Tópicos Especiais em Ciências de Dados II
Atividade Prática

Curso: CST em Ciência de Dados
Período Letivo: 2023 - 2° SEMESTRE - CST EM CIÊNCIA DE DADOS - 3/A
Nome: Bruna da Cunha Christiano Gomes
RGM: 1831980759

Proposta

Realizar experimento a partir da execução do algoritmo Kmeans utilizando uma das pastas de texto da base Reuters e analisar os clusters gerados, verificando se os assuntos estão relacionados entre si.

Contexto

Na plataforma Hadoop, utilizando uma das pastas com textos da Base Reuters C50train, após implementação e execução do algoritmo Kmeans com Mahout, considerando:

- Distância Euclidiana como métrica de distância;
- Valor máximo de 10 iterações do algoritmo;
- Geração de 10 clusters do modelo a ser gerado pelo algoritmo.

Clusters

Cluster VL-1866	
pla	=> 15.974252700805664
troops	=> 13.125126838684082
hong	=> 11.30401611328125
kong	=> 10.972043991088867
handover	=> 10.358711242675781
dickson	=> 9.945155143737793
send	=> 9.209076881408691
cui	=> 8.649324417114258
6,000	=> 8.558053016662598
courts	=> 8.37457275390625

Cluster VL-1782	
mci	=> 17.108234405517578
telecom	=> 13.894393920898438
t	=> 11.063084602355957
distance	=> 10.838738441467285
warchest	=> 9.945155143737793
british	=> 8.928696632385254
amp	=> 8.73509693145752
kagan	=> 8.173480033874512
schawb	=> 7.725433826446533
incorprated	=> 7.725433826446533

Cluster VL-1236	
prefix	=> 16.631319046020508
cruickshank	=> 14.875503540039062
numbers	=> 13.705758094787598
phoneday	=> 13.380844116210938
oftel	=> 11.679633140563965
020	=> 10.925413131713867
prefixes	=> 10.925413131713867
codes	=> 9.371740341186523
numbering	=> 8.708802223205566
ignored	=> 8.47231674194336

Cluster VL-440	
porterbrook	=> 15.450867652893066
stagecoach	=> 14.875503540039062
swebus	=> 13.380844116210938
bus	=> 13.268099784851074
souter	=> 10.925413131713867
hannah	=> 10.518568992614746
pounds	=> 9.508763313293457
13.5	=> 8.649324417114258
million	=> 8.149345636367798
pence	=> 7.7960686683654785

Cluster CL-1511	
its	=> 1.696800750059511
he	=> 1.6049599973031028
from	=> 1.601269574833646
percent	=> 1.5726499146527932
has	=> 1.5454719730600732
would	=> 1.5241503964207896
million	=> 1.5214095160791195
have	=> 1.5169029292971965
year	=> 1.5062150011319957
which	=> 1.4502068078186643

Cluster VL-202	
fairfax	=> 19.740516026814777
murdoch	=> 19.097466786702473
turner	=> 16.493302981058758
warner	=> 13.998514811197916
packer	=> 12.495693524678549
courier	=> 11.067626317342123
cnn	=> 10.638603687286377
attainable	=> 10.518568992614746
kerry	=> 8.851698875427246
news	=> 8.490553538004557

Cluster VL-2064	
tech	=> 13.45728588104248
211	=> 12.852301597595215
ballot	=> 12.678278923034668
cramer	=> 10.518568992614746
measure	=> 10.362813949584961
doerr	=> 10.202996253967285
lawsuits	=> 9.39264965057373
valley	=> 9.39264965057373
silicon	=> 9.280865669250488
activism	=> 9.222738265991211

Cluster VL-945	
csu	=> 11.559046745300293
m2	=> 10.202996253967285
gdp	=> 10.183777809143066
inflation	=> 9.823558807373047
21.8	=> 9.727152824401855
slowed	=> 9.640506744384766
crowns	=> 9.527206420898438
slowing	=> 9.335855484008789
maly	=> 8.964897155761719
cnb	=> 8.746894836425781

Cluster VL-953	
vauxhall	=> 14.352546374003092
astra	=> 12.024346987406412
reilly	=> 10.638603687286377
ellesmere	=> 9.491629918416342
motors	=> 7.702427228291829
6317	=> 7.437751770019531
model	=> 7.263563474019368
modernise	=> 6.521461009979248
4,200	=> 6.426150798797607
motor	=> 6.30958875020345

Cluster VL-567	
deposits	=> 15.174803733825684
inflation	=> 11.623391151428223
icbc	=> 10.925413131713867
m2	=> 10.202996253967285
relaxation	=> 9.371740341186523
monetary	=> 9.175860404968262
tight	=> 8.762789726257324
bank	=> 8.706725120544434
pose	=> 8.315020561218262
cuts	=> 8.229031562805176

Análise

Os clusters foram analisados a partir da percepção do contexto dos grupos formados pelo algoritmo k-means, verificando se os assuntos realmente possuem alguma relação entre si dentro de cada cluster.
De fato, cada cluster gerado pelo algoritmo aparenta ter uma temática específica com base nos termos mais relevantes.
Os percentuais apresentados dos principais termos em cada cluster também fornecem informações sobre a importância relativa de cada termo dentro do cluster específico. A análise desses percentuais pode ajudar a entender quais termos têm maior influência na formação do cluster.

Cluster VL-1866:

- **Principais Termos:**
  - pla, troops, hong, kong, handover, dickson, send, cui, 6,000, courts
  - O termo "pla" tem um peso percentual significativo de 15.97%, indicando que é um termo crucial para este cluster.

