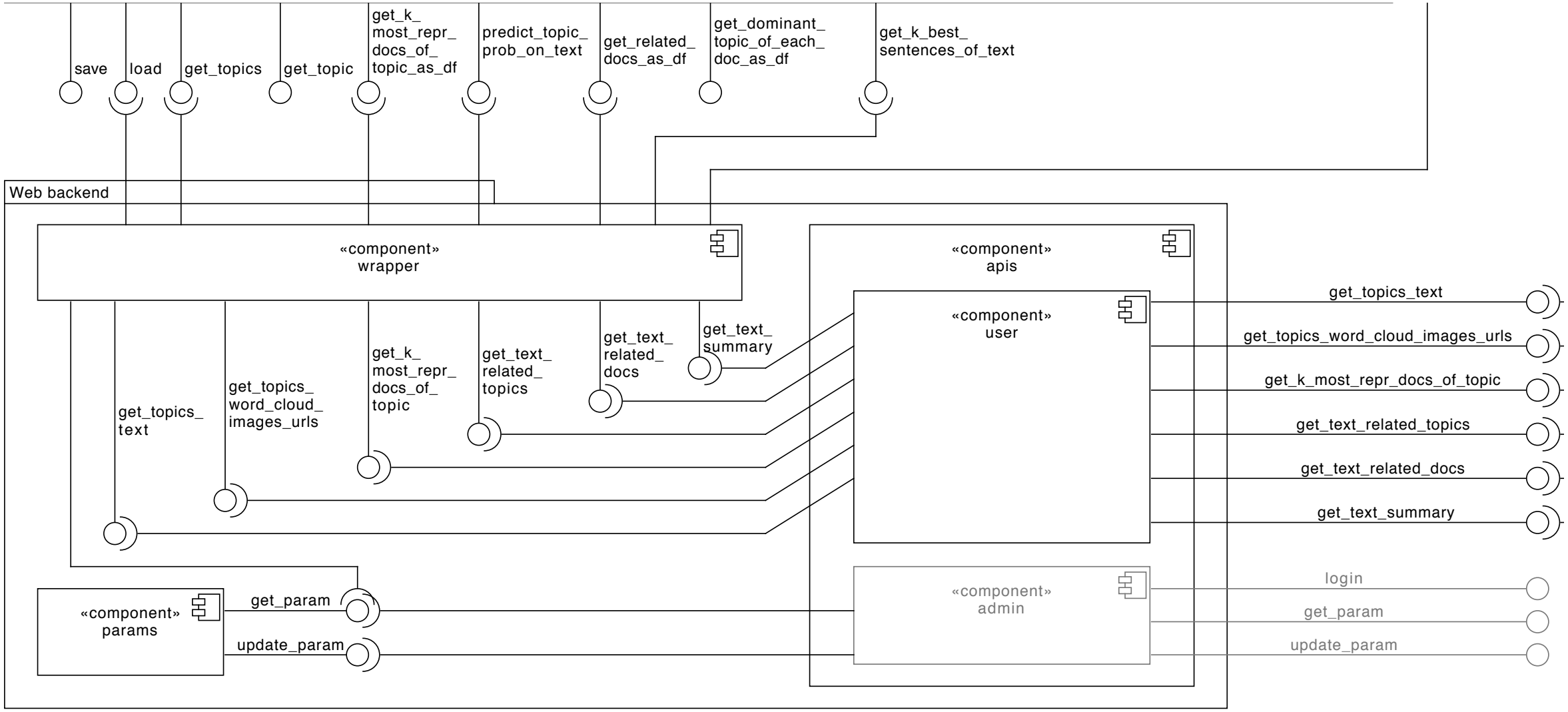
Web frontend

- Implementado en **Angular**
- Consume la **API REST**
- Permite la interacción con el sistema

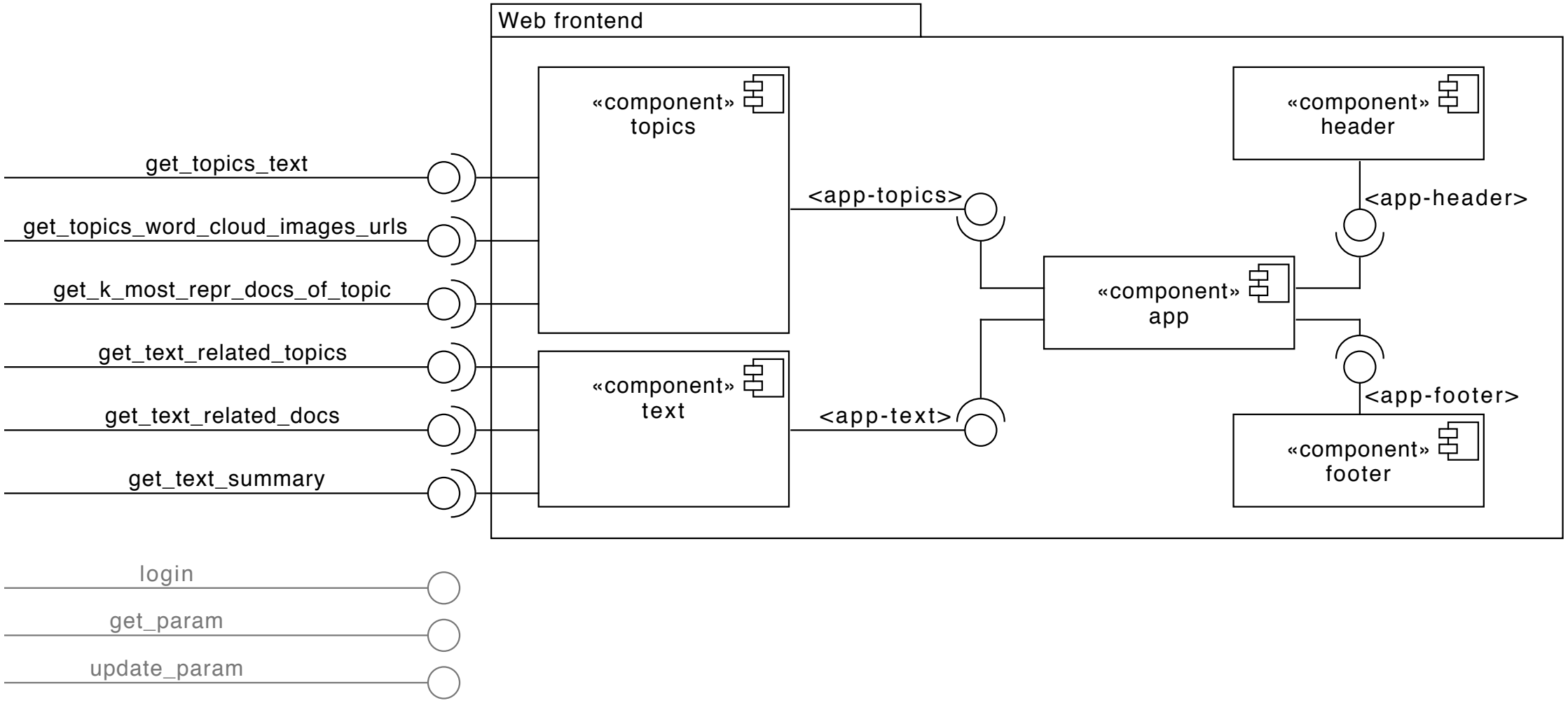
Componentes (Topics and Summary)



