Universidade do Minho
Escola de Engenharia

Conceção e implementação de um sistema de recomendação.

Grupo 9

Sistemas Baseados em similaridades

Autores:

João Sá

Tiago Silva

Índice

1. Introdução

Este trabalho tem como finalidade a Conceção e implementação de um Sistema de Recomendação.

Foi escolhida uma base de dados constituída por:

- 139 Anedotas
- 40863 Utilizadores

Cada utilizador avaliou diferentes anedotas e o objetivo deste trabalho é criar um sistema de recomendação que consiga recomendar piadas para um novo utilizador. O sistema irá ser implementado usando técnicas de filtragem colaborativa e filtragem por conteúdo.

O sistema de recomendação irá estar integrado em uma plataforma web de modo a facilitar a interação do utilizador com o sistema tornando a experiência também mais apelativa.

2. Data treatment

2.1. Datasets Iniciais

O tratamento dos dados foi realizado na plataforma ”KNIME”. O dataset estava dividido em dois ficheiros com formato “CSV”, no primeiro “CSV” tínhamos o Id das piadas e o texto das piadas como mostra a figura 1.

joke_id	joke_text
1	Q. What's O. J. Simpson's web address? A. Slash, slash, backslash, slash, slash, escape.
2	How many feminists does it take to screw in a light bulb? That's not funny.
3	Q. Did you hear about the dyslexic devil worshiper? A. He sold his soul to Santa.
4	They asked the Japanese visitor if they have elections in his country. Every morning, h...
5	Q: What did the blind person say when given some matzah? A: Who the hell wrote this?
6	Q. What is orange and sounds like a parrot? A. A carrot.
7	How many men does it take to screw in a light bulb? One. Men will screw anything.
8	A dog walks into Western Union and asks the clerk to send a telegram. He fills out a for...
9	Q: If a person who speaks three languages is called trilingual, and a person who speak...
10	What's the difference between a Macintosh and an Etch-a-Sketch? You don't have to s...
11	What's the difference between a used tire and 365 used condoms? One's a Goodyear, ...
12	A duck walks into a pharmacy and asks for a condom. The pharmacist asks, Would you ...
13	Q: What is the Australian word for a boomerang that won't come back? A: A stick.
14	What do you get when you run over a parakeet with a lawnmower? Shredded tweet.
15	Two kindergarten girls were talking outside: one said, You won't believe what I saw on ...
16	A guy walks into a bar and sits down next to an extremely gorgeous woman. The first t...
17	Bill Clinton returns from a vacation in Arkansas and walks down the steps of Air Force ...

Fig1. Dataset com as piadas

No segundo “CSV” tínhamos o Identificador do Usuário, Identificador da piada e o Rating da piada desse Usuário como mostra a figura 2.

user_id	joke_id	Rating
18947	76	2.125
12558	120	4.062
37394	139	5.469
40107	2	-7.281
12402	85	-5.344
36740	3	1.594
6959	104	-0.219
39399	25	-0.5
36438	3	-3.094
5573	93	-4.688
20840	96	7.031
8791	116	6.375
7267	11	-0.625
20819	43	7.688
25239	113	-9.812
4199	101	8.438
24141	110	-0.438

Fig2. Dataset com o rating dos usuários

2.2. Criação de géneros nas piadas

Começamos por juntar ambos os “CSV”, para fazermos isso utilizamos o Node ”Joiner”. Após juntar ambos os “CSV” dividiu-se as piadas em três categorias:

- Grande
- Média
- Pequena

Para isso utilizamos o node “Java Snippet” onde determinamos o tamanho de cada uma das piadas como mostra a figura 3.

```
// Enter your code here:
out_ = c_joke_text.length();
```

Fig.3 java snippet para obter o número de caracteres da piada

Após termos o tamanho de cada piada decidimos que o Length abaixo de 200 as piadas eram curtas, Length entre 200 e 700 as piadas eram médias e acima de 700 as piadas eram grandes. Para isso utilizamos três Rule Engines, com isto ficamos com três colunas (Short, Medium, Big) para avaliar cada uma das piadas.

