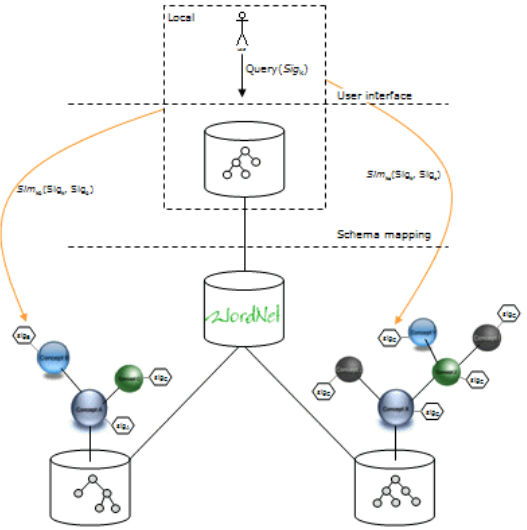
Seminario de Tesis I

Propuesta de Tesis

Calificación Automática de Ensayos Utilizando Técnicas de Text Mining



Caballero Ortiz, José Alberto

9 Diciembre 2006

Tesista

- Caballero Ortiz, José Alberto
  - Especialidad: Ing. de Sistemas
  - Ciclo: IX
  - Correo Electrónico: jkb_llero@yahoo.com

Título

Calificación Automática de Monografías Utilizando Técnicas de Text Mining

Justificación del Problema

- Presiones de recursos y tiempo dentro de la calificación de trabajos estudiantiles.
- Alta cuota de subjetividad en examinadores humanos.
- Tendencia del mercado a utilizar grupos de examinadores (incremento de costos).
- Implementaciones exitosas en Universidades norteamericanas y exámenes internacionales.
- Posibilidad de incrementar el volumen de trabajos corregidos y disminuir el tiempo de respuesta.

Ámbito de la investigación

- El ámbito de aplicación puntual es en la calificación de un grupo de trabajos presentados en todas las secciones asignadas a un curso de humanidades de una universidad local, utilizando a los profesores asignados a su corrección como el pool de examinadores.
- La aplicabilidad se encuentra orientada a trabajos escritos en español, con posibilidad de ampliar las funcionalidades del sistema con la inclusión de diccionarios y bibliotecas.

El Problema

- La necesidad por parte de las instituciones educativas de contar con procesos ágiles de calificación de trabajos que busquen ser lo más objetivos posibles, entregando resultados confiables en un tiempo prudente, siendo capaces de atender eficientemente a un gran número de estudiantes.
- Variables involucradas
  - Cantidad de recursos obtenidos para realizar la evaluación del documento (jurados).
  - Tiempo transcurrido entre la evaluación y la entrega de resultados

Objetivo

- Demostrar que un sistema de calificación automatizado haciendo uso de las herramientas que disponemos hasta el momento, permitiendo obtener resultados confiables para el ámbito analizado acorde con el juicio de examinadores humanos.
- Estandarización de los criterios de calificación requeridos para el análisis.
- Disminución de los recursos necesarios para la realización de estas tareas, asociado al uso de este sistema en el futuro.

Antecedentes

- Indique las referencias bibliográficas, por ejemplo:

  – Reimer (2002), desarrolla un experimento de calificación automatizada incluyendo un conjunto de criterios ajustables, obteniendo resultados sensibles a dichos criterios.

  – Boring (2000), desarrolla un conjunto de experimentos para comparar mecanismos de calificación analíticos y holísticos con el sistema LSA.

Tipo de Investigación

- Tipo de Investigación
  - Correlacional: Relaciona cambios en los valores de las variables dependientes, como en la eficiencia del proceso y la eficiencia del sistema, con las técnicas usadas para la calificación y los criterios de evaluación.

- Tipo de Diseño
  - Experimental: Por la disponibilidad de los datos es posible realizar experimentos independientes para cada uno de los casos

DISEÑO DEL EXPERIMENTO

Objeto de la Investigación

- El individuo de análisis será una palabra, siendo entendida como un concepto que aporta valor al texto y permite la aplicación de criterios para distinguirlo de otros y lograr clasificarlo.
- La realización del tratamiento de un ensayo independiente se entenderá como una repetición del experimento para la recopilación de resultados.

