

***- Sistemas de recomendación -
- Modelos basados en el conocimiento -***

Índice de usuario

- Análisis de varios documentos de ejemplo con la aplicación construida.....1
- Análisis de documents01.txt.....4
 - Análisis de documents01.txt.....4
 - Análisis de d1.txt.....6
 - Análisis de d2.txt.....7
- Conclusiones.....8

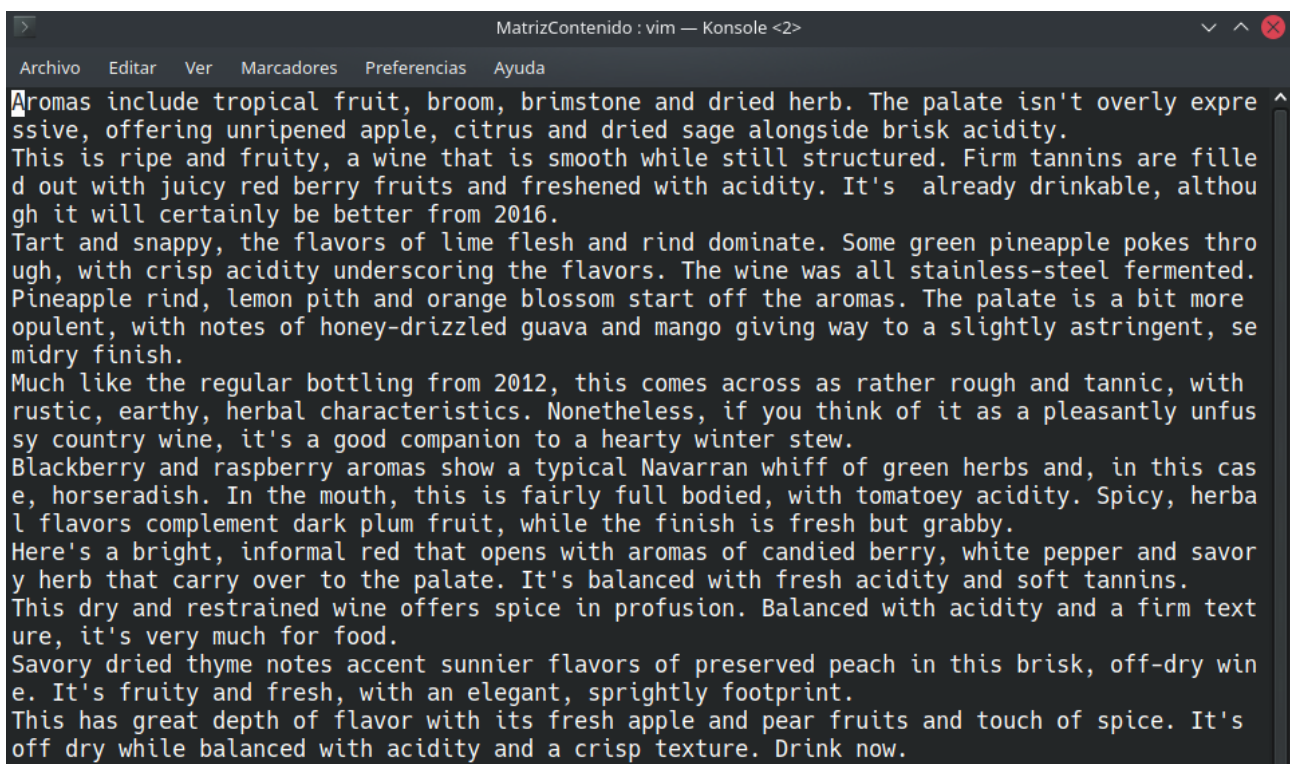
- Análisis de varios documentos de ejemplo con la aplicación construida.

