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

23 de Noviembre de 2021

Agenda

1. Introducción
2. Fundamentación
3. Marco teórico
4. Problema de investigación y propuesta
5. Experimentos
6. Resultados
7. Conclusiones

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.

Área temática II

Area temática

- Miles de nuevas preguntas son formuladas diariamente en sitios de CQA como Yahoo! Answers, Stackexchange, Stackoverflow, o Quora.

Área temática II

Area temática

- Miles de nuevas preguntas son formuladas diariamente en sitios de CQA como Yahoo! Answers, Stackexchange, Stackoverflow, o Quora.
- Muchas de las preguntas no están respondidas correctamente o no tienen respuestas.
  - Pequeño número de expertos entre la gran población de usuarios.
  - La pregunta es difícil de ubicar dentro del sitio.
  - Respuestas maliciosas.

Área temática II

Area temática

- Miles de nuevas preguntas son formuladas diariamente en sitios de CQA como Yahoo! Answers, Stackexchange, Stackoverflow, o Quora.
- Muchas de las preguntas no están respondidas correctamente o no tienen respuestas.
  - Pequeño número de expertos entre la gran población de usuarios.
  - La pregunta es difícil de ubicar dentro del sitio.
  - Respuestas maliciosas.
- Es de interés buscar si esa misma pregunta ha sido formulada por otro usuario previamente, y que tenga la respuesta buscada.

Área temática II

Area temática

- Miles de nuevas preguntas son formuladas diariamente en sitios de CQA como Yahoo! Answers, Stackexchange, Stackoverflow, o Quora.
- Muchas de las preguntas no están respondidas correctamente o no tienen respuestas.
  - Pequeño número de expertos entre la gran población de usuarios.
  - La pregunta es difícil de ubicar dentro del sitio.
  - Respuestas maliciosas.
- Es de interés buscar si esa misma pregunta ha sido formulada por otro usuario previamente, y que tenga la respuesta buscada.
- Preguntas que poseen la misma respuestas están formuladas de forma diferente en el sentido léxico.

  ¿Cómo elijo una revista para publicar mi artículo? y *¿Dónde publico mi artículo?*

Área temática III

Area temática (cont.)

- Es necesaria una medida de similaridad que tenga en cuenta características léxicas y semánticas.

Área temática III

Area temática (cont.)

- Es necesaria una medida de similaridad que tenga en cuenta características léxicas y semánticas.
- La tarea de recomendar preguntas similares en sitios de CQA puede ser llevada a cabo por un RS.

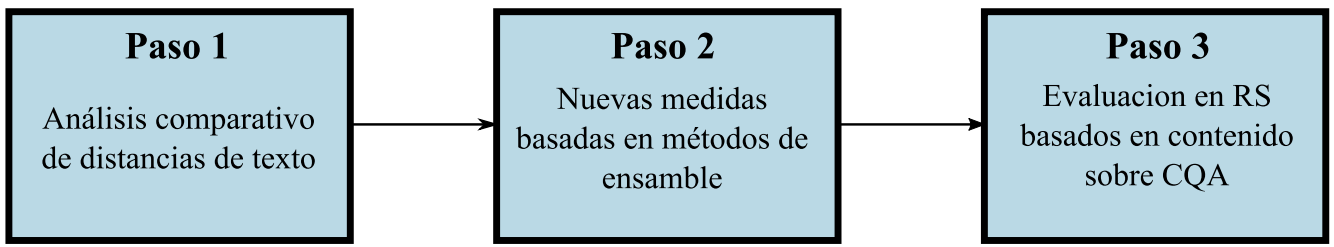
Área temática III

Area temática (cont.)

- Es necesaria una medida de similaridad que tenga en cuenta características léxicas y semánticas.
- La tarea de recomendar preguntas similares en sitios de CQA puede ser llevada a cabo por un RS.
- Se diseñó e implementó una arquitectura Big Data para crear una medida de similaridad que alimente a un RS para sitios de CQA.

Tema específico I

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



Este trabajo se enfoca en el **Paso 2** del pipeline.

Tema específico II

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

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

Esta situación plantea la necesidad de considerar una arquitectura Big Data.

Objetivo general

Objetivo general

El presente trabajo de investigación tiene como objetivo construir una **arquitectura Big Data** que incluye la posibilidad de ser aplicada a grandes conjuntos de datos en el ámbito de **CQA** y, a partir de esta arquitectura, implementar, evaluar y realizar un **análisis comparativo con el estado del arte de una nueva medida de similaridad entre textos** que pueda ser utilizada en **Sistemas de Recomendación**.

Objetivos específicos

Objetivos específicos

- Diseñar y desarrollar una **arquitectura Big Data** para cálculo de similaridad en grandes volúmenes de datos.

Objetivos específicos

- Diseñar y desarrollar una **arquitectura Big Data** para cálculo de similaridad en 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

Objetivos específicos

- Diseñar y desarrollar una **arquitectura Big Data** para cálculo de similaridad en 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.
- **Proponer una nueva medida** que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data.

Objetivos específicos

Objetivos específicos

- Diseñar y desarrollar una **arquitectura Big Data** para cálculo de similaridad en 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.
- **Proponer una nueva medida** que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data.
- **Evaluar el comportamiento** de una medida de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos.

Objetivos específicos

Objetivos específicos

- Diseñar y desarrollar una **arquitectura Big Data** para cálculo de similaridad en 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.
- **Proponer una nueva medida** que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data.
- **Evaluar el comportamiento** de una medida de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos.
- **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

- Encontrar la medida de representación adecuada para alimentar un RS de calidad.

Motivación de la tesis

Motivación

- Encontrar la medida de representación adecuada para alimentar un RS de calidad.
- Las medidas de similaridad del estado del arte tienen conocidos problemas (invariantes respecto al orden, consideran solo características sintácticas, bajo desempeño en términos de velocidad).

Motivación de la tesis

Motivación

- Encontrar la medida de representación adecuada para alimentar un RS de calidad.
- Las medidas de similaridad del estado del arte tienen conocidos problemas (invariantes respecto al orden, consideran solo características sintácticas, bajo desempeño en términos de velocidad).
- Creación de un método novedoso que combine medidas de similaridad existentes y pueda aplicarse en Sistemas de Recomendación.

