ATIVIDADE PRÁTICA - TÓPICOS ESPECIAIS EM CIÊNCIAS DE DADOS

####### PROPOSTA DA ATIVIDADE - RESUMO ######################

"É proposta uma atividade mais prática, considerando que vocês já possuem instalada a plataforma Hadoop, bem como o mahout, portanto, vocês poderão fazer os experimentos aqui propostos, onde é disponibilizada uma base de textos da Reuters.

A ideia da atividade é vocês executarem o algoritmo kmeans usando uma das pastas com os textos, e analisar qual o resultado do algoritmo. Observem os clusters gerados, e se de fato os assuntos possuem relação entre si."

####### APOSTILAS DE REFERÊNCIA DO EAD POR DISCIPLINA ######################

- **Big Data**
  - **Unid_IV - Big e IoT, Projeto Hadoop e Subprojetos**
    - (Introdução ao Projeto Hadoop)
  - **Unid_V - Outros Projetos Importantes do Hadoop, sua Instalação e Execução**
    - (Aqui fala sobre o hadoop e o mahout)

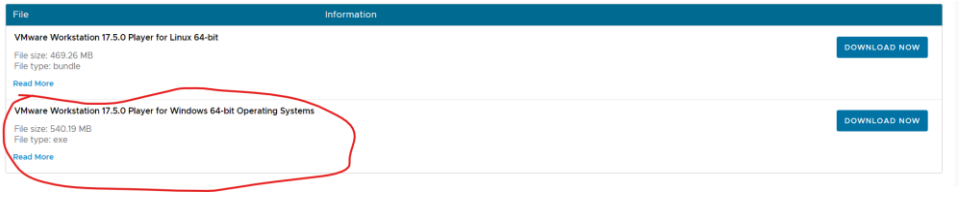
- **Tópicos Especiais em Ciências de Dados**
  - **Unid_III - Análise de Textos e Prática de Kmeans**
    - (Essa explica sobre a análise que é solicitada na atividade, a partir da execução do algoritmo de clustering)
  - **Unid_V - Algoritmos e Big Data na Prática com Apache Mahout**
    - (Essa apostila tem a proposta da atividade, mas considera que você já tem o Haddop e o Mahout instalado na máquina)

####### CONFIGURAÇÕES PARA EXECUTAR A ATIVIDADE ######################

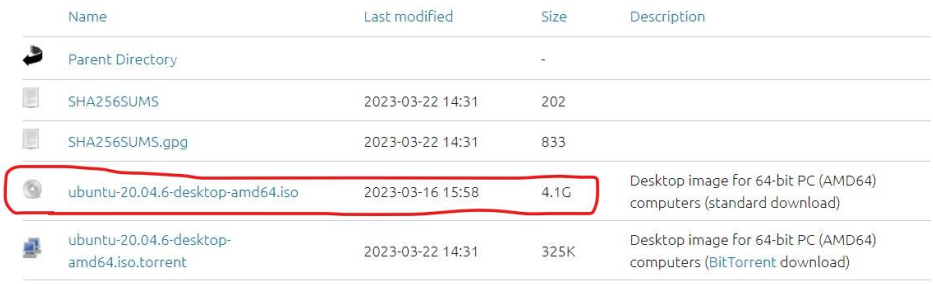
- **Eu segui todos os passos desse vídeo e você vai fazer o mesmo:**
  - #72 - Instalando Apache Hadoop e Apache Mahout
    - https://www.youtube.com/watch?v=hTQmAlSUaTE

- **Esse é o github do autor do vídeo com, praticamente, o mesmo projeto proposto pela atividade.**
  - https://github.com/netocosta/HadoopMahout#readme
    - Deixa aberto o arquivo README.md junto com o vídeo, pois nesse arquivo tem as linhas de comando.

