Desarrollo de una medida de similaridad para Sistemas de Recomendación en sitios de Community Question Answering. Análisis desde un enfoque Big Data y usando un método de ensamble de clustering

Ing. Federico Tesone

Tesis de Maestría

Maestría en Ingeniería en Sistemas de Información

Director: Dr. Guillermo Leale
Co-directora: Dra. Soledad Ayala

15 de Octubre de 2021

Agenda

Agenda

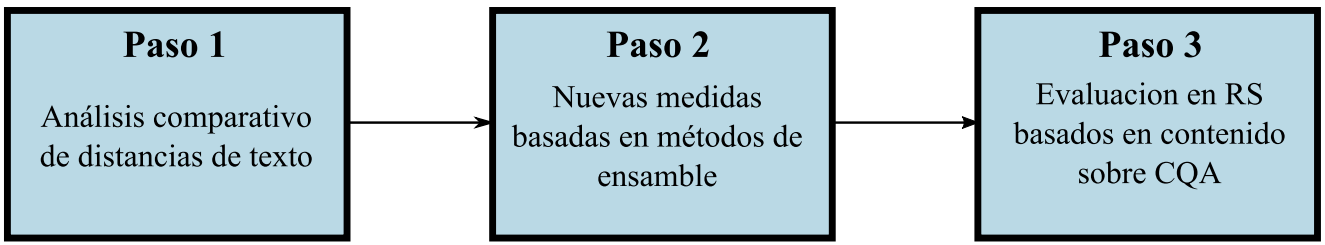
Área temática

Este trabajo se basa en 5 pilares teóricos:

- Sistemas de Recomendación.
- Sitios de Community Question Answering (CQA).
- Medidas de similaridad.
- Ensamble de Clustering.
- Big Data.

Tema específico I

Pipeline para un RS basado en contenido de CQA y en una nueva medida de similaridad.



Tema específico II

Considerando el conjunto completo de datos Quora (404301 pares de preguntas, es decir, 808602 preguntas totales), deberiamos realizar:

$\frac{n(n+1)}{2} = 326919001503$ calculos de distancias, donde $n = 808602$

Objetivo general

Objetivo general

Construir una arquitectura Big Data que incluye la posibilidad de ser aplicada a grandes conjuntos de datos de preguntas en el ámbito de CQA y, a partir de esta arquitectura, implementar y evaluar nuevas medidas de similaridad entre textos que puedan ser utilizadas en sistemas de recomendación.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.
- Proponer una nueva medida que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data y que sea extensible a otras medidas existentes en el estado del arte.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.
- Proponer una nueva medida que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data y que sea extensible a otras medidas existentes en el estado del arte.
- Brindar conclusiones, pautas y recomendaciones para trabajar con medidas de comparación de textos en grandes volúmenes de datos en sitios de CQA utilizando arquitecturas basadas en Big Data.

Agenda

Motivación de la tesis

Motivación

- Las medidas de similaridad del estado del arte tienen conocidos problemas.

Motivación de la tesis

Motivación

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que pueda aplicarse como entrada para un RS.

Motivación de la tesis

Motivación

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que pueda aplicarse como entrada para un RS.
- Es difícil identificar un algoritmo de Clustering que pueda manejar todos los tipos de formas y tamaños de cluster.

Motivación de la tesis

Motivación

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que pueda aplicarse como entrada para un RS.
- Es difícil identificar un algoritmo de Clustering que pueda manejar todos los tipos de formas y tamaños de cluster.
- Arquitectura de software que soporte el procesamiento del método propuesto de una forma eficiente y escalable.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.
- Aumento en la calidad de la base de conocimiento del sitio.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.
- Aumento en la calidad de la base de conocimiento del sitio.
- Explorar desafíos tecnológicos que sirvan como fuente de futuras investigaciones o desarrollos de software.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.
- Aumento en la calidad de la base de conocimiento del sitio.
- Explorar desafíos tecnológicos que sirvan como fuente de futuras investigaciones o desarrollos de software.
- Aplicación de los conocimientos obtenidos en este trabajo para mejorar otros sitios basadas en busquedas textuales.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.
- Presentación en congresos tales como AGRANDA, CONAIISI, o RecSys.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.
- Presentación en congresos tales como AGRANDA, CONAIISI, o RecSys.
- Elaborar material de estudio relacionado con la temática para materias de grado, cursos, o disertaciones.

Agenda

Sitios de CQA

Sitios de Community Question Answering

Los sitios de *Community Question Answering* CQA, son un tipo especial de sitios web de *Question Answering* (QA), los cuales permiten a los usuarios registrados responder a preguntas formuladas por otras personas.

Sitios de CQA

¿Por qué los sitios de CQA todavía no alcanzan a satisfacer las expectativas de los usuarios?

- Baja probabilidad de encontrar al experto.

Sitios de CQA

¿Por qué los sitios de CQA todavía no alcanzan a satisfacer las expectativas de los usuarios?

- Baja probabilidad de encontrar al experto.
- Respuestas de baja calidad.

Sitios de CQA

¿Por qué los sitios de CQA todavía no alcanzan a satisfacer las expectativas de los usuarios?

- Baja probabilidad de encontrar al experto.
- Respuestas de baja calidad.
- Preguntas archivadas y poco consultadas: muchas preguntas de los usuarios son similares.

Sistemas de Recomendación

Un RS es un conjunto de herramientas de software que sugiere ítems a un usuario, quien posiblemente utilizará algunos de ellos.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- FIltrado Colaborativo.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- FIltrado Colaborativo.
- Demográficos.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- FIltrado Colaborativo.
- Demográficos.
- Basados en conocimiento.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- FIltrado Colaborativo.
- Demográficos.
- Basados en conocimiento.
- Basados en comunidades (sociales).

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- FIltrado Colaborativo.
- Demográficos.
- Basados en conocimiento.
- Basados en comunidades (sociales).
- Sistemas Híbridos.

Big Data I

Big Data

"Conjuntos de datos cuyo tamaño está más allá de la habilidad de las herramientas software de base de datos para capturar, almacenar, gestionar y analizar los datos" (Manyika et al., 2011).

"Big Data son activos de información caracterizados por su alto volumen, velocidad y variedad que demandan formas innovadoras y rentables de procesamiento de información para mejorar la compresión y la toma de decisiones" (consultora Gartner).

Big Data II

Conceptos importantes relativos a este trabajo:

- Map-reduce.

Big Data II

Conceptos importantes relativos a este trabajo:

- Map-reduce.
- Arquitectura Hadoop (HDFS).

Big Data II

Conceptos importantes relativos a este trabajo:

- Map-reduce.
- Arquitectura Hadoop (HDFS).
- Apache Spark.

Information retrieval

Information retrieval

Information retrieval (IR), traducido a menudo como "recuperación de información", se define la acción de como encontrar material (generalmente documentos) de una naturaleza desestructurada (generalmente texto) que satisfaga una necesidad de información de grandes colecciones (generalmente almacenadas en computadoras) (Schütze et al. 2008).

Similaridad

Las medidas de similaridad son interés poder cuantificar la relación entre objetos.

1. La función de similaridad es definida satisfaciendo las condiciones:
   1. Simetría,
   $$S(x_i, x_j) = S(x_j, x_i);$$
   2. Positividad,
   $$0 \leq S(x_i, x_j) \leq 1, \quad \forall x_i, x_j.$$

Similaridad

Las medidas de similaridad son interés poder cuantificar la relación entre objetos.

1. La función de similaridad es definida satisfaciendo las condiciones:
   1. Simetría,
   $$S(x_i, x_j) = S(x_j, x_i);$$
   2. Positividad,
   $$0 \leq S(x_i, x_j) \leq 1, \quad \forall x_i, x_j.$$

2. Es posible transformar una medida de similaridad $S(x_i, x_j)$ en una de distancia $D(xi, xj)$ que cumpla $0 \leq D(x_i, x_j) \leq 1$, en el intervalo $[0, 1]$. Aplicando $D(x_i, x_j) = 1 - S(x_i, x_j)$.

Modelo de espacio vectorial

En el modelo de *espacio vectorial*, un texto es representado como un vector de términos. Si las palabras son elegidas como términos, entonces cada palabra del vocabulario sería una *dimensión* independiente en el espacio vectorial (Singhal et al., 2001).

Típicamente, el ángulo entre los dos vectores es usado como medida de divergencia entre los mismos, y el coseno del ángulo es usado como similaridad numérica.

Distancia del coseno I

Siendo $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$ dos documentos en forma de vectores:

$$d_c(\vec{D}_i, \vec{D}_j) = 1 - cos(\vec{D}_i, \vec{D}_j) = 1 - \frac{\vec{D}_i' \vec{D}_j}{\sqrt{\vec{D}_i' \vec{D}_i}\sqrt{\vec{D}_j' \vec{D}_j}}.$$

Particularmente en IR, es de interés el intervalo $[0, 1]$, entonces la distancia del coseno se puede derivar de la fórmula del *producto escalar*:

$$\vec{D}_i . \vec{D}_j = \left\|\vec{D}_i\right\| . \left\|\vec{D}_j\right\| .cos(\theta),$$

siendo $\left\|\overrightarrow{D_i}\right\|$ y $\left\|\overrightarrow{D_j}\right\|$ los módulos de los vectores $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$ respectivamente, y θ el ángulo formado entre ellos.

Distancia del coseno II

Entonces, la similaridad entre dos vectores puede medirse como:

$$cos(\theta) = \frac{\vec{D}_i . \vec{D}_j}{\left\|\vec{D}_i\right\| . \left\|\vec{D}_j\right\|},$$

donde d_i y d_j son los componentes de los vectores $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$ respectivamente.

Medidas de Similaridad

Medidas de similaridad utilizadas

- Term Frequency (TF)
- Term Frequency - Inverse Docuement Frequency (TF-IDF).
- Word2Vec
- FastText
- Semantic Distance

Term Frequency (TF)

Caracteristicas de Term Frequency:

- También conocido en la literatura como *Bag of words* (bolsa de palabras).
- El orden exacto de los términos es ignorado, pero se basa en el número de ocurrencias de cada uno de ellos en un documento.
- Cada documento corresponde a un vector y cada término a una dimensión.
- Se mide el grado de similaridad de dos documentos utilizando el coseno del ángulo.

"Mary is quicker than John" y *"John is quicker than Mary"*

Term Frequency Inverse Document Frequency (TF-IDF)

Caracteristicas de TF-IDF:

- Se define *document frequency* df_t como el número de documentos en una colección que contienen el término t.
- *Inverse document frequency*, o IDF, es un indicador basado en la cantidad de documentos que contienen (o son indexados por) un término en cuestión.
- Intuición: si un término de búsqueda se encuentra en muchos documentos, no es un buen discriminador, y se le debe asignar menor peso que a un término que se encuentra en pocos documentos.

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_j)$$

Word2Vec

Caracteristicas de Word2Vec:

- Modelos basados en redes neuronales con una capa oculta para computar representaciones de palabras como vectores continuos en grandes conjuntos de datos.
- Dos modelos: Skip-gram y Continuous Bag of Words.
- Las entradas y salidas de la red neuronal son palabras representadas como *one-hot* vector.
- Los pesos de la capa oculta se van ajustando utilizando un clasificador de regresión Softmax.
- Estos pesos resultantes dan como resultado a la representación vectorial de palabras utilizadas para el cálculo de similaridad de este trabajo.

