Uma análise da similaridade de cossenos: comparação de ações de empresas listadas na bolsa de valores do Brasil (B3)

Flavia do Valle
Yasmine Moura

Abstract

A análise minuciosa de setores e empresas é crucial para a tomada de decisões de investimento no mercado de ações. Com a evolução da inteligência artificial e ferramentas de análise de dados, o uso de tecnologia avançada na análise de ações correlacionadas tornou-se ainda mais relevante no setor financeiro. A metodologia adotada neste estudo é baseada na medida de similaridade por cosseno para identificar papéis com semelhanças no mercado de ações brasileiro. Com esse objetivo, o algoritmo de similaridade aplicada em Python foi empregado para transformar os dados das ações em vetores e calcular os ângulos de cosseno entre os pares de vetores. Os resultados apontaram os 10 menores ângulos de cosseno como indicadores das ações mais semelhantes, destacando o potencial da metodologia proposta para determinar as semelhanças em pesquisas financeiras e análises de investimentos, bem como para a seleção de ativos diversificados.
Palavras-chave: ações; similaridade; cosseno; ângulo; vetor.

1 Introdução

O mercado de ações representa uma das formas mais relevantes de investimento adotadas em escala global. Entretanto sua análise é considerada um desafio constante no universo financeiro, demandando a cuidadosa consideração de diversos fatores, como histórico do setor e da empresa, margens líquidas e brutas, liquidez média diária das ações, entre outros importantes indicadores de potencial retorno do investimento.

Visando alcançar uma eficiente gestão de riscos, a criação de uma carteira de investimentos diversificada é uma estratégia amplamente utilizada e, para tanto, a análise de títulos com traços semelhantes tem se tornado uma prática cada vez mais frequente no processo decisório de investimentos. Conforme a tecnologia de análise de dados e inteligência artificial avança, a utilização de medidas de similaridade na realização de pesquisas de investimentos eficazes tem crescido significativamente.

Além de ser amplamente empregada em ciência de dados, em áreas como mineração de texto, recuperação de informação e classificação de documentos. a similaridade por cosseno é uma das ferramentas disponíveis que quantifica a semelhança entre dois vetores. A relevância da similaridade por cosseno para a análise de investimentos se deve ao fato de que essa medida é aplicável a espaços vetoriais de alta dimensionalidade e esparsos, como é comum em dados financeiros.

No mercado financeiro, aplicação desta metodologia permite identificar ações e mensurar sua similaridade com base em suas caracteristicas específicos. Por exemplo, é possivel realizar a comparação do desempenho de duas ações em relação a um período de tempo determinado ou a um conjunto de indicadores financeiros relevantes. Além disso, essa abordagem pode ser particularmente útil em mercados mais voláteis, permitindo que o investidor diversifique sua carteira de forma mais eficiente e reduza os riscos de perda.

Para realizar a análise de similaridade por cosseno, foi necessário construir um modelo vetorial que representasse adequadamente as características das ações em questão. Nesse sentido, optou-se por empregar a linguagem de programação Python, cujos recursos permitiram a conversão dos dados das ações em vetores, possibilitando o uso da função cosine similarity para o cálculo dos ângulos de cosseno entre os pares de vetores. Assim, pode-se empregar modelos vetoriais que permitiram a identificação de ações com semelhanças no mercado de ações brasileiro.

Inicialmente, serão fornecidos os conceitos teóricos do método baseado na medida de similaridade por cosseno. Em seguida, serão detalhados os procedimentos metodológicos aplicados, que envolveram o uso de modelos vetoriais e técnicas de mineração de dados para comparar as características das ações. Os resultados obtidos serão expostos, evidenciando os 10 menores ângulos de cosseno, ou seja, as ações mais semelhantes que, por sua vez, podem ser utilizadas em pesquisas financeiras para a identificação de oportunidades de investimento. .

