

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

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

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.

Area temática

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

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.

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.
- Es de interés buscar si esa misma pregunta ha sido formulada por otro usuario previamente, y que tenga la respuesta buscada.

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.
- 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 estan formuladas de forma difente en el sentido léxico.

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.
- 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 estan formuladas de forma difente en el sentido léxico.
- Es necesaria una medida de similaridad que tenga en cuenta características léxicas y semánticas.

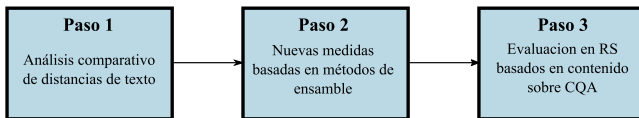
Area temática

- 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.
- 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 estan formuladas de forma difente en el sentido léxico.
- 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.

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.
- 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 estan formuladas de forma difente en el sentido léxico.
- 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.

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



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

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

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

Objetivos específicos

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

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.
- Evaluar el comportamiento de **medidas de similaridad de texto** del estado del arte respecto, en particular en el **ámbito de CQA**.

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.
- Evaluar el comportamiento de **medidas de similaridad de texto** del estado del arte respecto, 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.

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.
- Evaluar el comportamiento de **medidas de similaridad de texto** del estado del arte respecto, 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.
- **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

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

Motivación

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

Motivación

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

Motivación

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes mediante Ensamble de Clustering, que pueda aplicarse como entrada para un RS.
- Arquitectura de software que soporte el procesamiento del método propuesto de una forma eficiente y escalable.

Agenda

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

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.

Funciones de un Sistema de Recomendación

- Aumentar el ratio de conversión en un sitio o aplicación.
- Aumentar satisfacción y fidelidad del usuario.

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).

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 $D(x_i, x_j)$ 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.

Similaridad del coseno I

La distancia del coseno se puede derivar de la fórmula del *producto escalar*:

$$\vec{D}_i \cdot \vec{D}_j = \|\vec{D}_i\| \cdot \|\vec{D}_j\| \cdot \cos(\theta),$$

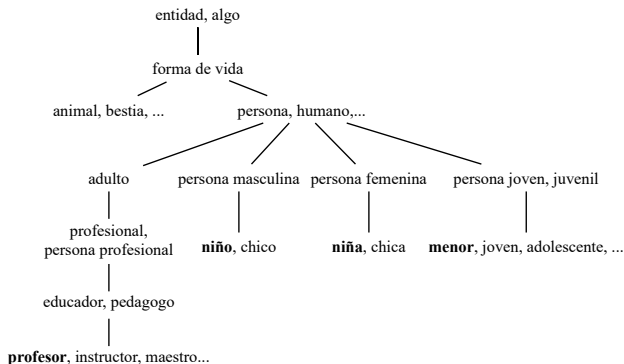
siendo $\|\vec{D}_i\|$ y $\|\vec{D}_j\|$ los módulos de los vectores \vec{D}_i y \vec{D}_j respectivamente, y θ el ángulo formado entre ellos.

La similaridad entre dos vectores puede medirse como:

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

donde d_i y d_j son los componentes de los vectores \vec{D}_i y \vec{D}_j respectivamente.

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 subsumer de las mismas.

Similaridad en taxonomías II

Se define $p(t)$ como la **probabilidad** de un concepto t , formalizando:

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

Se define el **contenido de información** $I(t) = 0$ como:

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

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

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

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

Medidas de similaridad utilizadas

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

Term Frequency (TF)

Características 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 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)

Características 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.

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

Características 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.

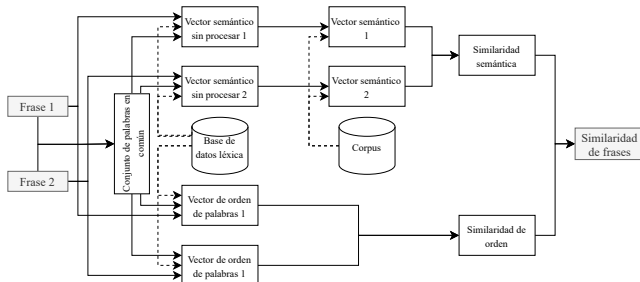
Características de FastText:

- Librería 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

Características 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 frases

Un conjunto de palabras comunes T es formado por la unión de las dos frases T_1 y T_2 en comparación. Luego, un vector semántico léxico \check{s} es derivado de este conjunto. Por cada una de las palabras w_i en T (*ejemplo para T_1*):

- **Caso 1.** Si w_i aparece en la T_1 , 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 similaridad entre palabras.

