Amanda G Dias
10275312

Visualização Computacional – Projeto Prático

Problema abordado

Como parte do meu trabalho de mestrado, é necessária a visualização de documentos semelhantes a um documento exemplo. A coleção de documentos em questão é de uma base de dados que contém 21578 documentos e um vocabulário de 42734 palavras, que foi disponibilizada pelos parceiros do projeto, juntamente com o algoritmo que calcula a frequência de cada termo dentro de cada documento.

Para calcular a similaridade entre documentos, foi utilizado a frequência do termo inverso de cada documento (Tf-idf) e os resultados desse cálculo foram salvos no arquivo svm.fil. Este arquivo será utilizado pelo código Exemplo.R desenvolvido para este trabalho para mapear os documentos na visualização.

O objetivo deste trabalho é verificar se existe a formação de grupos de documentos semelhantes de forma a garantir que os documentos próximos são realmente semelhantes e entender o motivo pelo qual foram colocados próximos um do outro.

Pré-processamento dos dados

O arquivo svm.fil apresenta em cada linha o seguinte formato:

Nome do documento termoID1:frequência termoID2:frequência termoID3:frequências....

Desta forma, foi necessário organizar estes dados de forma a gerar uma matriz nxm onde n representa o número de termos e m o número de documentos, para que a matriz fosse aceita na função de distância.

Como mapear todos os 21578 documentos é custoso e como o usuário deseja encontrar apenas documentos semelhantes ao documento exemplo, foi necessária uma filtragem inicial dos documentos. Para isso, foi feito um ranking dos k (neste caso, utilizei k=500) documentos mais semelhantes ao documento exemplo (para este trabalho, assumiu-se que o documento 00001 era o documento exemplo) a partir do cálculo da dissimilaridade do cosseno do documento exemplo para todos os outros. Esta medida de distância foi utilizada, pois é conhecida por ser adequada para dados com muitas dimensões, como é o caso de documentos (onde cada termo representa uma dimensão).

A seguir, foi calculado a dissimilaridade do cosseno entre os documentos do ranking, resultando em uma matriz k x k que será utilizada para calcular as coordenadas dos documentos no plano 2D a partir da técnica de redução de dimensionalidade tSNE. Utilizei esta técnica, pois sua implementação já está pronta no R e porque é uma das técnicas mais clássicas de redução de dimensionalidade, sendo interessante para uma análise inicial dos dados.

Mapeamentos visuais

Amanda G Dias
10275312

Para cumprir os objetivos deste trabalho, escolhi fazer duas visualizações em que o usuário pode interagir para explorar os dados. A primeira visualização é um *scatterplot* (Figura 1) onde cada círculo representa um documento e estes estão posicionados de acordo com a similaridade entre eles. Os círculos foram mapeados por cores, onde verde representa os 10 primeiros documentos no *ranking* dos mais similares, laranja os documentos entre as posições 11 e 299 no *ranking* e azul os acima de 500.