Empezaremos indicando que se han usado los ficheros de ejemplo proporcionados por el profesor para esta actividad en el repositorio de GitHub :

<https://github.com/cexposit/ull-gco/tree/main/examples-documents>

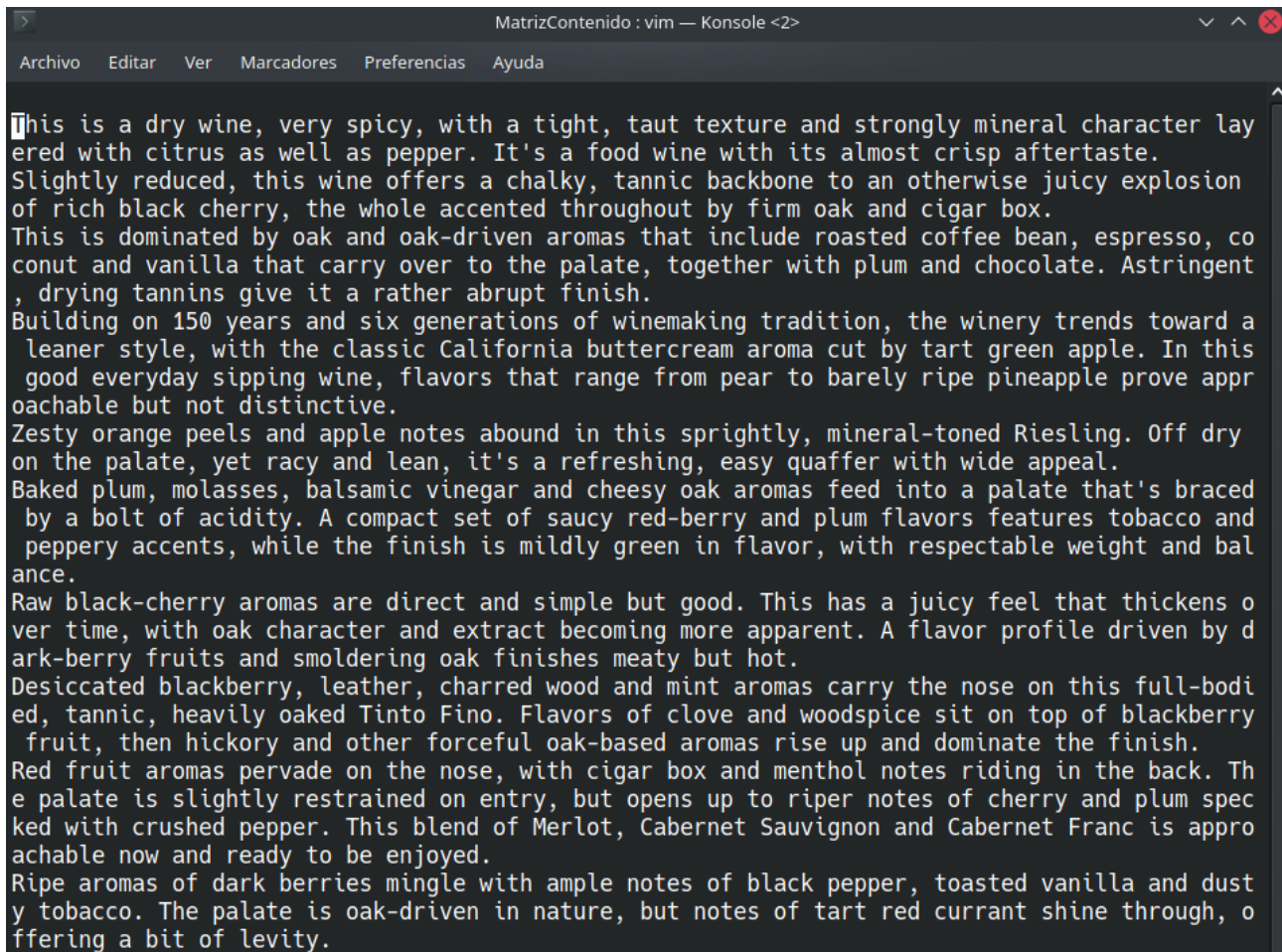
▪ Tenemos 3 ejemplos, en inglés, sobre valoraciones de distintos tipos de vinos. Los contenidos de cada fichero incluyen distintos tipos de *documentos*, líneas de texto seguidas hasta encontrar el primer salto de línea, que conforman cada una de las valoraciones mencionadas anteriormente :

- Fichero **documents01.txt** :

A screenshot of a Vim editor window titled "MatrizContenido : vim — Konsole <2>". The window has a menu bar with "Archivo", "Editar", "Ver", "Marcadores", "Preferencias", and "Ayuda". The main text area displays a wine review document. The text is as follows:

Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity. This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016. Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented. Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semi-dry finish. Much like the regular bottling from 2012, this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics. Nonetheless, if you think of it as a pleasantly unfussy country wine, it's a good companion to a hearty winter stew. Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby. Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins. This dry and restrained wine offers spice in profusion. Balanced with acidity and a firm texture, it's very much for food. Savory dried thyme notes accent sunnier flavors of preserved peach in this brisk, off-dry wine. It's fruity and fresh, with an elegant, sprightly footprint. This has great depth of flavor with its fresh apple and pear fruits and touch of spice. It's off dry while balanced with acidity and a crisp texture. Drink now.