FastText

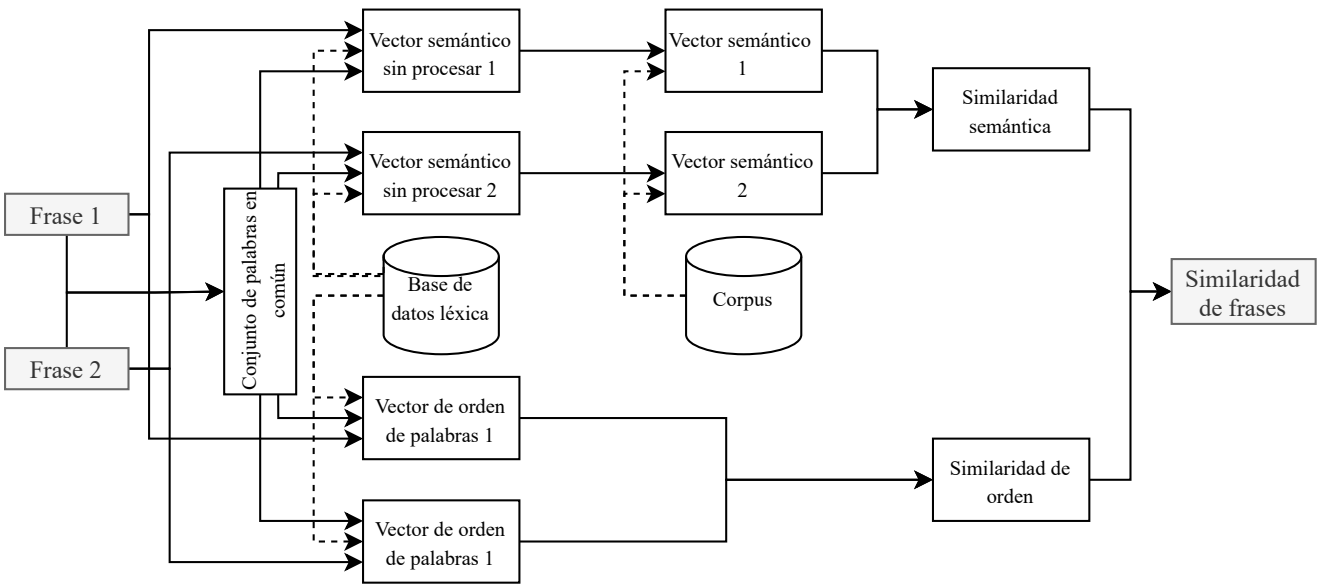
Caracteristicas de FastText:

- Libreria open-source desarrollada por Facebook.
- Basado en Skip-gram pero utilizando un modelo sub-palabra.
- Cada palabra es representada como una bolsa de *n-gramas*.
- Mayor precisión en diferentes medidas de rendimiento.

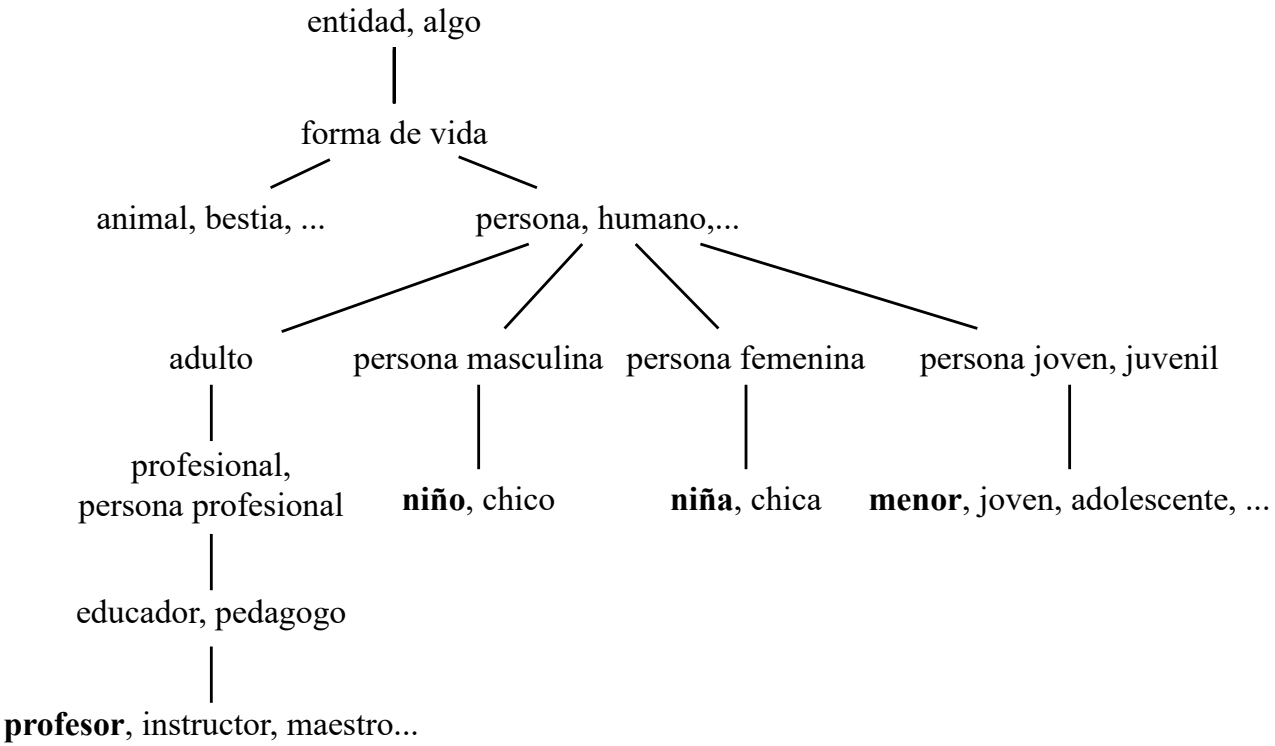
Semantic Distance I

Caracteristicas de Semantic Distance:

- La distancia semántica usada en este trabajo está basada en *redes semánticas* y *estadísticas de corpus* (Li et al., 2006).
- Enfocado en textos de distancia corta.
- Tiene en cuenta la *información semántica* y la *información del orden* de las palabras implicadas en las frases involucradas.



Similaridad semántica entre palabras



Semantic Distance III

Similaridad semántica entre frases

Este método semántico usa únicamente vectores semánticos formados por las frases en comparación. El valor de una entrada del vector semántico es calculado de la siguiente forma:

- **Caso 1**. Si w_i aparece en la frase, s_i es 1.
- **Caso 2**. Si w_i no está contenida en T_1, se calcula una similaridad semántica entre w_1 y cada palabra en T_1 utilizando el método de la sección anterior.

Semantic Distance IV

Similaridad semántica entre frases (cont.)

Se ponderan cada una de las palabras basadas en su contenido de información:

$$s_i = \check{s} \cdot I(w_i) \cdot I(\widetilde{w}_i),$$

La similaridad semántica entre dos frases es definida como el coeficiente del coseno entre los dos vectores:

$$S_s = \frac{s_1 . s_2}{||s_1||.||s_2||}.$$

Semantic Distance V

Similaridad de orden entre frases

Consideremos dos frases T_1 y T_2, por ejemplo:

- **T1**: A quick brown dog jumps over the lazy fox.
- **T2**. A quick brown fox jumps over the lazy dog.

$$r_1 = \{\, 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9 \,\},$$

$$r_2 = \{\, 1\ 2\ 3\ 9\ 5\ 6\ 7\ 8\ 4 \,\}.$$

Se propone entonces una medida de similaridad de orden entre frases de la siguiente manera:

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|}.$$

Semantic Distance VI

Similaridad total entre frases:

$$S(T_1, T_2) = \delta S_s + (1 - \delta) S_r,$$

$$S(T_1, T_2) = \delta \frac{s_1 . s_2}{\|s_1\| \, \|s_2\|} + (1 - \delta) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|},$$

donde $0 \leq \delta \leq 1$ decide la contribución relativa de cada una de las medidas de similaridad.

Clustering

El Clustering o *análisis cluster* tiene por objetivo agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos.

Ensamble de Clustering

Ensamble de Clustering

El *Ensamble de Clustering* es un método para extraer clusters consistentes dadas particiones variadas de entrada.

- Combina resultados de distintos algoritmos de Clustering con distintas formas de cluster.

Ensamble de Clustering

Ensamble de Clustering

El *Ensamble de Clustering* es un método para extraer clusters consistentes dadas particiones variadas de entrada.

- Combina resultados de distintos algoritmos de Clustering con distintas formas de cluster.
- Aprovecha la variabilidad agregada para encontrar una estructura *inter-patrón*.

Ensamble de Clustering

Ensamble de Clustering

El *Ensamble de Clustering* es un método para extraer clusters consistentes dadas particiones variadas de entrada.

- Combina resultados de distintos algoritmos de Clustering con distintas formas de cluster.
- Aprovecha la variabilidad agregada para encontrar una estructura *inter-patrón*.
- Identificación de clusters subyacentes con formas, tamaños y densidades arbitrarias.

Combinación de Evidencias

Combinar los resultados de múltiples ejecuciones de clustering dentro de una misma partición de datos viendo cada uno de esos resultados como una evidencia independiente de la organización de los mismos.

Tomado las co-ocurrencia de pares de patrones en el mismo cluster, las N particiones de datos para n patrones, son mapeadas en una *matriz de co-asociación* $n \times n$:

$$C(i,j) = \frac{n_{ij}}{N},$$

donde n_{ij} es el número de veces que el par de patrones (i,j) es asignado al mismo cluster entre las N particiones de datos.

Agenda

Hipótesis I

Hipótesis del trabajo de tesis

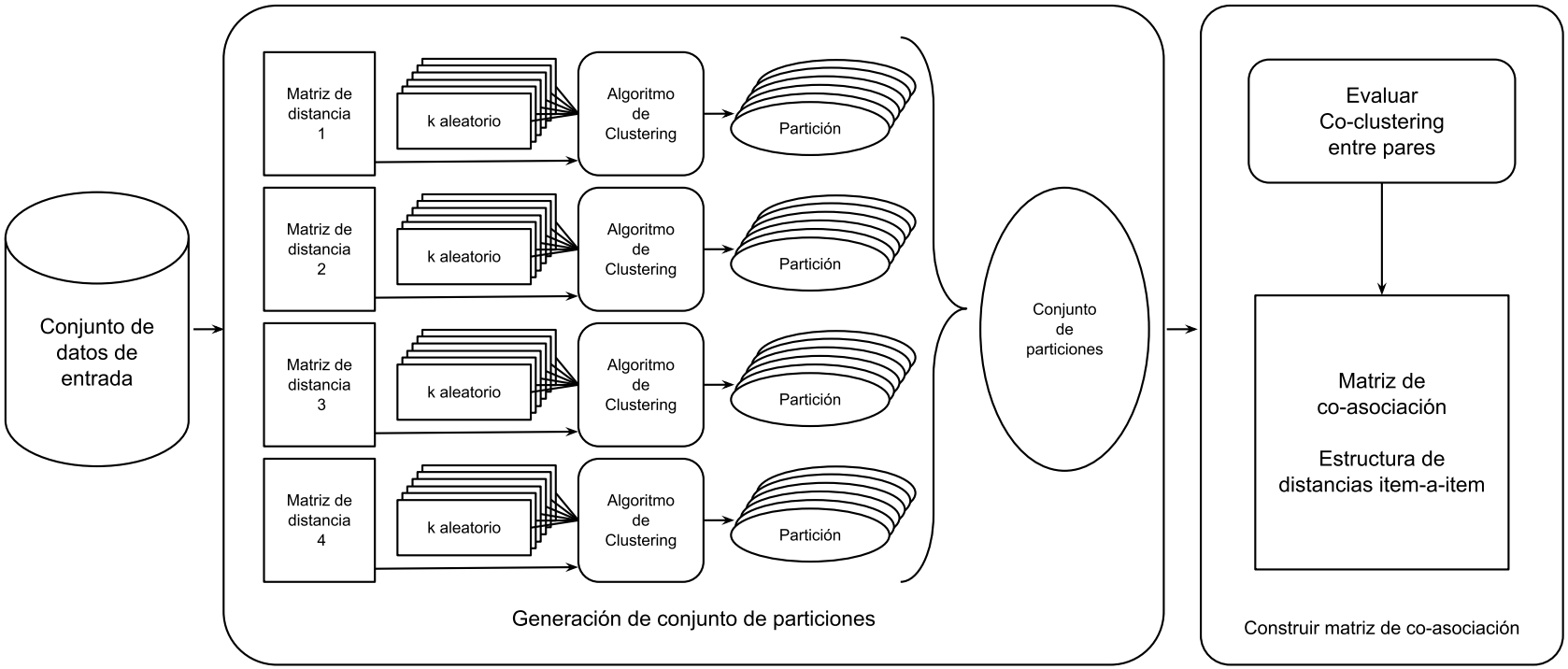
A partir del relevamiento del estado del arte se establece la hipótesis de que los algoritmos de cálculo de similaridad de texto en sitios de CQA, con el fin de participar del proceso inherente a la aplicación de Sistemas de Recomendación con gran volumen de datos, pueden ser mejorados en cuanto a medidas de rendimiento y de desempeño si se aplica un método de ensamble de clustering mediante una arquitectura Big Data apropiada.