Motivación de la tesis

Motivación

- Encontrar la medida de representación adecuada para alimentar un RS de calidad.
- Las medidas de similaridad del estado del arte tienen conocidos problemas (invariantes respecto al orden, consideran solo características sintácticas, bajo desempeño en términos de velocidad).
- Creación de un método novedoso que combine medidas de similaridad existentes y pueda aplicarse en Sistemas de Recomendación.
- Arquitectura de software que soporte el procesamiento del método propuesto de una forma eficiente y escalable.

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.

Sistemas de Recomendación

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

- **Ejemplos**
  - Recomendación de preguntas similares en el sitio web Quora.
  - Sitios conocidos como Netflix o Amazon.

Sistemas de Recomendación

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

- **Ejemplos**
  - Recomendación de preguntas similares en el sitio web Quora.
  - Sitios conocidos como Netflix o Amazon.

- **Funciones de un Sistema de Recomendación**
  - Aumentar el ratio de conversión en un sitio o aplicación.
    - Aumentar la cantidad de productos que usuario compra en Amazon, sobre la cantidad de visualizados.
  - Aumentar satisfacción y fidelidad del usuario.

Big Data

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).

Se necesitan **herramientas de software** que nos permitan procesar este volumen de información de manera eficiente y eficaz.

Similaridad

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

- Se utilizan dos tipos de medidas de similaridad en este trabajo:
  1. Basadas en espacios vectoriales.
  2. Basadas en taxonomías.

Similaridad

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

- Se utilizan dos tipos de medidas de similaridad en este trabajo:
  1. Basadas en espacios vectoriales.
  2. Basadas en taxonomías.

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

  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

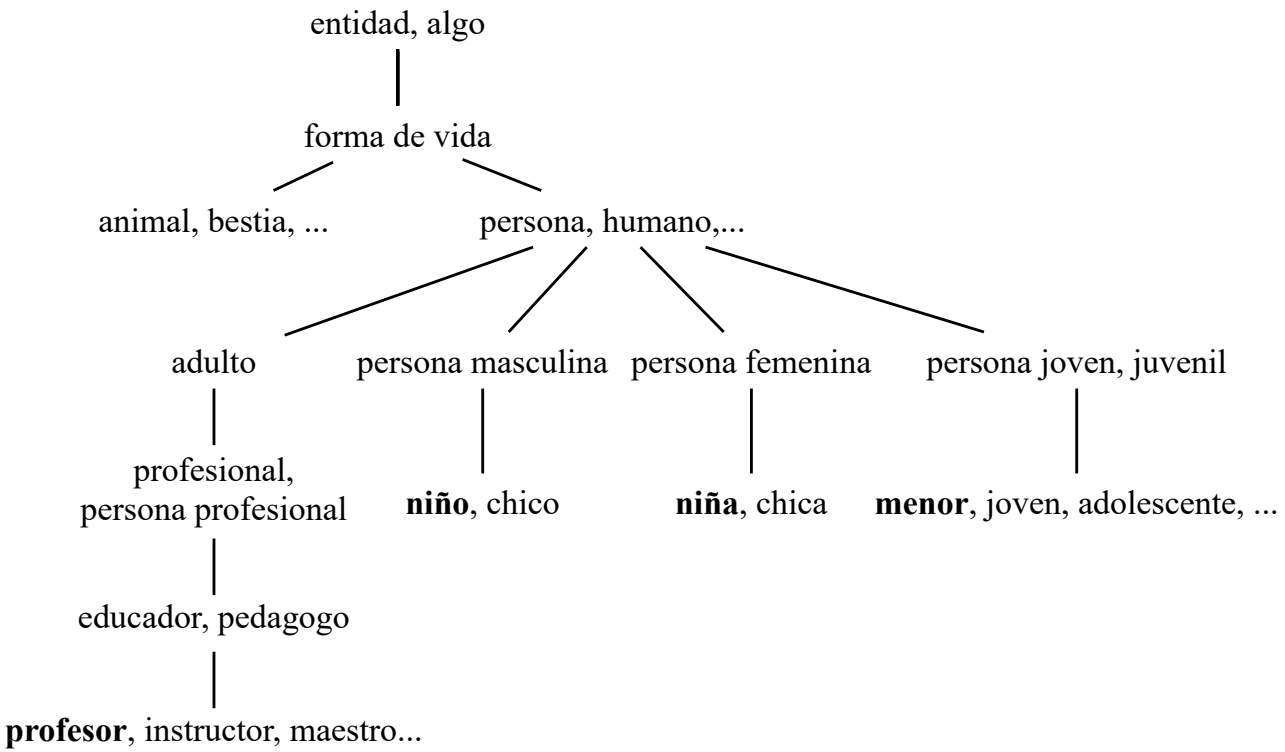
En el modelo de *espacio vectorial*, un texto es representado como un vector de términos. Si las palabras son elegidas como términos, entonces cada palabra del vocabulario sería una *dimensión* independiente en el espacio vectorial (Singhal et al., 2001).

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.

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

siendo θ el ángulo entre los vectores $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$.

Similaridad en taxonomías I



La similaridad entre palabras se define como $S(w_1, w_2) = f(l, h)$, donde l es el camino más corto entre w_1 y w_2, y h es la profundidad del subsumidor de las mismas.

Similaridad en taxonomías II

Se define $p(t)$ como la **probabilidad** de un concepto t. Se entiende como:

$$p(t) = \frac{freq(t)}{N},$$

Se define el **contenido de información** $I(t)$ como:

$$I(t) = -\log p(t).$$

Entonces la **similaridad** entre dos términos (i, j) se define como:

$$S_R = I(ms(t_i, t_j)),\ (Resnik,\ 1995)$$

$$S_L(t_i, t_j) = \frac{2S_R(t_i, t_j)}{I(t_i) + I(t_j)}.\ (Lin\ et\ al.,\ 1998)$$

Medidas de Similaridad

Medidas de similaridad utilizadas

- Term Frequency (TF)
- Term Frequency - Inverse Document 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).
- Cada documento corresponde a un vector y cada término a una dimensión.
- El orden exacto de los términos es ignorado, pero se basa en el número de ocurrencias de cada uno de ellos en un documento.
- Se mide el grado de similaridad de dos documentos utilizando el coseno del ángulo entre dos vectores.

