Proyecto: **ADR Retrieval System**
Centro: **Universidad de la Habana**

Ejecutores:
Rolando Sánchez Ramos C-311
David Manuel García C-311
Andry Rosquet Rodríguez C-311

Date: 16 de diciembre de 2022.

Resumen. En este documento se pretende hablar sobre el proyecto de Sistemas de Recuperación de Información donde se implementó un buscador que puede utilizar entre 3 modelos especializados en esta labor. Además de ello se trabajo con técnicas de retroalimentación y de evaluación de rendimiento mediante métricas.

Palabras claves: SRI, recuperación, información, modelos, retroalimentación, métricas

1. Introducción

Con este informe se tratará de brindar una explicación relacionada al proyecto ADR-Retrieval-System perteneciente a la asignatura System-Recovery-Information. Se abordarán los modelos implementados y algunas otras particularidades. Además se hará un análisis objetivo del producto obtenido y se explicará su funcionamiento.

2. Desarrollo

2.1. Modo de Ejecución:

Para echar andar el buscador solamente es necesario ejecutar el main.py en el directorio inicial del proyecto haciendo simplemente *python main.py*.

2.2. Funcionamniento y ejemplos:

A continuación en la *Fig*1 vemos una imagen de la pantalla general del buscador:

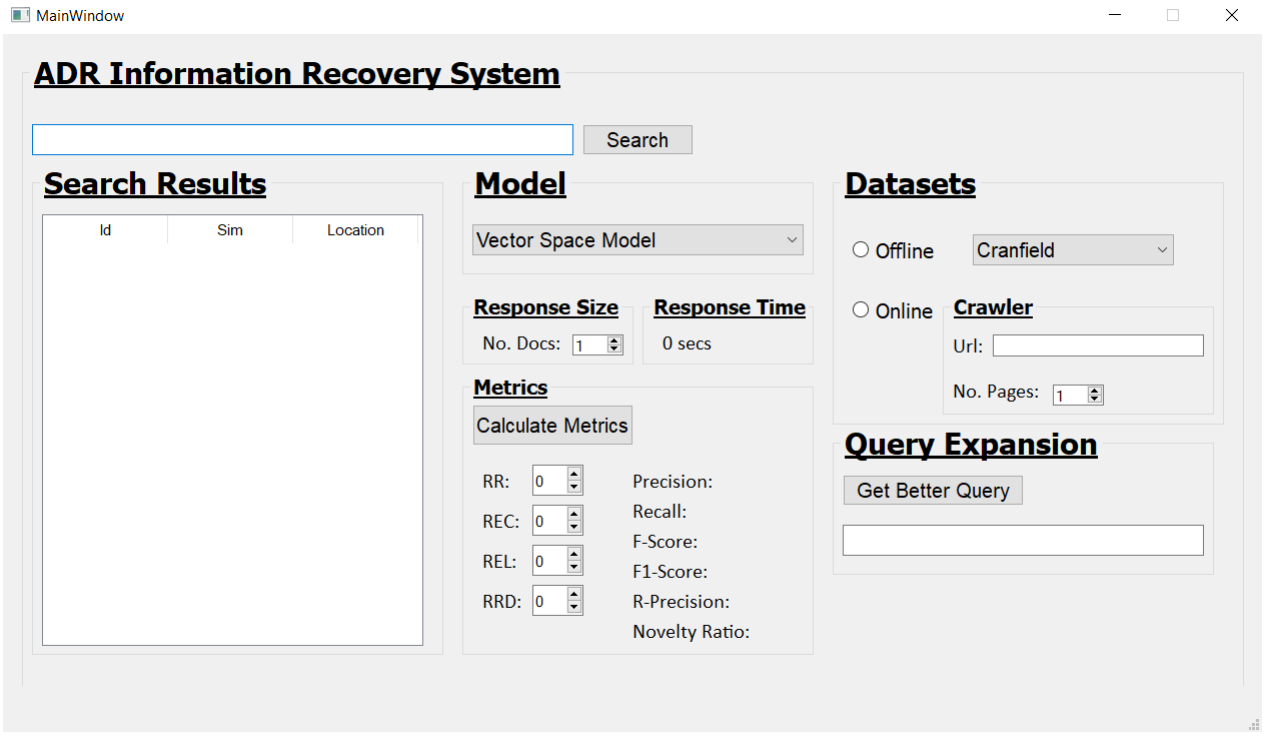
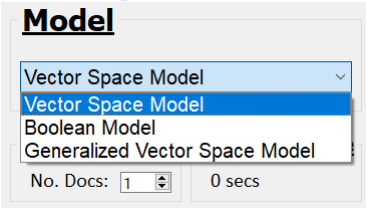


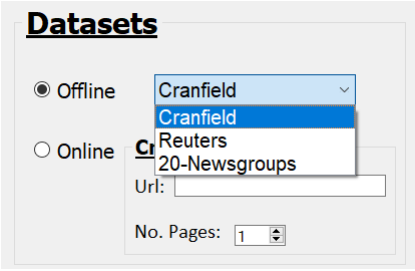
*Fig*1.**Pantalla General.**

Observar el apartado Model en la $Fig2$, donde es posible escoger el modelo que desea utilizar para las búsquedas.



