Abstract

Sería bueno delimitar mejor el dominio ¿Las tesis de grado eran todas del mismo dominio? y aclarar mejor el objetivo ¿EL objetivo era mejorar un buscador de tesis? ¿Enocntrar tesis con temáticas similares?. Al comentar los resultados sería conveniente incluir alguna métrica que demuestre porqué un método es mejor que otro.

En el resumen normalmente no se hacen llamadas a citas, porque muchas veces el abstract se indexa o lista, y el resto del trabajo dodne están las referencias detalladas puede no estar fácilmente disponible. En todo caso podés poner como “..se implementaron agoritmos en Maskana, un sistema para xxx”

Introducción

Párrafo 1, dice “trataen”, corregir

Sección **Planteamiento del problema**

Dice “Grupo de Investigación Aplicada en Sistemas

- GRIAS” ¿Qué es esto? ¿Una dependencia de la universidad, de una facultad, de un departamento? Aclarar y/o poner un URl si tiene dirección web. Corregir “Çonstrucción”

¿MASKANA tiene alguna referencia? Si no tiene paper o documento asociado, pero tiene presencia web, poner el URL. Por ejemplo, [http://grias.udenar.edu.co/gri as/?p=399](http://grias.udenar.edu.co/gri%20as/?p=399), estaría bien. El algún lugar, más adelante en la tesis, vas a tener que describir un poco mejor que es Maskana.

Pág. 11. “y entidades del conocimiento (NER).”. Ner ya estaba definido, directamente podés usar la abreviatura.

El párrafo que empieza con “"Los conceptos son elementos esenciales para el reconocimiento del mundo…” es una cita textual muy larga que deberías comenzar de otra manera. Sería mejor que introduzcas vos una idea y después digas como menciona xxx en [5] “Los conceptos son elementos esenciales… “. No estoy del todo de acuerdo con las definiciones y explicaciones de este párrafo, me parece que en el área de NPL los significado que usamos para estos términos son un tanto diferentes. Para nosotros una relación semántica es una relación entre coneptos del tipo “Camión es un tipo de vehiculo”. Pero bueno, esto depende de vos. No es algo que te pueda pedir que corrijas.

Después siguen dos párrafos más con citas textuales, están debidamente referenciadas, pero es mejor que el texto lo elabores vos y cites cuando sea necesario. ¿Qué elaboras vos a partir de textos de otros? No está bien poner uno a continuación del otro sin encadenarlos entre ni relacionarlos con tu argumentación.

Más adelante decís “Estos antecedentes proponen diferentes alternativas de minería de textos pero ninguno de ellos aplicado al dominio de trabajos de grado.” Está bien, pero tenés que defender mejor la idea de porqué puede haber diferencias al aplicar esas técnicas a tesis de grado ¿Hay algo en su escritura, en como se estructuran, etc. que generen dudas con respecto a su aplicación en tesis de grado?. Esto no invalida la tesis, para nada: tienen un corpus grande de trabajos de grado y quieren mejorar el acceso.

**Trabajos Relacionados**

**“**el NER desempeña un papel muy importante en diversos problemas relacionados a la búsqueda automática y la categorización de textos.” ¿Por qué? Explicar mejor, pero brevemente. no hace falta extenderse.

En el párrafo siguiente, cuando comentás la referencia [10] habría que explicar con qué se entrena red ¿son textos con palabras clave? La referencia [10] es un review bastante general y presenta varias técnicas ¿A cuál te referís? ¿Puede ser que esté mal la cita?

En los párrafos que siguen la descripción de los métodos propuestos es muy general. Eso es’ta bien para un paper, pero en una tesis que que explicar un poco más. aunque sin extenderte mucho. simplemente dejar más clara la idea de lo que cada método propuesto hace.

**Objetivos**

**Objetivos específicos**

Poner en orden. Por ejemplo, “Construir, limpiar y transformar el repositorio de documentos de trabajos de grado de la universidad de Nariño.”debería ir antes de “Apropiar el conocimiento en algoritmos de minería de texto, algoritmos de agrupación y aprendizaje automático”.

“Descubrir las relaciones conceptuales entre los trabajos de grado y evaluar los resultados.” ¿Qué tipo de relaciones conceptuales buscás? ¿Sobre qué métrica se evaluan los resultados? ¿Es para los diferentes métodos?

¿Cómo se relaciona “Descubrir las relaciones conceptuales entre los trabajos de grado y evaluar los resultados” con “Apropiar el conocimiento en algoritmos de minería de texto, algoritmos de agrupación y aprendizaje automático” . En una primera lectura rápida parecen dos formas diferentes de decirlo mismo.