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

("Mary es más rápida que John" y "John es más rápido que 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: 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

Características de Word2Vec:

- Modelos basados en redes neuronales con una capa oculta para computar representaciones de palabras como vectores.
- 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 las palabras de un vocabulario.

FastText

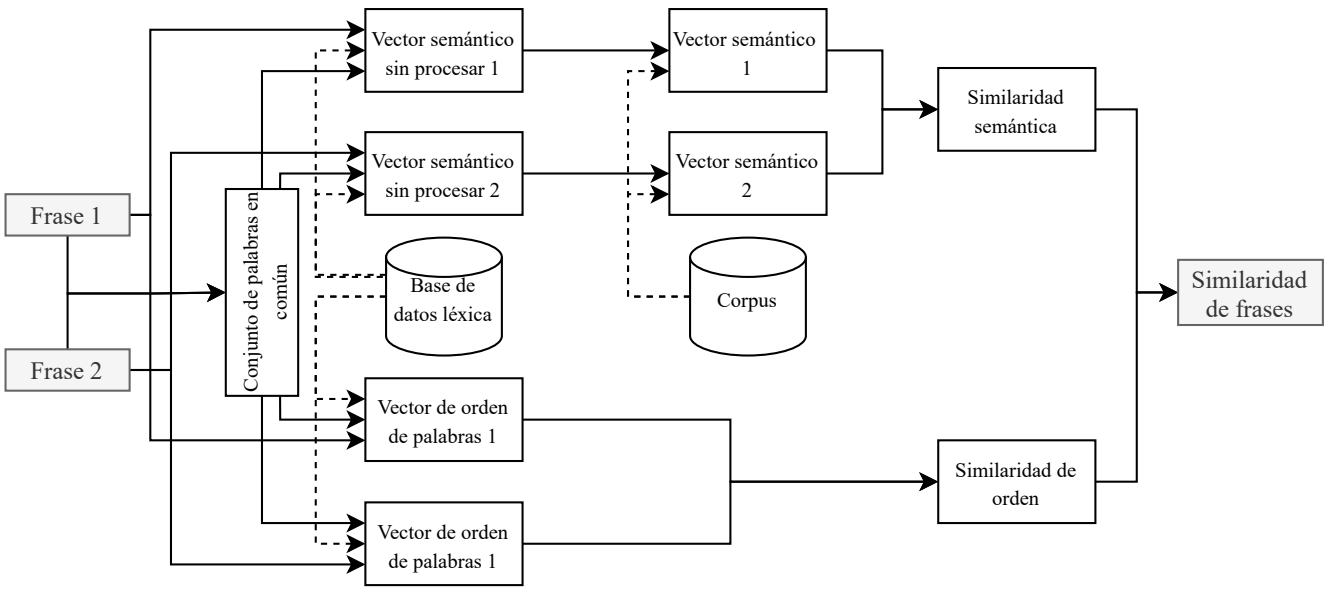
Caracteristicas de FastText:

- Librería open-source desarrollada por Facebook.
- Utiliza 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.



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 clusters de distintas formas.
  - Distintos algoritmos de Clustering.
  - Distintos parámetros para el mismo algoritmo.
  - Distintas medidas de distancia.

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 clusters de distintas formas.
  - Distintos algoritmos de Clustering.
  - Distintos parámetros para el mismo algoritmo.
  - Distintas medidas de distancia.
- 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 clusters de distintas formas.
  - Distintos algoritmos de Clustering.
  - Distintos parámetros para el mismo algoritmo.
  - Distintas medidas de distancia.
- Aprovecha la variabilidad agregada para encontrar una estructura *inter-patrón*.
- Identificación de clusters subyacentes con formas, tamaños y densidades arbitrarias.

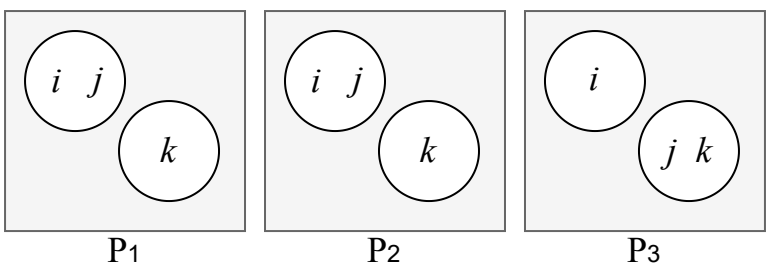
Combinación de Evidencias

Toma la co-ocurrencia de pares de patrones en el mismo cluster a lo largo de las N particiones de datos para n patrones, para luego mapearlas a 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.

Ejemplo. $N = 3$ particiones de datos.



$$C(i,j) = \frac{2}{3} \Rightarrow \begin{bmatrix} n_{11} & \cdots & n_{1n} \\ \vdots & n_{ij} = 2/3 & \\ \vdots & & \ddots \\ n_{n1} & & n_{nn} \end{bmatrix}$$

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.