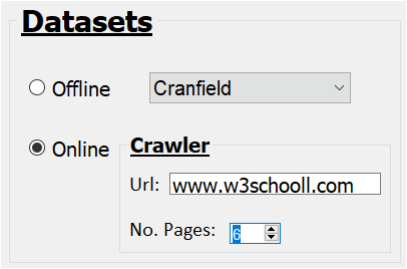
$Fig2$.**Selector de Modelo.**

En la $Fig3$ se puede apreciar el selector del dataset a utilizar.



$Fig3$.**Selector de dataset.**

En la $Fig4$ encontramos como podremos seleccionar entre una búsqueda local o utilizar alguna url y el funcionamiento de Crawler.



$Fig4$.**Escoger entre online o local.**

Se podrá establecer una cantidad máxima de documentos deseada en la respuesta de nuestro buscador en Response Size y a su derecha se mostrará el tiempo de respuesta a la consulta realizada. Las consultas podrán escribirse en el campo superior izquierdo y luego clickeando en el botón Search, como a continuación ($Fig5$):

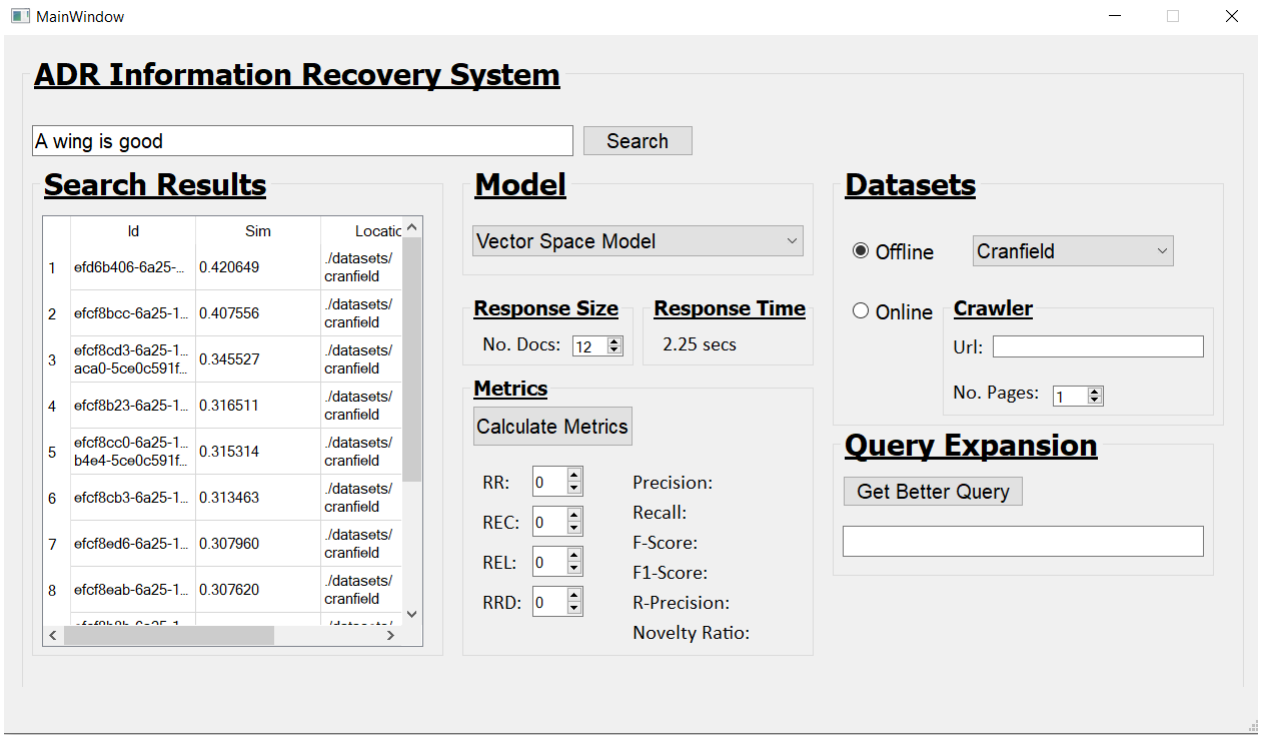


*Fig*5. **Ejemplo de consulta con modelo vectorial simple.**

Se podrá obtener una consulta expandida en el apartado Query Exansion (*Fig*6).

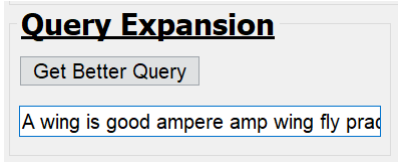


*Fig*6. **Apartado de expansión de consulta.**

Lo anterior puede utilizarse en el buscador y lo esperado es que mejore los resultados de la consulta no exandida como se muestra en la *Fig*7.