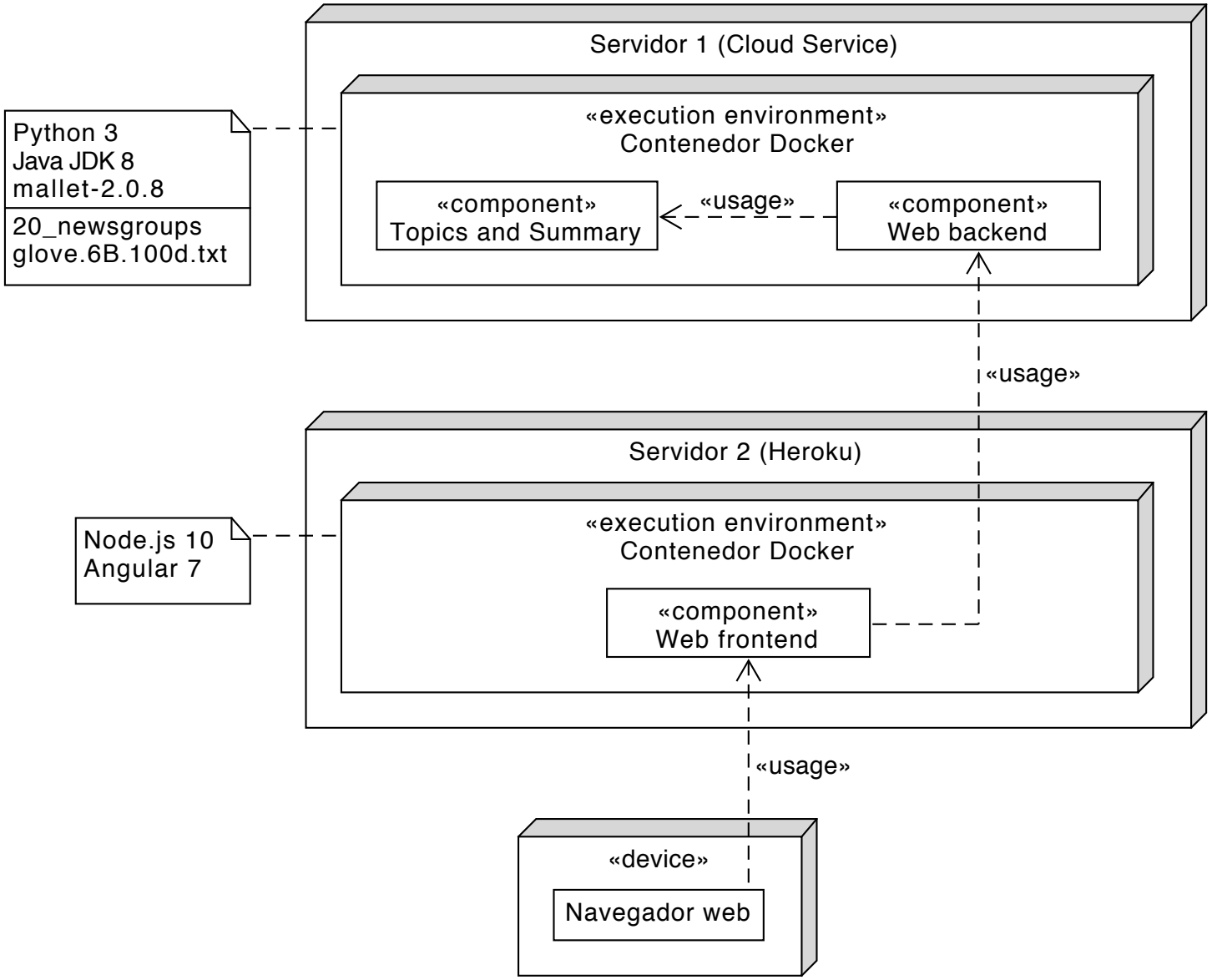
Componentes (Web backend)



Componentes (Web frontend)