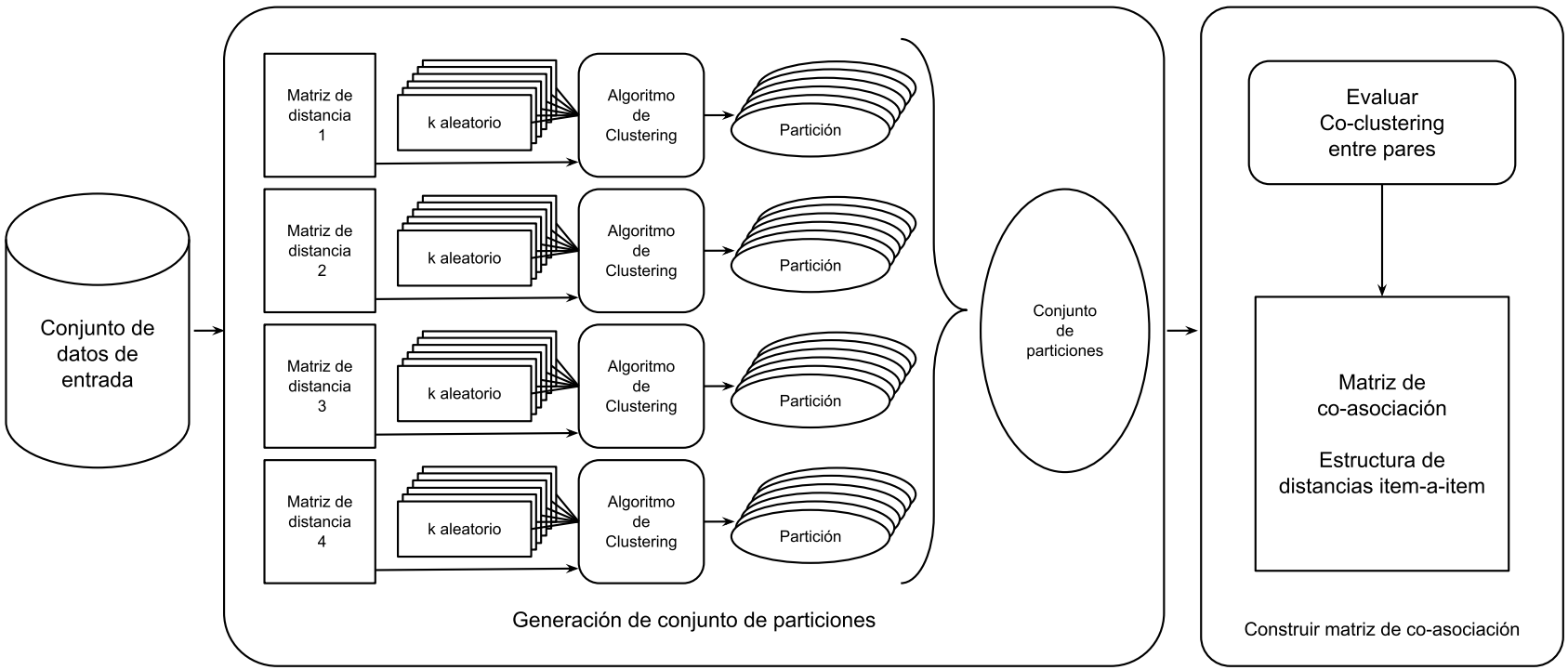
Hipótesis II

Hipótesis del trabajo de tesis (cont.)

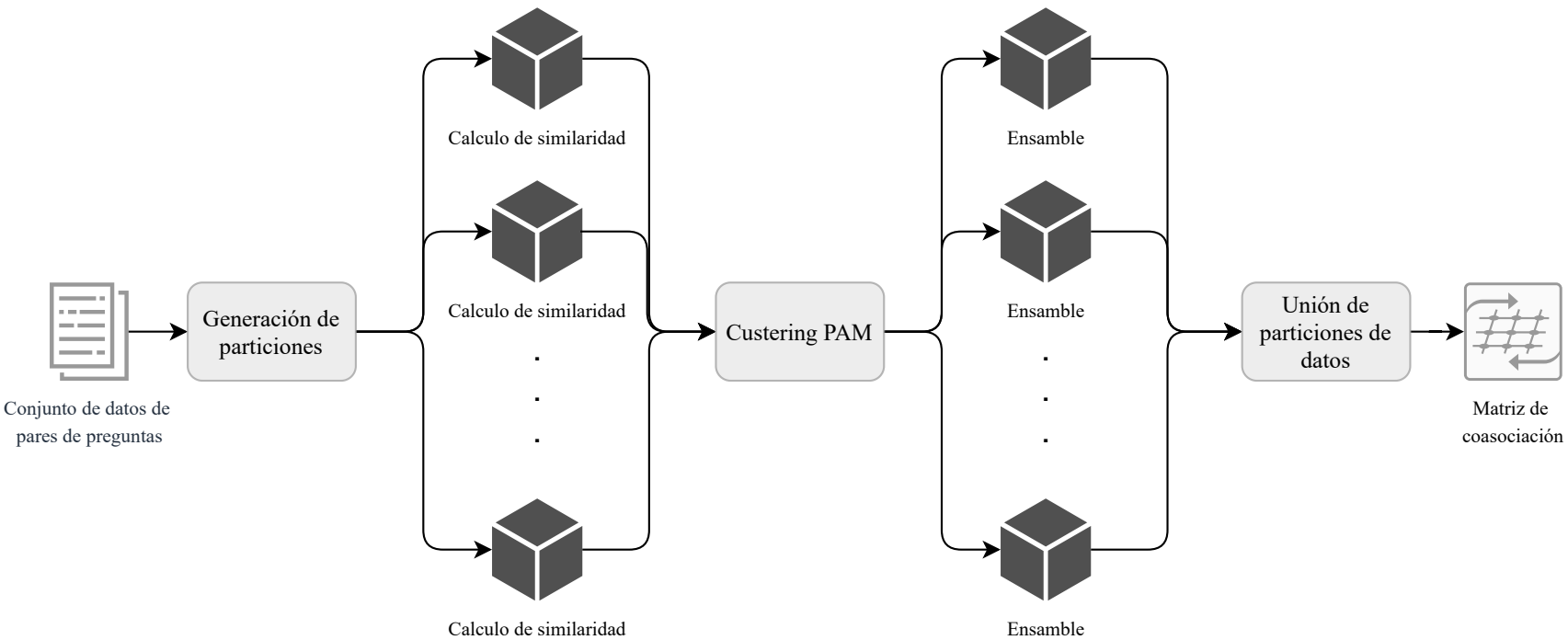
Por tal motivo, y como respuesta a la hipótesis planteada, se presenta:

- Un **desarrollo de un nuevo método de cálculo de similaridad de texto** basado en una **arquitectura Big Data**.
- Una **aplicación del método a un gran conjunto de datos reales** con el fin de verificar la eficiencia y eficacia del procedimiento.
- Un **análisis comparativo** del método presentado con los algoritmos para cálculo de similaridad de texto del estado del arte.