Población

- Clasificación de los individuos de la población, palabras o conceptos, basada en la temática del texto:
  - Stop Words
  - Términos del Dominio
  - Consideración de Sinonimia y palabras relacionadas

- Clasificación basada en el idioma utilizado, en este caso se tendrá como parámetro el idioma español.

Muestra

- El muestreo para la obtención de datos será no probabilístico, específicamente del tipo opinático o intencional, en el cual los criterios de selección de un individuo de la población se encuentran a criterio del investigador.

- Generación de Valores de Variables
  - Variables Independientes:
    - Técnicas Utilizadas: Documentación Existente (marco teórico instrumental)
    - Conjunto de Criterios Utilizados: Antecedentes y mecanismos utilizados (marco teórico conceptual)

- Obtención de datos:
  - Uso de software comercial para tratamiento de texto con formato.
  - Uso de parsers para extracción de palabras.
  - Software estadístico para cálculo de valores de variables independientes.
  - Apoyo en software orientado a técnicas.

Variables

Variables independientes: •Técnicas utilizadas para la extracción de patrones. •*Criterios utilizados para la medición*	**EXPERIMENTO**	**Variables dependientes:** **Desviación de Calificaciones:** Esta variable contempla el grado de cercanía entre la calificación obtenida por el sistema automatizado y las obtenidas por el modelo.
Instrumento de medición Las variables son caracterizadas mediante criterios de presencia/ ausencia, el conjunto de valores son extraídos de conceptos del dominio y marco teórico instrumental y establecidos según el criterio del investigador.		**Instrumento de medición** Recopilación de datos de calificaciones de examinadores para cada monografía analizada, calificaciones realizadas por el sistema y almacenada en la base de datos de registros.

Diseño Experimental

- Variables Independientes:
  - Técnicas Utilizadas: Se manipulan dichas variables incluyendo técnicas diferentes para la realización de comparaciones o mediante la inclusión o exclusión de dichas técnicas en el modelo.
  - Criterios Utilizados en la Calificación: Se cuantifican las variables dependientes, es decir
    - Correlación de los resultados respecto a la media de examinadores humanos.
    - Sensibilidad de Resultados

    Para distintos criterios incluidos o excluidos de acuerdo al objetivo del experimento.
  - Repetición del experimento para la aplicación de distintas técnicas y el establecimiento de distintos criterios.