Hipótesis II

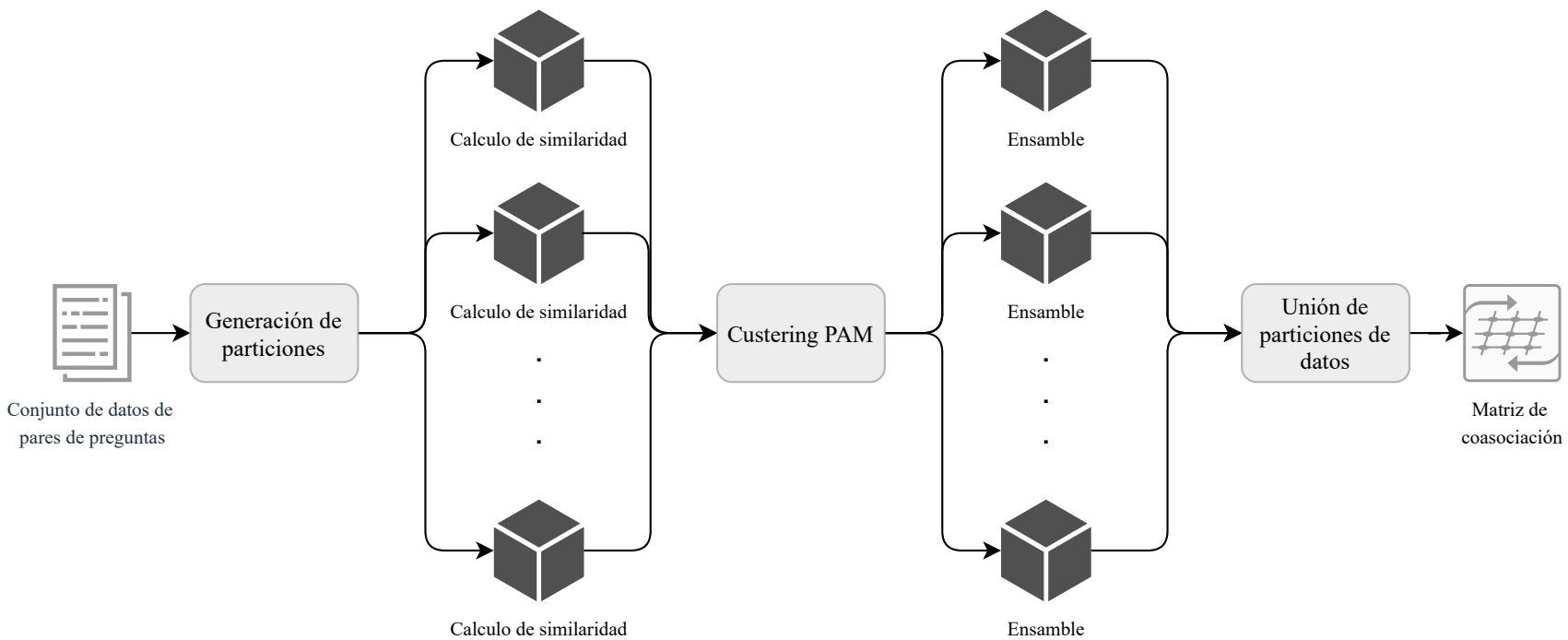
Hipótesis del trabajo de tesis (cont.)

Por tal motivo, y como respuesta a la hipótesis planteada, se presenta el desarrollo de un nuevo método de cálculo de similaridad de texto basado en una arquitectura Big Data. Este método aprovecha las características de adimensionalidad y variabilidad de datos propias del Ensamble de Clustering. El método se aplica a un gran conjunto de datos reales con el fin de verificar la eficiencia y eficacia del procedimiento. Asimismo, se realiza un análisis comparativo del método presentado con los algoritmos para cálculo de similaridad de texto del estado del arte.