Luego el vector semántico de cada frase es:

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

Entonces, la similaridad semántica entre dos frases es definida como el coeficiente del coseno entre los dos vectores.

Semantic Distance III

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.

Se define un conjunto T con la union de las palabras en T_1 y T_2 y luego se definen los vectores de orden:

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

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

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

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

Similaridad total entre frases:

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

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

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

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.

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.
- Aprovecha la variabilidad agregada para encontrar una estructura *inter-patrón*.

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.
- Aprovecha la variabilidad agregada para encontrar una estructura *inter-patrón*.
- Identificación de clusters subyacentes con formas, tamaños y densidades arbitrarias.

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

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

Hipótesis I

Hipótesis del trabajo de tesis

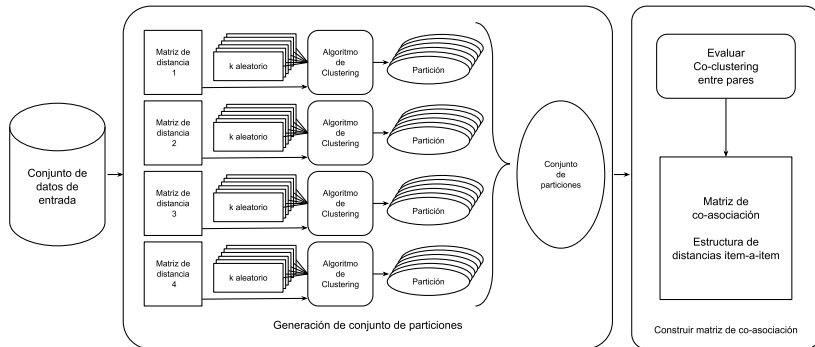
A partir del relevamiento del estado del arte se establece la hipótesis de que los **algoritmos de cálculo de similaridad de texto en sitios de CQA**, con el fin de participar del proceso inherente a la aplicación de **Sistemas de Recomendación** con gran volumen de datos, pueden ser mejorados en cuanto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

Hipótesis del trabajo de tesis (cont.)

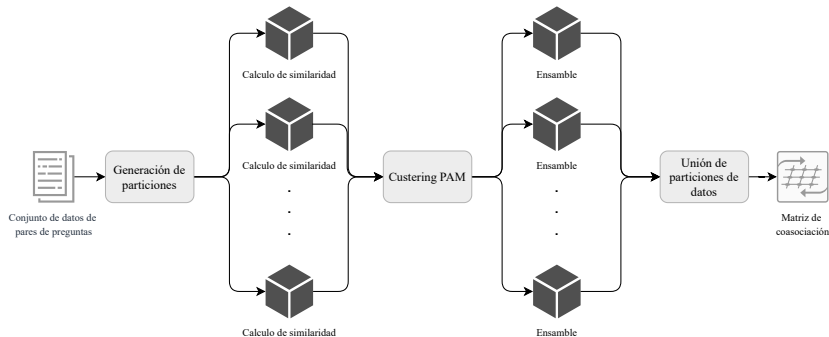
Por tal motivo, y como respuesta a la hipótesis planteada, se presenta:

- 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.

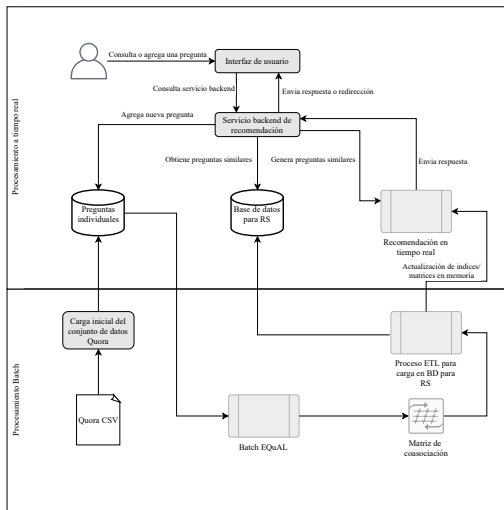
El método propuesto



Arquitectura de procesamiento de datos



Implementación en un sistema de recomendación de tiempo real



Agenda

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

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.

