Lab3: User Relevance Feedback

Q1 2022-23

Integrants:
Pablo Montón Gimeno
Cristian Sánchez Estapé

Queries y resultados asociados

Dadas las características de la tarea a tratar y del enunciado, se ha seguido de manera rigurosa el esquema sugerido para la implementación del *URF*. Debe decirse, pero, que con cierta confusión dada la naturaleza de *Rocchio* (el hecho de que el resultado de la fórmula sean términos, y cómo estos se relacionan con los documentos involucrados, que añaden palabras, ha sido una complicación) y de los valores de sus parámetros, valores que hemos determinado en base a las experiencias de otros usuarios (registradas en portales como *Wikipedia)*. Además, cabe destacar que la implementación realizada es sumamente flexible: si en una aplicación de *Rocchio* cualquiera se detecta una palabra nueva más relevante que alguna de las que ya estaban incluídas, ésta se incorporará, desplazando la que fuera menos importante en ese momento del flujo de ejecución del programa.

Con nuestra implementación, hemos conseguido que, mediante la entrada de un índice y un conjunto de palabras $\mathcal{Q}$, genere una lista con los $\mathcal{K}$ documentos más relevantes que incluyen dichos términos. Los resultados de una query de ejemplo serían los siguientes:

INPUT	
mind sadness death	
PATH	**SCORE**
./20_newsgroups/talk.politics.guns/0016434	295.62
./20_newsgroups/talk.politics.mideast/0017077	244.82
./20_newsgroups/talk.politics.guns/0016630	240.31
./20_newsgroups/alt.atheism/0000308	235.01
./20_newsgroups/alt.atheism/0000936	175.76

Figura 1: query de los términos "mind sadness death"

Tal y como se puede apreciar en la figura 1, los resultados registran tanto el *path* al documento concreto como su *score* asociado. Cabe destacar que dicho *score* viene definido tanto por el *tf-idf* de los términos involucrados en cada paso de aplicación de *Rocchio* como del peso inicial de cada palabra introducida. En este caso, vemos que, para lo que el usuario buscaba, el resultado ha sido considerablemente bueno. Como este ejemplo, se han registrado 8 resultados (incluido éste) sin uso de operadores **^**/**~**, y 6 resultados con dichos operadores. Con ambos casos, se ve claramente cómo el uso de **^**/**~** influye en la puntuación que cada documento recibe, aunque también es palpable cómo ésta no necesariamente debe ser más grande por el uso de dichos operadores (es especialmente visible en el caso de **~**, que involucra la aplicación de la *Levenshtein Edit Distance* y que, como en el caso de las búsquedas "city" y "city~2", lo demuestra).

En otro orden de cosas, los resultados de las queries generalmente tienen sentido: hemos realizado 4 *queries* distintas, y todas ellas o bien con términos distintos, o bien con pesos variantes. De todas estas, 3 de ellas generan resultados virtualmente sólidos en su totalidad, puesto que las palabras que se han añadido (a lo largo de las iteraciones realizadas) encajan temáticamente y contextualmente con las de la búsqueda original. Sin embargo, una de ellas arroja resultados un poco difusos para lo que es la búsqueda en sí,

ya que éstos parecen alejarse de lo que originalmente se pretendía encontrar. Las búsquedas han sido:

INPUT	
basketball baseball match	
PATH	**SCORE**
./20_newsgroups/rec.sport.hockey/0010108	179.5
./20_newsgroups/rec.sport.hockey/0010091	170.41
./20_newsgroups/rec.sport.hockey/0010687	134.07
./20_newsgroups/rec.sport.hockey/0010630	120.88
./20_newsgroups/misc.forsale/0006076	118.73
baseball, basketball, match, players, cuba, alchemy.chem.utoronto.ca, nba, defect, 1993apr6.141557.8864, nhl, adobe.com, perfectly, genesis, they, crawled, expos, gerald, game, mattel, league, situation, crawl, czech, lt8dl1inn2u2, sca's	

Figura 2: query de los términos "basketball baseball match"

Como se puede ver en este caso, la búsqueda es claramente de deportes, y parte de los resultados encontrados encajan con ello (palabras como "nhl", "game", "league" y "players" encaja, y otras como "czech" o "cuba" también lo pueden hacer si su relación con la *query* viene dada por un equipo, una liga, una competición, etc.). Sin embargo, encontramos resultados como "adobe.com" o "lt8dl1inn2u2" que, si bien pueden guardar cierta relación con la búsqueda original, no parecen, a priori, ser demasiado coherentes o relevantes para el usuario.