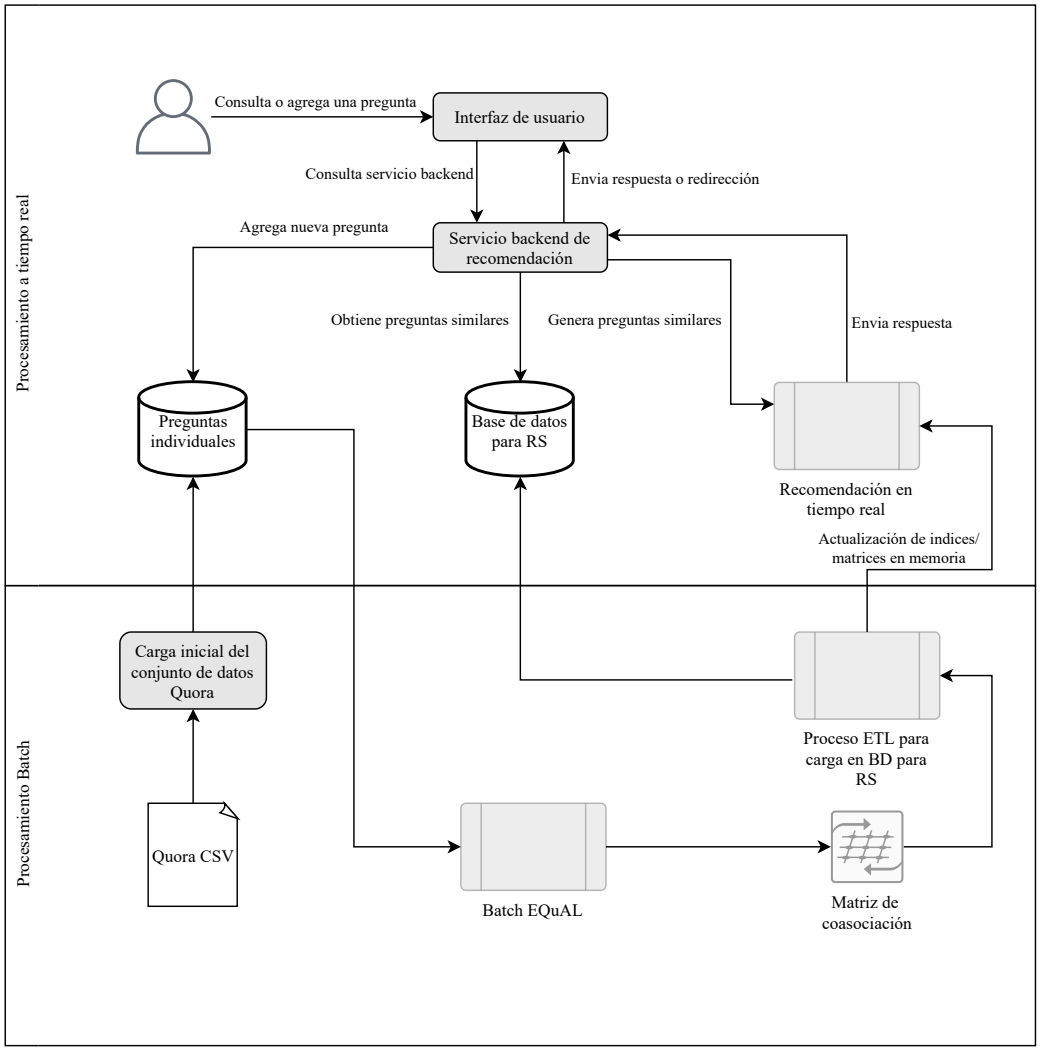
El método propuesto



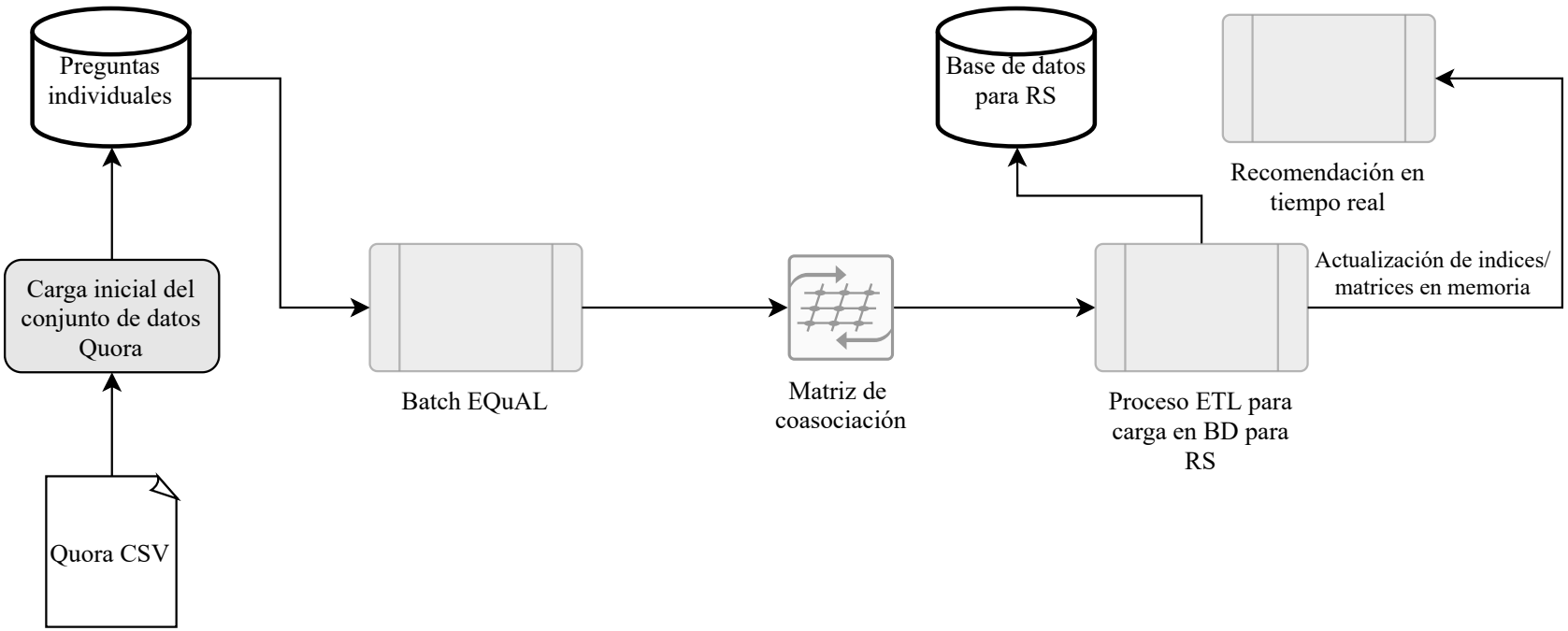
Arquitectura de procesamiento de datos



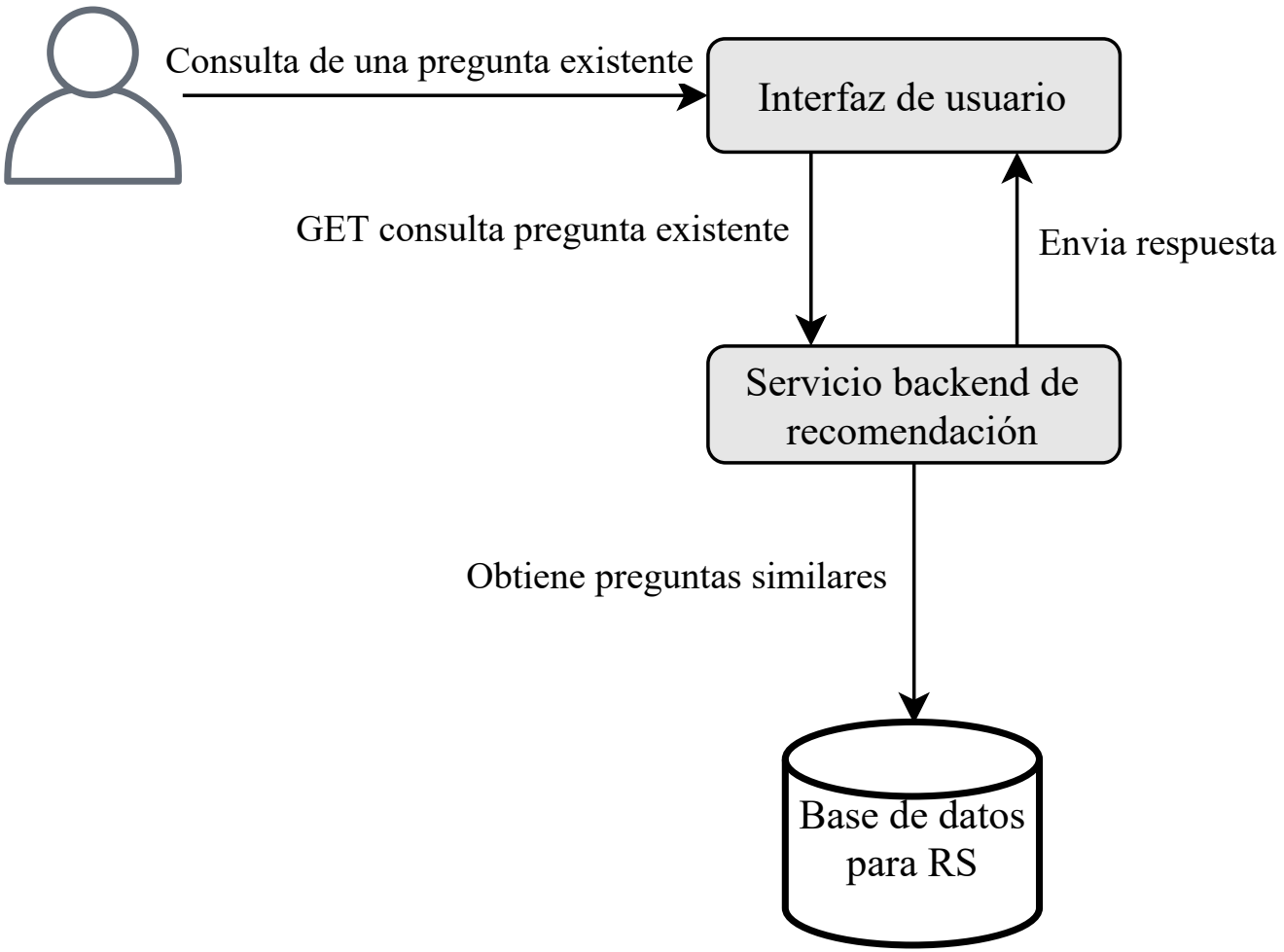
Implementación en un sistema de recomendación de tiempo real



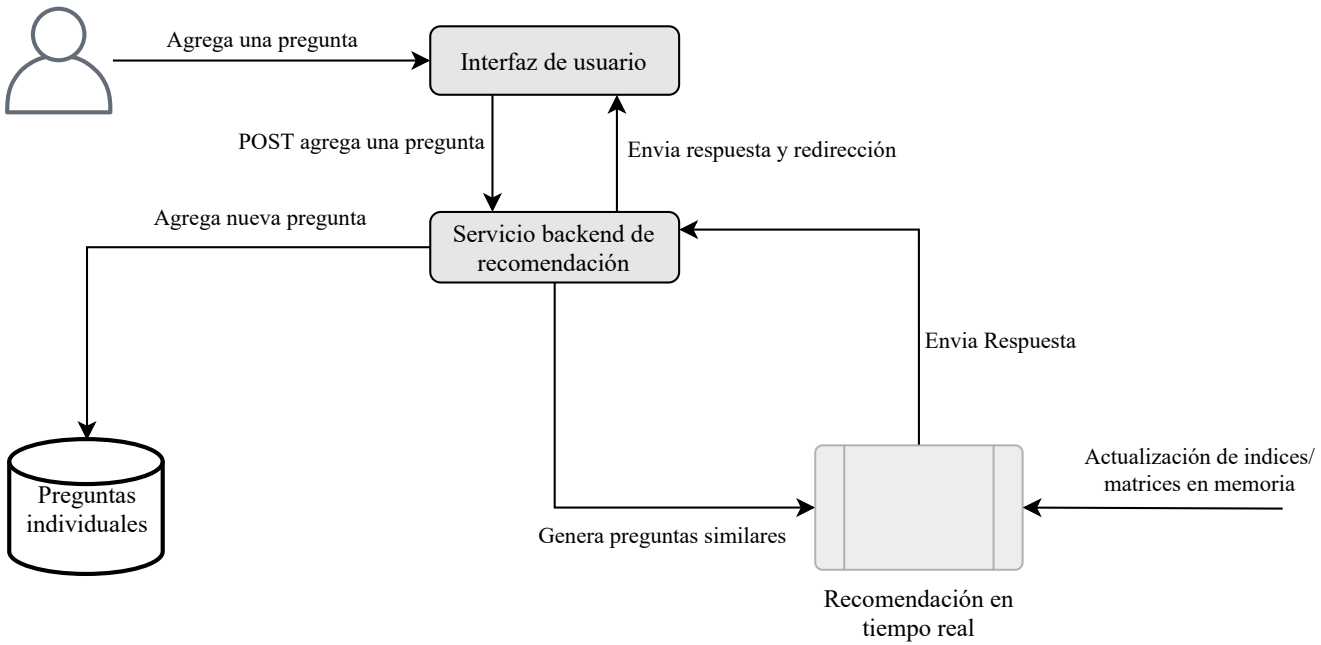
Procesamiento fuera de línea



Consulta de una pregunta existente



Agregar una nueva pregunta



Agenda

Estado del arte

			Predicho		Exactitud	Error
			0	**1**		
TF	**Real**	**0**	0.4355	0.1953	0.6776	0.3224
		1	0.1271	0.2421		
TF/IDF	**Real**	**0**	0.4477	0.1831	0.6685	0.3315
		1	0.1484	0.2208		
Word2Vec	**Real**	**0**	0.4343	0.1965	0.6788	0.3212
		1	0.1247	0.2445		
FastText	**Real**	**0**	0.5033	0.1275	0.6725	0.3275
		1	0.2	0.1692		
Semantic Distance	**Real**	**0**	0.4877	0.1431	**0.6797**	**0.3203**
		1	0.1772	0.192		

Preprocesamiento y muestreo del conjunto de datos

- Preprocesamiento:
  1. Convertir el texto en minúscula.
  2. Eliminar fórmulas; las cuales están encerradas entre etiquetas [math][/math] y [code][/code].
  3. Reemplazar números por letras.
  4. Eliminar caracteres especiales, ya que los datos deben ser uniformes.

Preprocesamiento y muestreo del conjunto de datos

- Preprocesamiento:
  1. Convertir el texto en minúscula.
  2. Eliminar fórmulas; las cuales están encerradas entre etiquetas [math][/math] y [code][/code].
  3. Reemplazar números por letras.
  4. Eliminar caracteres especiales, ya que los datos deben ser uniformes.

- Muestreo:
  - Generación pseudoaleatoria con criterios de aceptación.
  - Garantizar subconjuntos estadistícamente significativos.

Generación de particiones I

Ejemplo de la estructura de los subconjuntos de muestreo:

question_pair_id	question_1	question_2
123004	question_0	question_2
98776	question_1	question_3

Combinación de todas las preguntas individuales de una muestra:

sequence_id_1	question_id_1	sequence_id_2	question_id_2
0	question_0	1	question_1
0	question_0	2	question_2
0	question_0	3	question_3
1	question_1	2	question_2
1	question_1	3	question_3
2	question_2	3	question_3

Generación de particiones II

Cálculo de similaridad Ejemplo de la estructura de matriz de similaridad en formato de tabla.

sequence_id_1	question_id_1	sequence_id_2	question_id_2	similarity
0	question_0	1	question_1	similarity_01
0	question_0	2	question_2	similarity_02
0	question_0	3	question_3	similarity_03
1	question_1	2	question_2	similarity_12
1	question_1	3	question_3	similarity_13
2	question_2	3	question_3	similarity_23

También se puede ver como una matriz triangular superior:

$$\begin{bmatrix} 0 & similarity\_01 & similarity\_02 & similarity\_03 \\ 0 & 0 & similarity\_12 & similarity\_13 \\ 0 & 0 & 0 & similarity\_23 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Clustering y etiquetado

- Por cada una de las matrices de similaridad se realizan varias ejecuciones de clustering PAM.

Clustering y etiquetado

- Por cada una de las matrices de similaridad se realizan varias ejecuciones de clustering PAM.
- Por cada una de las ejecuciones PAM se proporciona un k inicial.

Clustering y etiquetado

- Por cada una de las matrices de similaridad se realizan varias ejecuciones de clustering PAM.
- Por cada una de las ejecuciones PAM se proporciona un k inicial.
- Los resultados poseen la siguiente estructura:

run_uuid	question_id	assigned_medoid
63815467136575428551131593057064980770	336	856
63815467136575428551131593057064980770	342	856
63815467136575428551131593057064980770	26	358
63815467136575428551131593057064980770	1364	437

Ensamble de Clustering I

Para explicar el procedimiento de **Ensamble de Clustering**, se consideran 3 resultados de ejecuciones ejemplo:

run_uuid	question_id	cluster_id
run_uuid_1	1	1
run_uuid_1	2	1
run_uuid_1	3	1
run_uuid_1	4	4

run_uuid	question_id	cluster_id
run_uuid_2	1	1
run_uuid_2	2	2
run_uuid_2	3	1
run_uuid_2	4	2

Ensamble de Clustering II

run_uuid	question_id	cluster_id
run_uuid_3	1	3
run_uuid_3	2	2
run_uuid_3	3	3
run_uuid_3	4	2

Luego, el resultado de todas las ejecuciones se agrupa por pregunta individual, de la siguiente forma:

question_id	tuples
1	[(run_uuid_1,1),(run_uuid_2,1),(run_uuid_3,3)]
2	[(run_uuid_1,1),(run_uuid_2,2),(run_uuid_3,2)]
3	[(run_uuid_1,1),(run_uuid_2,1),(run_uuid_3,3)]
4	[(run_uuid_1,4),(run_uuid_2,2),(run_uuid_3,2)]

Se genera un conjunto de datos intermedio con la interseccion de los conjuntos para la combinación de todas las preguntas individuales:

Ensamble de Clustering III

question_id_1	question_id_2	tuples
1	2	[(run_uuid_1,1)]
1	3	[(run_uuid_1,1),(run_uuid_2,1),(run_uuid_3,3)]
1	4	[]
2	3	[(run_uuid_1,1)]
2	4	[(run_uuid_2,2)]
3	4	[]

Por ejemplo: pregunta 1 = [**(run_uuid_1,1)**,(run_uuid_2,1),(run_uuid_3,3)] y la pregunta 2 = [**(run_uuid_1,1)**,(run_uuid_2,2),(run_uuid_3,2)].

Ensamble de Clustering IV

Se cuenta la cantidad de veces que una pregunta coincide con otra para una misma ejecución.

$$len(set(tuples\_1).intersection(set(tuples\_2)))/total\_runs$$

Respondiendo a la formula de Ensamble de Clustering de Acumulación de Evidencias.

$$C(i,j) = \frac{n_{ij}}{N}.$$

Ensamble de Clustering V

Y se genera la siguiente estructura como resultado ($total\_runs = 3$):

question_id_1	question_id_2	similarity
1	2	0.3333
1	3	1.0
1	4	0
2	3	0.3333
2	4	0.3333
3	4	0

La estructura resultante es una ***matriz de co-asociación***.

Método de validación

Validando los resultados

- Generación de resultados estadísticamente significativos se ejecutó el proceso completo de modo iterativo, variando dos parámetros principales:
  1. El tamaño de la muestra.
  2. El número de clusters k.