Short	Medium	Big
0	1	0
0	1	0
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0

Fig4. Binary Enconding no género das piadas

Aplicamos uma técnica denominada de “Binary Encoding” como mostra a fig.4. Depois de categorizarmos cada umas das piadas como pequena, média ou grande tínhamos outro objetivo para tratar.

2.3. Categorização da variável Rating

A variável Rating tinha valores entre -10 e 10, para simplificarmos o nosso dataset decidimos categorizar a nossa variável em três valores diferente (gostou, mais ou menos e não gostou) e finalmente determinarmos a opinião de cada utilizador para cada uma das piadas.

Para isso utilizamos diferentes nodes, começamos por determinar que todas as piadas com Rating acima de 3 o utilizador gostou, Rating [-3, 3] considerou razoável e Rating abaixo de -3 não gostou onde no final tínhamos o id de cada usuário e a opinião de cada piada.

Para dividirmos o dataset em três grupo utilizamos o node "Row Filter". Na figura 5 temos o "Row Filter" utilizado para determinar as piadas que cada utilizador gostou, o processo foi similar para todas as categorias.

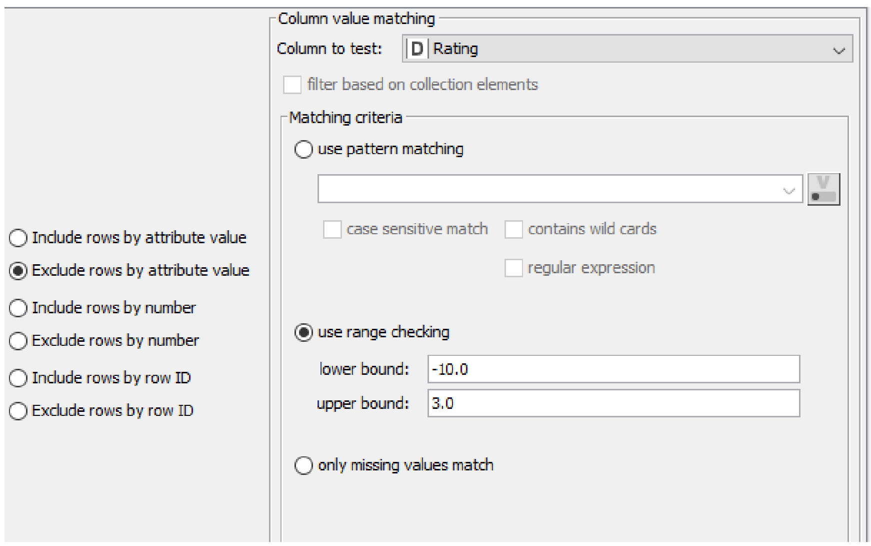


Fig.5 Obtenção de piadas gostadas por cada utilizador

Para cada um dos "CSV" deixamos como 0 todas as anedotas que não correspondiam, por exemplo se o utilizador não tinha gostada de uma piada o valor ficava a 0. Na figura abaixo conseguimos obter uma melhor perceção do resultado.

I user_id	S combined string
1	44,117,26,19,59,17,43,110,21,66,62,81,99,79,56,82,25,83,109,98,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
2	7,26,4,25,62,43,11,2,19,73,16,20,52,21,79,3,39,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
4	39,10,66,9,33,77,35,29,62,87,11,36,12,63,25,81,26,22,55,83,7,86,1,68,82,37,6,72,92,85,0,0,0,0,0,0,
5	36,25,4,9,21,43,26,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
6	59,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
7	83,88,46,28,59,12,99,36,78,32,82,1,81,43,42,66,40,15,11,98,77,73,58,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
8	1,104,5,3,8,2,7,103,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
9	26,106,9,53,98,108,11,72,52,35,55,80,81,56,97,96,43,40,22,79,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10	10,9,17,40,7,56,79,26,110,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
11	100,8,9,93,101,104,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Fig6. Dataset tratado com as piadas gostadas para cada utilizador

Tal como mostra na figura, temos o id de cada utilizador e o id de cada piada que o utilizador gostou, todas as outras piadas deixamos com o valor igual a zero. Repetimos o processo para as piadas que cada utilizador achou mais ou menos e para as piadas que cada utilizador não gostou.