- Fichero **documents02.txt** :



The image shows a screenshot of a Vim editor window titled "MatrizContenido : vim — Konsole <2>". The window has a menu bar with options: Archivo, Editar, Ver, Marcadores, Preferencias, and Ayuda. The main text area contains several paragraphs of wine descriptions. The text is as follows:

This is a dry wine, very spicy, with a tight, taut texture and strongly mineral character layered with citrus as well as pepper. It's a food wine with its almost crisp aftertaste. Slightly reduced, this wine offers a chalky, tannic backbone to an otherwise juicy explosion of rich black cherry, the whole accented throughout by firm oak and cigar box. This is dominated by oak and oak-driven aromas that include roasted coffee bean, espresso, coconut and vanilla that carry over to the palate, together with plum and chocolate. Astringent, drying tannins give it a rather abrupt finish.

Building on 150 years and six generations of winemaking tradition, the winery trends toward a leaner style, with the classic California buttercream aroma cut by tart green apple. In this good everyday sipping wine, flavors that range from pear to barely ripe pineapple prove approachable but not distinctive.

Zesty orange peels and apple notes abound in this sprightly, mineral-toned Riesling. Off dry on the palate, yet racy and lean, it's a refreshing, easy quaffer with wide appeal.

Baked plum, molasses, balsamic vinegar and cheesy oak aromas feed into a palate that's braced by a bolt of acidity. A compact set of saucy red-berry and plum flavors features tobacco and peppery accents, while the finish is mildly green in flavor, with respectable weight and balance.

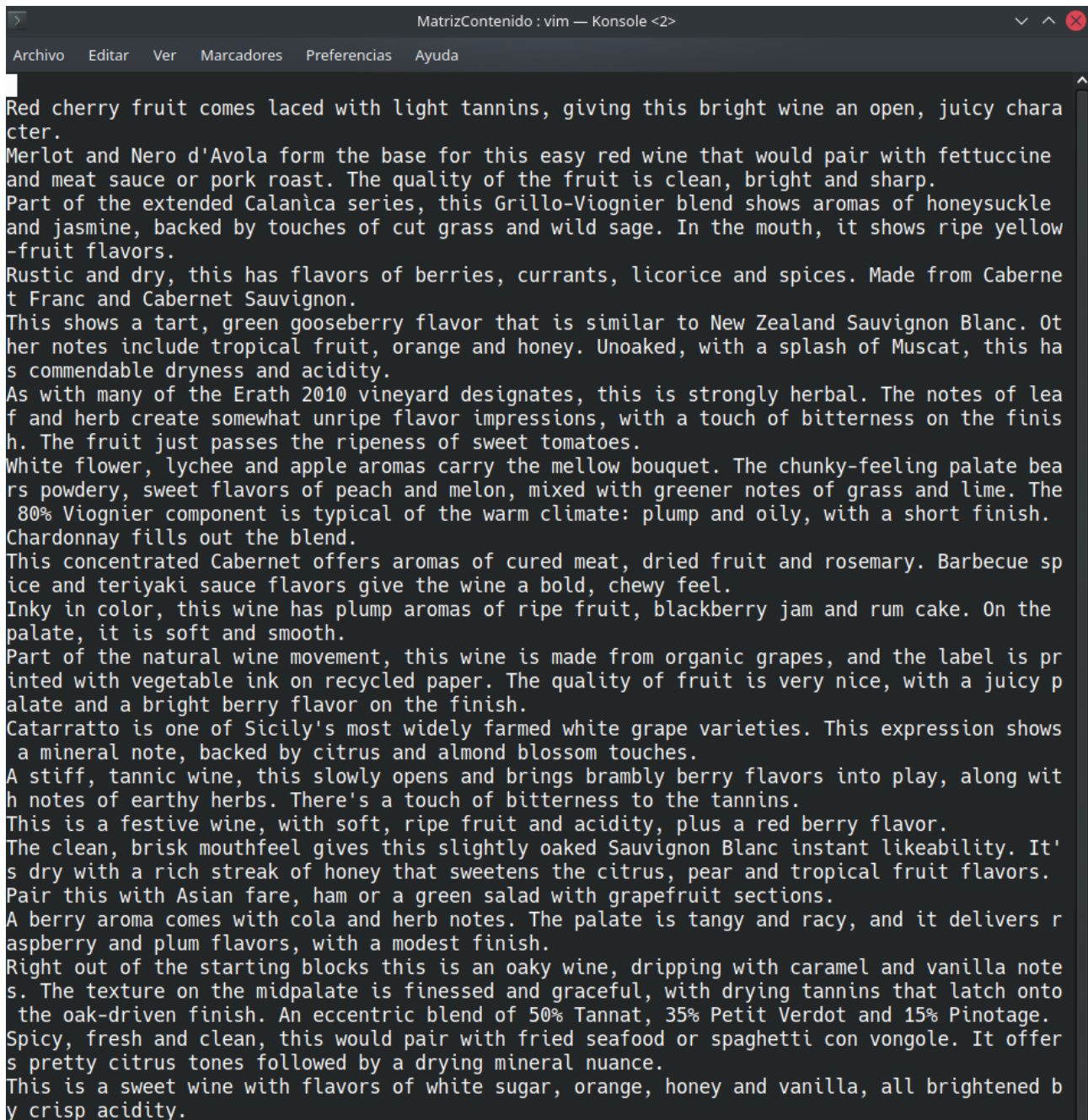
Raw black-cherry aromas are direct and simple but good. This has a juicy feel that thickens over time, with oak character and extract becoming more apparent. A flavor profile driven by dark-berry fruits and smoldering oak finishes meaty but hot.