Método de validación

Validando los resultados

- Generación de resultados estadísticamente significativos se ejecutó el proceso completo de modo iterativo, variando dos parámetros principales:
  1. El tamaño de la muestra.
  2. El número de clusters k.
- Como experimentos para este trabajo se realizaron ejecuciones con conjuntos de datos aleatorios de 100, 500, 1000, 1500 y 2000 pares de preguntas.

Método de validación

Validando los resultados

- Generación de resultados estadísticamente significativos se ejecutó el proceso completo de modo iterativo, variando dos parámetros principales:
  1. El tamaño de la muestra.
  2. El número de clusters k.
- Como experimentos para este trabajo se realizaron ejecuciones con conjuntos de datos aleatorios de 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Para cada tamaño de muestra, se realizaron 10 muestras aleatorias manteniendo un k fijo.

Método de validación

Validando los resultados

- Generación de resultados estadísticamente significativos se ejecutó el proceso completo de modo iterativo, variando dos parámetros principales:
  1. El tamaño de la muestra.
  2. El número de clusters k.
- Como experimentos para este trabajo se realizaron ejecuciones con conjuntos de datos aleatorios de 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Para cada tamaño de muestra, se realizaron 10 muestras aleatorias manteniendo un k fijo.
- Dando un total de 5 (distintos tamaños de muestra) $\times$ 10 (cantidad de ejecuciones por tamaño de muestra) $=$ 50 matrices de co-asociación resultado, para cada valor de k dado.

Matrices de confusión

Matriz de confusión utilizada

		Predicho	
		0	**1**
Real	**0**	a	b
	1	c	d

Los indicadores de rendimiento seleccionados para ser evaluados en los experimentos realizados con fines comparativos, son los siguientes:

- **Exactitud:** $(a + d)/(a + b + c + d)$.
- **Error:** $(b + c)/(a + b + c + d)$.

En estos indicadores se cumple la condición $a + b + c + d = 1$.

Preparación de los datos I

Muestras de pares de preguntas que se utilizó como entrada del método EQuAL:

sequence_id	question_pair_id	question_1	question_2	equal
0	123004	question_10	question_11	1
1	98776	question_11	question_21	0

Matriz de co-asociación generada por el proceso EQuAL:

question_id_1	question_id_2	question_1	question_2	similarity
question_10	question_11	contenido	contenido	0.857
question_10	question_20	contenido	contenido	0.210
question_10	question_21	contenido	contenido	0.126
question_11	question_20	contenido	contenido	0.006
question_11	question_21	contenido	contenido	0.368
question_20	question_21	contenido	contenido	0.146

Preparación de los datos II

Se filtra la tabla anterior con los pares de preguntas que se encuentran en el conjunto de datos de entrada:

question_id_1	**question_id_2**	**question_1**	**question_2**	**similarity**
question_10	question_11	contenido	contenido	0.857
question_11	question_21	contenido	contenido	0.368

Elección del umbral correcto I

La similaridad S entre un par de preguntas (q_1, q_2) es igual o superior a cierto umbral t se considera que son iguales (valor 1) y distintas si sucede lo contrario (valor 0). De esta forma

$$f(x) = \begin{cases} 1 & si \quad S(q_1, q_2) \geq t \\ 0 & si \quad S(q_1, q_2) < t \end{cases} .$$

Cual de los valores de umbral t tiene mejor rendimiento?

Elección del umbral correcto II

Se toma valores potenciales de umbral con intervalos 0,05 y se forma un arreglo como [0,05, 0,1, 0,15, ..., 0,90, 0,95] y se itera sobre cada uno de ellos. Por cada uno de los valores en el arreglo:

1. Se consideran todos los pares de preguntas tomados para realizar la comparación, provenientes de la matriz de co-asociación.
2. Por cada uno de los valores de similaridad, se asigna 1 si son mayores o iguales al umbral, 0 si pasa lo contrario.
3. Si los valores asignados en el paso anterior coinciden con el valor real, se asigna un valor *true* (verdadero), si no coinciden, se asigna *false* (falso).
4. Se calcula la proporción de pares de preguntas asignadas con *true*, es decir, que el valor real coincide con el predicho, y se obtiene la *exactitud* del método.

Elección del umbral correcto III

Ejemplo de comparación de valores reales y predichos para construcción de matrices de confusión:

sequence_id	question_pair_id	question_1	question_2	real	predicted	equal
0	123004	question_10	question_20	1	1	true
1	98776	question_11	question_21	1	1	true
2	14422	question_12	question_22	1	0	false
3	12321	question_13	question_23	1	1	true
4	999	question_14	question_24	0	1	false
5	7448	question_15	question_25	0	0	true
6	69553	question_16	question_26	0	0	true
7	2447	question_17	question_27	1	1	true

Elección del umbral correcto IV

Matriz de confusión obtenida:

		Predicho	
		0	**1**
Real	**0**	0.25	0.125
	1	0.125	0.5

- **Exactitud:** 0,75.

Agenda

Análisis del método propuesto I

Metodología utilizada

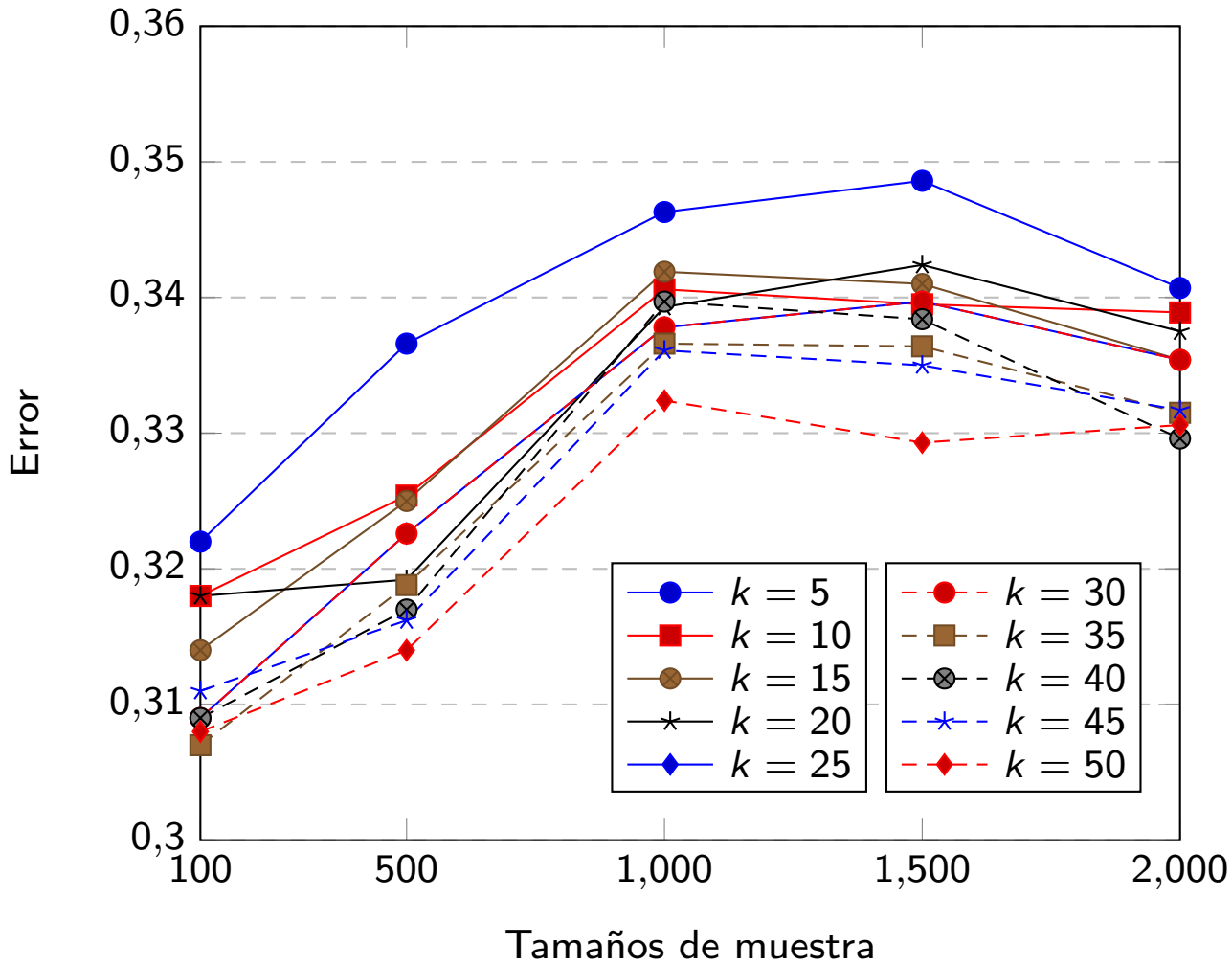
- Se comparan distintas ejecuciones del método EQuAL.
- Se varia el tamaño de muestra en particular: 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Por cada uno de los tamaños de muestra se varia el número de clusters k distinto: 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50.

Análisis del método propuesto II

k / Tam. muestra	100	500	1000	1500	2000	Media	Varianza
5	0.322	0.3366	0.3463	0.3486	0.3407	0.33884	0.0004429
10	0.318	0.3254	0.3406	0.3395	0.3389	0.33248	0.0004162
15	0.314	0.325	0.3419	0.341	0.3354	0.33146	0.0005621
20	0.318	0.3192	0.3393	0.3424	0.3375	0.33128	0.0005489
25	0.309	0.3226	0.3378	0.3397	0.3354	0.3289	0.0006738
30	0.304	0.3218	0.3364	0.3384	0.3353	0.32718	0.0008430
35	0.307	0.3188	0.3366	0.3364	0.3315	0.32606	0.0006635
40	0.309	0.317	0.3397	0.3384	0.3296	0.32674	0.0007216
45	0.311	0.3162	0.3361	0.335	0.3317	0.326	0.000536
50	0.308	0.314	0.3324	0.3293	0.3306	0.32286	0.0004917
Media	0.312	0.32166	0.33871	0.33887	0.33466		
Varianza	0.0003	0.0003736	0.0001299	0.0002258	0.0001248		

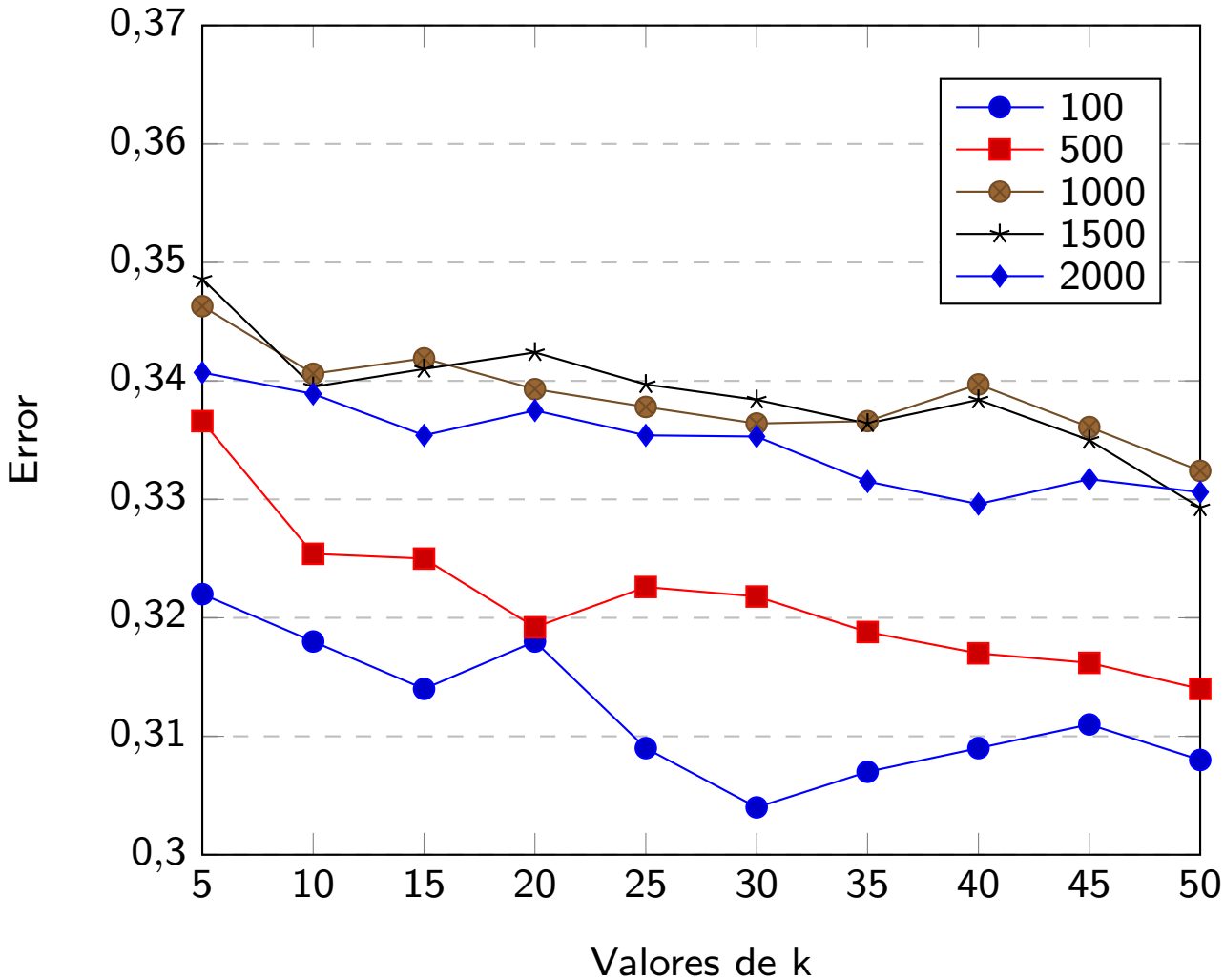
Análisis del método propuesto III

Errores de los valores de k para los distintos tamaños de muestra:



Análisis del método propuesto IV

Errores de los distintos tamaños de muestra para los valores de k.