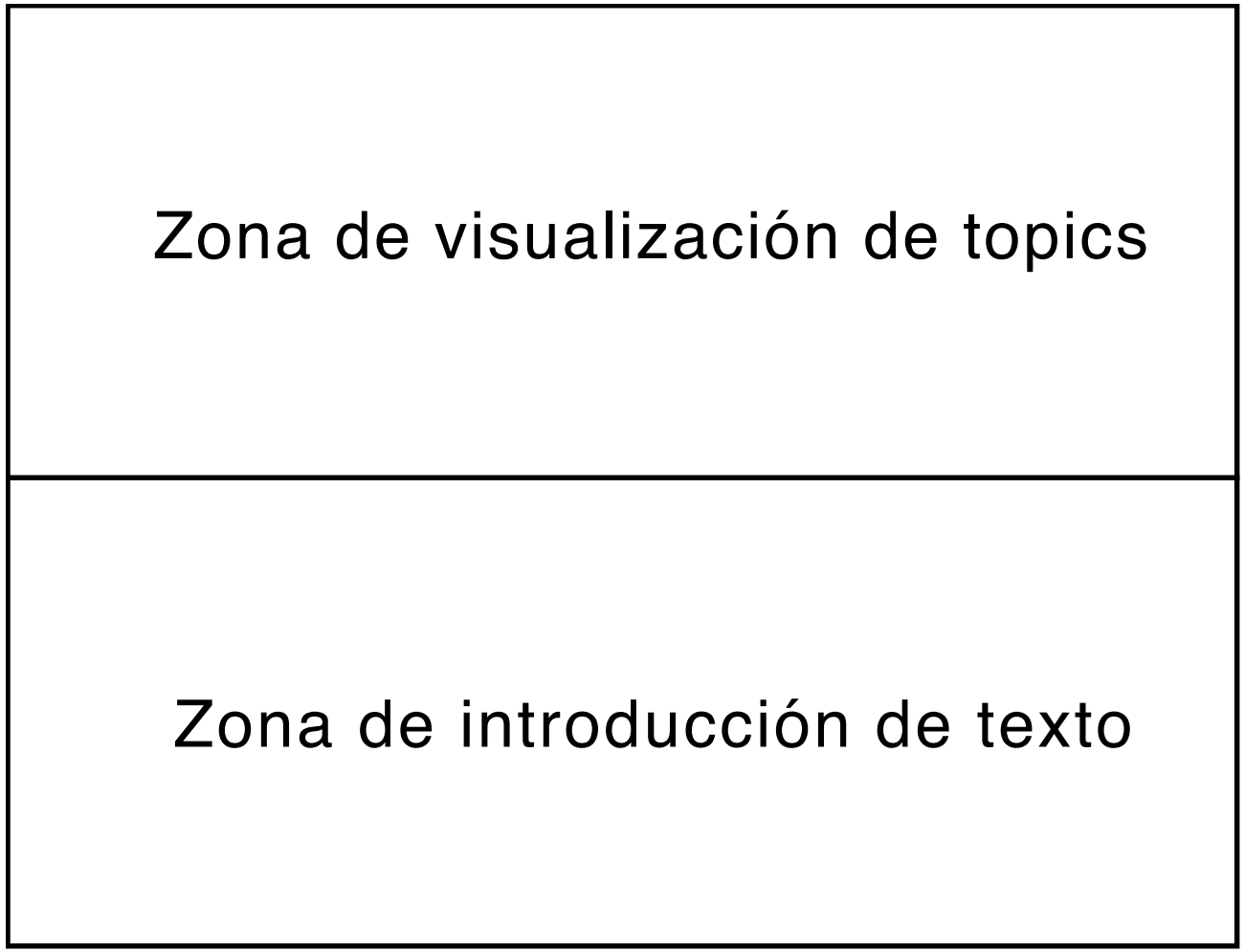


Despliegue

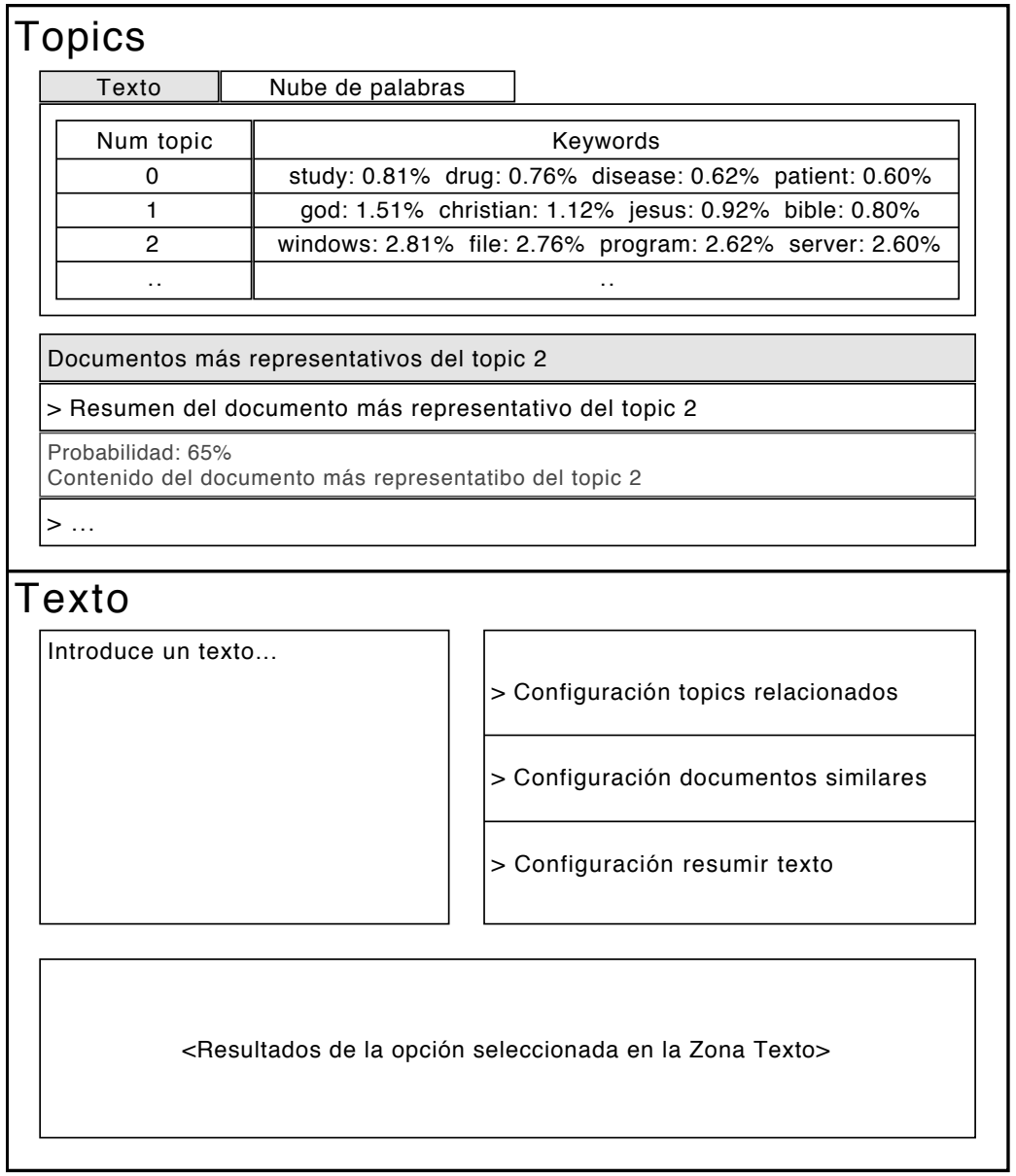

Servidor 1 (Cloud Service)
«execution environment»
Contenedor Docker
«component»
Topics and Summary
«usage»
«component»
Web backend
Python 3
Java JDK 8
mallet-2.0.8
20_newsgroups