Com isto já organizamos os nossos dados conforme a opinião de cada utilizador e categorizámos cada piada pelo tamanho. O atributo tamanho será utilizado no início do nosso sistema de recomendação quando não tivermos qualquer opinião do novo utilizador, vamos deixar o utilizador escolher se pretende piadas longas, medias, curtas ou diferentes combinações entre eles. Esta técnica permite evitar o chamado “Cold Start”.

2.4. Ordenação de piadas por género e rating

Para terminarmos o nosso trabalho no KNIME e o tratamento dos dados decidimos também obter a média de avaliações para cada uma das piadas e assim conseguimos descobrir quais as piadas curtas, médias ou grandes mais bem pontuadas.

O objetivo deste processo é na escolha das primeiras piadas o nosso sistema de recomendação escolher aleatoriamente entre as melhores piadas.

Para a realização deste processo utilizados o node” Group By” onde fizemos a média do Rating para cada piada como mostram as figuras abaixo.

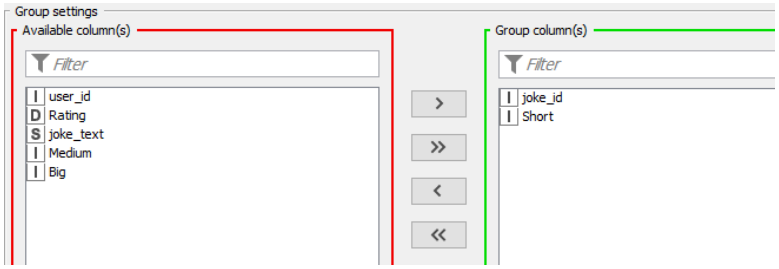


Fig.1. Calcular para cada Joke_Id

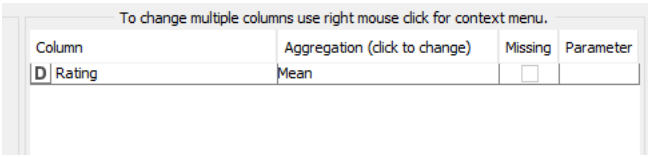


Fig.8. Média de Rating para cada Joke_Id

joke_id	Short	Mean(Rating)
1	1	-2.247
2	1	-1.933
3	1	-0.672
4	1	-0.605
5	1	-1.416
6	1	-1.613
7	1	0.707
8	0	-0.083
9	1	0.767
10	1	-0.447
11	1	3.024
12	1	1.231
13	1	0.986
14	1	-1.019
15	1	1.253
16	0	2.461
17	0	2.962

Fig.9 Resultado final

Para terminarmos utilizamos o node “Sorter” para ordenar a média de Rating de cada uma das piadas.

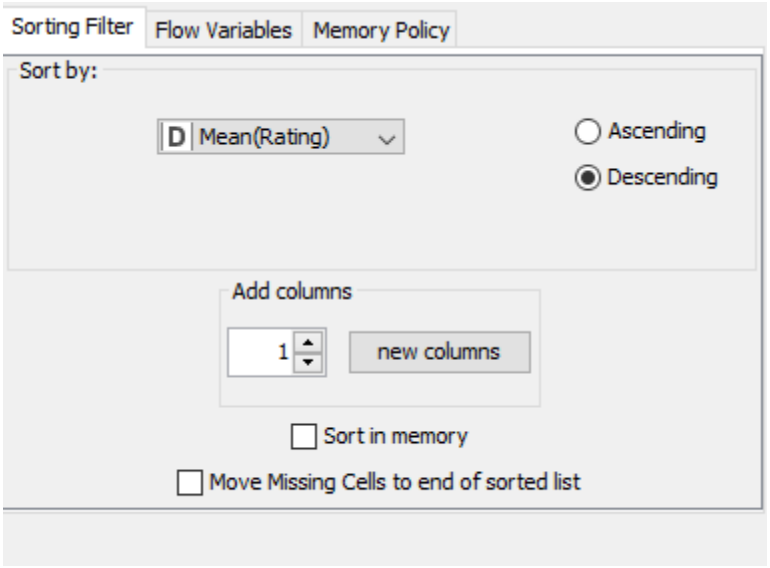


Fig.2 Ordenação da média de Rating

Realizamos este processo para os três tipos de piada (curta, média ou grande) e no final exportamos os três “CSV” para utilizarmos no python na realização do nosso sistema de recomendação.