Matrices de confusión

Matrices de confusión utilizadas

		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 (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 triangular superior:

$$\begin{bmatrix} 0 & similarity_01 & similarity_02 & similarity_03 \\ 0 & 0 & similarity_12 & similarity_13 \\ 0 & 0 & 0 & similarity_23 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

- 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

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.

$$\text{len}(\text{set}(\text{tuples_1}).\text{intersection}(\text{set}(\text{tuples_2}))) / \text{total_runs}$$

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*.

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 .

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 .
- Para cada tamaño de muestra, se realizaron N muestras aleatorias manteniendo un k fijo.

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 .
- Para cada tamaño de muestra, se realizaron N muestras aleatorias manteniendo un k fijo.
- Por ejemplo, para 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.

Preparación de los datos 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

Preparación de los datos 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

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

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

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?

Elección del umbral correcto III

Primero, consideraciones de la elección del umbral correcto:

- ➊ Un valor de umbral alto, reducirá la cantidad de falsos positivos, ya que menos pares de preguntas van a ser consideradas como iguales (clase 1).
- ➋ Además, la cantidad de falsos negativos aumentará, ya que el proceso será restrictivo en cuanto a clasificar una instancia como positiva.
- ➌ Por el otro lado, sucederá todo lo contrario en cuanto se reduzca el valor de umbral.
- ➍ Por este motivo, se realiza un proceso iterativo para identificar el valor de umbral óptimo t^* .

Agenda

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

Metodología utilizada

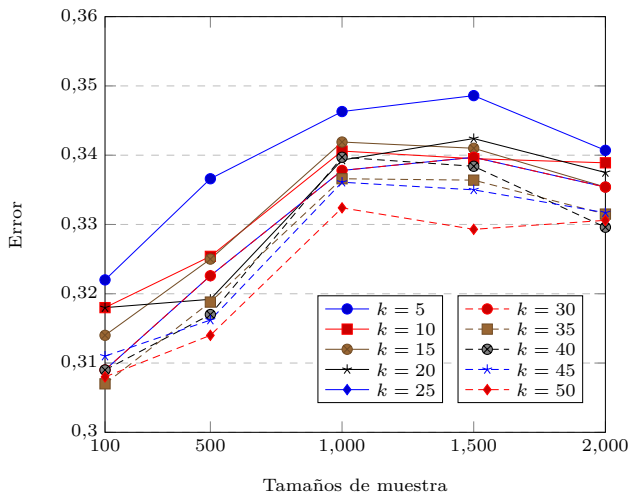
- Se comparan distintas ejecuciones del método EQuAL.
- Se varia el tamaño de muestra en particular: 100, 500, 1000, 1500 y 2000 pares de preguntas.
- Por cada uno de los tamaños de muestra se varia el número de clusters k distinto: 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50.
- Se realizaron 10 ejecuciones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

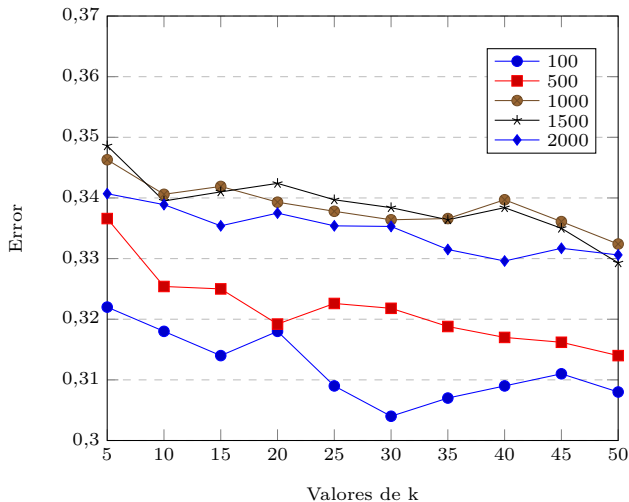
Análisis del método propuesto III

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 .