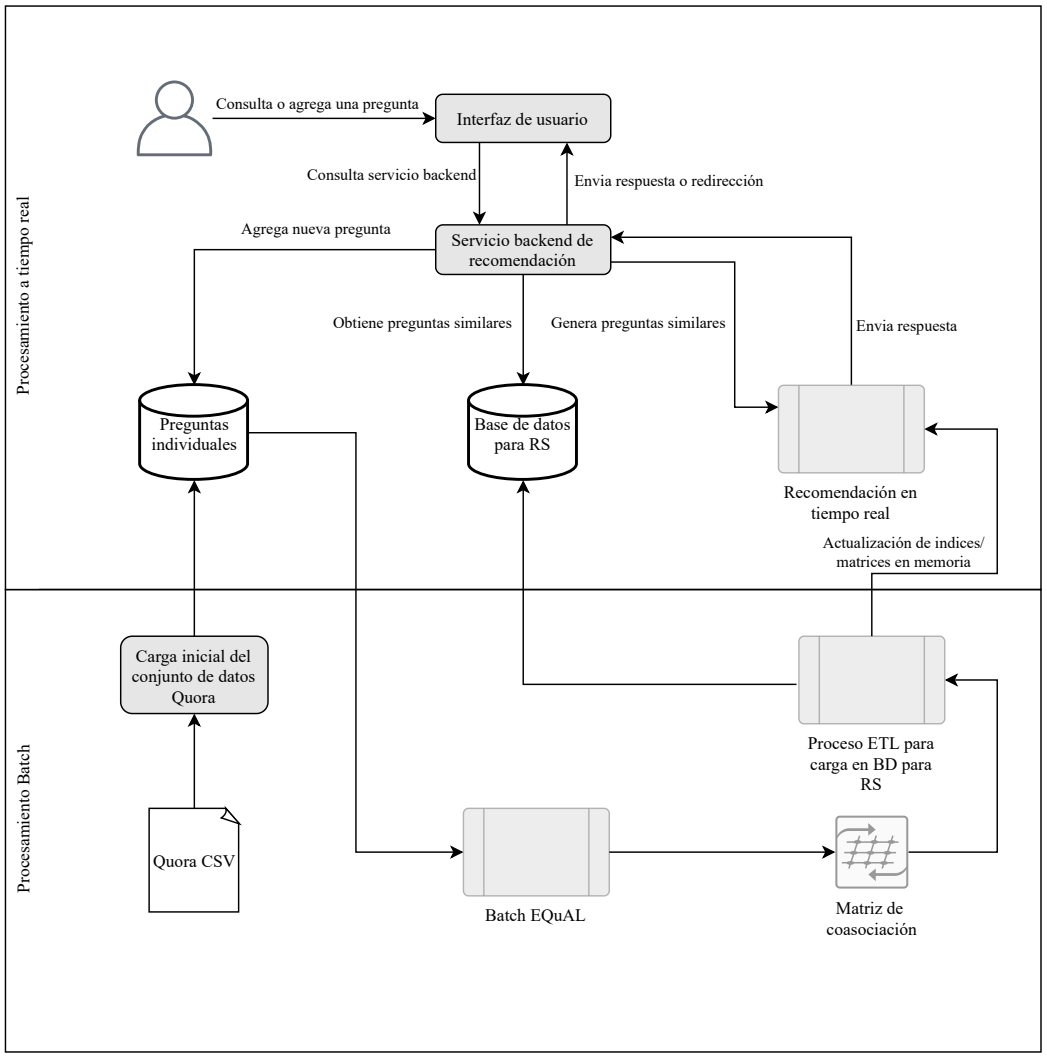
El método propuesto (EQuAL)



Arquitectura de procesamiento de datos



Implementación en un sistema de recomendación de tiempo real



Agenda

Objetivos de experimentación

En este capítulo se describirá:

- **Resultados obtenidos** en el trabajo del estado del arte.
- **Método de experimentación** en forma de detallada.
- **Método de validación y presentación** de resultados.

Los experimentos se realizaron con las siguientes tecnologías:

- Código de experimentación en **Python**, con la utilización de **Apache Spark** y librerías de soporte.
- Validación estadística del método EQuAL en **R**.

Matrices de confusión

Matrices de confusión utilizadas

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

- a = Verdaderos Negativos.
- b = Falsos Positivos.
- c = Falsos Negativos.
- d = Verdaderos Positivos.

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

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

En estos indicadores se cumplen la condiciones $a+b+c+d=1$ y $error = 1 - exactitud$.

Estado del arte

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

Nota: el conjunto de datos original tiene $36,9\,\%$ pares de preguntas de clase 1 y $63,1\,\%$ de clase 0.

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 estadísticamente significativos.

Generación de particiones I

Ejemplo de la estructura de los subconjuntos de muestreo:

question_pair_id	question_id_1	question_id_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 $N \times N$ 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 (Partición Alrededor de Medoides).

Clustering y etiquetado

- Por cada una de las matrices de similaridad se realizan varias ejecuciones de clustering PAM (Partición Alrededor de Medoides).
- 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 (Partición Alrededor de Medoides).
- 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

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

Ensamble de Clustering II

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, por ejemplo:

- pregunta 1 = [**(run_uuid_1,1)**, $(run\_uuid\_2, 1)$, $(run\_uuid\_3, 3)$],
- pregunta 2 = [**(run_uuid_1,1)**, $(run\_uuid\_2, 2)$, $(run\_uuid\_3, 2)$].

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	[]

Ensamble de Clustering III

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 IV

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 I

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

sequence_id	question_pair	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

Método de validación II

Muestra de datos original

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

Se filtra la matriz de co-asociación 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

¿Cómo sabemos cuando este valor de similaridad nos indica que dos preguntas son iguales (1) o distintas (0)?

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

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

Elección del umbral correcto II

Asignación de resultados utilizando un umbral de 0,65:

question_id_1	question_id_2	question_1	question_2	similarity	equal
question_10	question_11	contenido	contenido	0.857	1
question_11	question_21	contenido	contenido	0.368	0

Asignación de resultados utilizando un umbral de 0,9:

question_id_1	question_id_2	question_1	question_2	similarity	equal
question_10	question_11	contenido	contenido	0.857	0
question_11	question_21	contenido	contenido	0.368	0

¿Cómo elegimos los valores de umbral t tienen **mejor rendimiento**?

Proceso iterativo para identificar el valor de **umbral óptimo** $t*$.