Desiccated blackberry, leather, charred wood and mint aromas carry the nose on this full-bodied, tannic, heavily oaked Tinto Fino. Flavors of clove and woodspice sit on top of blackberry fruit, then hickory and other forceful oak-based aromas rise up and dominate the finish.

Red fruit aromas pervade on the nose, with cigar box and menthol notes riding in the back. The palate is slightly restrained on entry, but opens up to riper notes of cherry and plum speckled with crushed pepper. This blend of Merlot, Cabernet Sauvignon and Cabernet Franc is approachable now and ready to be enjoyed.

Ripe aromas of dark berries mingle with ample notes of black pepper, toasted vanilla and dusty tobacco. The palate is oak-driven in nature, but notes of tart red currant shine through, offering a bit of levity.

- Fichero **documents03.txt** :



The screenshot shows a terminal window with the title 'MatrizContenido : vim — Konsole <2>'. The menu bar includes 'Archivo', 'Editar', 'Ver', 'Marcadores', 'Preferencias', and 'Ayuda'. The text content is as follows:

Red cherry fruit comes laced with light tannins, giving this bright wine an open, juicy character.

Merlot and Nero d'Avola form the base for this easy red wine that would pair with fettuccine and meat sauce or pork roast. The quality of the fruit is clean, bright and sharp.

Part of the extended Calanica series, this Grillo-Viognier blend shows aromas of honeysuckle and jasmine, backed by touches of cut grass and wild sage. In the mouth, it shows ripe yellow-fruit flavors.

Rustic and dry, this has flavors of berries, currants, licorice and spices. Made from Cabernet Franc and Cabernet Sauvignon.

This shows a tart, green gooseberry flavor that is similar to New Zealand Sauvignon Blanc. Other notes include tropical fruit, orange and honey. Unoaked, with a splash of Muscat, this has commendable dryness and acidity.

As with many of the Erath 2010 vineyard designates, this is strongly herbal. The notes of leaf and herb create somewhat unripe flavor impressions, with a touch of bitterness on the finish. The fruit just passes the ripeness of sweet tomatoes.

White flower, lychee and apple aromas carry the mellow bouquet. The chunky-feeling palate bears powdery, sweet flavors of peach and melon, mixed with greener notes of grass and lime. The 80% Viognier component is typical of the warm climate: plump and oily, with a short finish. Chardonnay fills out the blend.

This concentrated Cabernet offers aromas of cured meat, dried fruit and rosemary. Barbecue spice and teriyaki sauce flavors give the wine a bold, chewy feel.