2 Revisão da Literatura

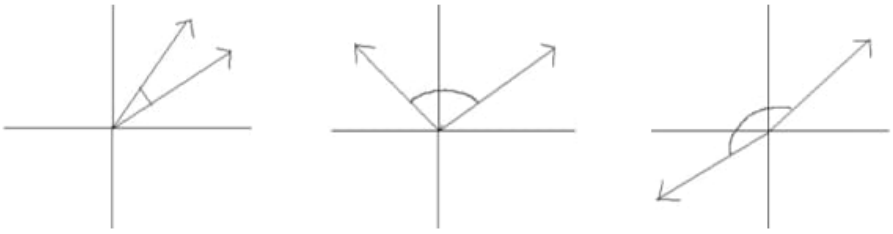
Utilizada para medir o grau de similaridade entre dois vetores em espaços de alta dimensionalidade e esparsos, a medida de similaridade do cosseno é uma métrica comumente empregada na comparação direta entre dois vetores. Ela pressupõem o cálculo do cosseno do ângulo formado entre eles como uma aproximação de sua similaridade. (SALAZAR, 2012)

Diferentemente de outras métricas de similaridade, o cosseno não leva em conta a magnitude dos vetores, mas apenas a orientação dos elementos em um espaço vetorial normalizado. Por isso, a similaridade entre dois vetores é determinada pelo ângulo formado entre eles, independentemente da frequência dos elementos nos vetores. (MAIA, 2017) A fórmula utilizada para calcular a similaridade do cosseno pode ser escrita como:

$$sim(S_i, S_j) = \frac{\sum_{l=1}^{t} w_{l,i} \times w_{l,j}}{\sqrt{\sum_{l=1}^{t} w_{l,i}^2} \times \sqrt{\sum_{l=1}^{t} w_{l,j}^2}}$$

2.1 Análise da Similaridade por Cosseno

Através da equação mencionada, a medida de similaridade do cosseno fornece valores na faixa de 0 a 1. Quando o ângulo entre os vetores diminui, o valor do cosseno tende a 1, indicando que a distância entre eles é menor. Abaixo são apresentadas as figuras que ilustram como a pontuação de similaridade do cosseno é expressa:



É possível observar que a medida de similaridade do cosseno retorna valores próximos a 1 quando dois vetores estão na mesma direção, com um ângulo próximo a 0 grau. Quando os vetores são quase ortogonais, com um ângulo próximo de 90 graus, a medida de similaridade do cosseno se aproxima de 0. Por fim, quando os vetores estão em posições opostas, com um ângulo próximo de 180 graus, a medida de similaridade do cosseno se aproxima de -1. (MAIA, 2017)

Essa interpretação dos valores de similaridade do cosseno é fundamental para compreender as comparações de similaridade realizadas em várias áreas, incluindo na análise e pesquisa financeiras.

2.2 Aplicações no Mercado Financeiro e de Capitais

A similaridade do cosseno é uma medida que pode ser aplicada em diversas áreas da Ciência e Análise de Dados, incluindo o mercado financeiro e de capitais, convertendo-se em uma ferramenta valiosa para:

- Análise de padrões de investimentos em ações e carteiras de investimento;
- Mensuração de desempenho de diferentes ativos financeiros;
- Análise de portfólio de fundos de investimento, ajudando a identificar oportunidades de diversificação e minimizar riscos em uma carteira de investimentos;
- Identificações de padrões em transações financeiras a fim de prevenir fraudes;
- Análise de risco de crédito: comparar a semelhança entre perfis de crédito de clientes, ajudando a avaliar a probabilidade de default e determinar limites de crédito.

3 Procedimentos Metodológicos

A metodologia utilizada neste estudo incluiu a coleta de dados por meio da plataforma de assessoria de investimentos “Eu Quero Investir” como fonte de dados.No que tange ao processo de limpeza dos dados, foram eliminadas palavras irrelevantes (Stop Words) e linhas contendo dados nulos e de qualidade baixa.