2.5. Sumário

Recapitulando o tratamento de dados realizado no KNIME:

- Categorizar cada uma das piadas como pequena, média ou grande;
- Categorizar o Rating como gostou, mais ou menos, não gostou;
- Obtermos a média de Rating de cada uma das piadas e ordenarmos da melhor para a pior;

No final ficámos com 6 “CSV”:

- Piadas que cada utilizador gostou;
- Piadas que cada utilizador achou mais ou menos;
- Piadas que cada utilizador não gostou;
- Top piadas Pequenas;
- Top piadas Médias;
- Top piadas Grandes;

3. Implementação do sistema de recomendação

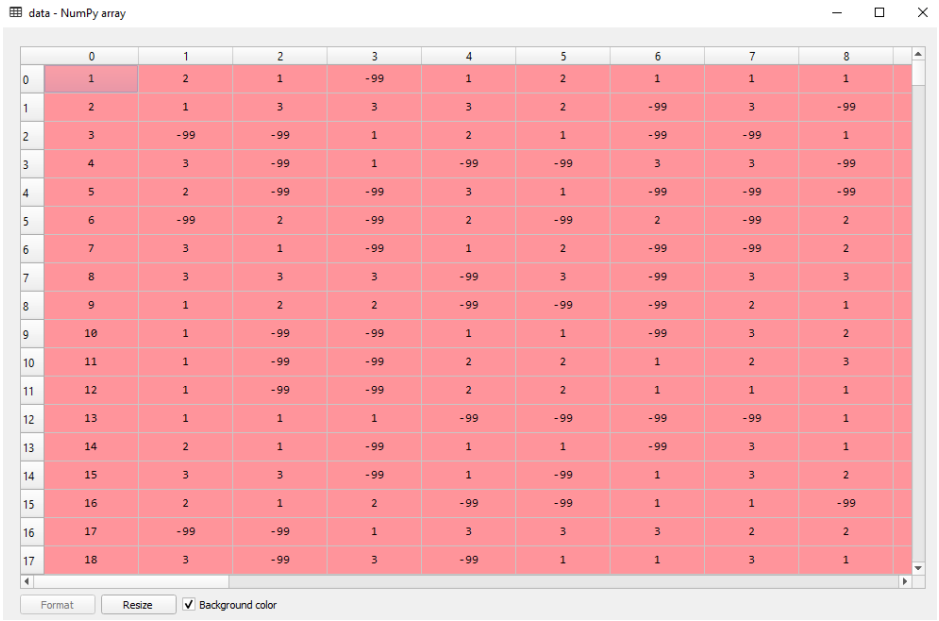
3.1. Tecnologias utilizadas

Para implementar o sistema de recomendação utilizou-se a linguagem de programação Python e o IDE “Spyder”. Foram usadas as librarias comuns a projetos desta categoria como numpy, pandas, scipy e matplotlib.

Para implementação da plataforma web utilizou-se a framework “FLASK” devido a sua simplicidade e facilidade de uso que no entanto permite atingir os nossos objetivos.

3.2. Dataset final

Usando os datasets preparados do “Knime” que continham as piadas gostadas por cada utilizador criou-se uma matriz [40863, 140] sendo que cada linha representa um utilizador e cada coluna representa o rating desse utilizador para cada uma das 139 piadas, com exceção da primeira coluna que contem o id de cada utilizador.