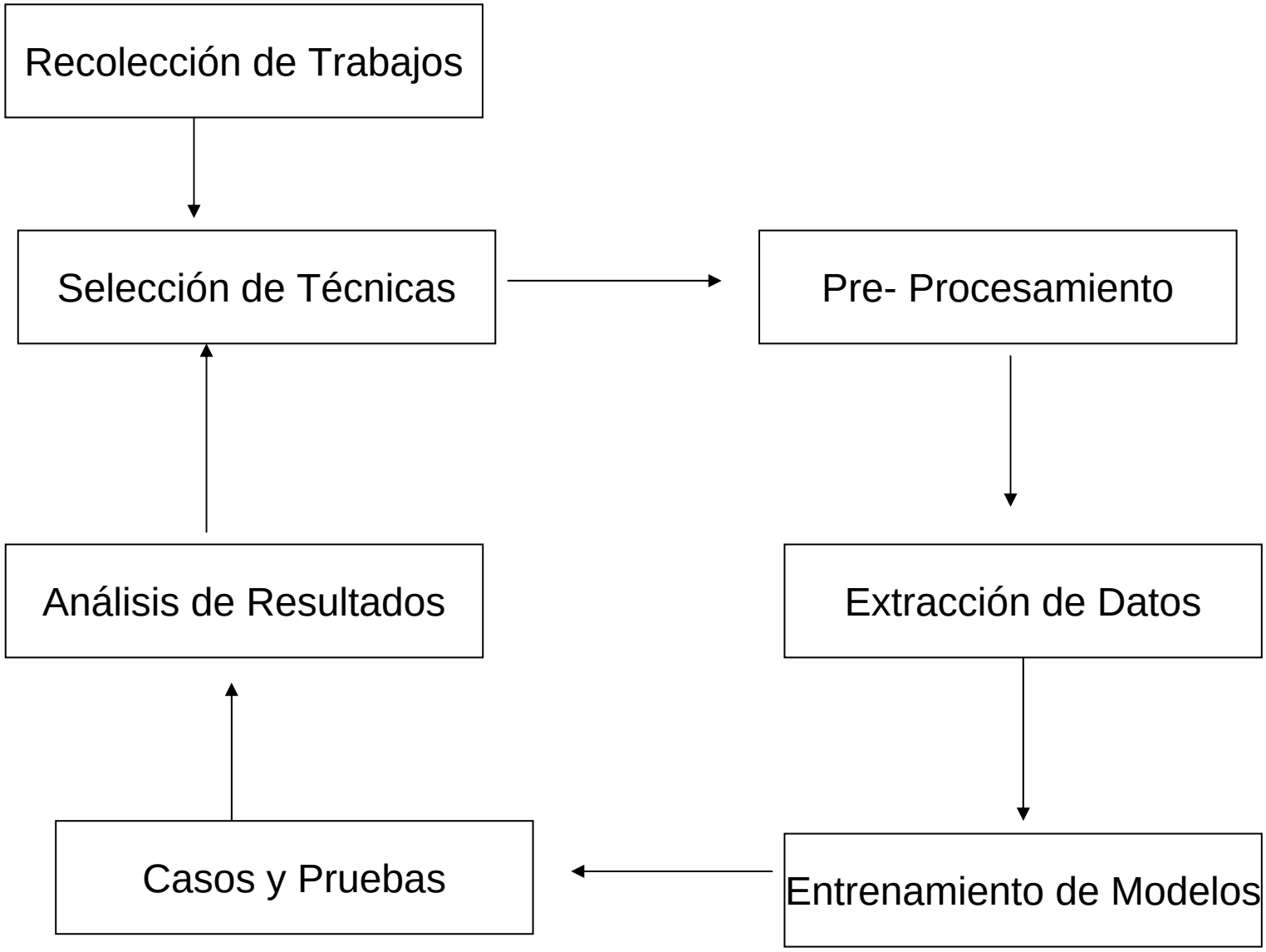
Hipótesis

- La utilización de técnicas automatizadas de calificación ofrece una desviación menor al 20% con respecto a la media obtenida haciendo uso de un grupo de examinadores humanos.

Diseño del Experimento

- Cada revisión de monografía se considera un experimento independiente.
- Trabajaremos sobre la base de grupos de experimentos asociados a las técnicas utilizadas.
- Los resultados de estos grupos de experimentos serán almacenados en la base de datos de registros.

Modelo de Solución



Grupos de Experimentos

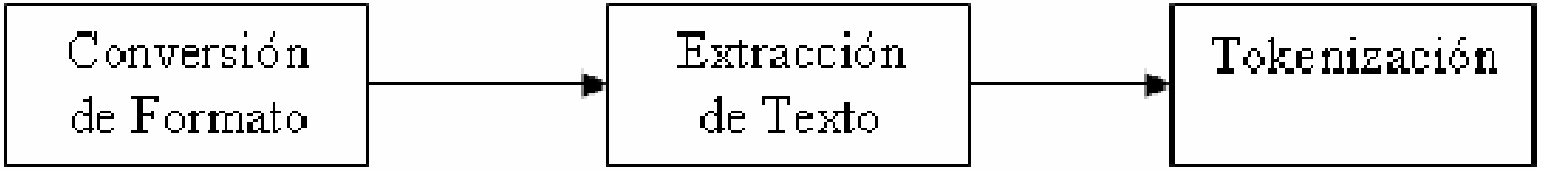
Grupo de Experimentos	Técnica de Pre-Procesamiento	Técnica de Procesamiento	Calificador	Nro. De Realizaciones
1	Ninguna	Ninguna	Humano	1
2	Lematización	Reglas de Asociación	Máquina	Ilimitadas
3	Stemming	Reglas de Asociación	Máquina	Ilimitadas
4	Stemming	Enfoque "Bolsa de Palabras	Máquina	Ilimitadas
5	Stemming	Reglas de Asociación	Máquina	Ilimitadas
6	Stemming	Categorización de Textos	Máquina	Ilimitadas

Recolección de Datos y Selección de Técnicas

- Los trabajos serán seleccionados desde la plataforma informática designada al curso, aunque sería deseable que se encuentren en un formato de lenguaje de etiquetas, puede trabajarse con archivos de MS Word.
- Las técnicas serán seleccionadas de acuerdo al orden establecido en el diseño del experimento, teniendo en cuenta que según la eficacia de dichas técnicas puede modificarse dicho orden incluyendo técnicas nuevas.