Agenda

Análisis del método propuesto I

Metodología utilizada

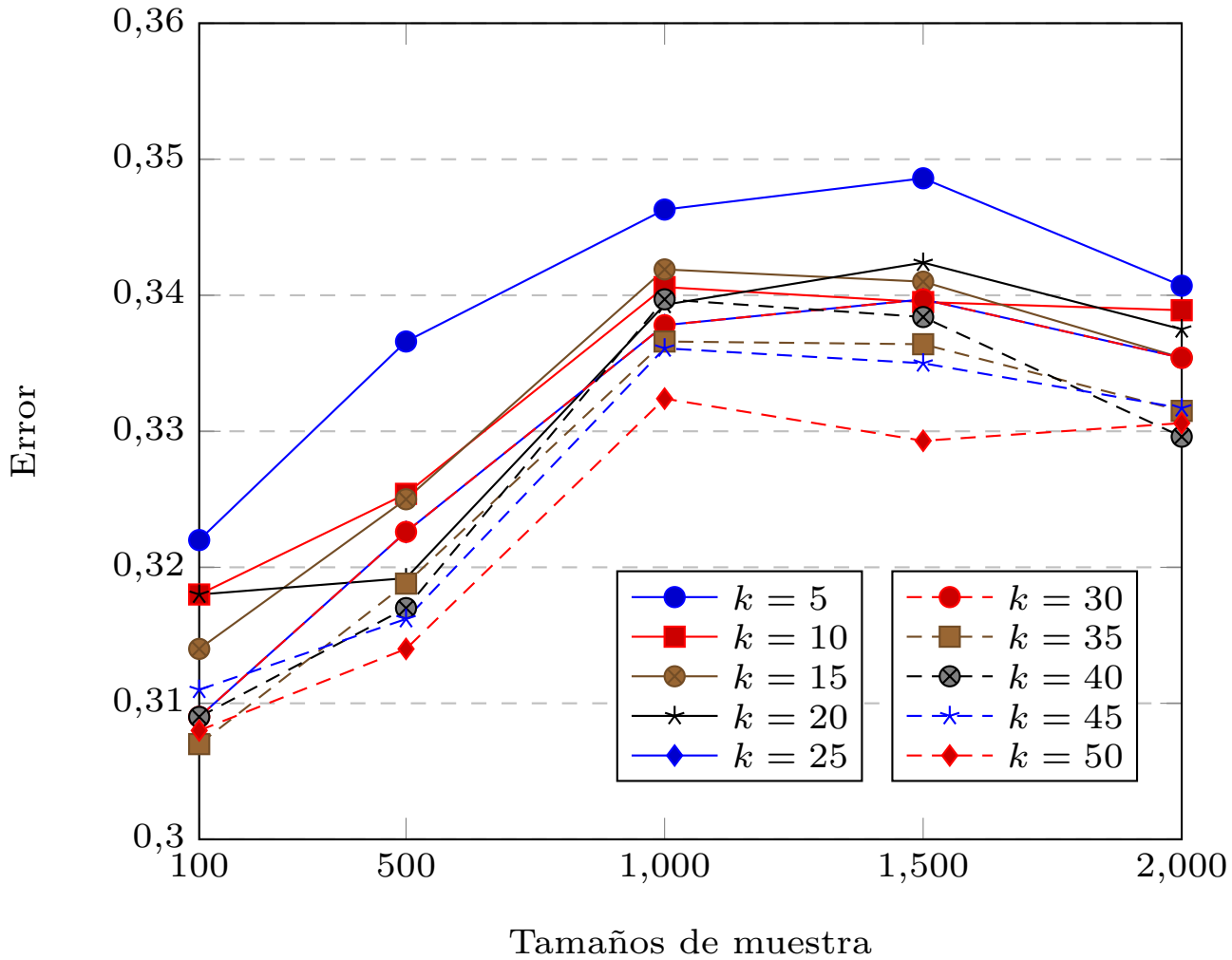
- Se comparan distintas ejecuciones del método EQuAL.
- Se varía el tamaño de muestra en particular: 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Por cada uno de los tamaños de muestra se varía el número de clusters k distinto: 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50.
- Se realizaron 10 ejecuciones con una muestra aleatoria distinta para cada una de las combinaciones de k y tamaño de muestra.
- El resultado de cada una de las ejecuciones es una matriz de confusión.