INPUT	
god^5 jesus death	
PATH	**SCORE**
./20_newsgroups/talk.religion.misc/0019535	323.03
./20_newsgroups/talk.religion.misc/0019686	308.08
./20_newsgroups/talk.religion.misc/0019235	308.08
./20_newsgroups/talk.religion.misc/0019516	291.46
./20_newsgroups/talk.religion.misc/0019905	286.15
death, god jesus, eternal, hell, atheists, die, unfair, atterlep, vela.acs.oakland.edu, interpreters, believe, problem, expecting, bible, condemn, christians, literal, forever, fail, threads, fires, conciousness, penalty, deterrent	

Figura 3: query de los términos "god jesus death^10"

Para esta *query*, se puede apreciar claramente cómo los términos añadidos encajan completamente con el contexto de la búsqueda. Además, vale la pena destacar que ello venga dado tanto por el peso de los términos que se introducen (en este caso, *death* será sumamente relevante por su peso asociado) como por el tipo de terminología introducida (claramente todas las palabras son de carácter contextualmente religioso, con lo cual es menos difuso que el vocabulario deportivo préviamente usado, por ejemplo).

Por todo ello, en lo que a *precision* y *recall* respecta, resulta razonable afirmar que ambas métricas aumentan: teniendo en cuenta que se parte de una *query* original $\mathcal{Q}$, y que, de entrada, en dicho conjunto (asumimos) no todos los documentos incluidos pertenecen al conjunto de documentos relevantes $\mathfrak{R}$, a medida que se aplica Rocchio, los resultados generados convergen (teóricamente) cada vez más a lo que desea el usuario. Esto es lo mismo que decir que tanto *precision* como *recall* aumentan porque, principalmente, se trabaja sobre un conjunto de documentos $\mathcal{K}$ de tamaño siempre igual, y tras cada iteración en que se afina la *query* $\mathcal{Q}$, conseguimos resultados que se acercan más a lo que se busca. Cabe decir que ello es una aseveración sustentada en la hipótesis (razonablemente justificada) que la métrica de *score* de los documentos sea rigurosa y debidamente computada (sabemos que ésta métrica se sustenta en el *tf-idf*, que sabemos con certeza que sirve, de manera orientativa al menos, para identificar la relevancia de un documento según sus vocablos).

Finalmente, como breve apéndice, cabe añadir que los experimentos básicos aplicados a los índices *novels* y *arxivs* añaden información relativa al script desarrollado: mientras que *arxivs* no presenta resultados demasiado alejados a lo constatado con *news*, el índice *novels* tiene, en media, unos *scores* relevantemente pequeños en comparación con los otros casos, y aunque se usen operadores **^**/**~**, se siguen manteniendo por debajo de lo que se puede ver. Con esto, si lo comparamos con la totalidad de los experimentos realizados para cada índice, podemos intuir que la justificación de dicho fenómeno es debido al tipo de documentos albergados en el índice *novels*, que más concretamente, para este laboratorio, podría venir a decir que los documentos (contextualmente) literarios o con un uso del lenguaje principalmente estilístico podrían tender a diluir métricas como el *tf-idf*, con lo que es más difícil encontrar palabras eminentemente más importantes que otras y, por lo tanto, conseguir *scores* más elevados.

Experimentación de datos

El programa, para calcular el valor de la fórmula de *Rocchio* y para una query dada, utiliza unos parámetros que a continuación cambiaremos para así observar como varían los documentos relevantes. En el fichero ***data***, anexado a este documento, contiene todos los datos de experimentación de la práctica, y más concretamente los referentes a las pruebas para cambiar los parámetros de [*nrounds*, *k*, *R*, *α*, *β*]. Para cada parámetro, utilizamos varios valores de éste para así intentar encontrar el valor más óptimo, aquél que haga que el programa funcione de una manera más rápida, obteniendo unos valores que arrojen resultados coherentes en el contexto de la *query*. Cabe añadir que, mientras se varía el valor de un parámetro, los restantes se mantienen a un valor fijo (que en el excel adjunto a este informe vienen indicados como ***BASIC VALUES***).