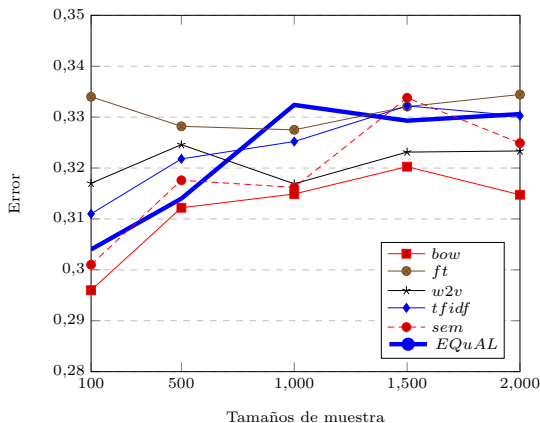
Análisis del método propuesto y algoritmos del estado del arte I

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.



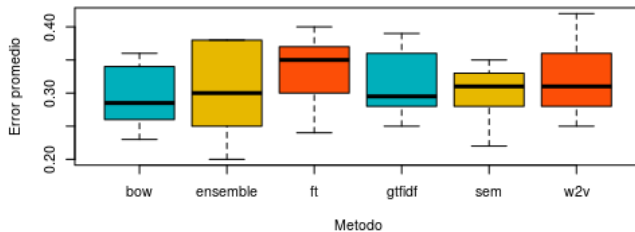
Análisis de varianza del método propuesto.

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

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

Otras observaciones de interés II

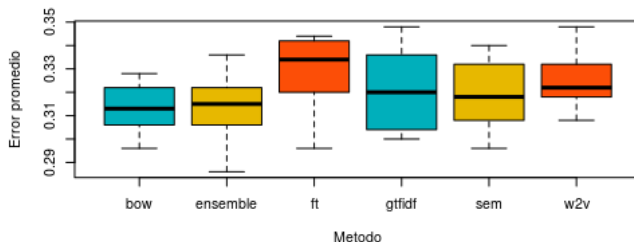
Tamaño de muestra de 100 pares de preguntas.



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

Otras observaciones de interés III

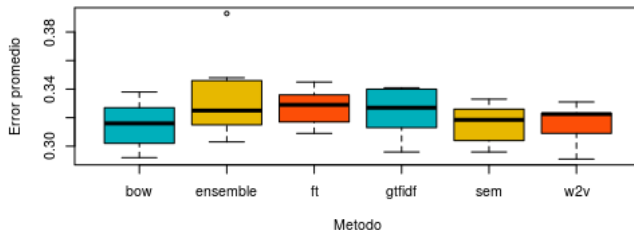
Tamaño de muestra de 500 pares de preguntas.



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

Otras observaciones de interés IV

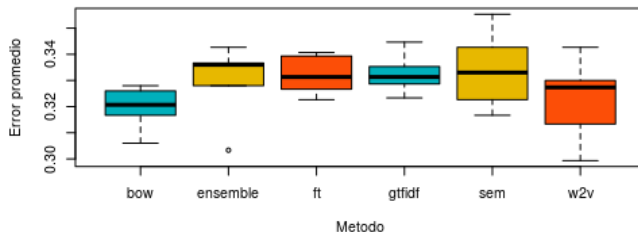
Tamaño de muestra de 1000 pares de preguntas.



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

Otras observaciones de interés V

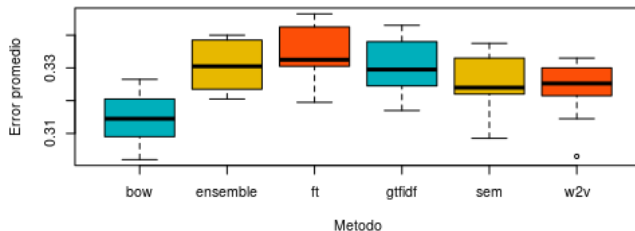
Tamaño de muestra de 1500 pares de preguntas.



- En todos los casos, los intervalos de confianza incluyen al valor 0 y $p\text{-adj} > \alpha$.
- 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 $p\text{-adj} > \alpha$.