Pre- Procesamiento

- Se busca la obtención de texto plano desde el archivo en su formato nativo.
- Se usa un lenguaje intermedio (generalmente de etiquetas como HTML).
- Luego se separa dicho texto en palabras (tokenización)



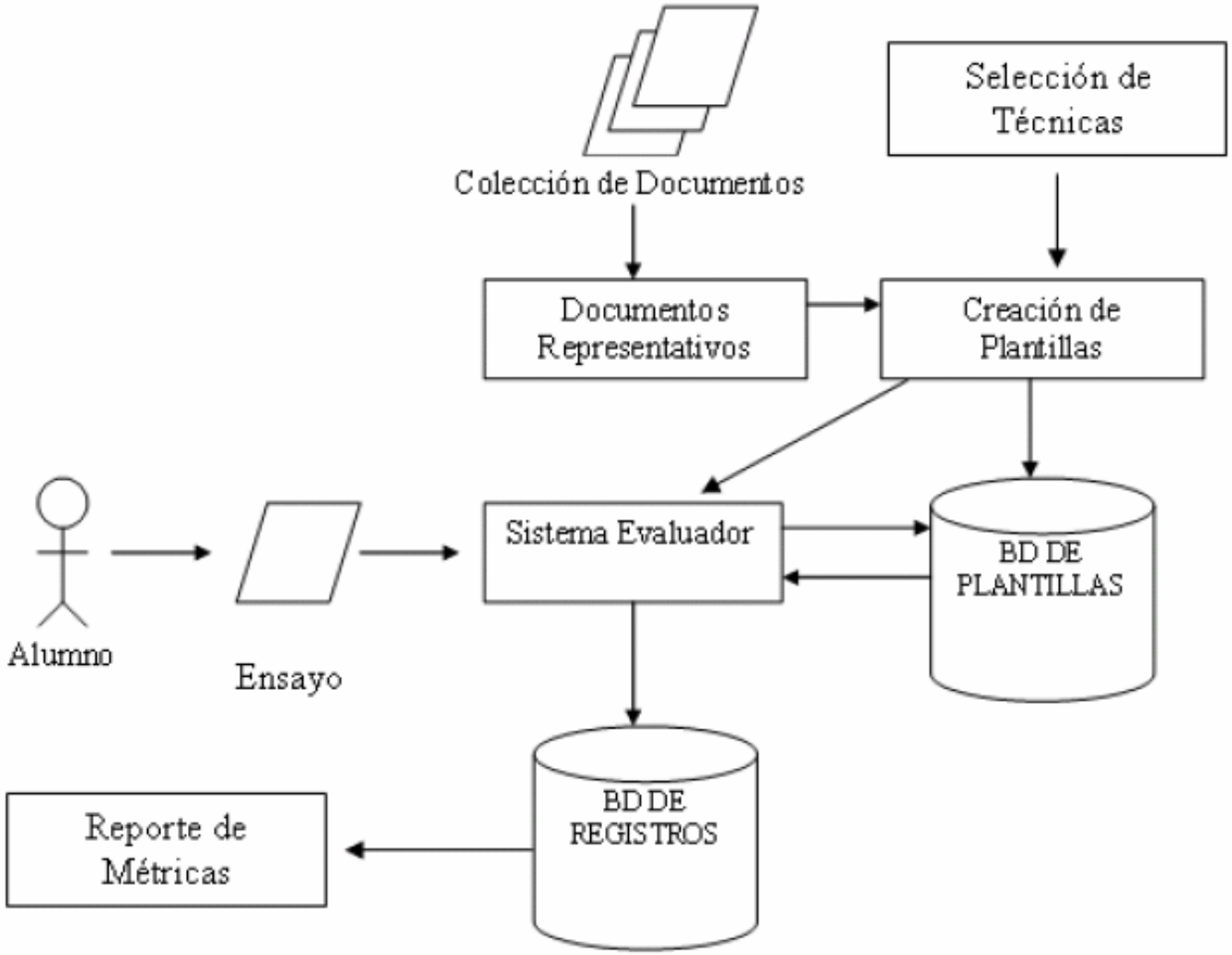
Extracción de Datos

- Se eliminan palabras que no aportan valor al texto o que no sirven para los criterios de clasificación aceptados.
- Se buscan palabras de significados similares o raíces comunes para ser tratadas como un solo individuo.
- Finalmente se realiza en conteo de frecuencias para cada concepto seleccionado.