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		

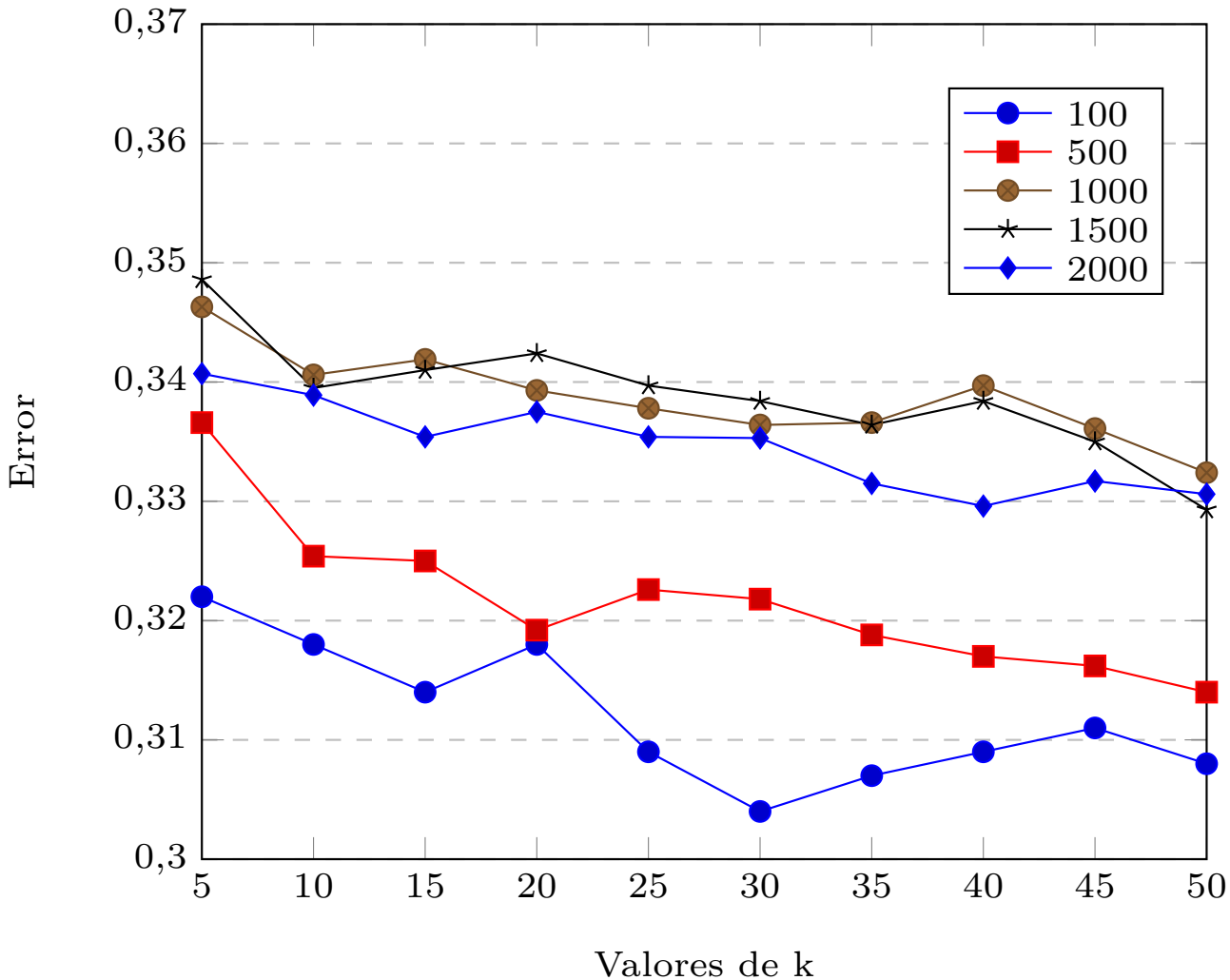
Análisis del método propuesto III

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.



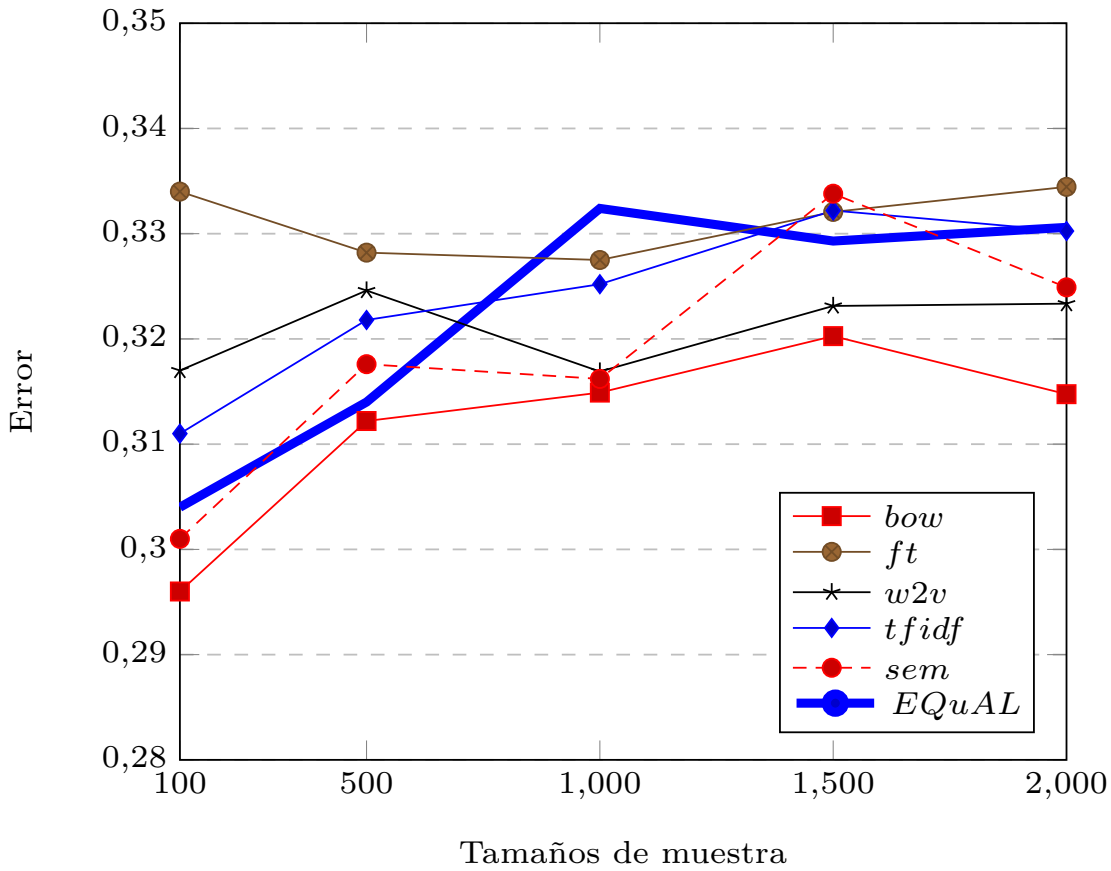
Análisis del método propuesto y algoritmos del estado del arte I