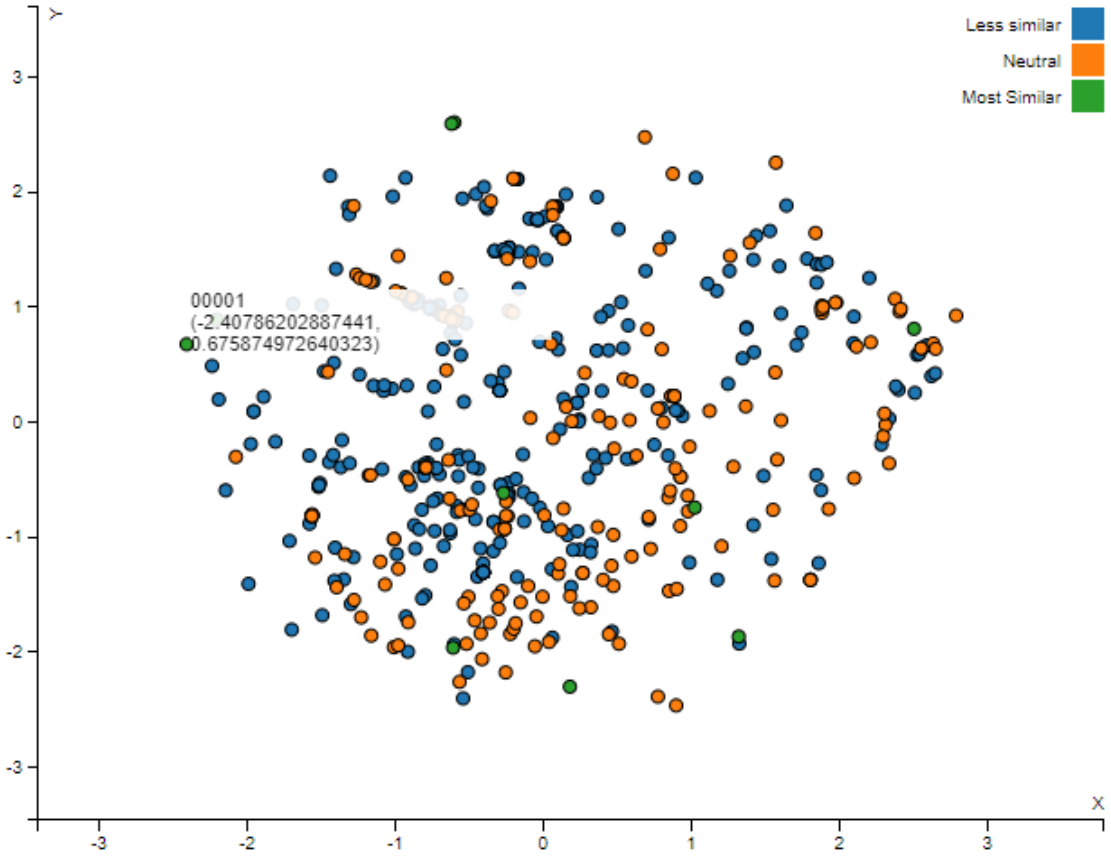

Figura 1. Os documentos são representados por círculos e suas cores o nível de similaridade com o documento exemplo. É possível passar o mouse por cima dos círculos para obter a informação do nome do documento e suas coordenadas. O documento em evidência é o documento exemplo 00001.

A partir desta visualização, é possível explorar esta coleção para verificar se os documentos próximos são realmente similares. Para isto, escolhi um grupo de documentos, representados na Figura 2 e identifiquei três documentos denominados 5027, 2082 e 5493. Após ler os documentos em questão pude observar que todos dividem o assunto de financias no âmbito empresarial. Outros documentos próximos também são parecidos no sentido de que apresentam tabelas com algumas informações financeiras de empresas.

Amanda G Dias
10275312

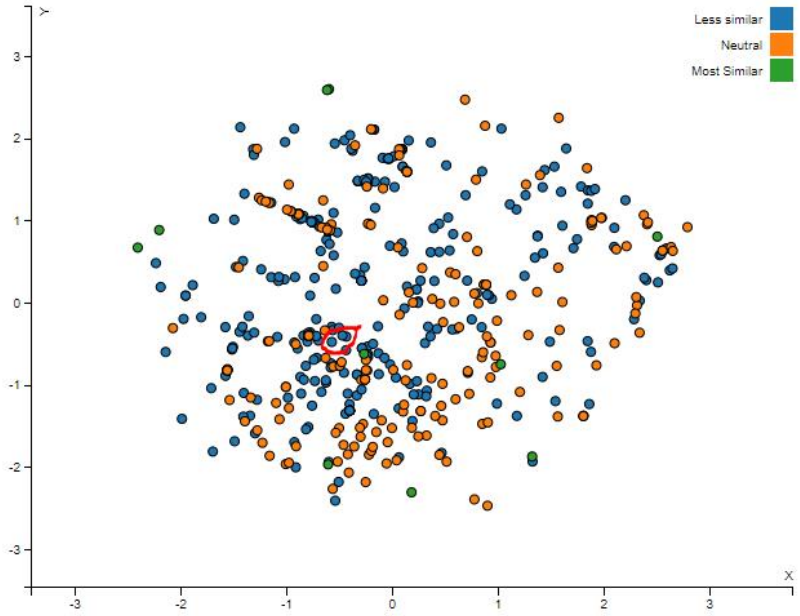

Figura 2. Os três documentos dentro da marcação em vermelho foram lidos e constatou-se que eles de fato dividem certa semelhança.