Em concordancia com o objetivo deste de analisar informações de papéis de empresas listadas na Bolsa de Valores do Brasil (B3), foi realizado o levantamento de informações sobre os papéis de empresas selecionadas, conforme exibido no Quadro 1 abaixo:

	TICKER	Setor	SubSetor	Segmento	PREÇO	Cotação Máxima 12 meses	Queda do Máximo	MARGEM BRUTA	MARG. LIQUIDA	LIQUIDEZ MEDIA DIARIA
0	VALE3	Materiais Básicos	Mineração	Minerais Metálicos	69.61	96.18	-0.276253	0.4370	3.10	38.91
1	PETR4	Petróleo, Gás e Biocombustíveis	Petróleo, Gás e Biocombustíveis	Exploração, Refino e Distribuição	22.84	30.20	-0.243709	0.5042	1.54	31.47
2	ITUB4	Financeiro	Intermediários Financeiros	Bancos	24.86	30.37	-0.181429	0.1970	5.45	15.59
3	BBDC4	Financeiro	Intermediários Financeiros	Bancos	13.59	20.79	-0.346320	0.2247	6.49	14.43
4	B3SA3	Financeiro	Serviços Financeiros Diversos	Serviços Financeiros Diversos	11.54	15.75	-0.267302	0.5911	10.97	3.58

Como observado, vários aspectos foram considerados, incluindo o setor de atividade, subsetor, segmento, preço das ações, cotação máxima nos últimos 12 meses, margens, liquidez média diária, dentro outros.

Em seguida, foi feito uso da linguagem de programação Python para a implementação do método de similaridade. As bibliotecas e ferramentas apresentadas na Figura 2 foram empregadas para realizar a análise e tratamento dos dados:

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt
import numpy as np
import math
pd.set_option('display.max_columns', 100)
pd.set_option('display.max_rows', 80)
```

Posteriormente, os dados das ações foram transformados em vetores numéricos para que os ângulos de cosseno pudessem ser calculados para cada par de vetores.