	0	1	2	3	4	5	6	7	8
0	1	2	1	-99	1	2	1	1	1
1	2	1	3	3	3	2	-99	3	-99
2	3	-99	-99	1	2	1	-99	-99	1
3	4	3	-99	1	-99	-99	3	3	-99
4	5	2	-99	-99	3	1	-99	-99	-99
5	6	-99	2	-99	2	-99	2	-99	2
6	7	3	1	-99	1	2	-99	-99	2
7	8	3	3	3	-99	3	-99	3	3
8	9	1	2	2	-99	-99	-99	2	1
9	10	1	-99	-99	1	1	-99	3	2
10	11	1	-99	-99	2	2	1	2	3
11	12	1	-99	-99	2	2	1	1	1
12	13	1	1	1	-99	-99	-99	-99	1
13	14	2	1	-99	1	1	-99	3	1
14	15	3	3	-99	1	-99	1	3	2
15	16	2	1	2	-99	-99	1	1	-99
16	17	-99	-99	1	3	3	3	2	2
17	18	3	-99	3	-99	1	1	3	1

Fig.11 Dataset Final

3.3. Filtragem por conteúdo

Filtragem por conteúdo baseia-se na premissa de utilizar os meta dados do conteúdo de modo a fazer recomendações ao utilizador. No nosso sistema esta técnica é utilizada no nosso sistema de modo a evitar um "cold start". Quando o utilizador inicialmente abre o nosso website são feitas três questões de modo a perceber de que tipo de piadas o utilizador gosta, Curtas, médias ou longas, fig12. As opções escolhidas serão o tipo de piadas inicialmente apresentadas ao utilizador.

As piadas fornecidas são escolhidas de entre as mais bem cotadas pelos utilizadores em média de modo a tentar fornecer as piadas que mais agradam no geral, fig 13.