glove.6B.100d.txt
«usage»
Servidor 2 (Heroku)
«execution environment»
Contenedor Docker
«component»
Web frontend
Node.js 10
Angular 7
«usage»
«device»
Navegador web

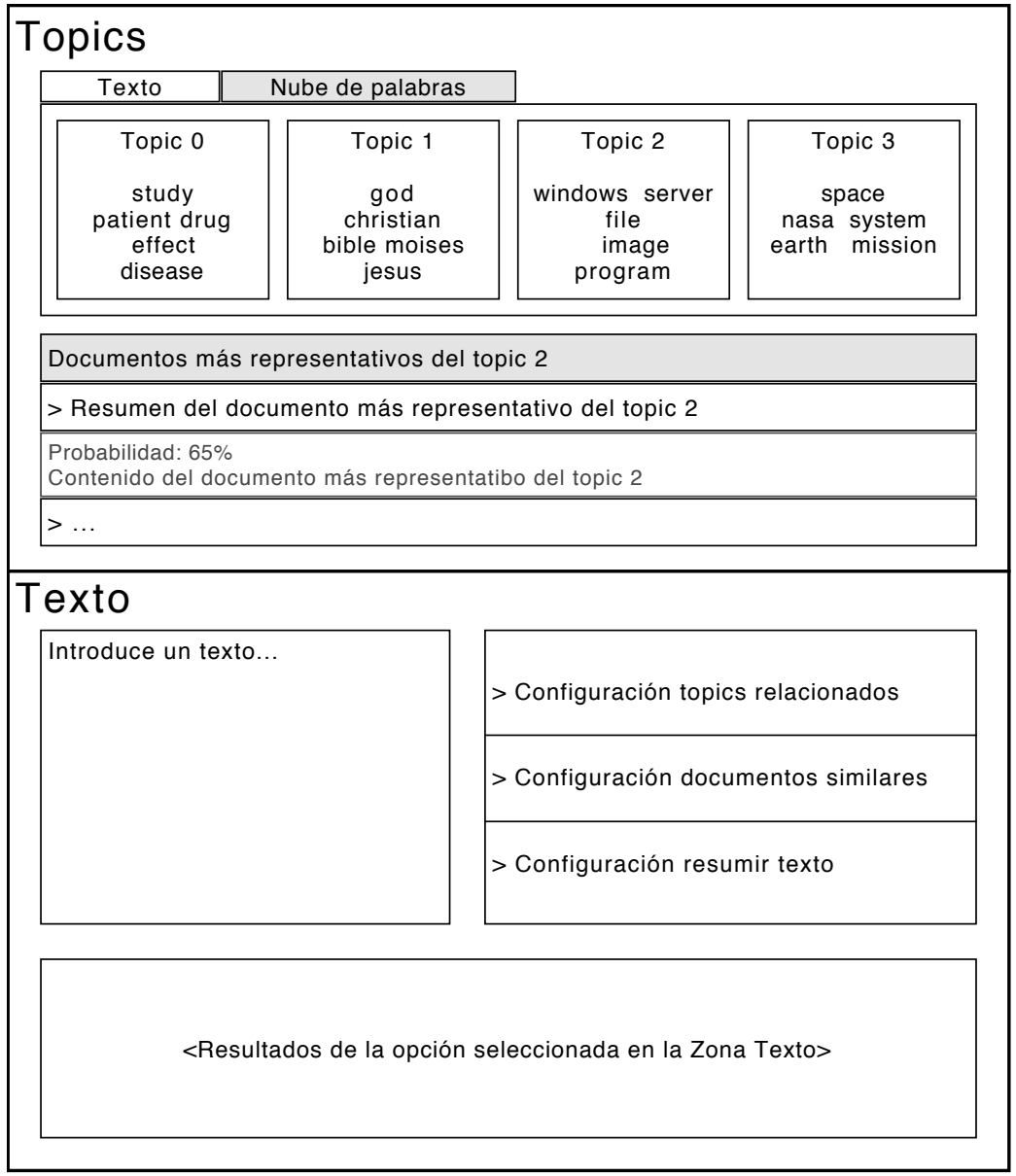
6.
EVOLUCIÓN DE LA INTERFAZ GRÁFICA

Evolución de la interfaz (fase análisis)