Inky in color, this wine has plump aromas of ripe fruit, blackberry jam and rum cake. On the palate, it is soft and smooth.

Part of the natural wine movement, this wine is made from organic grapes, and the label is printed with vegetable ink on recycled paper. The quality of fruit is very nice, with a juicy palate and a bright berry flavor on the finish.

Catarratto is one of Sicily's most widely farmed white grape varieties. This expression shows a mineral note, backed by citrus and almond blossom touches.

A stiff, tannic wine, this slowly opens and brings brambly berry flavors into play, along with notes of earthy herbs. There's a touch of bitterness to the tannins.

This is a festive wine, with soft, ripe fruit and acidity, plus a red berry flavor.

The clean, brisk mouthfeel gives this slightly oaked Sauvignon Blanc instant likeability. It's dry with a rich streak of honey that sweetens the citrus, pear and tropical fruit flavors.

Pair this with Asian fare, ham or a green salad with grapefruit sections.

A berry aroma comes with cola and herb notes. The palate is tangy and racy, and it delivers raspberry and plum flavors, with a modest finish.

Right out of the starting blocks this is an oaky wine, dripping with caramel and vanilla notes. The texture on the midpalate is finessed and graceful, with drying tannins that latch onto the oak-driven finish. An eccentric blend of 50% Tannat, 35% Petit Verdot and 15% Pinotage.

Spicy, fresh and clean, this would pair with fried seafood or spaghetti con vongole. It offers pretty citrus tones followed by a drying mineral nuance.

This is a sweet wine with flavors of white sugar, orange, honey and vanilla, all brightened by crisp acidity.

▪ Por otra parte se ha tomado el fichero **documents01.txt** y se ha modificado en dos ficheros (**d1.txt** y **d2.txt**) para usarse como ejemplos de *juguete* y que se aprecie un poco mejor los resultados.

- Fichero **d1.txt**. Sólo tiene dos documentos exactamente iguales salvo que ambos difieren en una palabra en cada una (**tropic/tropical**) :

```
MatrizContenido : vim — Konsole <2>
Archivo  Editar  Ver  Marcadores  Preferencias  Ayuda
Aromas include tropic fruit, broom, brimstone and dried herb.
Aromas include tropical fruit, broom, brimstone and dried herb.
```

- Fichero **d2.txt**. También dos documentos iguales salvo que se ha eliminado una palabra en el primer documento (**tropical**) :

```
MatrizContenido : vim — Konsole
Archivo  Editar  Ver  Marcadores  Preferencias  Ayuda
Aromas include fruit, broom, brimstone and dried herb.
Aromas include tropical fruit, broom, brimstone and dried herb.
```

- Descritos los ficheros a usar, haremos algunos análisis.