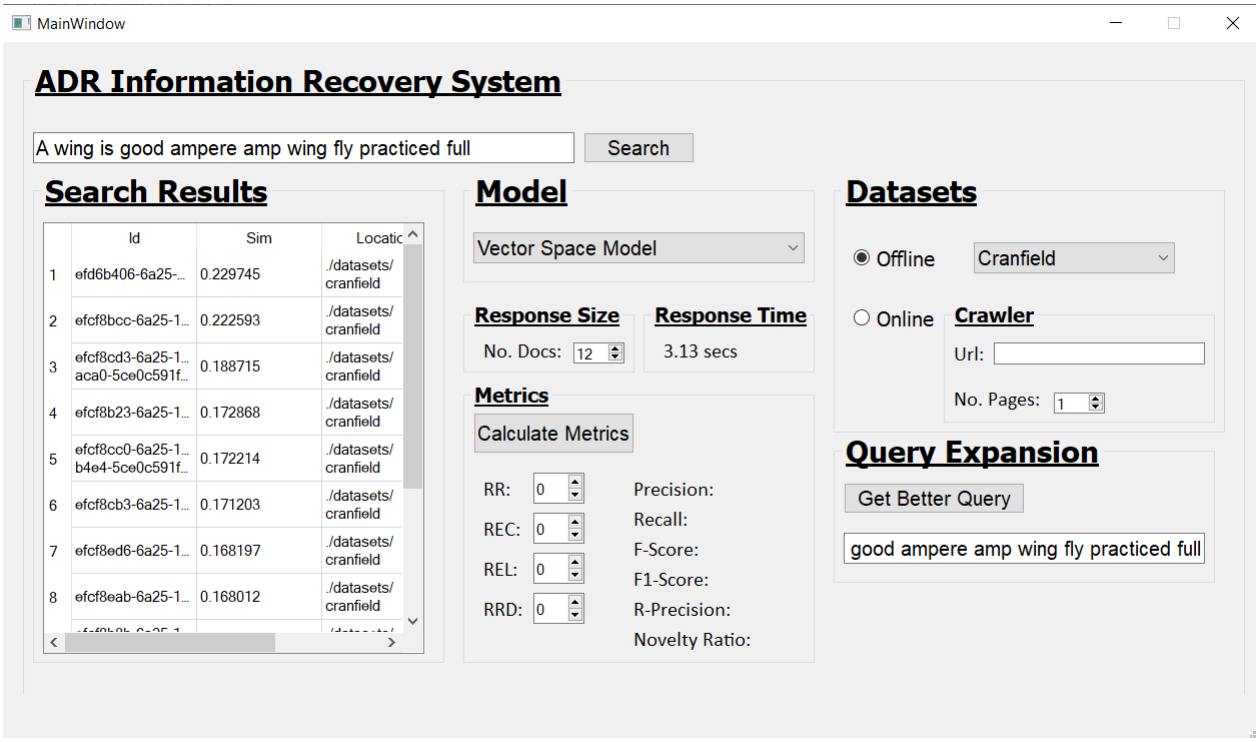


*Fig*7.**Ejemplo con expansión de consulta.**

Por último también podremos hacer nuestro propio cálculo de métricas analizando la relevancia de los documentos obtenidos con la consulta y cuantos en la colección lo eran. Esto mediante las métricas como en la *Fig*8:

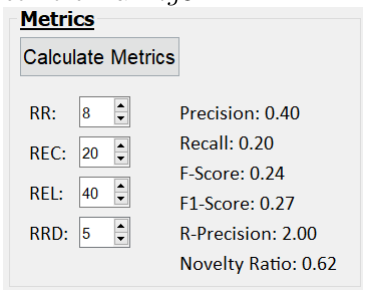


*Fig*8.**Cálclo de Métricas.**

A continuación, en la *Fig*9 y *Fig*10, un par de ejemplos. El primero un ejemplo de búsqueda con el modelo booleano y el segundo con el vectorial generalizado.