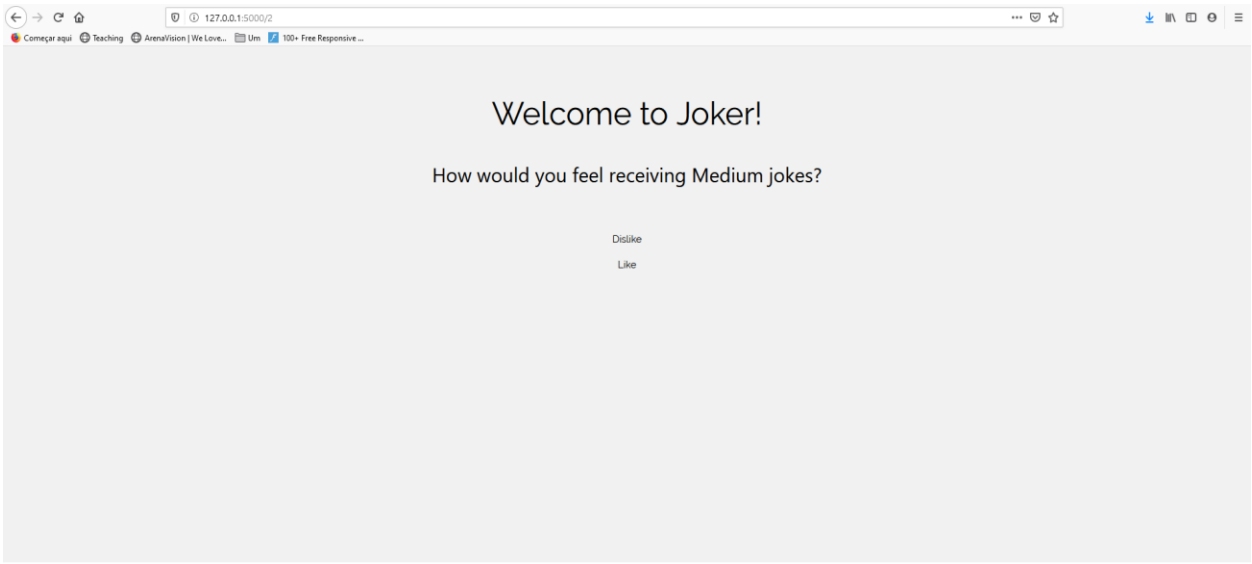


Fig12. Exemplo de questão ao utilizador

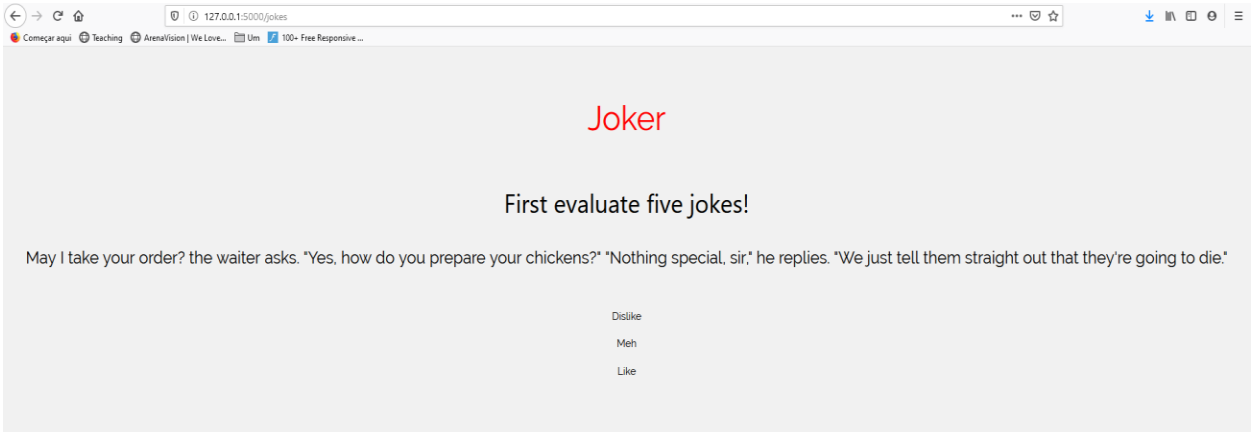


Fig13.Primeiras cinco piadas utilizando filtragem por conteúdo

3.4. Filtragem colaborativa

Filtragem colaborativa têm com premissa que utilizadores semelhantes tem gostos semelhantes e, portanto, é encontrado um ou vários utilizadores semelhantes ao utilizador e este recebe as recomendações baseadas nos gostos dos utilizadores semelhantes

No nosso sistema esta técnica começa a ser implementada após o utilizador avaliar as primeiras cinco piadas pois consideramos que com esta informação já se consegue encontrar utilizadores relativamente semelhantes

3.4.1. Métricas utilizadas

Para medir a distância entre dois utilizadores foi utilizada a distância euclidiana, apenas piadas que foram avaliadas pelos dois utilizadores contam para a distância pois só deste modo podemos avaliar a diferença de gostos.

Como o nosso dataset de utilizadores tem cerca de 40000 utilizadores diferentes consegue-se medir a distância do utilizador atual e cada um dos utilizadores do datasets com um desempenho bastante satisfatório, daqui retorna uma lista ordenada contendo todas as distâncias, figura 14.