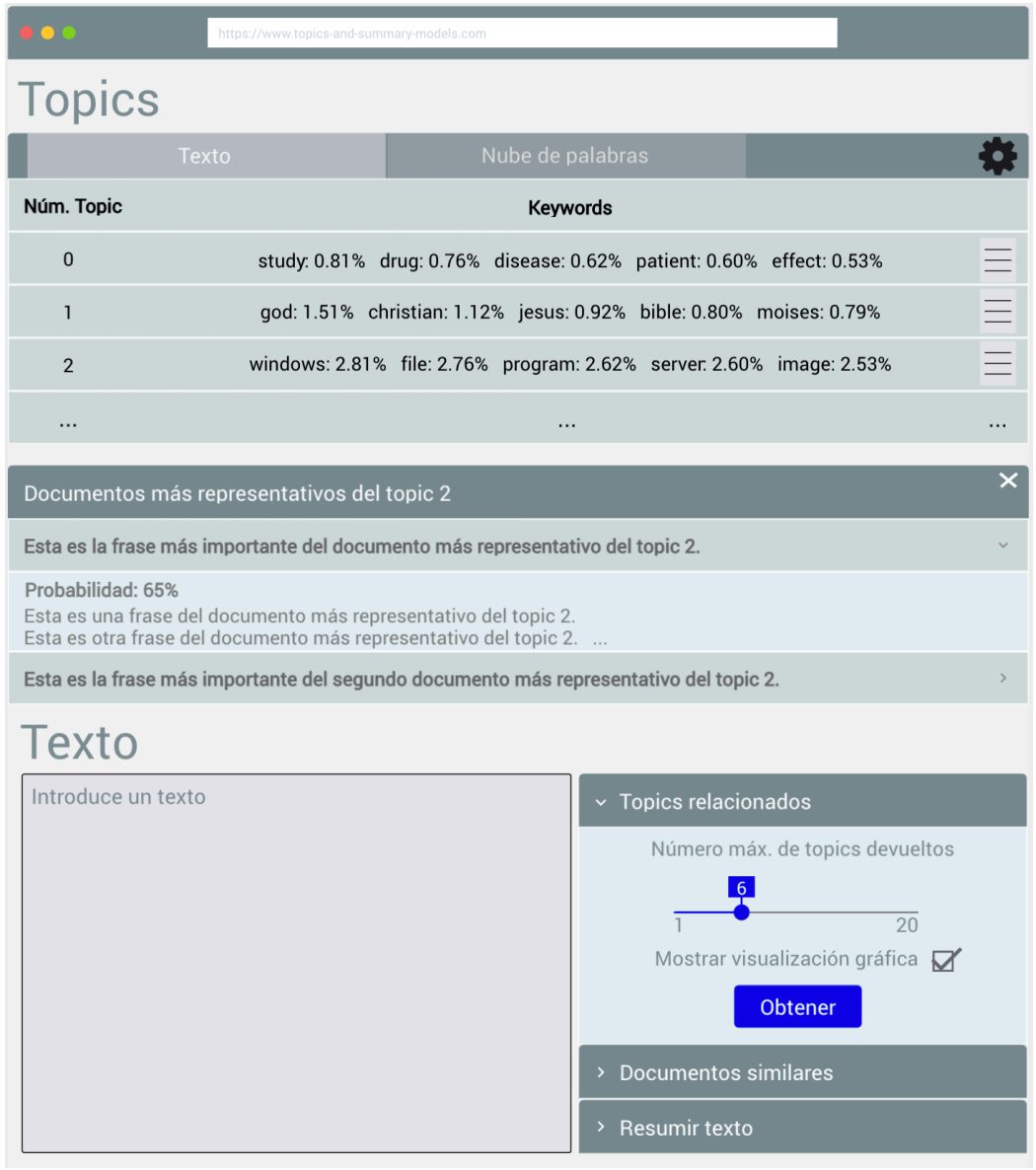


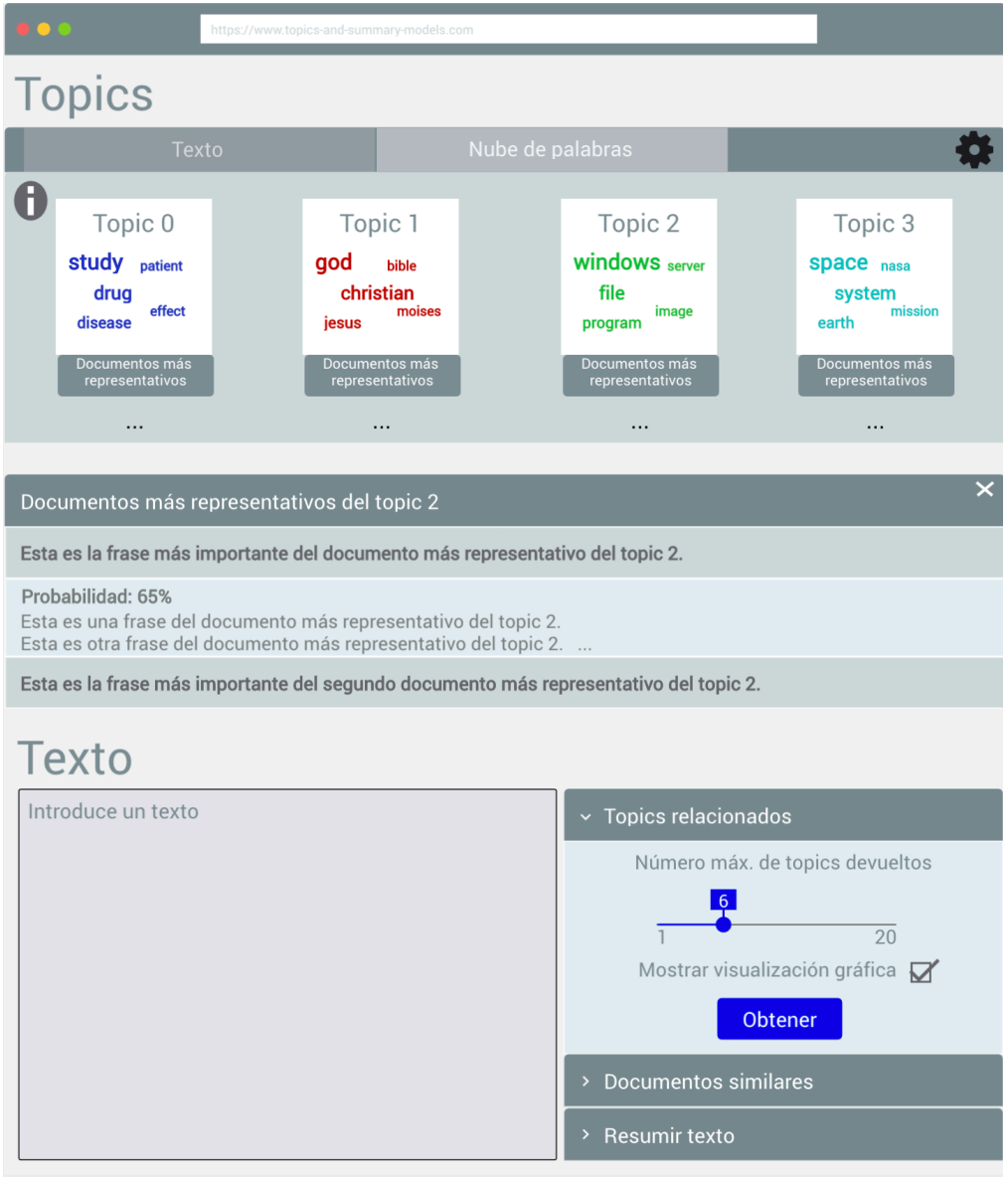
Evolución de la interfaz (fase análisis)