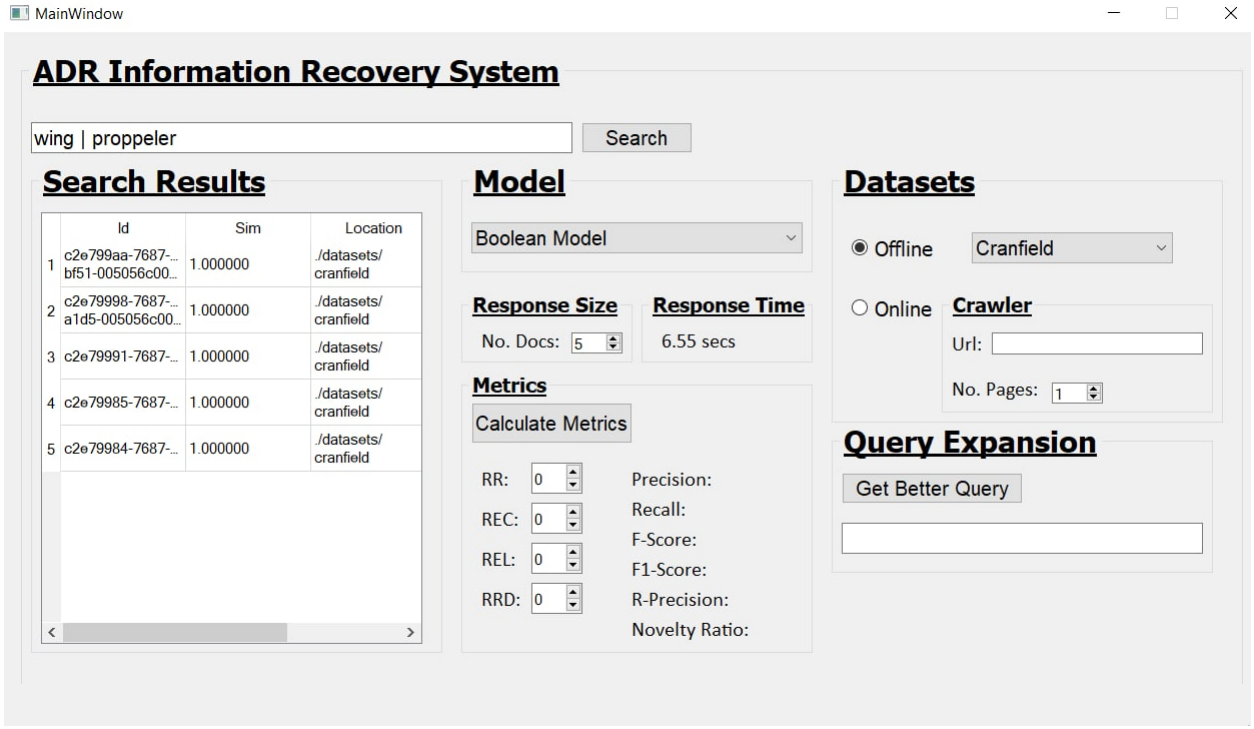


*Fig*9.**Búsqueda del Modelo Booleano.**

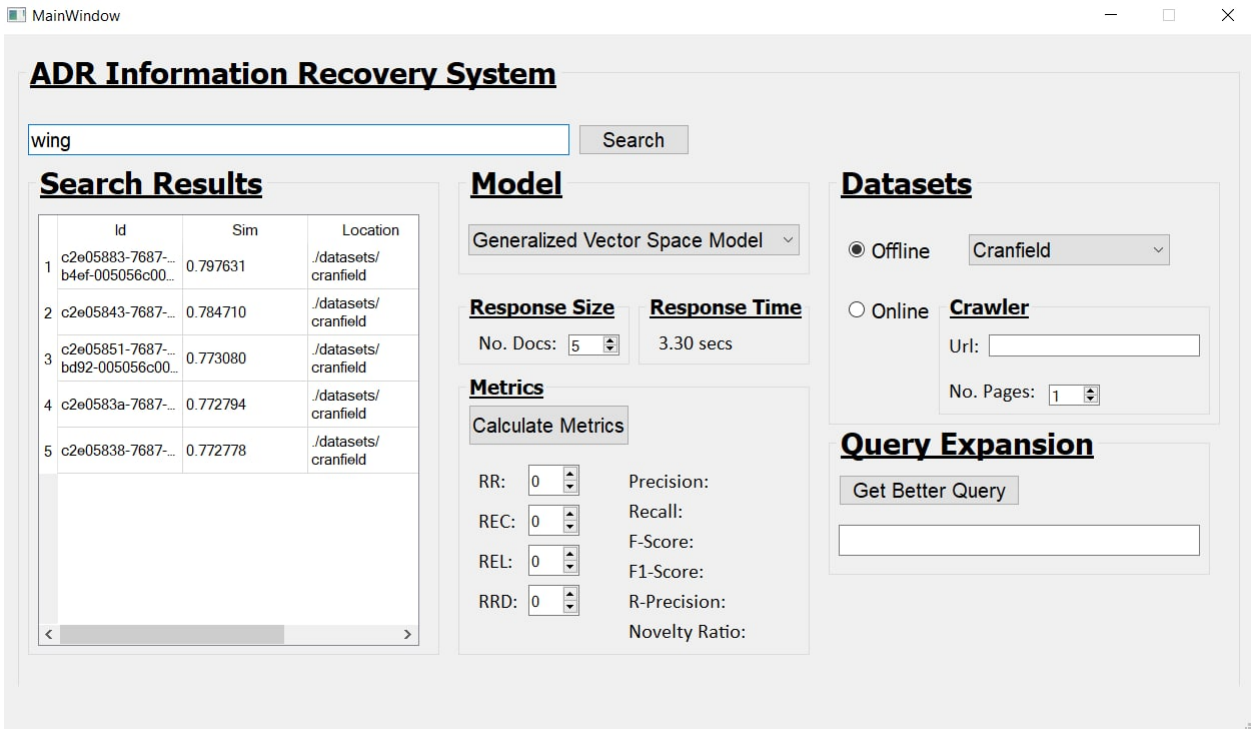
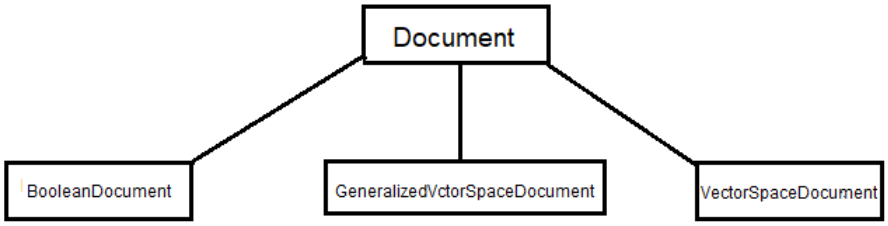