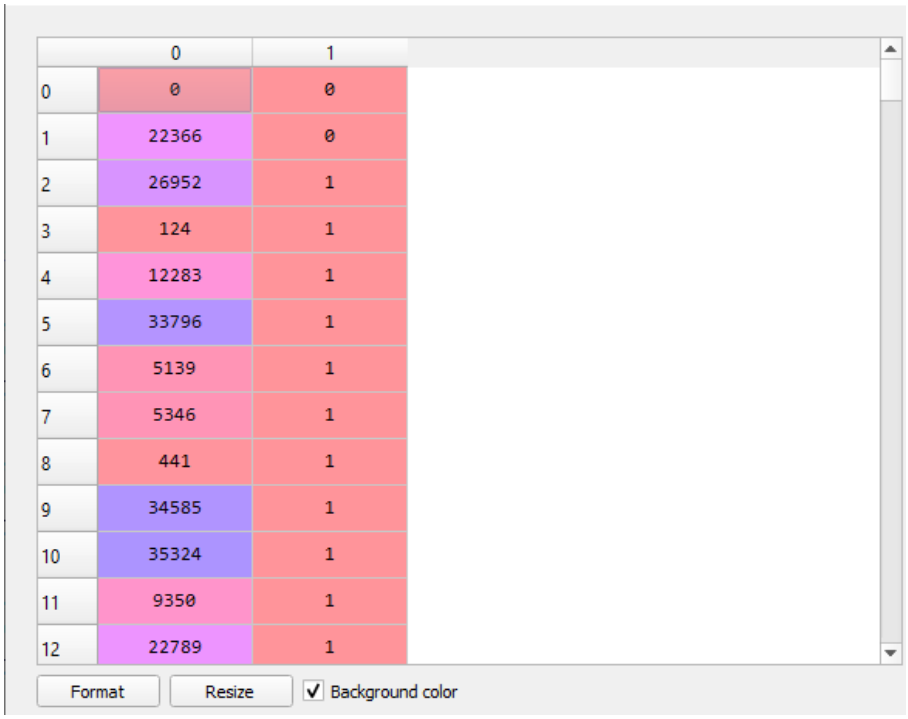


Fig.14 Lista de distâncias

Neste caso em específico pode-se observar que o utilizador da base de dados mais próximo do utilizador atual é o 22366 que tiveram exatamente os mesmos gostos. Este processo é repetido para cada nova piada que o utilizador avalie e a distância entre dois utilizadores

apenas é validade se tiverem pelo menos cinco piadas em comum para evitar que utilizadores que só tem uma piada em comum tenham uma distância de zero.

3.4.2. Seleção de piada

Tendo encontrado o utilizador mais próximo o sistema escolhe aleatoriamente uma das piadas com rating positivo avaliadas pelo utilizador encontrado e recomenda-a. Esta é avaliada e o processo repete-se

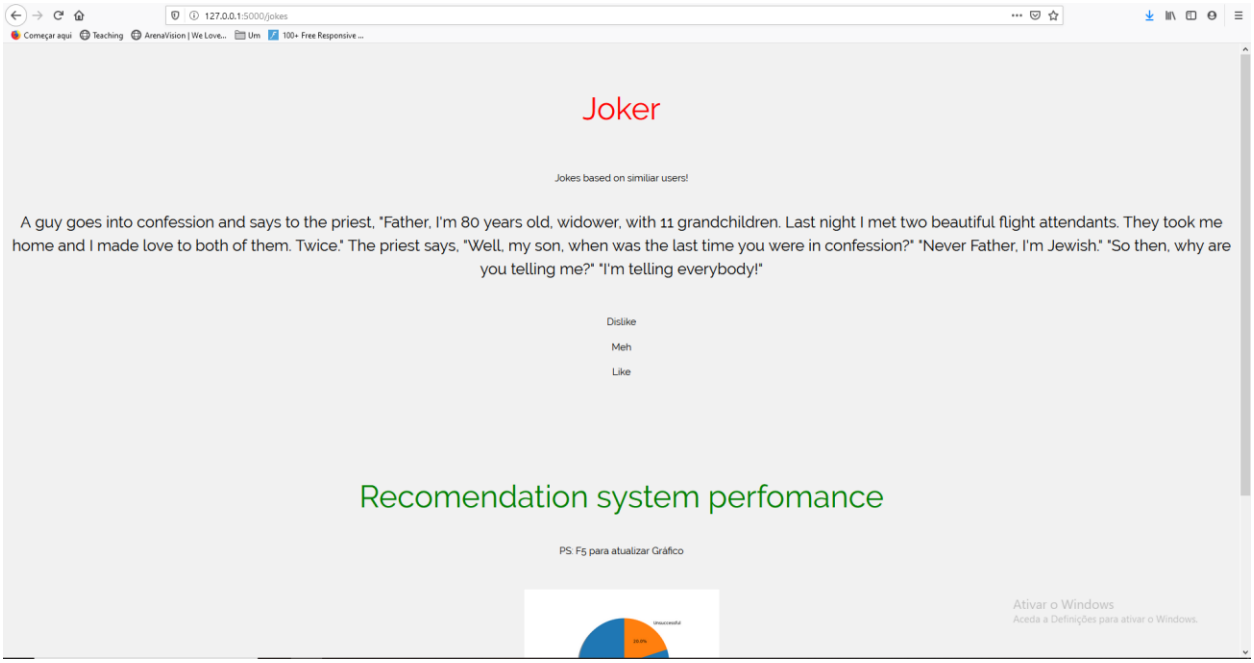


Fig.15 Filtragem Colaborativa

3.4.3 Feedback Sistema

De modo a dar feedback das recomendações dadas ao utilizador está incluído no website uma secção que mostra um gráfico com a taxa de sucesso de recomendações. Deste modo tornamos a experiência mais dinâmica e interessante para o utilizador e estes têm noção da qualidade das recomendações que recebem, figura 16.