▪ Análisis de documents01.txt

◦ Ejecutamos el programa desde la terminal, haciendo un **\$./CalcSim documents01.txt** y vemos el resultado de los cálculos de la similaridad coseno en la matriz del final del análisis:

```
Archivo  Editar  Ver  Marcadores  Preferencias  Ayuda
Documento 10 :
-----
[1] - { acidity, (1, 0.154902, 0.154902)}
[2] - { apple, (1, 0.69897, 0.69897)}
[3] - { balanced, (1, 0.522879, 0.522879)}
[4] - { crisp, (1, 0.69897, 0.69897)}
[5] - { depth, (1, 1, 1)}
[6] - { drink, (1, 1, 1)}
[7] - { dry, (1, 0.69897, 0.69897)}
[8] - { flavor, (1, 1, 1)}
[9] - { fresh, (1, 0.39794, 0.39794)}
[10] - { fruits, (1, 0.69897, 0.69897)}
[11] - { great, (1, 1, 1)}
[12] - { has, (1, 1, 1)}
[13] - { its, (1, 0.221849, 0.221849)}
[14] - { its, (1, 1, 1)}
[15] - { now, (1, 1, 1)}
[16] - { off, (1, 0.69897, 0.69897)}
[17] - { pear, (1, 1, 1)}
[18] - { spice, (1, 0.69897, 0.69897)}
[19] - { texture, (1, 0.69897, 0.69897)}
[20] - { touch, (1, 1, 1)}
[21] - { while, (1, 0.522879, 0.522879)}

- Datos de similaridades coseno -
-----
      (D1) (D2) (D3) (D4) (D5) (D6) (D7) (D8) (D9) (D10)
(D1) | 1 | | 0.0403429 | | 0.0483263 | | 0.0897462 | | 0 | | 0.119041 | | 0.191885 | | 0.0567605 | | 0.119228 | | 0.0959427 |
(D2) | 0.0403429 | 1 | | 0.0806858 | | 0.0487368 | | 0.104044 | | 0.122321 | | 0.200233 | | 0.189535 | | 0.129766 | | 0.160186 |
(D3) | 0.0483263 | | 0.0806858 | 1 | | 0.0897462 | | 0.0415444 | | 0.130986 | | 0.0479714 | | 0.113521 | | 0.119228 | | 0.0959427 |
(D4) | 0.0897462 | | 0.0487368 | | 0.0897462 | 1 | | 0 | | 0.121626 | | 0.0890871 | | 0 | | 0.0481125 | | 0.0445435 |
(D5) | 0 | | 0.104044 | | 0.0415444 | | 0 | | 1 | | 0.0341117 | | 0.0412393 | | 0.146385 | | 0.0890871 | | 0.0412393 |
(D6) | 0.119041 | | 0.122321 | | 0.130986 | | 0.121626 | | 0.0341117 | | 1 | | 0.118166 | | 0.0466055 | | 0.0850896 | | 0.118166 |
(D7) | 0.191885 | | 0.200233 | | 0.0479714 | | 0.0890871 | | 0.0412393 | | 0.118166 | | 1 | | 0.169031 | | 0.154303 | | 0.190476 |
(D8) | 0.0567605 | | 0.189535 | | 0.113521 | | 0 | | 0.146385 | | 0.0466055 | | 0.169031 | | 1 | | 0.121716 | | 0.338062 |
(D9) | 0.119228 | | 0.129766 | | 0.119228 | | 0.0481125 | | 0.0890871 | | 0.0850896 | | 0.154303 | | 0.121716 | | 1 | | 0.102869 |
(D10) | 0.0959427 | | 0.160186 | | 0.0959427 | | 0.0445435 | | 0.0412393 | | 0.118166 | | 0.190476 | | 0.338062 | | 0.102869 | 1 |

mrrp@portatil:~/Escritorio/MatrizContenido$ ./CalcSim documents-01.txt
```


Nota : La matriz es simétrica, esto quiere decir que los datos por encima de la diagonal son los mismos que los que están por debajo de la diagonal, sólo están en posiciones en los que los documentos que representan las filas representan los de las columnas y viceversa al buscar el mismo valor por encima o debajo de la diagonal. También fijarse que como en la diagonal se está haciendo la similitud con el mismo documento, su resultado es una similitud perfecta, un 1 (esto no nos servirá para nada en los análisis, por cierto). Los valores de la matriz son valores reales que irán desde el 0 (0% de similaridad de términos entre dos documentos) al 1 (100% de similaridad de términos entre dos documentos). Así un 1 entre dos documentos significa que ambos tienen los mismos términos y que estos aparecen las mismas veces en ellos, mientras que un 0 significa que no tienen ni un sólo término en común entre ellos.

- Apreciamos los datos que destacan, resaltados en colores :

- Los mismos colores pertenecen a la misma celda (además se resaltó con el mismo color a cuáles documentos en la fila y columna pertenece)

- Los valores en verde y en azul pertenecen a los valores máximos de similitud/similaridad posibles, que son entre el documento 8º y el 10º (ó entre el 10º y el 8º, como queramos verlo). El valor es **0,338062**, es decir, aproximadamente un **33,8062%** de similitud de términos entre ellos. Aproximadamente un tercio de los términos del documento 8 aparecen en el documento 10.