*Fig*10.**Búsqueda con Vectorial Generalizado.**

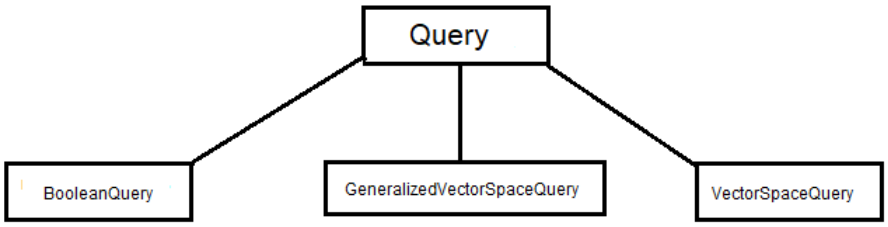
2.3. Modelos y otras características:

Los modelos implementados para nuestro sistema de recuperación de información son el Booelano, Vectorial y Vectorial Generalizado (el resumen de estos modelos se presentará en el informe definitivo). La implementación de estos se encuentra en el directorio llamado retrieval_models, cada uno de ellos se encuentra en una carpeta con su propio nombre. Cada modelo se encuentra dividido en tres archivos cuyos formatos son (NOMBRE será el específico de cada modelo):

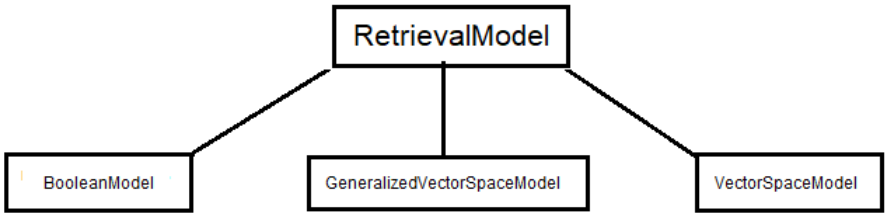
- NOMBRE_document: define una clase NOMBREDocument la cual se encarga de crear el vector del documento deseado con un vocabulario ya definido. Para cada modelo el vector se forma de manera específica, ejemplo: el booleano crea un vector binario pero los vectoriales forman uno basado en la frecuencia de cada término.



- NOMBRE_query: define una clase NOMBREQuery que mediantes sus métodos y propiedades específicas le otorga un formato a la consulta para que sea posible aplicarle la función de similitud con un documento.



- NOMBRE_model: define una clase NOMBREModel la cual mediante sus métodos internos permite calcular la función de similitud entre un documento y la consulta. Recordar que en función del modelo, la función de similitud y operaciones necesarias serán o no diferentes.



Abordemos otros elementos importantes del proyecto, estos los encontramos en el directorio utils. En este se llevaron a cabo implementaciones como:

- El algoritmo de Rocchio que sirve para aplicarle retroalimentación a los modelos vectoriales y otorgarle importancia a la información brindada por unos documentos u otros. Este se encuentra implementado en vector_feedback.py donde además de el método rocchio_algorithm, se implementó retroalimentación básica para modelos vectoriales en el método classical_vector_feedback.
- El archivo query_expansion, como dice su nombre, contiene la implementación de expansión de consultas; utilizada para obtener consultas más específicas y de la cual se esperan reultados más correctos o precisos.
- También se calculan las métricas basadas en la opinión de un experto, entre estas presición, recobrado, medida-f, r-presición y otras. Se pueden encontrar en metrics.py.

2.4. ¿Mejoras?

El proyecto no es perfecto y por tanto es propicio a mejoras. Es fácil notar que la interfaz del usuario no es la más simple ni la más vistosa, tomando como referencia otros buscadores famosos, por tanto este sería un buen aspecto a tener en cuenta. Notar que la ejecución de las consultas con modelos como el vectorial generalizado tardan un tiempo inadecuado, mayormente en la primera ejecución. Lo anterior responde a que es un modelo de recuperación de información bastante costoso y susceptible a ordenadores con pocas capacidades. Teniendo en cuenta todo lo anterior la ejecución de búsquedas con cualquier modelo resulta bastante costoso por tanto es un aspecto mejorable.