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 II

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



Otras observaciones de interés I

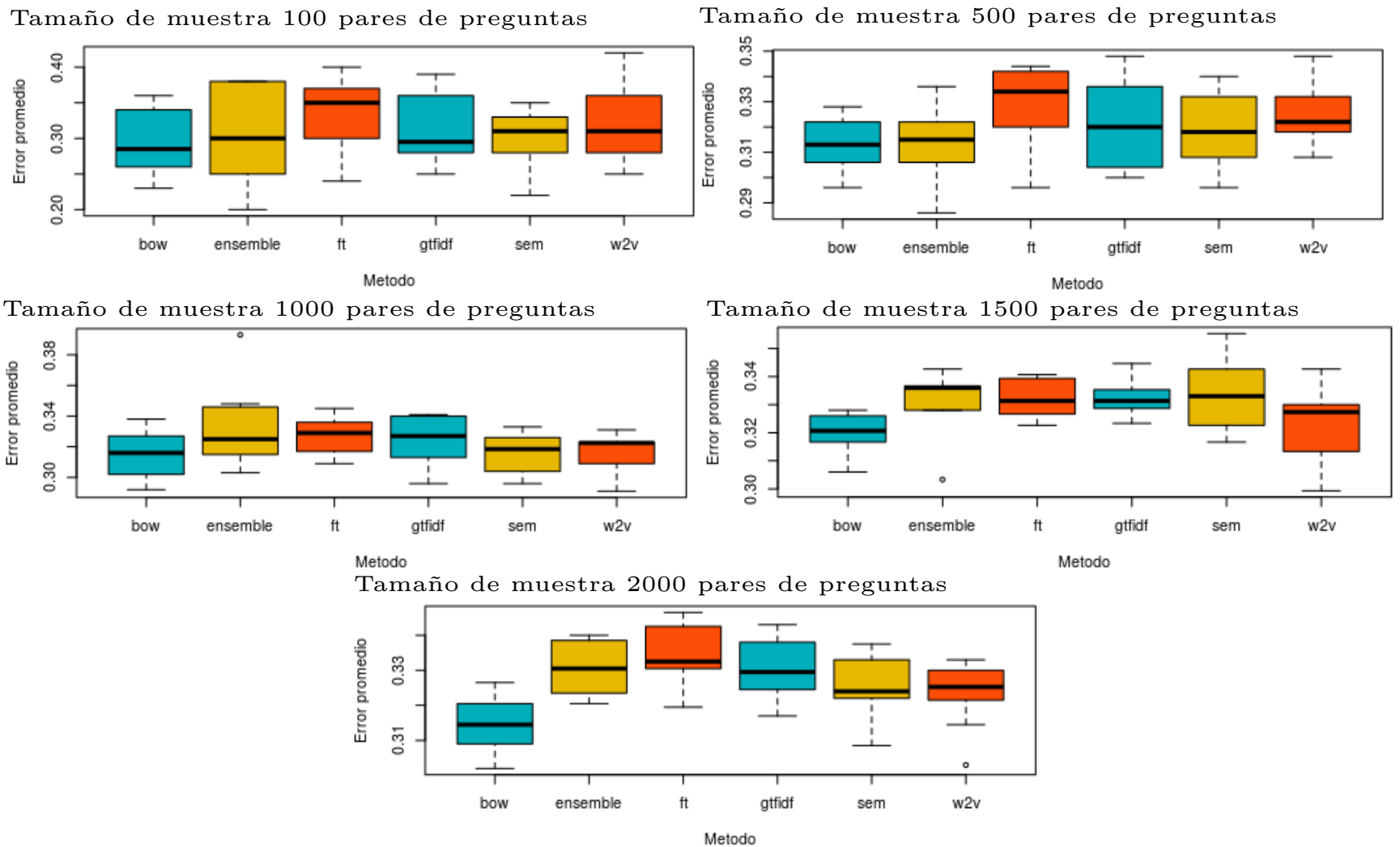
Análisis de varianza del método propuesto.

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

Diagramas de caja y bigote para tamaño de muestra 100, 500, 1000, 1500 y 2000 pares de preguntas ($\alpha = 0,05$).



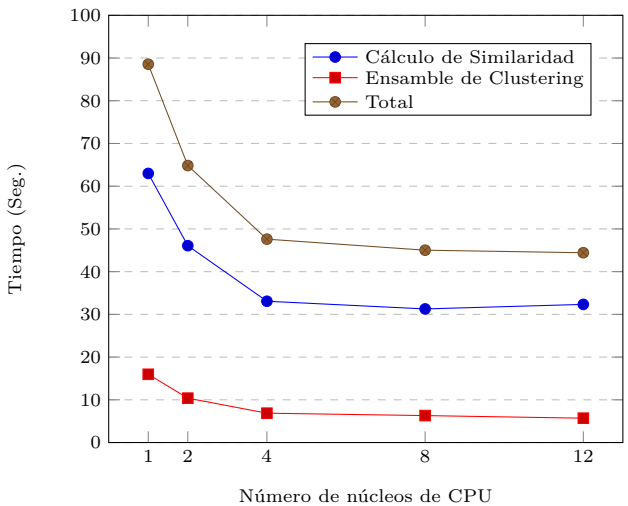
Análisis de desempeño I

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

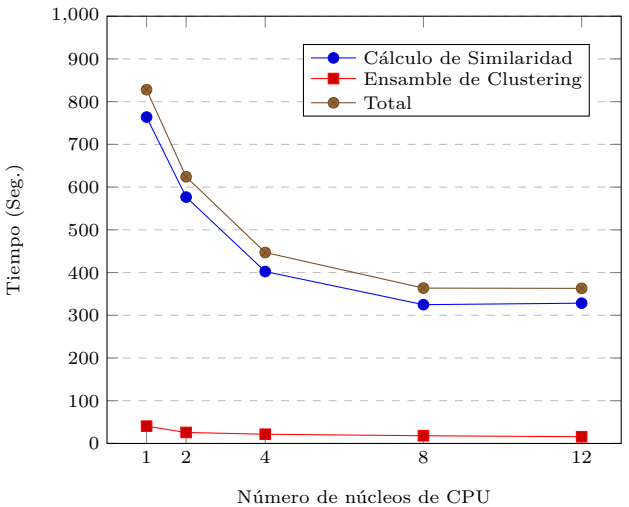
- Cluster Hadoop en localhost.
- Tamaños de muestra 100, 500 y 1000 pares de preguntas.
- En cada una de las ejecuciones, utilizan dos técnicas de similaridad (TF y TFIDF), para luego ensamblarlas.
- Cantidad de núcleos CPU asignados: 1, 2, 4, 8 y 12. Esta configuración puede ser extrapolada fácilmente a un cluster de computadoras para escalar horizontalmente.

Análisis de desempeño II

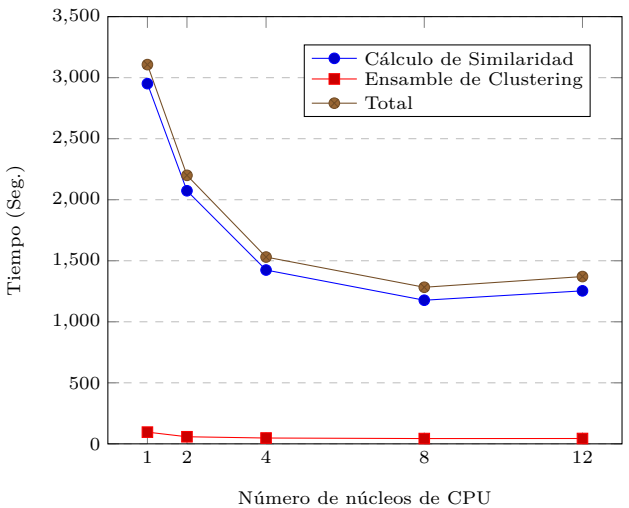
Tamaño de muestra 100.



Tamaño de muestra 500.



Tamaño de muestra 1000.



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.

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.
- 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.
- 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 desarrolló 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 aprendizajes 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 aprendizajes 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 aprendizajes 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 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!