Análisis del método propuesto y algoritmos del estado del arte I

Metodología utilizada

- Se realizaron 10 ejecuciones de una muestra aleatoria para cada uno de los algoritmos del estado del arte.
- Se varia el tamaño de muestra en particular: 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Por cada uno de los tamaños de muestra se varia el número de clusters k distinto: 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50.

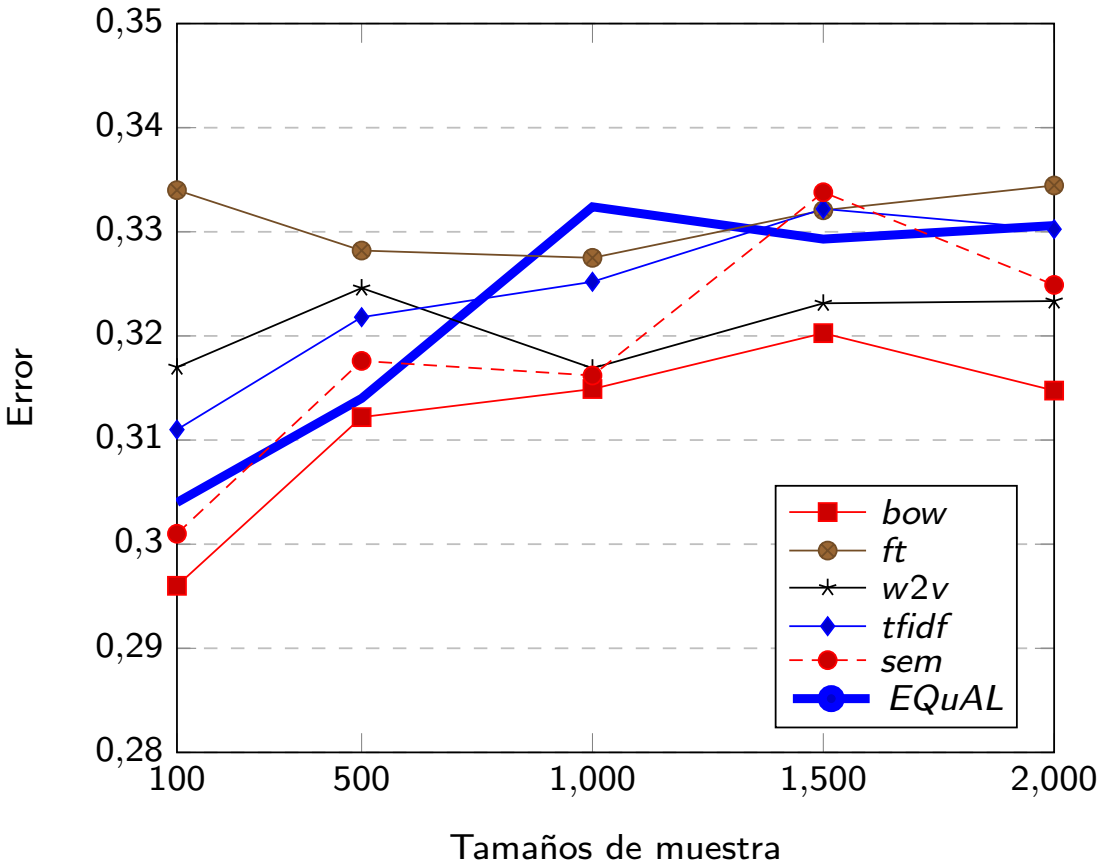
Análisis del método propuesto y algoritmos del estado del arte II

Error en los algoritmos del estado del arte vs. método EQuAL por tamaño de muestra, media y varianza.

	100	500	1000	1500	2000	Media	Varianza
bow	0.296	0.3122	0.3149	0.3202667	0.31475	0.3116233	0.0003396
ft	0.334	0.3282	0.3275	0.3320667	0.33445	0.3312433	0.0000418
w2v	0.317	0.3246	0.3169	0.3231333	0.32335	0.3209967	0.0000558
gtfidf	0.311	0.3218	0.3252	0.3322	0.33025	0.32409	0.0002815
sem	0.301	0.3176	0.3162	0.3338	0.3249	0.3187	0.0005872
EQuAL	0.308	0.314	0.3324	0.3293	0.3306	0.32286	0.0004917

Análisis del método propuesto y algoritmos del estado del arte III

Errores de los tamaños de muestra para el método EQuAL y los algoritmos del estado del arte.



Otras observaciones de interés I

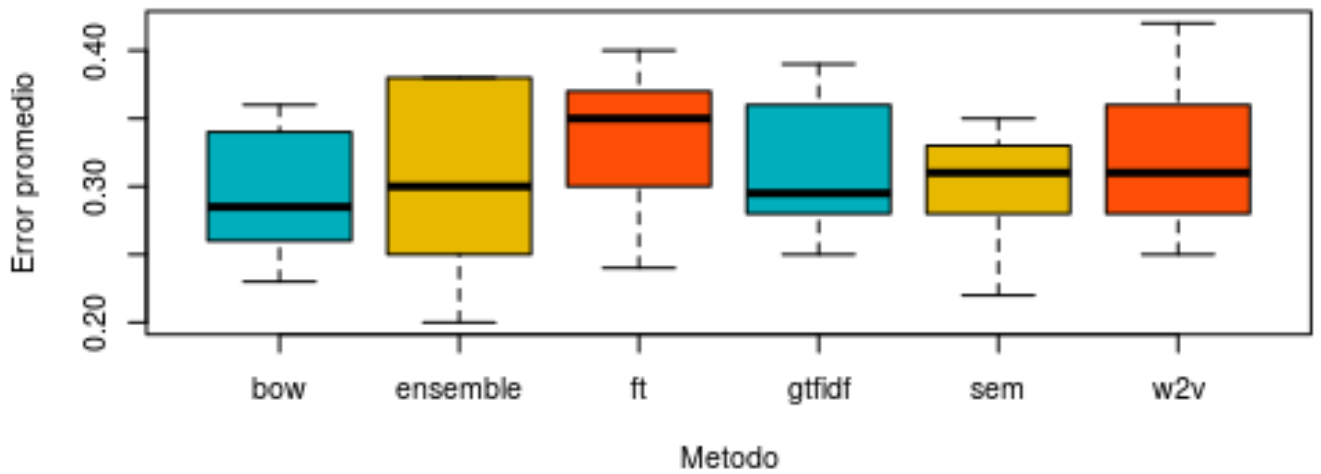
Análisis de varianza del método propuesto.

Se denomina μ_0 a la esperanza de los errores del método EQuAL y se denominan $\mu_i, i = 1, ..., 5$ a las esperanzas de los errores de los métodos, TF, TF-IDF, FastText, Word2Vec y Semantic Distance, respectivamente. Se plantean las siguientes hipótesis:

- H_0: $\mu_0 - \mu_i = 0, i = 1, ..., 5.$
- H_1: $\mu_0 - \mu_i \neq 0, i = 1, ..., 5.$

Otras observaciones de interés II

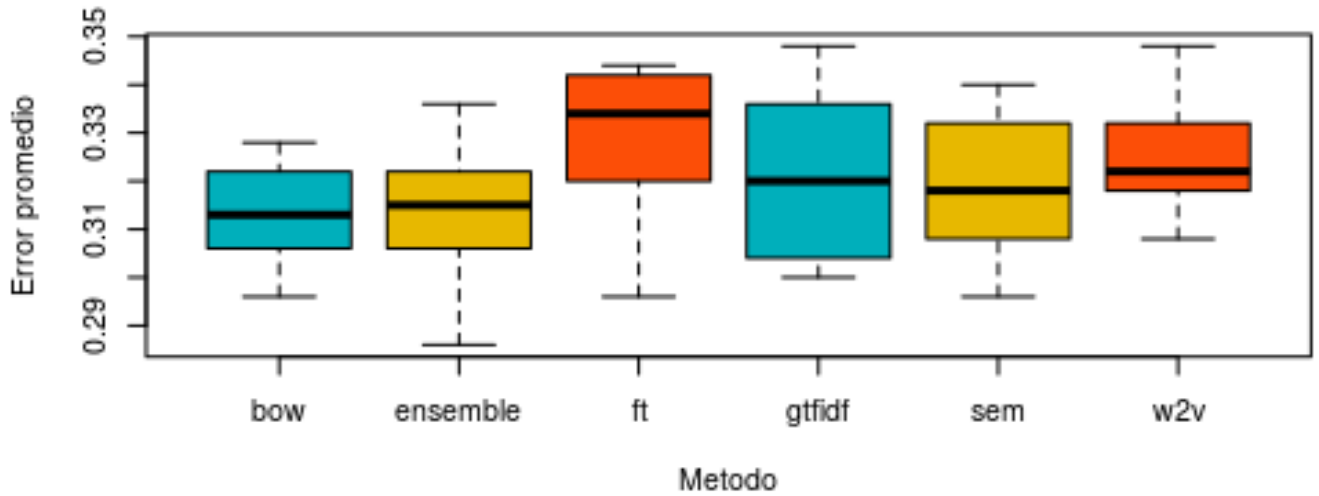
Tamaño de muestra de 100 pares de preguntas.



- En todos los casos, los intervalos de confianza incluyen al valor 0 y $p-adj > \alpha$.
- El método EQuAL posee una media de error que no posee diferencias significativas a todos los métodos.

Otras observaciones de interés III

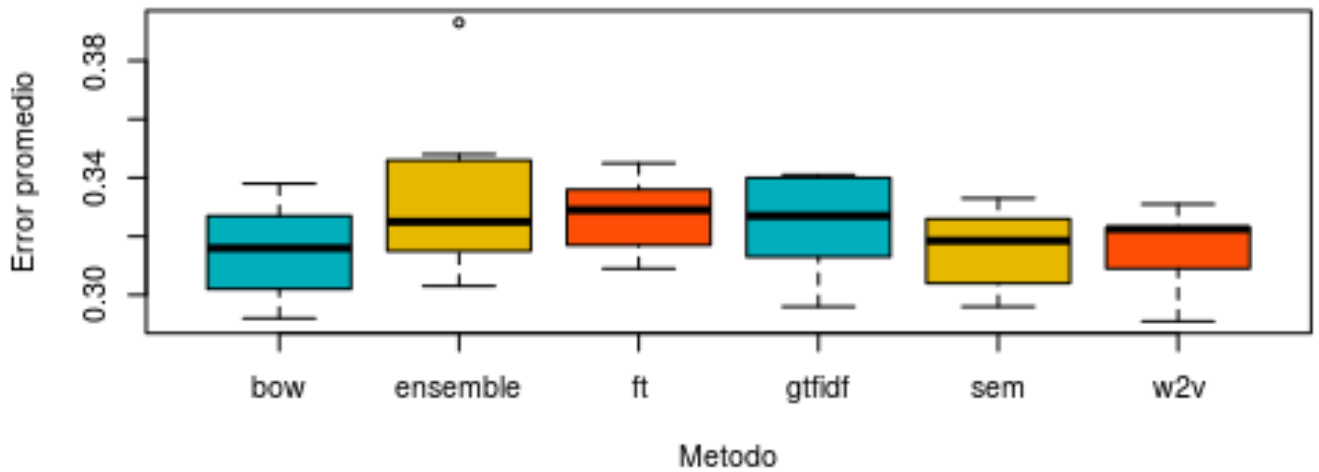
Tamaño de muestra de 500 pares de preguntas.