3. Resultados y Conclusiones

Los 3 modelos funcionan tan lento como amplio el espacio de búsqueda, en especial el Vectorial Generalizado. Se les realizaron pruebas con diferentes colecciones de docuementos, cada una arrojó resultados diferentes. Notamos que en espacios de búsqueda más reducidos los modelos obtienen resultados mucho mejores, ejemplo, para la colección de cranfield con 1400 documentos los modelos vectoriales tienen una precisión, alrededor de 0.007000000000000002, dependiendo de la calidad de la consulta y la cantidad de documentos relevantes que esta posea. El recobrado se comporta un poco mejor de manera general (oscila alrededor de 0.03922214471352403) para la mayor parte de las consultas, también depende en gran medida de la cantidad k de documentos que se designe recuperar. El modelo booleano no ha podido ser probado correctamente debido al formato de las consultas, esto provoca que encuentre muchos documentos que contienen los términos y quizás no era un documento seleccionado como relevante. Para las colecciones vaswani y trec-covid que poseen un gran número de documentos los modelos brindan peores resultados debido al gran espacio de búsqueda y a la gran cantidad de términos, el vectorial generalizado solo fue testeado 1 vez con estas colecciones debido al tiempo que tardó, en solo procesar los vectores de correlación entre términos para cada uno de ellos (ki). Finalmente se obtuvo una Medida-F de 0.008289312401129892 y de F1 de 0.01117802268269257, como promedio. Para ver información más detallada de los resultados abrir los archivos en formato txt junto al informe en el repositorio de Git Hub o ver las imágenes (Fig 11, Fig 12, Fig 13, Fig 14 y Fig 15). A modo de conclusión los modelos son muy mejorables basándonos en los experimentos realizados con dichas colecciones. Sin embargo en pruebas individuales realizads por el equipo los modelos no se comportan tan erróneos. Notamos que el Vectorial Generalizado es bastante probable mejore los resultados del vectorial simple pero es muy difícil comprobarlo debido al tiempo que tarda en ejecutar todos los cáculos necesarios. El modelo booleano encuentra resultados muy poderosos, sin embargo desprecia algunos parciales que son relevantes pero quizás no contiene todos los términos precisados en la consulta, además está sujeto a la correcta especificación de as necesidades en la consulta. El

proyecto brinda una perspectiva de lo importante que son los modelos de búsqueda y el trabajo constante en mejorarlos con la misión de obtener resutados más precisos y más rápidos.

```
GeneralizedVectorSpaceModel
Acuracy
{1: 0.14285714285714285, 2: 0.047619047619047616, 3: 0.047619047619047616, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0}
Recovery
{1: 0.10344827586206896, 2: 0.04, 3: 0.1111111111111111, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0}
F.Score
{1: 0.13274336283185842, 2: 0.045871559633027525, 3: 0.053763440860215055, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0}
F1.Score
{1: 0.12000000000000002, 2: 0.043478260869565216, 3: 0.06666666666666667, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0}
```

*Fig*11.**Test en la colección Trec-Covid con los primeros 5000 y se le pidió al modelo Vectorial Generalizado los 10 primeros.**

```
VectorSpaceModel, System Recuperation Model
Cranfield, search space
Response 50, recovered documents
Acuracy
{1: 0.02, 2: 0, 3: 0.02, 4: 0, 5: 0, 6: 0.02, 7: 0, 8: 0.02, 9: 0, 10: 0, 11: 0.02, 12: 0.02, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0,
19: 0, 20: 0.02}
Recovery
{1: 0.034482758620689655, 2: 0, 3: 0.1111111111111111, 4: 0, 5: 0, 6: 0.2, 7: 0, 8: 0.08333333333333333, 9: 0, 10: 0, 11: 0.125, 12: 0.16666666666666666, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.1}
F.Score
{1: 0.021834061135371178, 2: 0, 3: 0.023923444976076555, 4: 0, 5: 0, 6: 0.024390243902439025, 7: 0, 8: 0.023584905660377736, 9: 0, 10: 0, 11: 0.02403846153846154, 12: 0.024271844660194174, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.023809523809523808}
F1.Score
{1: 0.025316455696202253, 2: 0, 3: 0.033898305084745766, 4: 0, 5: 0, 6: 0.036363636363636366, 7: 0, 8: 0.03225806451612903, 9: 0, 10: 0, 11: 0.034482758620689655, 12: 0.03571428571428571, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.03333333333333333}
```