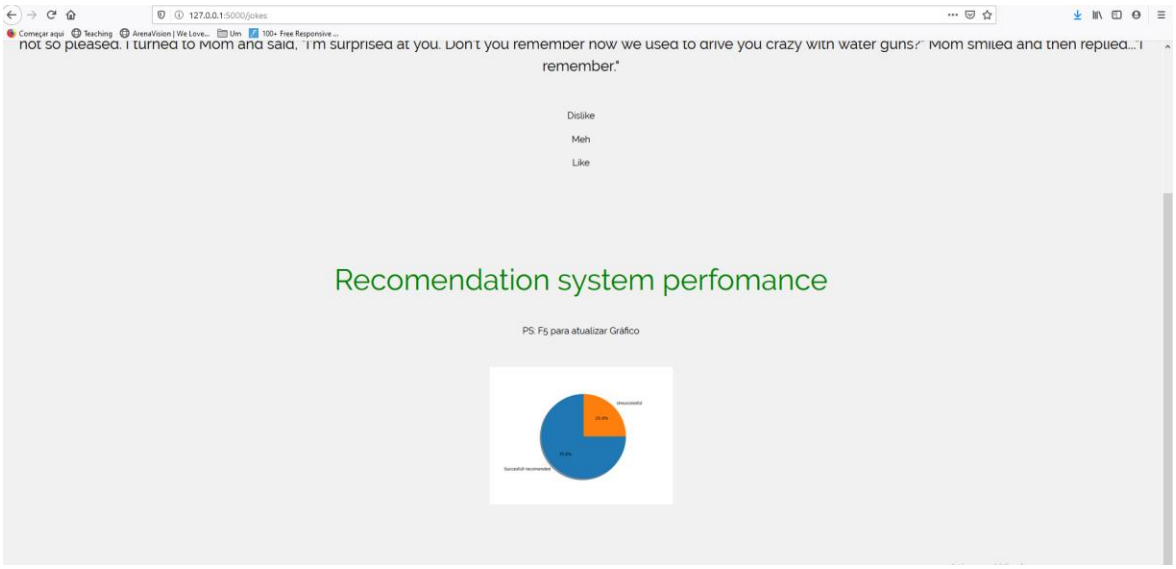


Fig16. System Perfomance

3.5 Futuro do sistema

A cada utilização por um novo utilizador, este é acrescentado á base de dados. Isto permite que o sistema se torne continuamente mais robusto e adaptável a um maior número de utilizadores, no entanto, no futuro seria necessário otimizar a filtragem colaborativa pois torna-se inviável comparar o utilizador atual com o crescente número de registos na base de dados.

4. Tutorial de utilização

4.1. Linha de Comandos

1º. Instalar todas as dependências

- Pip install numpy
- Pip install matplotlib
- Pip install os
- Pip install scipy
- Pip install pandas
- Pip install random
- pip install Flask

2º. Adicionar o projeto ao Flask

```
C:\path\to\app>set FLASK_APP=KNNProjectV-1.py
```

3º. Correr a aplicação

```
C:\path\to\app>python -m flask run
```

4º-Abrir a aplicação

Depois de correr a aplicação temos de abrir o browser e colocar o seguinte endereço:

- http://127.0.0.1:5000/

5. Conclusão

Em suma, neste trabalho o grupo percebeu os fundamentos para a criação de um sistema de recomendação e as dificuldades inerentes para a concretização do mesmo. Assim como perceber a importância de contornar o “cold start”.

Descobrir e moldar um dataset existente de modo a permitir aplicar um sistema de recomendação no mesmo provou-se um dos maiores desafios para a concretização deste trabalho. Aplicar várias técnicas de recomendação em conjunto foi também um desafio.

No futuro seria importante testar outros algoritmos para a filtragem colaborativa e aumentar a performance dos mesmos.