Evolución de la interfaz (fase diseño)



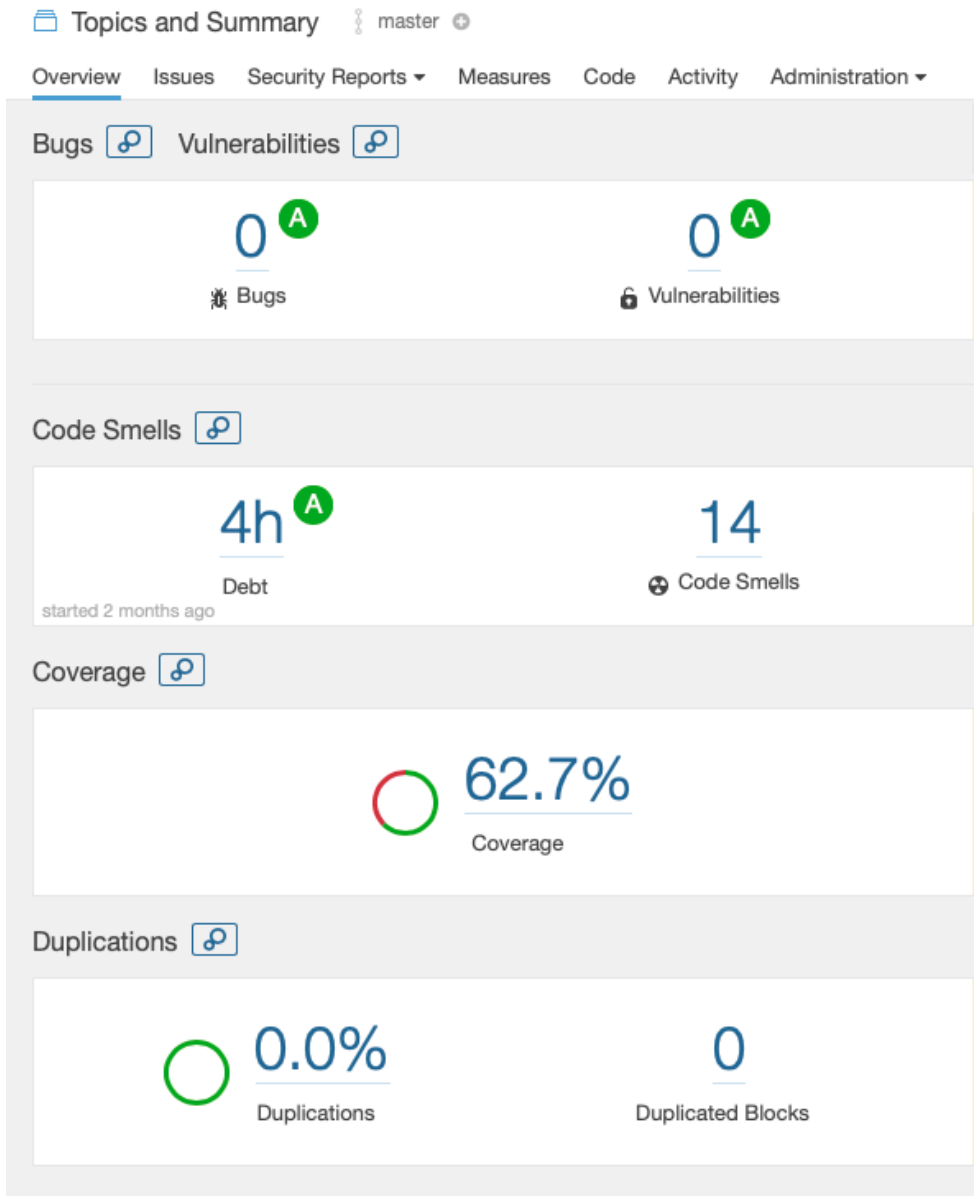


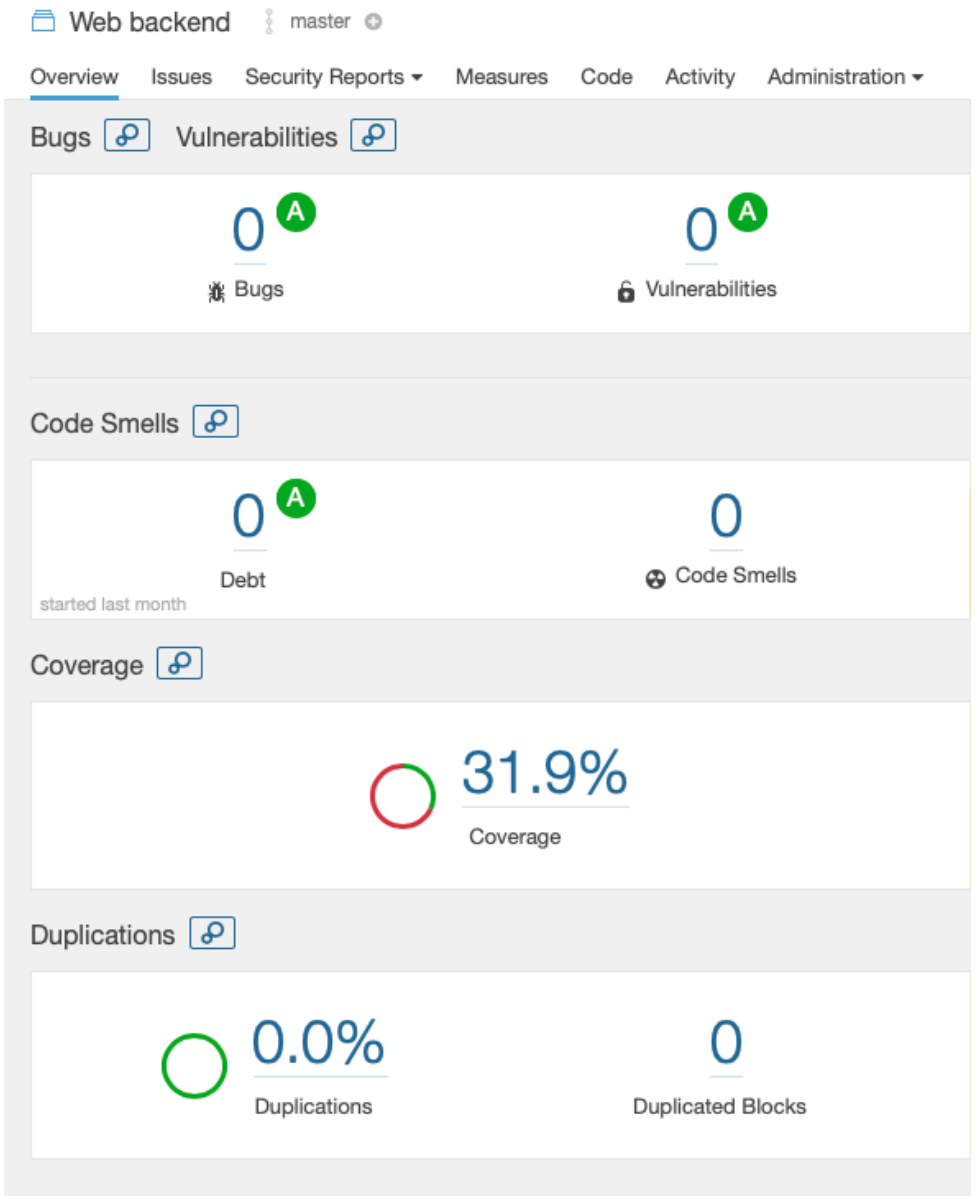
7.
PRUEBAS

Pruebas unitarias automatizadas

- Subsistema **Topics and Summary**
- Subsistema **Web backend**:
  - Componente params
  - Resumen de un texto cuando TextRank:
    - Converge
    - No converge

Análisis estático de código

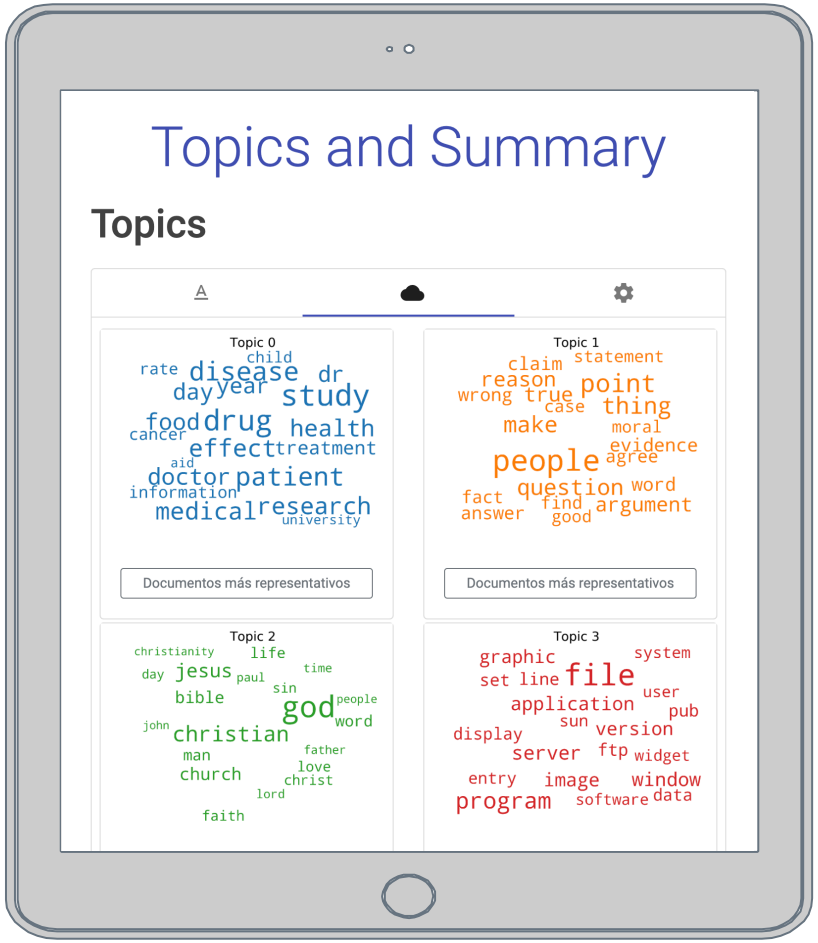




Pruebas de usabilidad

- Actividades guiadas y cuestionarios a **3 usuarios**
- **Resultados:**
  - Número errores y tiempo tareas **bajo**
  - **Más solicitado:** Notificación cuando aparece la información dinámicamente
- Posibles **cambios**:
  - **Aumentar** número mensajes **ayuda**
  - **Disminuir tiempo** generación **resúmenes**
  - Permitir **datasets español**
  - **Añadir scroll automático** que desplace al usuario a la información que aparece dinámicamente