    - Os termos "troops", "hong", "kong", e "handover" também têm pesos percentuais consideráveis, sugerindo que questões relacionadas a Hong Kong e transferência de poder podem ser temas importantes neste cluster.
- **Interpretação:** Este cluster parece estar relacionado a eventos geopolíticos ou políticos relacionados a Hong Kong, com termos como sugerindo eventos específicos, talvez relacionado a transferência de poder.

Cluster VL-1782:

- **Principais Termos:**
    - mci, telecom, t, distance, warchest, british, amp, kagan, schawb, incorprated
    - O termo "mci" tem um peso percentual notável de 17.11%, indicando sua importância neste cluster.
    - Outros termos como "telecom", "t", e "distance" também têm pesos significativos, sugerindo uma forte associação com o setor de telecomunicações.
- **Interpretação:** Este cluster pode estar relacionado ao setor de telecomunicações, com termos como "telecom" e "amp" sugerindo empresas, com destaque para a empresa MCI e temas relacionados com tecnologias específicas.

Cluster VL-1236:

- **Principais Termos:**
    - prefix, cruickshank, numbers, phoneday, oftel, 020, prefixes, codes, numbering, ignored
    - O termo "prefix" tem um peso percentual alto de 16.63%, indicando que é um termo-chave para este cluster.
    - Termos como "cruickshank", "numbers", e "phoneday" também têm pesos consideráveis, sugerindo uma possível associação com numeração ou códigos.
- **Interpretação:** Este cluster pode estar associado a questões relacionadas a códigos telefônicos e regulamentações de telefonia.

Cluster VL-440:

- **Principais Termos:**
    - porterbrook, stagecoach, swebus, bus, souter, hannah, pounds, 13.5, million, pence
    - Os termos "porterbrook" e "stagecoach" têm pesos percentuais elevados, indicando sua importância na formação deste cluster.
    - Outros termos relacionados ao setor de transporte, como "bus" e "swebus", também têm pesos significativos.
- **Interpretação:** Este cluster parece relacionado ao setor de transporte, possivelmente ônibus e empresas associadas, especialmente a empresas específicas como Stagecoach e Porterbrook.

Cluster CL-1511:

- **Principais Termos:**
    - its, he, from, percent, has, would, million, have, year, which
    - Os termos apresentados têm pesos percentuais relativamente baixos e podem indicar que este cluster é mais genérico, abrangendo uma variedade de temas.

- **Interpretação:** Termos genéricos que podem ser difíceis de interpretar sem mais contexto. Pode ser um cluster mais geral ou diversificado. Este cluster parece ser mais genérico e pode incluir documentos que não se encaixam claramente em categorias específicas, ou seja, pode não ter uma associação temática clara com base nos termos destacados.

Cluster VL-202:

- **Principais Termos:**
  - fairfax, murdoch, turner, warner, packer, courier, cnn, attainable, kerry, news
  - Termos como "fairfax", "murdoch", e "turner" têm pesos percentuais altos, sugerindo uma forte associação com mídia e notícias.
- **Interpretação:** Este cluster pode estar relacionado a notícias e mídia, com termos como "cnn" e "news" indicando temas de mídia, além de empresas do setor, com menção a figuras conhecidas como Murdoch e Turner.

Cluster VL-2064:

- **Principais Termos:**
  - tech, 211, ballot, cramer, measure, doerr, lawsuits, valley, silicon, activism
  - O termo "tech" tem um peso percentual elevado, indicando uma forte associação com tecnologia.
  - Outros termos como "211", "ballot", e "cramer" também têm pesos significativos.
- **Interpretação:** Este cluster parece estar relacionado a tecnologia, com termos como "tech", "silicon", "ballot" e "valley", além questões legais associadas.

Cluster VL-945:

- **Principais Termos:**
  - csu, m2, gdp, inflation, 21.8, slowed, crowns, slowing, maly, cnb
  - O termo "csu" tem um peso percentual elevado, indicando sua importância neste cluster.
  - Termos econômicos como "gdp", "inflation", e "slowdown" também têm pesos consideráveis.
- **Interpretação:** Este cluster pode estar relacionado a indicadores econômicos, como PIB, inflação e outros termos econômicos.

Cluster VL-953:

- **Principais Termos:**
  - vauxhall, astra, reilly, ellesmere, motors, 6317, model, modernise, 4,200, motor
  - O termo "vauxhall" tem um peso percentual notável, indicando uma forte associação com a marca automotiva.
  - Outros termos relacionados à indústria automotiva também têm pesos significativos.
- **Interpretação:** Este cluster parece estar relacionado à indústria automotiva, com termos como "vauxhall", "astra" e "motors".

Cluster VL-567:

- **Principais Termos:**
  - deposits, inflation, icbc, m2, relaxation, monetary, tight, bank, pose, cuts

    - Termos como "deposits", "inflation", e "icbc" têm pesos percentuais elevados, sugerindo uma forte associação com questões econômicas e bancárias.
- **Interpretação:** Este cluster pode estar relacionado a políticas monetárias e questões bancárias, com termos como "inflation" e "bank" indicando temas econômicos.

Conclusão

As análises apresentadas são interpretações baseadas nos termos mais relevantes em cada cluster. Ao analisar os percentuais dos principais termos em cada um destes, é possível identificar os termos mais influentes e compreender as principais temáticas ou tópicos associados a cada grupo. Isso auxilia na interpretação dos clusters e na compreensão das relações entre os documentos agrupados pelo algoritmo K-means.

É útil destacar que o conhecimento do conjunto de dados e de domínio adicional elevam a eficácia das interpretações no âmbito da situação.