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

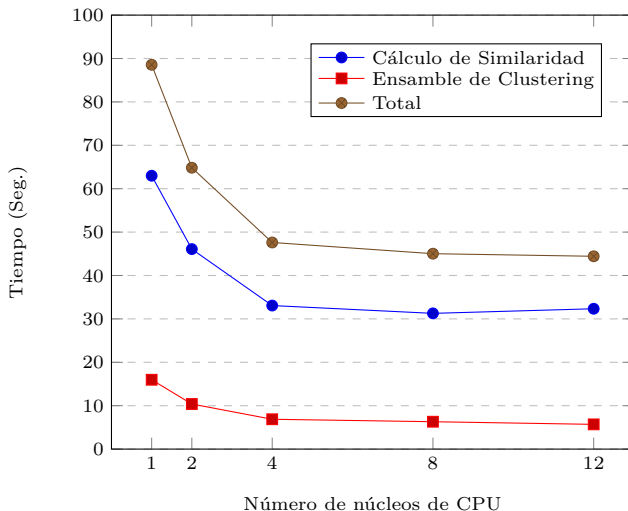
- El agregado de variabilidad de datos puede ser influyente en tamaños de muestra pequeños.
- El método EQuAL es dependiente a los métodos subyacentes.
- Debido a la naturaleza del conjunto de datos utilizado, no es posible identificar fácilmente cuál es la forma de los clusters en cuestión, y verificar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

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.

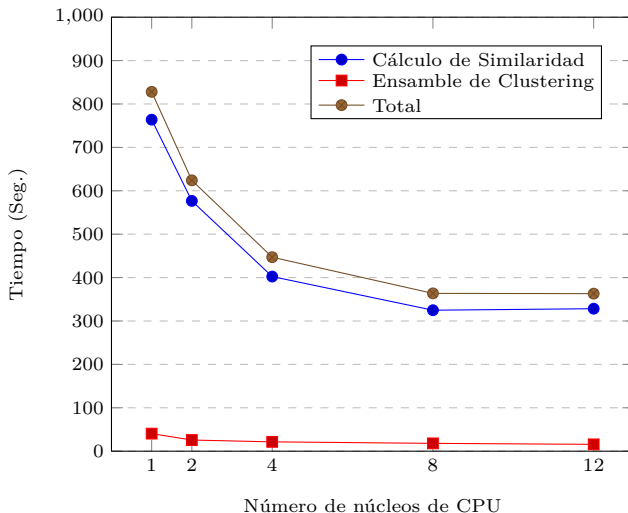
Análisis de desempeño II

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



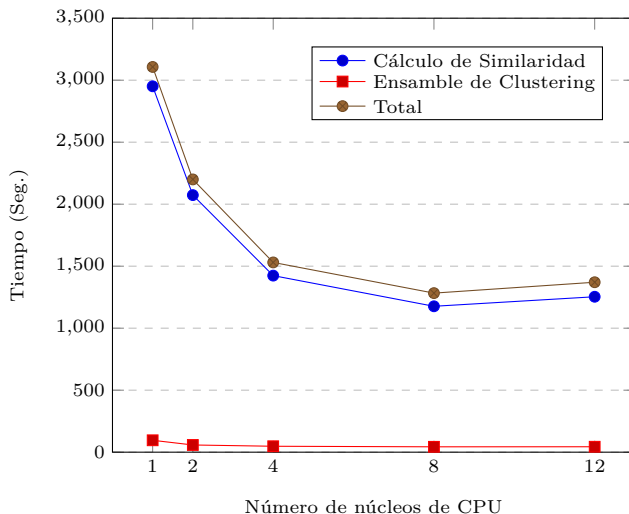
Análisis de desempeño III

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.

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

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

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.

El presente trabajo sirve como estado del arte para las siguientes líneas 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.

El presente trabajo sirve como estado del arte para las siguientes líneas 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.

El presente trabajo sirve como estado del arte para las siguientes líneas 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.

El presente trabajo sirve como estado del arte para las siguientes líneas 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.

El presente trabajo sirve como estado del arte para las siguientes líneas 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!