Pruebas de adaptabilidad



8.
DEMO

9.
CONCLUSIONES

Conclusiones

- El sistema **cumple los objetivos**
- Los **modelos** dan **buenos resultados**
  - Aunque el tiempo de generación de resúmenes es en ocasiones elevado con textos grandes
- El **frontend sigue** los **prototipos** diseñados
- La **arquitectura** en 3 subsistemas fue un **acierto**
- Ha supuesto un **gran aprendizaje**, ya que se ha:
  - Profundizado en Procesamiento de Lenguaje Natural
  - Pasado por todas las fases del desarrollo de software

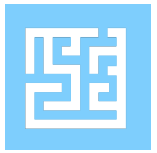
10.
AMPLIACIONES

Ampliaciones

Implementar las funcionalidades del administrador

Probar los modelos de extracción de topics con más datasets

Evaluar los resultados del algoritmo TextRank



Probar otros algoritmos de generación de resúmenes

Aumentar el número de mensajes de ayuda en la GUI

Generar los resúmenes de los documentos del dataset en batch

Permitir documentos en español

Añadir i18n al frontend

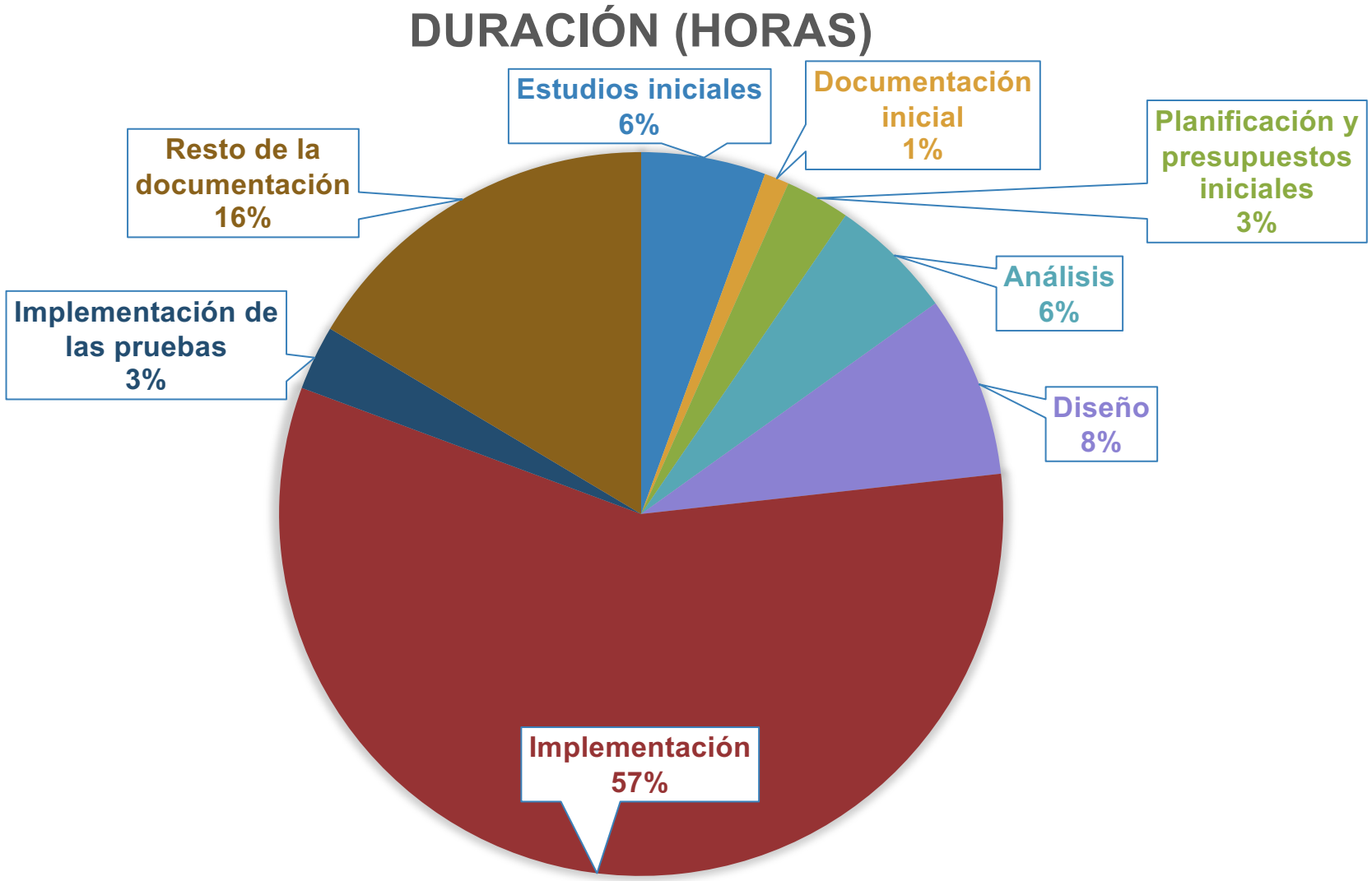
Hacer la interfaz web accesible

11.
PLANIFICACIÓN Y PRESUPUESTO FINALES

Planificación final

- **01 febrero 2019 – 07 julio 2019**
- Mismas etapas, salvo despliegue
- Duración total: ~~300h~~ **450h**

Planificación final



Presupuesto final

Presupuesto de costes final		
Cód.	***Partida***	***Total***
01	Estudios iniciales y planificación inicial	1.425,00 €
02	Análisis y diseño del sistema	2.306,25 €
03	Implementación y despliegue del sistema	10.312,45 €
04	Documentación	2.962,50 €
05	Otros costes	1.354,20 €
	Total Coste	**18.360,40 €**

Total Coste Presupuesto Inicial: 12.483,00 €
Incremento en el coste: 5.877,40€

¡Gracias!

¿Preguntas?

Créditos

Un agradecimiento especial a todas las personas que hicieron y lanzaron estos increíbles recursos de forma gratuita:

- Plantilla de la presentación de SlidesCarnival