“Elaboración del documento final de tesis.” Esto no va como objetivo específico

La sección “**marco teórico**” tiene queser parte de la introducción.

Pág. 16. Esta oración no se entiende: “Es así, como en los textos ya descritos se representa el lenguaje natural, los cuales son objeto de extracción de conocimiento”.

**1.2 Bag of Words (Bolsa de palabras, BoW)**

En general está bien, pero hay que redactarlo mejor y tratar de profundizar un poc más la explicación. Si facilita, se puede agregar un esquema.

Esta oración “El peso wij de un término ti en un documento dj es proporcional a la importancia de ese término” sin una formula y/o esquema, no aporta mucho. Si lo explicás má adelante, referí la sección donde se explica.

En el párrafo siguiente, sacar los paréntesis: “(stemming o raiz,lexeme o lematización,eliminar palabras muertas)”.

**Lematización**

La explicación está bien, pero esta frase: “así que en algunos casos obtendremos resultados inesperados” necesitaría un ejemplo.

“Como desventaja pueden “recortar” demasiado la raíz y encontrar relaciones entre palabras

que realmente no existen” Poner un ejemplo

**1.3 Term Frequency—Inverse Document Frequency (Término de Frecuencia - Frecuencia inversa del documento,TF-IDF)**

Un esquema con un ejemplo ayuda mucho a a entender esto.

**1.4 Topic Modeling Latent Dirichlet Allocation ( Modelado de temas, LDA)**

Redactar esto bien, y no repetir lo que ya esté explicad (agregar una referencia):

“El modelado de tópicos es una técnica que suele aplicarse para el agrupamiento y

clasificación de datos en una colección de textos [32].

previamente a la explicación del algoritmo como se define en [33], explicamos los conceptos

de palabra, documento, corpus y tópico.”

pág. 22. Dice:

“Un documento es una secuencia de N palabras denotadas por w = (w1, w2,

..., wN ), donde wn es la n-ésima palabra en la secuencia. Un corpus es una colección

de M documentos denotada por: D = w1, w2, ..., wM.”

No está bueno que uses W para las palabras y los documentos.

No se entiende muy bien el algoritmo

**1.5 Wordvec**

Unir los renglones:

“Propuesto por

Sahlgren [35] (hipótesis distributiva) las palabras con un significado similar ocurren en

similares contextos”.

Un esquema / figura facilitaría la explicación. La explicación de este método y el que sigue es un poco difícil de seguir.

Pág. 28. “es un vector codificado en caliente,…” ponelo en inglés e itálicas “*one-hot encoding*”, porque en español no lo usa casi nadie.

Pág. 32 “Do22vec”, corregir.

Pág. 32 “El modelo PV-DM es basado en el modelo Word2vec CBOW, el cual se miró que

consiste”. Redactar major.

Pág. 34 “El modelo de Distributed Bag of Words (DBOW) es la contra parte al modelo PV-DM.

El modelo DBOW no tiene en cuanta las palabras de contexto en la entrada,” Corregir: contraparte y “en cuenta”

Pág. 35. “El SRC es la función objetiva del k-media y el objetivo es minimizarla” debería ser “La SRC es la función objetiva del algortimo k-medias y el objetivo es minimizarla”.

Creo que una fórmula ayuda mucho a entender qué es lo que se hace con K-medias.

Hay bastantes errores de concordancia entre singular y plral, femenino-masculino. revisar con cuidado.

Metodología

**2.1 Comprensión del problema**

Decís: “Como resultado de la fase de compresión se construyó un marco teórico de diferentes técnicas de minería de textos para abordar el problema.”. Esto ya pertenece a tus resultados, y es un paso importante. no puede quedar solo mencionado. ¿Qué es este marco teórico? ¿Es algo que ya comentaste?. Si ya lo comentaste, hace una referencia aquí. También tené en cuenta que podés usar la metodología CRISP para hacer el trabajo, pero después, al escribir la tesis no es necesario seguirla paso a paso.

**2.2 Comprensión de los datos**

Este es otro punto importante de tu trabajo. No podés decir simplemente que armaste un repositorio de 10,000 documentos. Deberías incluir la distribución del tamaño de documentos, el tamaño mínimo, máximo, medo, mediano. Lo mismo, podrías ver como se distribuyen los términos lematizados.

**2.3 Preparación de los datos**

Aquí la descripción tiene que ser mucho más detallada que lo que mostrás. Deberías indicar cada paso, cómo lo hiciste, que resultados tuviste. Tendría que ser posible reproducir tu trabajo a partir de lo que contas (si se extiende mucho, puede ir en un anexo o un github). Hay que proveer detalles de cada paso, cuáles fueron los resultados, si resultaron satisfactorios, si hubo que hacer cambios, etc.