- En todos los casos, los intervalos de confianza incluyen al valor 0 y $p-adj > \alpha$.
- El método EQuAL posee una media de error que no posee diferencias significativas a todos los métodos.

Otras observaciones de interés IV

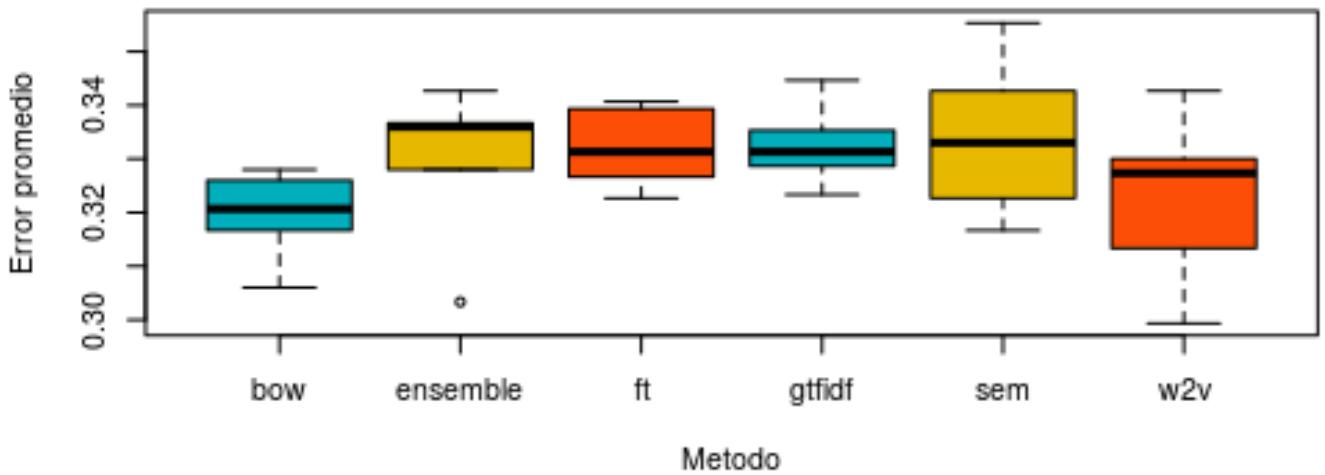
Tamaño de muestra de 1000 pares de preguntas.



- En todos los casos, los intervalos de confianza incluyen al valor 0 y $p-adj > \alpha$.
- El método EQuAL posee una media de error que no posee diferencias significativas a todos los métodos.

Otras observaciones de interés V

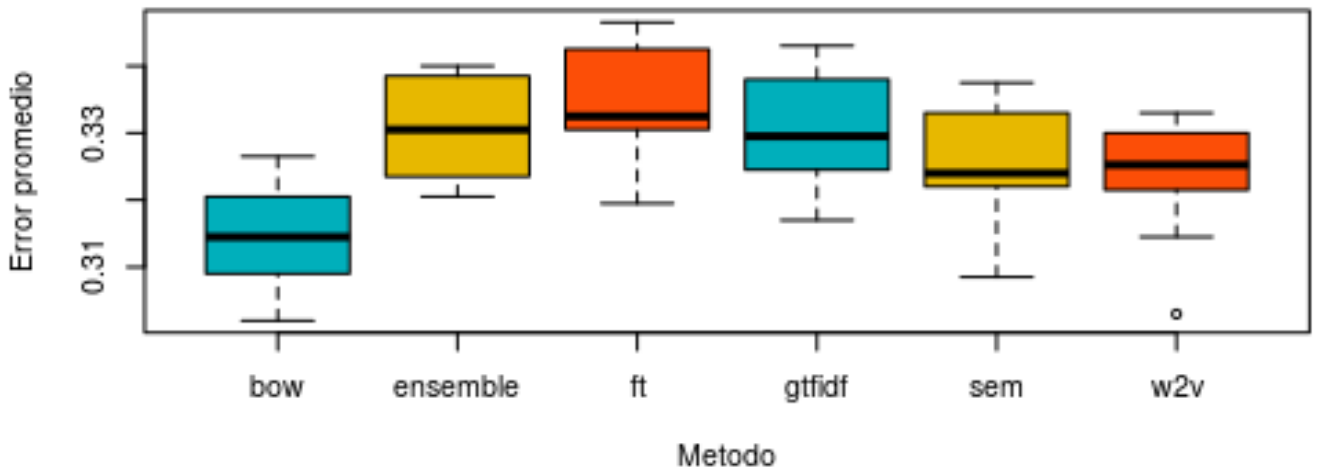
Tamaño de muestra de 1500 pares de preguntas.



- En todos los casos, los intervalos de confianza incluyen al valor 0 y $p-adj > \alpha$.
- El método EQuAL posee una media de error que no posee diferencias significativas a todos los métodos.

Otras observaciones de interés VI

Tamaño de muestra de 2000 pares de preguntas.



- El intervalo de confianza contra el método bow no incluye al cero, ya que este método tuvo muy buenos indicadores en este tamaño de muestra.
- En el resto de los casos, los intervalos de confianza incluyen al valor 0 y $p-adj > \alpha$.

Otras observaciones de interés VII

Resumen de resultados del análisis de varianza

- Se realizaron 25 intervalos de confianza, y en solo uno el método EQuAL se obtuvo una media de error más alta, ya que el método bow fue significativamente mejor al resto.
- El método EQuAL tiene un buen comportamiento en cuanto a medias de error a lo largo de todos los tamaños de muestra.
- Las esperanzas de error no tienen diferencias significativas con los métodos del estado del arte.
- Se puede concluir que el método EQuAL es apto para su implementación en RS.

Otras observaciones de interés VIII

Otras observaciones

- El agregado de variabilidad de datos puede ser influyente en tamaños de muestra pequeños.
- El método EQuAL es dependiente a los métodos subyacentes.
- Debido a la naturaleza del conjunto de datos utilizado, no es posible identificar fácilmente cuál es la forma de los clusters en cuestión, y verificar si el método EQuAL se adapta perfectamente al conjunto de datos.

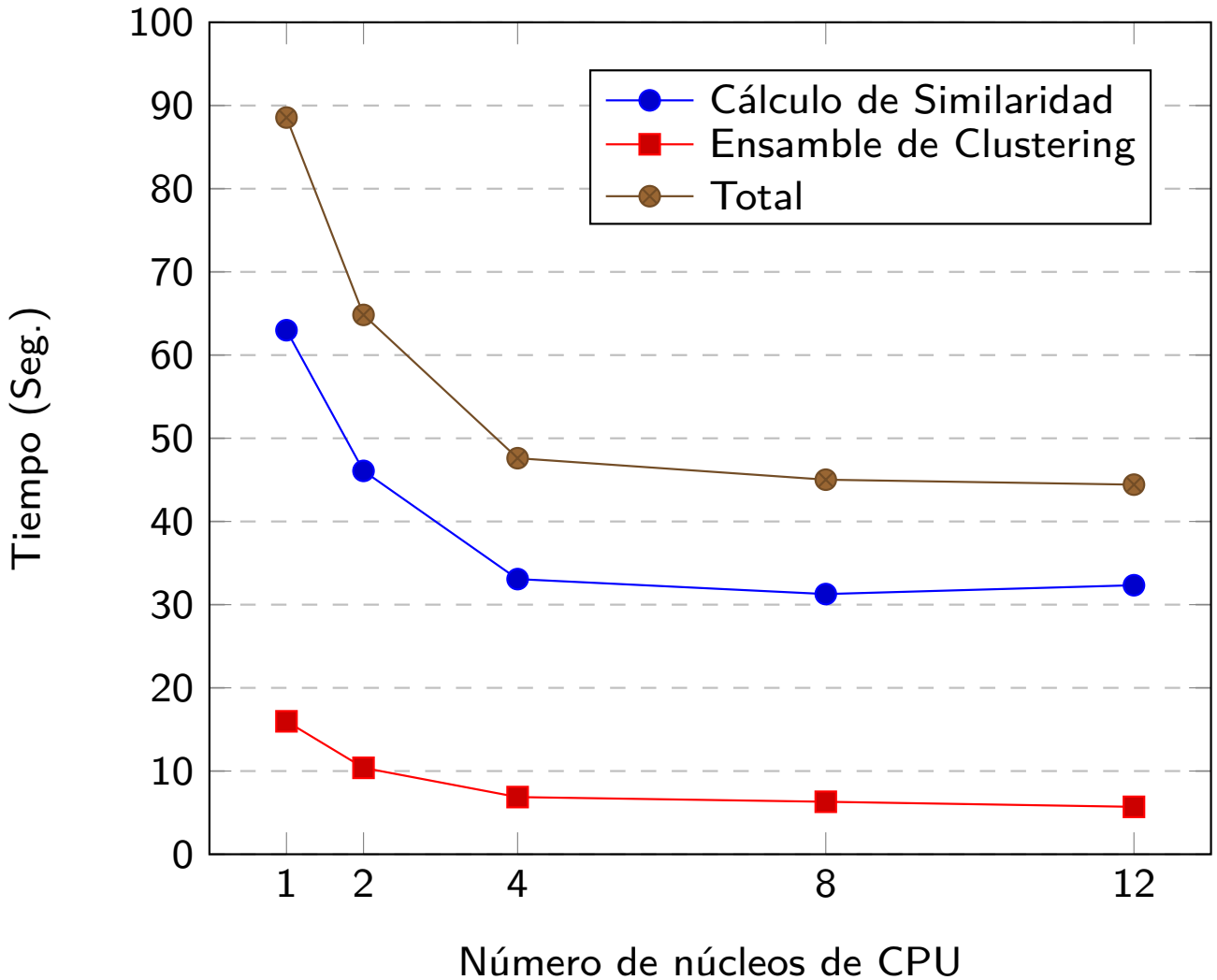
Análisis de desempeño I

Se realizó un **análisis de desempeño** con las siguientes características:

- Cluster Hadoop en localhost.
- Tamaños de muestra 100, 500 y 1000 pares de preguntas.
- En cada una de las ejecuciones, utilizan dos técnicas de similaridad (TF y TFIDF), para luego ensamblarlas.
- Cantidad de núcleos CPU asignados: 1, 2, 4, 8 y 12.