Além da exploração dos documentos, deseja-se saber o motivo pelo qual os documentos foram posicionados próximos. Para isso, criei uma outra visualização utilizando *barcharts*. Escolhi esta visualização, pois ela é simples de interpretar e codifica bem categorias e frequências. Para esta visualização, foi necessário criar uma matriz de termos e frequências para cada documento e, ao clicar em um documento na primeira visualização (*scatterplot*), um *barchart* correspondente àquele documento é inserido. Assim, o usuário pode visualizar vários *barcharts* referentes a diferentes documentos e explorar a frequência de termos de cada um. Esta interação pode ser visualizada na Figura 3.

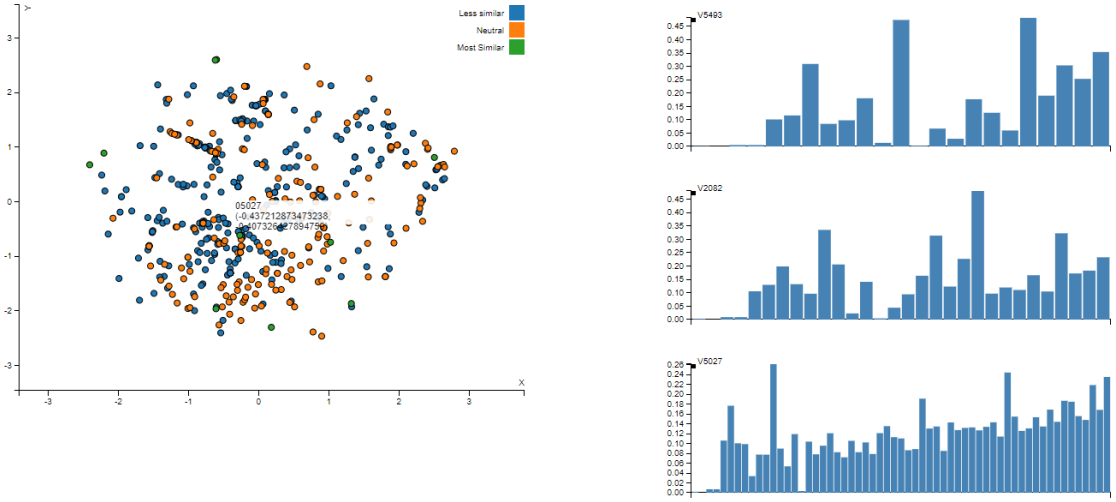

Figura 3. Clicando nos documentos da visualização a esquerda, faz surgir o gráfico de barras correspondente à direita onde cada barra corresponde aos termos daquele documento e a altura das barras a sua frequência. Neste caso, o usuário clicou nos círculos referentes aos documentos 5493, 2082 e 5027.

Ao passar o mouse por cima das barras é possível visualizar o Id do termo em questão e a frequência exata daquele termo. Este id do termo pode ser consultado no arquivo Terms onde

o Id é ligado com a linha do documento. No caso das figuras abaixo, após breve exploração, pude ver que o termo 33 está presente em todos os documentos selecionados e que este representa a palavra “year”.