*Fig*12.**Test en la colección Cranfield y se le pidió al modelo Vectorial los 50 primeros.**

```
Vaswani, search space primeros 4000 documentos
Response 20, recovered documents
GeneralizedVectorSpaceModel
Acuracy
{1: 0.05, 2: 0, 3: 0.05, 4: 0, 5: 0.05, 6: 0, 7: 0.05, 8: 0.05, 9: 0, 10: 0}
Recovery
{1: 0.034482758620689655, 2: 0, 3: 0.125, 4: 0, 5: 0.030303030303030304, 6: 0, 7: 0.1, 8: 0.14285714285714285, 9: 0, 10: 0}
F.Score
{1: 0.045871559633027525, 2: 0, 3: 0.056818181818181816, 4: 0, 5: 0.04424778761061947, 6: 0, 7: 0.05555555555555555, 8: 0.05747126436781609, 9: 0, 10: 0}
F1.Score
{1: 0.04081632653061224, 2: 0, 3: 0.07142857142857142, 4: 0, 5: 0.03773584905660377, 6: 0, 7: 0.06666666666666667, 8: 0.07407407407407407, 9: 0, 10: 0}
```

*Fig*13.**Test en la colección Vaswani, con los primeros 4000 y se le pidió al modelo Vectorial Generalizado los 20 primeros.**

```
Cranfield, search space
Response 20, recovered documents
GeneralizedVectorSpaceModel
Acuracy
{ 1: 0, 2: 0.05, 3: 0.05, 4: 0, 5: 0, 6: 0.02, 7: 0, 8: 0, 9: 0, 10: 0.05}
Recovery
{1: 0, 2: 0.04, 3: 0.2, 4: 0, 5: 0, 6: 0.083333333333, 7: 0, 8: 0, 9: 0, 10: 0.14285714285714285}
F.Score
{1: 0, 2: 0.047619047619047616, 3: 0.058823529411764705, 4: 0, 5: 0, 6: 0.02358490566037736, 7: 0, 8: 0, 9: 0, 10:0.05747126436781609}
F1.Score
{1: 0, 2: 0.044444444444444446, 3: 0.08, 4: 0, 5: 0, 6: 0.03225806451612903, 7: 0, 8: 0, 9: 0, 10:0.07407407407407407}
```

*Fig*14.**Test en la colección Cranfield y se le pidió al modelo Vectorial Generalizado los 20 primeros.**