Documento 8 :	Documento 10 :
-----	-----
1 - { acidity, (1, 0.154902, 0.154902)}	1 - { acidity, (1, 0.154902, 0.154902)}
2 - { balanced, (1, 0.522879, 0.522879)}	2 - { apple, (1, 0.69897, 0.69897)}
3 - { dry, (1, 0.69897, 0.69897)}	3 - { balanced, (1, 0.522879, 0.522879)}
4 - { firm, (1, 0.69897, 0.69897)}	4 - { crisp, (1, 0.69897, 0.69897)}
5 - { food, (1, 1, 1)}	5 - { depth, (1, 1, 1)}
6 - { for, (1, 1, 1)}	6 - { drink, (1, 1, 1)}
7 - { it's, (1, 0.221849, 0.221849)}	7 - { dry, (1, 0.69897, 0.69897)}
8 - { much, (1, 0.69897, 0.69897)}	8 - { flavor, (1, 1, 1)}
9 - { offers, (1, 1, 1)}	9 - { fresh, (1, 0.39794, 0.39794)}
10 - { profusion, (1, 1, 1)}	10 - { fruits, (1, 0.69897, 0.69897)}
11 - { restrained, (1, 1, 1)}	11 - { great, (1, 1, 1)}
12 - { spice, (1, 0.69897, 0.69897)}	12 - { has, (1, 1, 1)}
13 - { texture, (1, 0.69897, 0.69897)}	13 - { it's, (1, 0.221849, 0.221849)}
14 - { very, (1, 1, 1)}	14 - { its, (1, 1, 1)}
15 - { wine, (1, 0.30103, 0.30103)}	15 - { now, (1, 1, 1)}
	16 - { off, (1, 0.69897, 0.69897)}
	17 - { pear, (1, 1, 1)}
	18 - { spice, (1, 0.69897, 0.69897)}
	19 - { texture, (1, 0.69897, 0.69897)}
	20 - { touch, (1, 1, 1)}
	21 - { while, (1, 0.522879, 0.522879)}

- Los valores en rojo pertenece al cálculo de la similaridad que da el mínimo valor posible de similitud : el **0** (un **0%**). Se eligió un valor al azar, que representa aquí a los documentos 5 y 4. Los términos de ambos documentos se muestra a continuación :

Documento 4 :	Documento 5 :
-----	-----
1 - { aromas, (1, 0.39794, 0.39794)}	1 - { 2012, (1, 1, 1)}
2 - { astringent, (1, 1, 1)}	2 - { across, (1, 1, 1)}
3 - { bit, (1, 1, 1)}	3 - { bottling, (1, 1, 1)}
4 - { blossom, (1, 1, 1)}	4 - { characteristics, (1, 1, 1)}
5 - { finish, (1, 0.69897, 0.69897)}	5 - { comes, (1, 1, 1)}
6 - { giving, (1, 1, 1)}	6 - { companion, (1, 1, 1)}
7 - { guava, (1, 1, 1)}	7 - { country, (1, 1, 1)}
8 - { honey-drizzled, (1, 1, 1)}	8 - { earthy, (1, 1, 1)}
9 - { is, (1, 0.522879, 0.522879)}	9 - { good, (1, 1, 1)}
10 - { lemon, (1, 1, 1)}	10 - { hearty, (1, 1, 1)}
11 - { mango, (1, 1, 1)}	11 - { herbal, (1, 0.69897, 0.69897)}
12 - { more, (1, 1, 1)}	12 - { it, (1, 0.69897, 0.69897)}
13 - { notes, (1, 0.69897, 0.69897)}	13 - { it's, (1, 0.221849, 0.221849)}
14 - { off, (1, 0.69897, 0.69897)}	14 - { like, (1, 1, 1)}
15 - { opulent, (1, 1, 1)}	15 - { much, (1, 0.69897, 0.69897)}
16 - { orange, (1, 1, 1)}	16 - { nonetheless, (1, 1, 1)}
17 - { palate, (1, 0.522879, 0.522879)}	17 - { pleasantly, (1, 1, 1)}
18 - { pineapple, (1, 0.69897, 0.69897)}	18 - { rather, (1, 1, 1)}
19 - { pith, (1, 1, 1)}	19 - { regular, (1, 1, 1)}
20 - { rind, (1, 0.69897, 0.69897)}	20 - { rough, (1, 1, 1)}
21 - { semidry, (1, 1, 1)}	21 - { rustic, (1, 1, 1)}
22 - { slightly, (1, 1, 1)}	22 - { stew, (1, 1, 1)}
23 - { start, (1, 1, 1)}	23 - { tannic, (1, 1, 1)}
24 - { way, (1, 1, 1)}	24 - { think, (1, 1, 1)}
	25 - { unfussy, (1, 1, 1)}
	26 - { wine, (1, 0.30103, 0.30103)}
	27 - { winter, (1, 1, 1)}
	28 - { you, (1, 1, 1)}

Si comparamos todos los términos, no existe ni uno que aparezca en ambos documentos y por ello tenemos una similaridad del **0%**.