- ***nrounds:*** en clase aprendimos que, en un contexto óptimo, el valor de *nrounds*, en caso de aumentar, no arrojaba valores ciertamente mejores, ya que los documentos indicados no varían. Como podemos ver en los experimentos que hemos realizado, con valores (1,5,10,20) podemos observar ésta propiedad, ya que los documentos que obtenemos, para los valores a partir de 5, son idénticos. Por lo tanto lo ideal sería coger un número de *nrounds* cercano a 5, puesto que vemos que es donde el resultado se vuelve estático. Algo a destacar es que aquí se vuelve evidente cómo el

score depende de los *tf-idf* y el peso que *Rocchio* da a cada palabra (para un *nrounds* muy grande, el *score* se dispara).

- **α, β:** como hemos hablado anteriormente, los valores de α y β ya han sido investigados y experimentados por otros usuarios que han visto que los óptimos son α = 1; β = 0.8. Igualmente, hemos querido probar con otros valores para ver si podemos aseverar que éstos dan los resultados óptimos. Con los experimentos realizados, hemos visto que los valores mencionados, o similares, devuelven los mismos documentos, aunque generalmente con un orden variado. En base a los *scores* asociados a cada experimento, y al hecho de que α y β influya en esta misma métrica, suponemos que el valor ideal estándar es el valor básico ya elegido, puesto que, a parte de los resultados de los *scores*, tienen coherencia para ponderar los resultados de la manera en que se hace (con α = 1 garantizamos que los términos "originales" sean unitariamente importantes, mientras que con β = 0.8 definimos que los términos que puedan añadirse de los documentos relevantes sean menos importantes que los que ya teníamos originalmente).
- **R:** Al cambiar *R* básicamente lo que hacemos es aumentar los términos relevantes que al final se utilizan para calcular el *score*. Como vemos, el cambio de la cantidad total de palabras relevantes puede dar pie a que ciertos documentos (que originalmente no tenían relación o relevancia alguna para la búsqueda del usuario) ahora cobren importancia e influyan en el resultado final. Nosotros hemos pensado que un valor medio (similar a 25) puede dar unos resultados coherentes, ya que, por un lado, se añaden suficientes términos como para tratar de contextualizar la búsqueda del usuario, mientras que se evita la potencial influencia de documentos que no tienen nada que ver con dicha búsqueda.
- **K:** Al cambiar K, que se utiliza para limitar el número de documentos involucrados en el cálculo de la *query* de *Rocchio*, vemos cómo, al trabajar con dichos documentos seleccionados, hay un resultado similar a la experimentación de R, ya que al utilizar una mayor cantidad de documentos, podemos observar cómo se disuelve la *score* al tener una mayor selección de palabras a incorporar en la *query*. Éste fenómeno, al igual que en R, viene a indicar que, para valores elevados de K, habrá una mayor oferta de palabras a tener en cuenta, lo que implica que el conjunto de documentos relevantes para la búsqueda del usuario será menos precisa (a pesar de aumentar el *recall*), y por lo tanto las *scores* descenderán. En su defecto, para valores pequeños de *K*, (de acuerdo con lo comentado en este informe) tendremos una alta precisión y, en su defecto, un descenso del *recall*; pero en ambos casos lo que sí es cierto es que tendremos una disolución del contexto de la búsqueda, con lo que, para un valor de *K* intermedio (como el elegido) sería, según nuestro criterio, el más indicado.

En definitiva, vemos cómo la modificación de parámetros influye en lo que hasta ahora nos hemos referido por contexto, entendiendo éste como el medio abstracto y subyacente a lo que el usuario busca con su query. La modificación y afinamiento de los parámetros sirve para acotar dicho contexto y, por ello, creemos que los valores que se han marcado como **BASIC VALUES** y que hemos explorado son, al menos para el conjunto de muestras que hemos presentado, los más adecuados para garantizar una búsqueda de documentos adecuada y coherente con lo que (en principio) desea el usuario.