VectorSpaceModel
Trec-Covid
Response 30, recovered documents
Acuracy
{1: 0.02, 2: 0, 3: 0.02, 4: 0, 5: 0, 6: 0.02, 7: 0, 8: 0.02, 9: 0, 10: 0, 11: 0.02, 12: 0.02, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0,
19: 0, 20: 0.02, 21: 0, 22: 0, 23: 0.02, 24: 0, 25: 0, 26: 0.02, 27: 0, 28: 0, 29: 0.02, 30: 0, 31: 0, 32: 0.02, 33: 0, 34: 0, 35: 0, 36: 0, 37: 0, 38: 0, 39: 0, 40:
0.02, 41: 0.02, 42: 0, 43: 0, 44: 0, 45: 0.02, 46: 0.04, 47: 0.04, 48: 0, 49: 0, 50: 0, 51: 0.02, 52: 0, 53: 0, 54: 0, 55: 0.02, 56: 0, 57: 0.02, 58: 0, 59: 0.02, 60:
0, 61: 0, 62: 0, 63: 0, 64: 0, 65: 0, 66: 0, 67: 0, 68: 0, 69: 0, 70: 0, 71: 0, 72: 0.02, 73: 0, 74: 0.02, 75: 0.02, 76: 0.02, 77: 0.02, 78: 0, 79: 0, 80: 0, 81: 0,
82: 0, 83: 0.02, 84: 0, 85: 0.02, 86: 0, 87: 0.04, 88: 0, 89: 0, 90: 0.02, 91: 0, 92: 0.02, 93: 0, 94: 0.02, 95: 0, 96: 0, 97: 0, 98: 0, 99: 0, 100: 0.02}
Recovery
{1: 0.034482758620689655, 2: 0, 3: 0.1111111111111111, 4: 0, 5: 0, 6: 0.2, 7: 0, 8: 0.08333333333333333, 9: 0, 10: 0, 11: 0.125, 12: 0.16666666666666666, 13: 0, 14: 0,
15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.1, 21: 0, 22: 0, 23: 0.030303030303030304, 24: 0, 25: 0, 26: 0.14285714285714285, 27: 0, 28: 0, 29: 0.1, 30: 0, 31: 0, 32:
0.14285714285714285, 33: 0, 34: 0, 35: 0, 36: 0, 37: 0, 38: 0, 39: 0, 40: 0.07692307692307693, 41: 0.25, 42: 0, 43: 0, 44: 0, 45: 0.07692307692307693, 46: 0.125, 47:
0.13333333333333333, 48: 0, 49: 0, 50: 0, 51:
0.09090909090909091, 52: 0, 53: 0, 54: 0, 55: 0.09090909090909091, 56: 0, 57: 0.06666666666666667, 58: 0, 59: 0.2, 60: 0, 61: 0, 62: 0,
63: 0, 64: 0, 65: 0, 66: 0, 67: 0, 68: 0, 69: 0, 70: 0, 71: 0, 72: 0.05555555555555555, 73: 0, 74: 0.14285714285714285, 75: 0.16666666666666666, 76: 0.125, 77:
0.14285714285714285, 78: 0, 79: 0, 80: 0, 81: 0, 82: 0, 83: 0.2, 84: 0, 85: 0.2, 86: 0, 87: 0.2222222222222222,
88: 0, 89: 0, 90: 0.07142857142857142, 91: 0, 92: 0.07142857142857142, 93: 0, 94: 0.07692307692307693, 95: 0, 96: 0, 97: 0, 98: 0, 99: 0, 100: 0.1}
F.Score
{1: 0.021834061135371178, 2: 0, 3: 0.023923444976076555, 4: 0, 5: 0, 6: 0.024390243902439025, 7: 0, 8: 0.02358490566037736, 9: 0, 10: 0, 11: 0.02403846153846154, 12:
0.024271844660194174, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.023809523809523808, 21: 0, 22: 0, 23: 0.02145922746781116, 24: 0, 25: 0, 26:
0.024154589371980676, 27: 0, 28: 0, 29: 0.023809523809523808, 30: 0, 31: 0, 32: 0.024154589371980676, 33: 0, 34: 0, 35: 0, 36: 0, 37: 0, 38: 0, 39: 0, 40:
0.023474178403755867, 41: 0.024509803921568627, 42: 0, 43: 0, 44: 0,
45: 0.023474178403755867, 46: 0.046296296296296294, 47: 0.046511627906976744, 48: 0, 49: 0, 50: 0, 51: 0.023696682464454975, 52: 0, 53:
0, 54: 0, 55: 0.023696682464454975, 56: 0, 57: 0.023255813953488372, 58: 0, 59: 0.024390243902439025, 60: 0, 61: 0, 62: 0, 63: 0, 64: 0, 65: 0, 66: 0, 67: 0, 68: 0,
69: 0, 70: 0, 71: 0, 72: 0.022935779816513763, 73: 0, 74: 0.024154589371980676, 75: 0.024271844660194174, 76: 0.02403846153846154, 77: 0.024154589371980676, 78: 0, 79:
0, 80: 0, 81: 0, 82: 0, 83: 0.024390243902439025, 84: 0, 85: 0.024390243902439025, 86: 0, 87: 0.04784688995215311, 88: 0, 89: 0, 90: 0.02336448598130841, 91: 0, 92:
0.02336448598130841, 93: 0, 94: 0.023474178403755867, 95: 0, 96: 0, 97: 0, 98: 0, 99: 0, 100: 0.023809523809523808}
F1.Score
{1: 0.02531645569620253, 2: 0, 3: 0.03389830508474576, 4: 0, 5: 0, 6: 0.036363636363636364, 7: 0, 8: 0.03225806451612903, 9: 0, 10: 0, 11: 0.034482758620689655, 12:
0.03571428571428571, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, 18: 0, 19: 0, 20: 0.03333333333333333, 21: 0, 22: 0, 23: 0.024096385542168676, 24: 0, 25: 0, 26:
0.03508771929824561, 27: 0, 28: 0, 29: 0.03333333333333333, 30: 0, 31: 0, 32: 0.03508771929824561, 33: 0, 34: 0, 35: 0, 36: 0, 37: 0, 38: 0, 39: 0, 40:
0.031746031746031744, 41: 0.037037037037037035, 42: 0, 43: 0, 44: 0, 45: 0.031746031746031744, 46: 0.06060606060606061, 47: 0.06153846153846154, 48: 0, 49: 0, 50: 0,
51: 0.03278688524590164, 52: 0, 53: 0, 54: 0, 55: 0.03278688524590164, 56: 0, 57: 0.03076923076923077, 58: 0, 59: 0.036363636363636364, 60: 0, 61: 0, 62: 0, 63: 0, 64:
0, 65: 0, 66: 0, 67: 0, 68: 0, 69: 0, 70: 0, 71: 0, 72: 0.029411764705882353, 73: 0, 74: 0.03508771929824561, 75: 0.03571428571428571, 76: 0.034482758620689655, 77:
0.03508771929824561, 78: 0, 79: 0, 80: 0, 81: 0, 82: 0, 83: 0.036363636363636364, 84: 0, 85: 0.036363636363636364, 86: 0, 87: 0.024096385542168676, 88: 0, 89: 0, 90:
0.03508771929824561, 91: 0, 92: 0, 93: 0.03333333333333333, 94: 0, 95: 0, 96: 0.03508771929824561, 97: 0, 98: 0, 99: 0, 100: 0.03333333333333333}

*Fig*15.**Recorte del test en la colección Trec-Covid, completa, y se le pidió al modelo Vectorial los 30 primeros.**