Questão 1

Wellington Charles Lacerda Nobrega

18/04/2021

1. Estruturar uma fórmula para atribuir uma reputação a um estabelecimento j do setor k em um dia arbitrário t a partir de todas as avaliações individuais:

Primeiro, vamos definir $s_{t,j,k}$ como sendo o score médio da j-ésima firma individual do k-ésimo setor no dia t, assim, podemos calcular o score médio como:

$$s_{t,j,k} = \frac{\sum_{i=1}^{n_{t,j,k}} S_{t,j,k,i}}{n_{t,j,k}} \tag{1}$$

em que $n_{t,j,k}$ é o número de vendas da empresa j do setor k no dia t e $S_{t,j,k,i}$ denota um score individual recebido pela j-ésima firma do k-ésimo setor no dia t pela i-ésima venda. Assim, o score médio ($s_{t,j,k}$) retorna o valor médio das avaliações no dia da empreja j, do setor k no dia t.

2. Estruturar uma agregação para um determinado setor k em um dia qualquer t.

Vamos definir $K_{t,k}$ como sendo o indicador agregado para o setor k no dia t. Para a agregação, será realizado uma ponderação de acordo com a importância relativa de cada j-ésima firma sobre o total de vendas do setor k. Para isso, aplicaremos uma ponderação Ω_j para captar a importância relativa, onde, quanto maior for a participação da j-ésima firma sobre o total de vendas do k-ésimo setor, maior será o valor de Ω para essa firma.

$$K_{t,k} = \frac{\sum_{j=1}^{m_{t,k}} s_{t,j,k}.\Omega_j}{m_{t,k}} \tag{2}$$

onde m é a quantidade de empresas do setor k que foram avaliadas no dia t.

A ponderação Ω_j é dada por:

$$\Omega_j = \frac{n_{t,j,k}}{\sum_{j}^{m_{t,k}} n_{t,j,k}} \tag{3}$$

onde $n_{t,j,k}$ (já foi definida anteriormente) é o número de vendas da empresa j do setor k no dia t e $\sum_{j}^{m_{t,k}} n_{t,j,k}$ é o total de vendas do setor k. Destaca-se que $\sum \Omega_j = 1$.

3. Construir um método para um indicador agregado de reputação em um dia qualquer t.

Defina T_t como sendo o indicador que agrega o score setorial. Novamente, realizaremos uma ponderação de acordo com a importância relativa do k-ésimo setor sobre as vendas totais em um dia t, de acordo com:

$$T_t = \frac{\sum_{k=1}^{q_t} K_{t,k}.Z_k}{q_t} \tag{4}$$

em que q_t é a quantidade total de setores avaliados no dia t.

A ponderação Z_k é dada por:

$$Z_k = \frac{n_{t,k}}{n_t} \tag{5}$$

onde n_t é o total de vendas e $n_{t,k}$ denota o total de vendas do setor k no período t.

4. Como incorporar informações qualitativas dos *reviews*?

Nesta etapa do indicador é possível aplicar ténicas de Análise de Sentimentos para que seja possível parametrizar as informações qualitativas em informações quantitativas e, então, agregá-las ao índice. A Linguagem Natural de Processamento (NLP) é um ramo da linguística e computação que investiga os problemas de compreensão das línguas naturais humanas a partir do aprendizado de máquina. Essa etapa não é trivial, mas é extremamente fascinante. De uma forma estruturada:

Passo 1: O primeiro passo na incorporção das informações qualitativas no indicador é a limpeza das informações contidas nos *reviews*. É necessário que seja aplicado ténicas de *text mining* para a remoção de espaços duplos, *stop words*, pontuação, números, quebra de linhas e padronização de todos os caracteres para minúsculo. O objetivo final é deixar apenas as informações que de fato sejam relevantes e possam ser traduzidas em informações.

Passo 2: O segundo passo nesse processo é a escolha de um dicionário de palavras que seja capaz de capturar as informações relevantes no nosso *corpus* de *reviews*. Existem diversos dicionários de palavras pré-classificadas como sendo capazes de expressar o sentimento "negativo" ou "positivo". Por exemplo: *reviews* que contenham as palavras "ruim", "mal" ou "péssimo" podem estar associados a clientes insatisfeitos, gerando um sentimento negativo. Enquanto *reviews* que contenham palavras como "excelente", "ótimo" ou "bom" podem estar associados a clientes satisfeitos, gerando um sentimento positivo. Ademais, é necessário observar cuidadosamente palavras que possam efetivamente representar sentimento ambíguo quando avaliada em setores diferentes.

Passo 3: O terceiro passo seria a parametrização das informações qualitativas e posterior inclusão das informações no indicador de reputação. Para isso, é importante que a alternativa de parametrização seja independente de qualquer tipo de escala subjetiva. Ou seja, o algoritmo que transformará as informações qualitativas dos *reviews* em informações quantitativas deverá de alguma forma ser baseado na frequência ponderada com a qual as palavras vão aparecendo nos *reviews*. Uma alternativa, é a seguinte:

$$\text{sentimento}_t = \frac{\sum \text{Palavras Positivas} - \sum \text{Palavras Negativas}}{\sum \text{Palavras Positivas} + \sum \text{Palavras Neutras} + \sum \text{Palavras Negativas}} \tag{6}$$

A alternativa acima pertence ao intervalo -1 e 1, ou seja, $\text{sentimento}_t \in [-1, 1]$. Quando $0 \geq \text{sentimento}_t \leq 1$ tem-se um sentimento positivo; neutro quando $\text{sentimento}_t = 0$ e negativo quando $\text{sentimento}_t < 0$. Ao se ponderar o indicador com a abordagem de sentimentos, em tese, seríamos capazes de levar as informações qualitativas contidas nos *reviews* em consideração na construção do indicador de reputação.