Entrenamiento de Modelos y Análisis de Resultados

- Dependiendo de las técnicas utilizadas, el modelo requerirá cierto tiempo de entrenamiento, en algunos casos esto se podrá realizar con software comercial (Rules Asociation Mining → SQL Server 2005).
- El análisis de resultados se llevará a cabo mediante herramientas estadísticas que realizarán operaciones de selección sobre la base de datos de registros.

Visión General del Sistema



ANÁLISIS DE FACTIBILIDAD

Datos y Experimentos

- Las fuentes de datos se encontrarán principalmente en repositorios electrónicos de documentos como las comunidades de e-learning.
- Atributos importantes:
  - Formato de los Datos
  - Origen de los mismos
  - Originalidad
- El muestreo será por recolección directa del contenido de los documentos.
- Es posible repetir el experimento propuesto.

Costos

- Conjunto de gastos
  - Sueldo del investigador.
  - Inversión en Bibliografía, sobre todo relacionada a las técnicas por utilizar.
  - Gastos en servicios, producto del uso de equipos para el procesamiento de datos.

CONCEPTO	MONTO TOTAL
Sueldo del Tesista	$2500
Alquiler de Equipos de Cómputo	$120
Adquisición de Revistas y Bibliografía	$250
Costo de Servicios	$80
Costo de los Suministros	$150
TOTAL	3100

Plan de Trabajo

ACTIVIDAD	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Desarrollo del Software Asociado										
Realización de Pruebas Preliminares										
Recopilación de Datos y Pre-Procesamiento										
Entrenamiento de Modelos e Identificación de Patrones										
Ejecución y Registro de Experimentos Variando las Técnicas Usadas										
Ejecución y Registro de Experimentos Variando Parámetros de Evaluación										
Extracción de Medidas de Rendimiento										
Análisis de Resultados y Redacción de Informe final										

MARCO TEORICO

Conceptual

- Calificación Holística
  - Mayor orientación a la categorización de textos.
  - Uso de modelos por categoría.
  - Mayor orientación a calificación por expertos.

- Calificación Analítica
  - Establecimiento de un conjunto de criterios para el análisis de textos.
  - Posibilidad de inclusión o exclusión de nuevos criterios.
  - Establecimiento de ponderaciones y estandarización de mecanismos de calificación

Instrumental

- Pasos a seguir:
  - Extracción de palabras: Del contenido del archivo enviado (depende del formato).
  - Stop Words: Palabras que no aportan contenido deben ser retiradas.
  - Stemming: Extracción de raíces de palabras buscando coincidencias y sinonimia.
  - Contabilización de Frecuencias: Realización de gráficas y cálculos en función de la ocurrencia de palabras.
  - Aplicación de Técnicas de Data Mining.
    - Rules Asociation Mining.
    - Words Bag.
    - Categorización de Textos y Recuperación de Información.

CONCLUSIONES

Conclusiones

- Por la gran aceptación de las plataformas informáticas como repositorio de documentos, la data puede ser conseguida con facilidad, y existiendo software relacionado con la aplicación de técnicas, el tema tratado es viable.
- Las ventajas mencionadas en los sistemas de calificación y los tiempos de respuesta hacen que sea una alternativa interesante en la cual invertir.
- Los antecedentes y las herramientas estadísticas nos proporcionan un conjunto de variables, las cuales irán depurándose a medida que avance el proyecto.

Trabajos Futuros

- La realización de una etapa preliminar a la ejecución del experimento para la elección del conjunto de técnicas a utilizar, dicha etapa puede llevarse a cabo con un conjunto de mini- experimentos utilizando data de prueba, de tal forma.

- Coordinación con los docentes encargados del Área de humanidades para la utilización de portales educativos o grupos de interés para la publicación de sus documentos, permitiendo la sencilla extracción de los datos.
- Diseño y codificación del software necesario para el pre-procesamiento y tratamiento posterior de les ensayos recopilados; pudiendo en algunos casos obtener software comercial o de código abierto para la automatización de las operaciones.
- Adquisición y utilización de herramientas estadísticas que permitan el tratamiento de la información almacenada en la base de datos de registros y la obtención de las medidas de rendimientos para las distintas técnicas analizadas.