**2.4 Modelado**

…las dos primeras proveniente de la librería Sklearn ¿Referencia? ¿URL?

…la tercera técnica utilizada fue Doc2vec proveniente de la librería Gensim ¿Referencia? ¿URL?

**Modelos Bow y TF-IDF**

¿Cómo se eligieron los valores? ¿Hubo búsqueda de hiperparámetros? Pssdrías poner una tabla resumen con los diferentes modelos, sus características, resultados, etc.

pág. 44. “pero produce mejores resultados por lo cual se eligió (sg=1)” ¿Cómo definis mejor rendimiento (métrica)?

los resultados de la clusterización después los usas para asignar clases en el entrenamiento supervisado. Tiene que quedar mucho más claro como cerras esta sección de aprendizaje no supervisado y cuál es el producto con el que vas a avanzar en la etapa de entrenar clasificadores.

**Clasificadores**

Para todos los clasificadores, si hubo optimización de hiperpárametros, especificar.

Mostrar resultados comparativos de clasificadores, indicando métricas de performance (explicaciones como “el clasificador X dio mejores resultados” no son suficiente).

Quien lea la tesis debe entender cómo fue le procedimiento y entender a cada paso las decisiones que se tomaron y tener los resultados con suficiente detalle para juzgar cómo

Po rejemplo, para la red neuronal no hay información sobre la función de pérdida, sobre el algoritmo de entrenamiento (¿Backpropagation? o alguna otra variante de método del gradiente? ¿Batch único, minibatch? ¿Cantidad de epochs?

**2.5 Implementación**

Sería bueno para algunas partes claves de los algoritmos competir el código, con un anexo, github, o algun otro medio. También

**3.2 Evaluación de trabajos de grado relacionados conceptualmente propuestos**

**por Maskana.**

En esta sección tendrías que hacer un resumen profundizando en los resultados de todos los clusters. Parte se puede deducir del Silhouette, pero no todo. Sería bueno ver cuantos trabajos quedaron en cada cluster, cuál era la máxima, mínima, etc. distancias dentro de cada grupo. ¿Hay superposición entre grupos? En los casos donde haya superposición importante sería interesante hacer un análisis cualitativo, es decir, mirar qué trabajos entran en cada uno.

El análisis de tres grupos lo extendería a más grupos, pero en el cuerpo de la tesis pondría algo mucho más resumido que lo que estás poniendo. Los detalles irían a un anexo.

Me parece no haberlo visto, pero sería bueno tener claro cuántos trabajos usaste para la parte de análisis no supervisado, si usaste todos, cómo particionaste esos trabajos. Teneindo en cuenta que primero hacés un entrenamiento supervisado y después muchas pruebas de clasificación. Te recomendaría tener un dataset de entrenamiento, uno de prueba y otro de validación, que lo usas sólo al final.

**3.3 Interpretación de las categorías mediante grafos conceptuales.**

Trataría de buscar una forma más compacta de representar los resultados para cada grupo. Además, para varios mostrás los grafos, pero no comentás nada sobre ellos, cuando algunos muestran características interesantes. Por ejemplo, aparecen términos “hub”, en otros casos parecería haber comunidades de términos, etc. Entonces, si mostrás grafos, tenés que comentar algo más profundo sobre ellos.

Te insisto con lo de mostrar los resultados más compactos. Esta sección tiene casi 12 páginas. Queda muy desbalanceada con el resto de la tesis. Por otro lado, hay que tener en cuenta que varios de los cambios que te pido / sugiero van a resultar en otras secciones más extensas. Pero así y todo, esta sección es larga y un tanto árida.

**3.4 Evaluación de los dominios de conocimiento entrenados.**

Este análisis tendría que estar integrado después de la sección donde hablás sobre clasificadores.

Agregá la fórmula de F1 y Precisión que usaste.

Creo que este análisis, al ser de clasificación multiclase, se enriquecería bastante analizando las matrices de confusión.

¿Probaste algún ensamble de clasificadores? en los casos de clasificaciones multiclase suelen andar bien.

**4 CONCLUSIONES Y TRABAJOS FUTUROS**

Falta una discusión integradora de tus resultados, y donde compares tu trabajo con el que hicieron otros en el mismo campo o similares. Es decir, podría haber una sección “Discusión y trabajos futuros”, de una a tres páginas donde haces ese trabajo de integración. En la introducción presentás algunos trabajos interesantes, aquí deberías comparar tus resultados con lo de ellos.

finalmente podrías cerrar con unas conclusiones que sería un resumen y punteo de los hallazgos o logros más importantes.