▪ Análisis de d1.txt

- En este primer ejemplo de juguete los resultados son :


```

Aromas include tropic fruit, broom, brimstone and dried herb.
Aromas include tropical fruit, broom, brimstone and dried herb.

Documento 1 :
-----

|1| - { aromas, (1, 0, 0)}
|2| - { brimstone, (1, 0, 0)}
|3| - { broom, (1, 0, 0)}
|4| - { dried, (1, 0, 0)}
|5| - { fruit, (1, 0, 0)}
|6| - { herb, (1, 0, 0)}
|7| - { include, (1, 0, 0)}
|8| - { tropic, (1, 0.30103, 0.30103)}

Documento 2 :
-----

|1| - { aromas, (1, 0, 0)}
|2| - { brimstone, (1, 0, 0)}
|3| - { broom, (1, 0, 0)}
|4| - { dried, (1, 0, 0)}
|5| - { fruit, (1, 0, 0)}
|6| - { herb, (1, 0, 0)}
|7| - { include, (1, 0, 0)}
|8| - { tropical, (1, 0.30103, 0.30103)}

- Datos de similitudes coseno -
-----

      (D1) (D2)
(D1)  | 1 || 0.875 |
(D2)  | 0.875 || 1 |

mrrp@portatil:~/Escritorio/MatrizContenido$ ./CalcSim d1.txt

```

- Se aprecia que como cada documento tienen los mismos términos salvo uno en cada uno de ellos (términos **tropic** y **tropical**), se muestra que ambos documentos son iguales salvo en ($\frac{1}{8} = 0,125 \rightarrow$ Un 12,5%). Es decir, que son similares entre ellos en un 87,5% (un 0,875)

▪ Análisis de d2.txt

- En este segundo ejemplo de juguete los resultados son :

```

Aromas include fruit, broom, brimstone and dried herb.
Aromas include tropical fruit, broom, brimstone and dried herb.

Documento 1 :
-----

|1| - { aromas, (1, 0, 0)}
|2| - { brimstone, (1, 0, 0)}
|3| - { broom, (1, 0, 0)}
|4| - { dried, (1, 0, 0)}
|5| - { fruit, (1, 0, 0)}
|6| - { herb, (1, 0, 0)}
|7| - { include, (1, 0, 0)}

Documento 2 :
-----

|1| - { aromas, (1, 0, 0)}
|2| - { brimstone, (1, 0, 0)}
|3| - { broom, (1, 0, 0)}
|4| - { dried, (1, 0, 0)}
|5| - { fruit, (1, 0, 0)}
|6| - { herb, (1, 0, 0)}
|7| - { include, (1, 0, 0)}
|8| - { tropical, (1, 0.30103, 0.30103)}

- Datos de similitudes coseno -
-----

      (D1) (D2)
(D1)  | 1 || 0.935414 |
(D2)  | 0.935414 || 1 |

mrrp@portatil:~/Escritorio/MatrizContenido$ ./CalcSim d2.txt

```

- Vemos aquí que aumenta la similaridad respecto al ejemplo de *juguete* anterior porque hemos quitado el el documento 1 una palabra que no aparecía en el documento 2, esto ocasiona que ahora estemos en una situación en la sólo haya un término que esté en un sólo documento, en vez de dos términos como en **d1.txt**. Todo lo anterior se traduce en una mayor similaridad entre documentos, subiendo de un **87,5%** a un **93,5414%**.

- Conclusiones.

1º) Hemos visto que con esta herramienta/aplicación podremos analizar y calcular cuán similares pares de documentos son entre sí, a nivel de términos y veces que aparecen en cada uno de ellos.

2º) Cuantos más términos/palabras en común tengan dos documentos más alto será su similaridad. Y al revés, cuantos menos tengan en común más bajo será esta similaridad.

3º) También hemos deducido que hay más similitud entre pares de documentos cuanto más términos coincidan entre ellos y menos términos diferentes entre ellos aparezcan.

4º) Las utilidades de este tipo de cálculos de similaridad de términos en contenidos es una herramienta útil para poder recomendar obras relativamente iguales a los usuarios que consuman contenidos en una plataforma determinada.