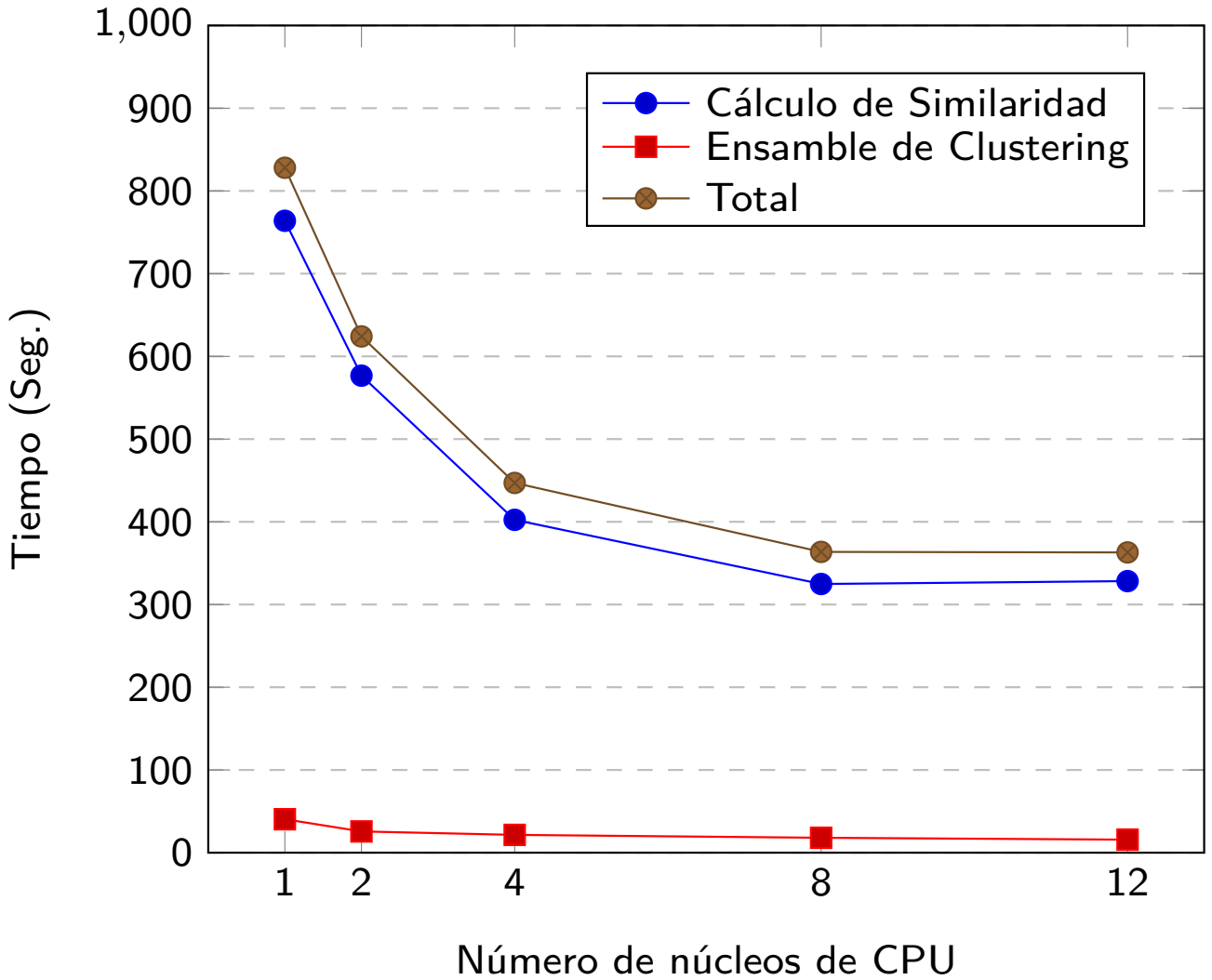
Análisis de desempeño II

Tamaño de muestra de 100 pares de preguntas y distintos núcleos de CPU.



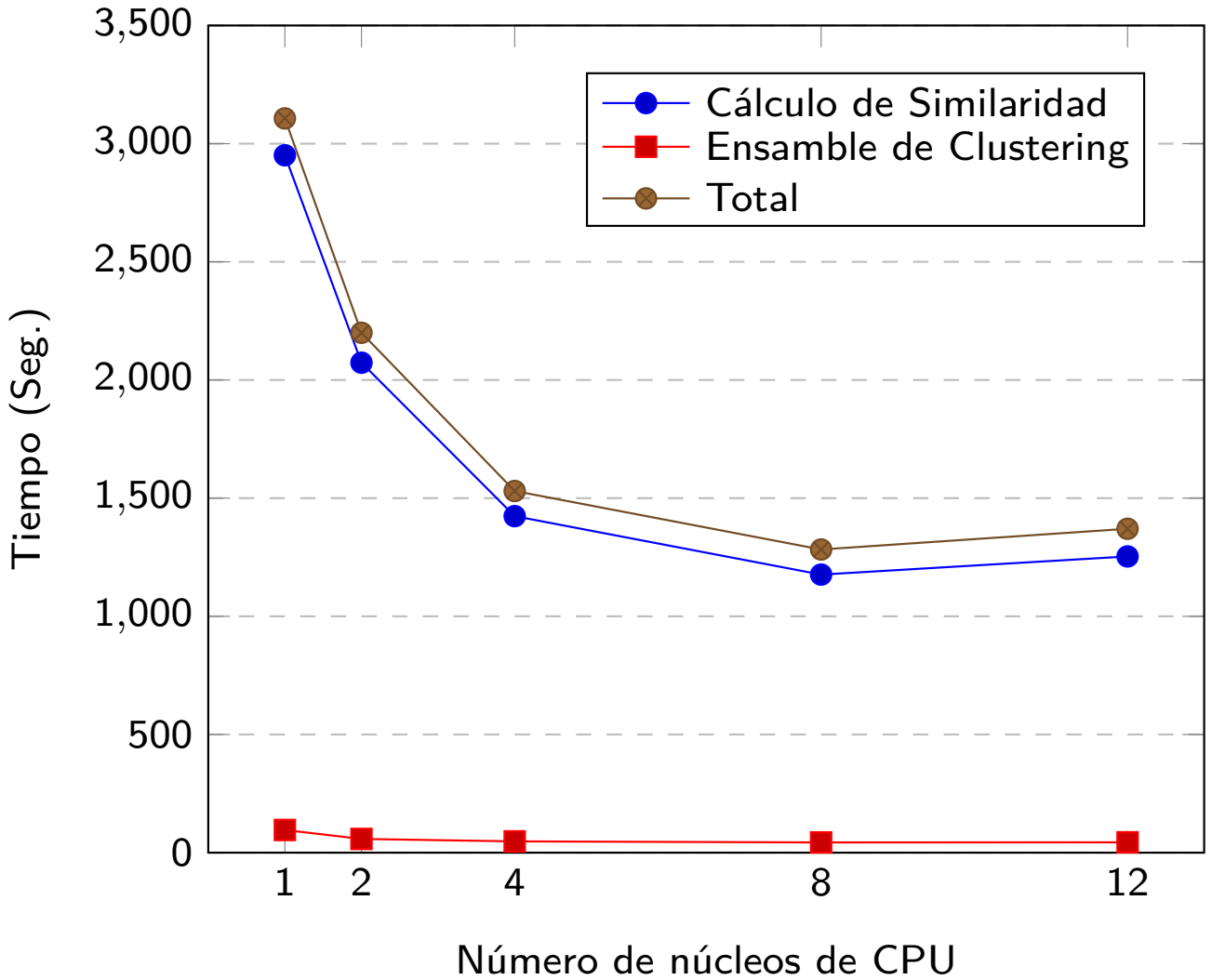
Análisis de desempeño III

Tamaño de muestra de 500 pares de preguntas y distintos núcleos de CPU.



Análisis de desempeño IV

Tamaño de muestra de 1000 pares de preguntas y distintos núcleos de CPU.



Resumen de resultados

- El método EQuAL tuvo buen rendimiento con tamaños pequeños de muestras y con un alto número de clusters.

Resumen de resultados

- El método EQuAL tuvo buen rendimiento con tamaños pequeños de muestras y con un alto número de clusters.
- Comparando el método EQuAL con los algoritmos del estado del arte, se concluye que posee indicadores aptos para su aplicación en RS, en cuanto a medias de error y varianza.

Resumen de resultados

- El método EQuAL tuvo buen rendimiento con tamaños pequeños de muestras y con un alto número de clusters.
- Comparando el método EQuAL con los algoritmos del estado del arte, se concluye que posee indicadores aptos para su aplicación en RS, en cuanto a medias de error y varianza.
- Es altamente probable que el método EQuAL arroje buenos resultados si los algoritmos subyacentes también lo hacen, y viceversa.

Resumen de resultados

- El método EQuAL tuvo buen rendimiento con tamaños pequeños de muestras y con un alto número de clusters.
- Comparando el método EQuAL con los algoritmos del estado del arte, se concluye que posee indicadores aptos para su aplicación en RS, en cuanto a medias de error y varianza.
- Es altamente probable que el método EQuAL arroje buenos resultados si los algoritmos subyacentes también lo hacen, y viceversa.
- Es posible adaptar el método al conjunto de datos y elegir los algoritmos subyacentes adecuados.

Resumen de resultados

- El método EQuAL tuvo buen rendimiento con tamaños pequeños de muestras y con un alto número de clusters.
- Comparando el método EQuAL con los algoritmos del estado del arte, se concluye que posee indicadores aptos para su aplicación en RS, en cuanto a medias de error y varianza.
- Es altamente probable que el método EQuAL arroje buenos resultados si los algoritmos subyacentes también lo hacen, y viceversa.
- Es posible adaptar el método al conjunto de datos y elegir los algoritmos subyacentes adecuados.
- Se desarrolló una arquitectura de software con enfoque Big Data que realiza los cálculos de similaridad y procesamiento del ensamble de clustering de manera escalable y adaptable.

Agenda

Contribuciones realizadas

- Se diseñó un método que utiliza una medida de similaridad de texto confiable y efectiva entre preguntas de un sitio de CQA.

Contribuciones realizadas

- Se diseñó un método que utiliza una medida de similaridad de texto confiable y efectiva entre preguntas de un sitio de CQA.

- Se diseñó y desarolló una arquitectura de software de procesamiento distribuido con un enfoque Big Data.

Futuras investigaciones

El presente trabajo sirve como estado del arte para las siguientes lineas de investigación/desarrollo:

- Continuar con el desarrollo para lograr un RS operativo en su totalidad utilizando el método propuesto basado en ensamble de clustering y similaridad entre ítems.

Futuras investigaciones

El presente trabajo sirve como estado del arte para las siguientes lineas de investigación/desarrollo:

- Continuar con el desarrollo para lograr un RS operativo en su totalidad utilizando el método propuesto basado en ensamble de clustering y similaridad entre ítems.
- Elaborar una arquitectura Big Data adaptable que mejore y optimice el funcionamiento de algunos aspectos.

Futuras investigaciones

El presente trabajo sirve como estado del arte para las siguientes lineas de investigación/desarrollo:

- Continuar con el desarrollo para lograr un RS operativo en su totalidad utilizando el método propuesto basado en ensamble de clustering y similaridad entre ítems.
- Elaborar una arquitectura Big Data adaptable que mejore y optimice el funcionamiento de algunos aspectos.
- Utilizar los resultados obtenidos en otros tipos de sitios donde se puedan aplicar RS basados en texto, tales como sitios de e-commerce, portales académicos o redes sociales.

Futuras investigaciones

El presente trabajo sirve como estado del arte para las siguientes lineas de investigación/desarrollo:

- Continuar con el desarrollo para lograr un RS operativo en su totalidad utilizando el método propuesto basado en ensamble de clustering y similaridad entre ítems.
- Elaborar una arquitectura Big Data adaptable que mejore y optimice el funcionamiento de algunos aspectos.
- Utilizar los resultados obtenidos en otros tipos de sitios donde se puedan aplicar RS basados en texto, tales como sitios de e-commerce, portales académicos o redes sociales.
- Continuar el desarrollo para crear un framework adaptable a distintas técnicas de distancias de texto.

Futuras investigaciones

El presente trabajo sirve como estado del arte para las siguientes lineas de investigación/desarrollo:

- Continuar con el desarrollo para lograr un RS operativo en su totalidad utilizando el método propuesto basado en ensamble de clustering y similaridad entre ítems.
- Elaborar una arquitectura Big Data adaptable que mejore y optimice el funcionamiento de algunos aspectos.
- Utilizar los resultados obtenidos en otros tipos de sitios donde se puedan aplicar RS basados en texto, tales como sitios de e-commerce, portales académicos o redes sociales.
- Continuar el desarrollo para crear un framework adaptable a distintas técnicas de distancias de texto.
- Crear y estructurar información para policy makers e instituciones de ciencia, tecnología, innovación y desarrollo con el objetivo de construir insumos para el diseño, implementación, ejecución y evaluación de políticas públicas y educativas.

¡Muchas gracias!