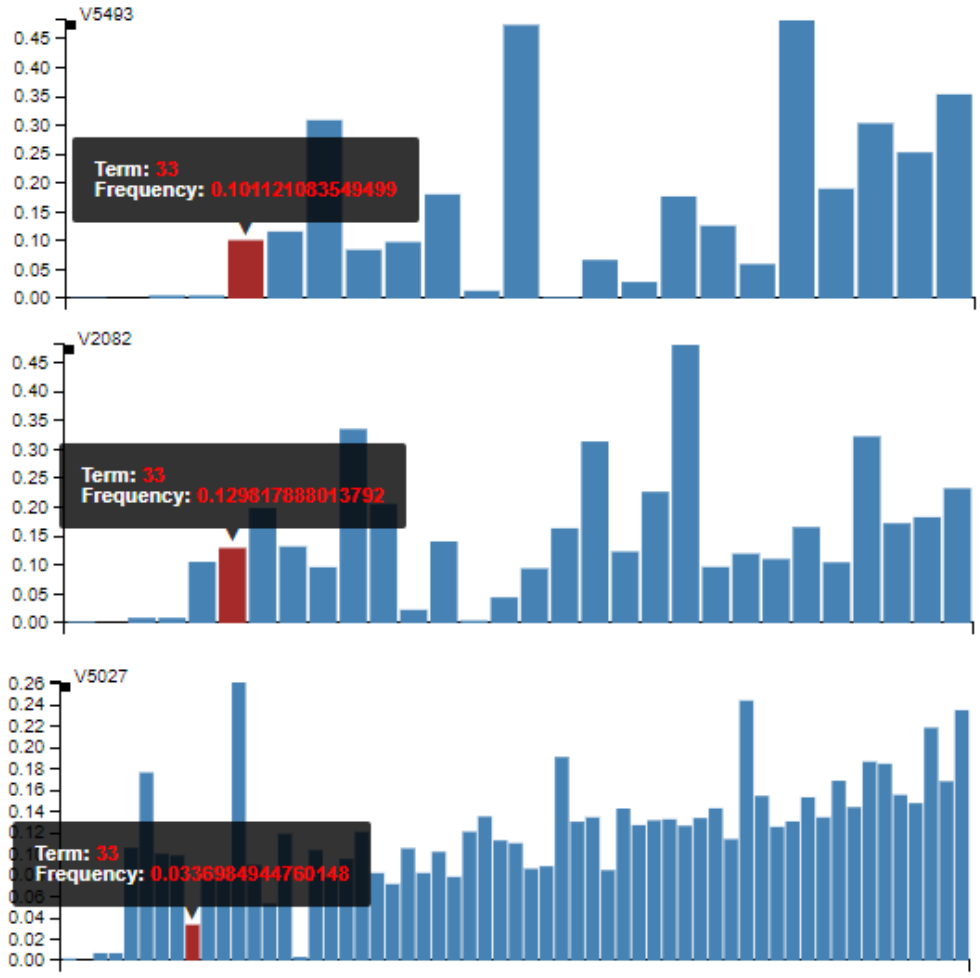

Figura 4. Ao passar o mouse por cima das barras, é possível ter a informação do Id do termo e a frequência exata. Os gráficos não são interligados, mas foi possível encontrar um termo em comum (Term: 33) de forma rápida explorando um documento por vez.

Termos que apresentam alta importância entre estes mesmo documentos são os termos “revenues” no primeiro, “best” no segundo e “sales” no terceiro. O que faz sentido, pois o primeiro documento fala sobre as receitas da Media General Inc, o segundo fala sobre as vendas da empresa Best Buy e o terceiro fala sobre as vendas de carros por empresas dos Estados Unidos.

Uma ferramenta interessante nesta visualização, seria ao selecionar um termo de um dos gráfico de barras, os outros gráficos também selecionassem o termo caso exista no documento. Desta forma, a exploração por termos em comum entre os documentos seria mais rápida.

Resultados e Conclusões

Amanda G Dias
10275312

O uso da visualização me possibilitou verificar se o algoritmo que calcula a similaridade entre os documentos está correto, através da análise dos documentos posicionados próximos um dos outros. Também foi possível verificar o motivo pelo qual os documentos foram posicionados juntos através da visualização da frequência dos termos no documento.

Algumas melhorias podem ser apontadas, como uma melhor escolha de cores para mapear elementos mais similares e menos similares, além de novas interações com o gráfico de barras para uma comparação mais rápida entre os documentos. Também seria interessante incluir a palavra em si ao invés do Id do termo.