```
(0, 119)      0.44461814808819566
(0, 58)       0.4288564127803777
(0, 65)       0.4630942331686446
(0, 82)       0.4774505536953834
(0, 94)       0.4195015847284563
(0, 68)       0.4313797799445089
(0, 48)       0.4563106749900521
(0, 75)       0.4884521888834135
(0, 111)      0.4884521888834135
(0, 35)       0.35849821714362
```

A partir da representação vetorial das informações, que variam entre [0,1], tornou-se possível empregar as medidas de similaridade do cosseno para avaliar a distância entre dois conjuntos de vetores.

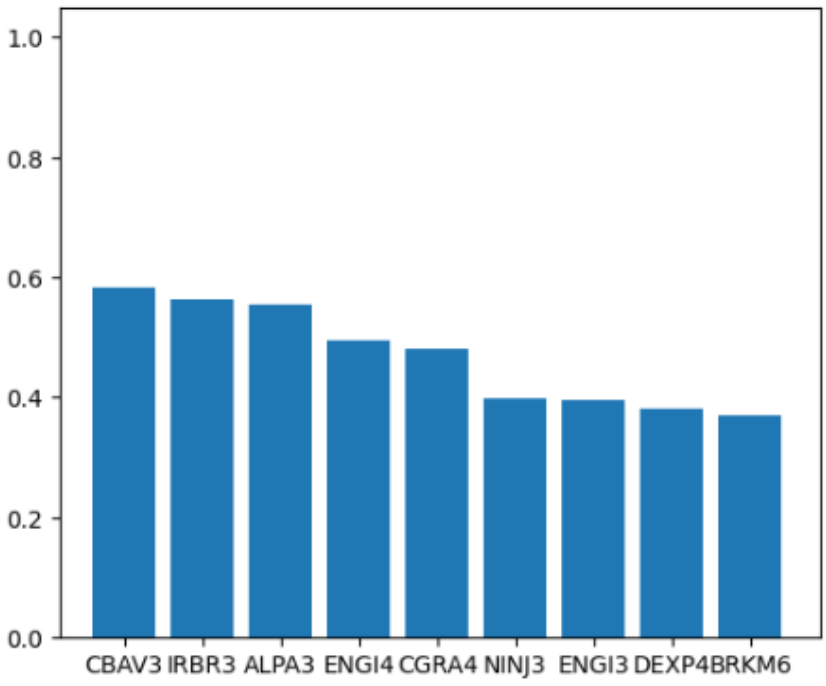
4 Análise e Resultados

Os resultados finais deste estudo indicam que a função recebe uma variável selecionada como entrada e gera uma matriz de similaridade entre cada par de ações como saída, com uma estrutura semelhante à matriz de correlação, observados na Tabela 1:

TICKER	ALPA4	Angulo
CBAV3	0.582626	54.364541
IRBR3	0.562701	55.757171
ALPA3	0.555905	56.226901
ENGI4	0.495194	60.317447
CGRA4	0.479532	61.345168
NINJ3	0.399235	66.469665
ENGI3	0.394890	66.740875
DEXP4	0.381799	67.554857
BRKM6	0.370015	68.283461

Com base na Tabela 1, buscou-se na matriz a seleção do papel ALPA4 da Alpargatas S.A., empresa brasileira especializada em calçados e vestuário, utilizando a indexação ou busca por colunas. Após a seleção, uma lista contendo os 10 menores ângulos de cosseno é gerada, indicando as ações mais semelhantes.

Similarmente, a representação gráfica da matriz de similaridade pode ser uma forma visual e intuitiva de entender quais ações são mais semelhantes entre si.



Neste caso, estamos interessados em encontrar as 10 ações mais semelhantes ao papel ALPA4. Considerando que valores dos ângulos de cosseno variam de 0 a 1, onde 0 representa a menor similaridade e 1 representa a maior, as 10 barras mais altas do gráfico representam as ações com maior similaridade. Assim, o gráfico nos permitirá visualizar rapidamente as ações que têm maior similaridade com o papel ALPA4, o que pode ser útil para análise de investimentos e tomada de decisões financeiras.

A representação das informações em forma de vetores numéricos permite uma análise mais objetiva e quantitativa, reduzindo a subjetividade na identificação de similaridades entre as ações.

A utilização de medidas de similaridade, como o cosseno, permite a comparação direta entre os conjuntos de vetores, possibilitando uma avaliação mais precisa da similaridade entre as ações. A geração da matriz de similaridade, que representa a similaridade entre cada par de vetores, é uma forma eficiente de resumir as informações e permitir uma análise mais abrangente da relação entre as ações.

A seleção do vetor de interesse, como o papel ALPA4, permite a identificação das ações mais semelhantes a ele com base nos valores de similaridade na matriz de similaridade. Essa abordagem permite uma análise mais focada e específica, reduzindo o esforço e tempo necessários para uma análise abrangente de outras ações.

5 Conclusão

Este estudo apresentou um novo método baseado na medida de similaridade do cosseno para identificar a similaridade entre ações financeiras com base em suas informações. Os resultados indicam que a abordagem proposta pode ser útil para a seleção de ativos diversificados e redução de riscos, o que pode ser valioso para para a análise de investimentos, mas deve ser considerada como uma parte de um conjunto maior de ferramentas e técnicas para a tomada de decisões de investimento.

É importante ressaltar que, apesar dos resultados promissores, esta pesquisa apresenta algumas limitações que devem ser consideradas. A similaridade do cosseno é uma das medidas mais populares para a avaliação de similaridade de dados multidimensionais. No entanto, é importante destacar que existem outras medidas que podem ser mais apropriadas para diferentes tipos de dados ou contextos. A escolha da medida de similaridade dependerá do objetivo da análise, das características dos dados e do método de análise utilizado.

Neste caso, a medida de similaridade do cosseno por ser baseada em informações numéricas, pode não capturar todos os aspectos relevantes das empresas ou setores em análise. Além disso, a abordagem utilizada não leva em consideração outras informações importantes, como notícias ou eventos externos que possam afetar o desempenho das ações. Dessa forma, sugere-se que futuras pesquisas explorem outras fontes de informação e outras medidas de similaridade para complementar a análise proposta neste estudo.