- **Mas já pode adiantar e fazer o download e instalação da VMware, máquina virtual que vai emular o Linux no seu Windows (estou supondo que seu SO é Windows 64bits).**
  - Pra isso você fará o download da VMware: https://customerconnect.vmware.com/en/downloads/details?downloadGroup=WKST-PLAYER-1750&productId=1377&rPId=111473



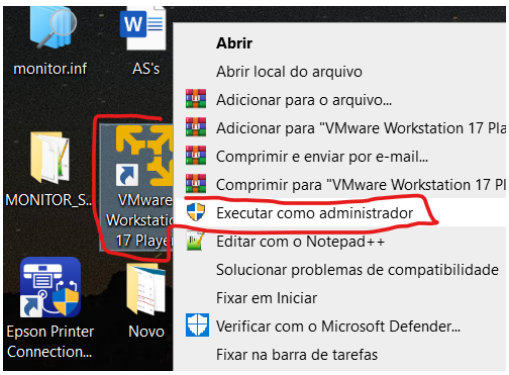
  - E do Ubuntu 20 (Linux): https://releases.ubuntu.com/focal/



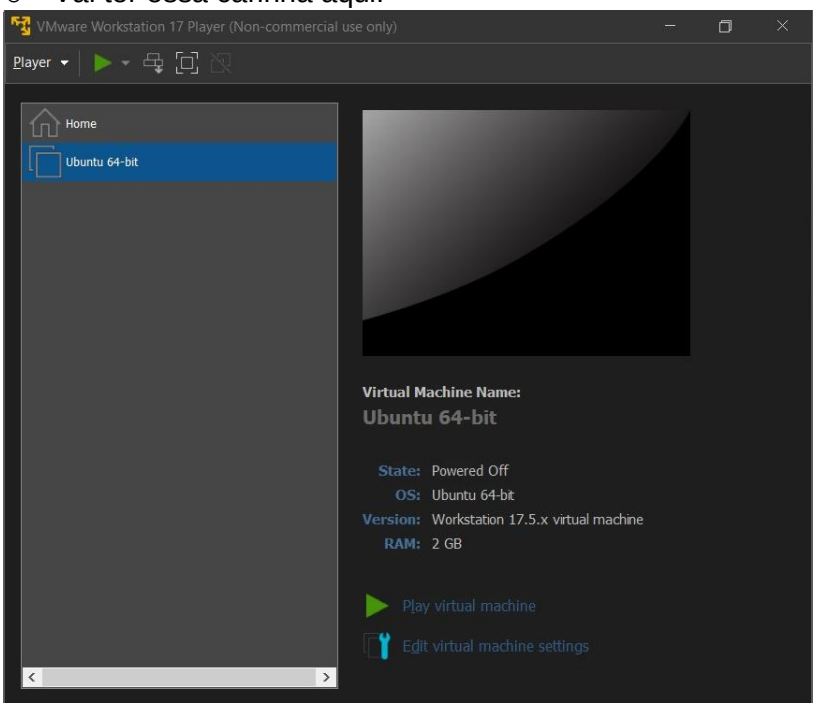
Name	Last modified	Size	Description
Parent Directory		-	
SHA256SUMS	2023-03-22 14:31	202	
SHA256SUMS.gpg	2023-03-22 14:31	833	
ubuntu-20.04.6-desktop-amd64.iso	2023-03-16 15:58	4.1G	Desktop image for 64-bit PC (AMD64) computers (standard download)
ubuntu-20.04.6-desktop-amd64.iso.torrent	2023-03-22 14:31	325K	Desktop image for 64-bit PC (AMD64) computers (BitTorrent download)

- **Você instala a VMWare, conforme ele te orienta no vídeo e depois você vai instalar o Ubuntu pela VMware, referenciando o arquivo de imagem do Ubuntu que você fez baixou.**
  - **NÃO CUSTA LEMBRAR:** Sempre abra, execute, instale... como Administrador (botão direito do mouse > Executar como Administrador).
  - **DICA IMPORTANTE:** Na hora da instalação do Ubuntu na VMware, tem a parte da configuração que você pode configurar a capacidade do HD. O Padrão é 20 GB, mas eu indico aumentar para no mínimo 35, pois na hora do processamento, pode ser que 20 GB não seja suficiente e trave o processo do algoritmo.

- **Depois da instalação do Ubuntu na VMWare:**
  - Caso você saia da VMware e execute ela novamente, não se esqueça de sempre executar como Administrador:

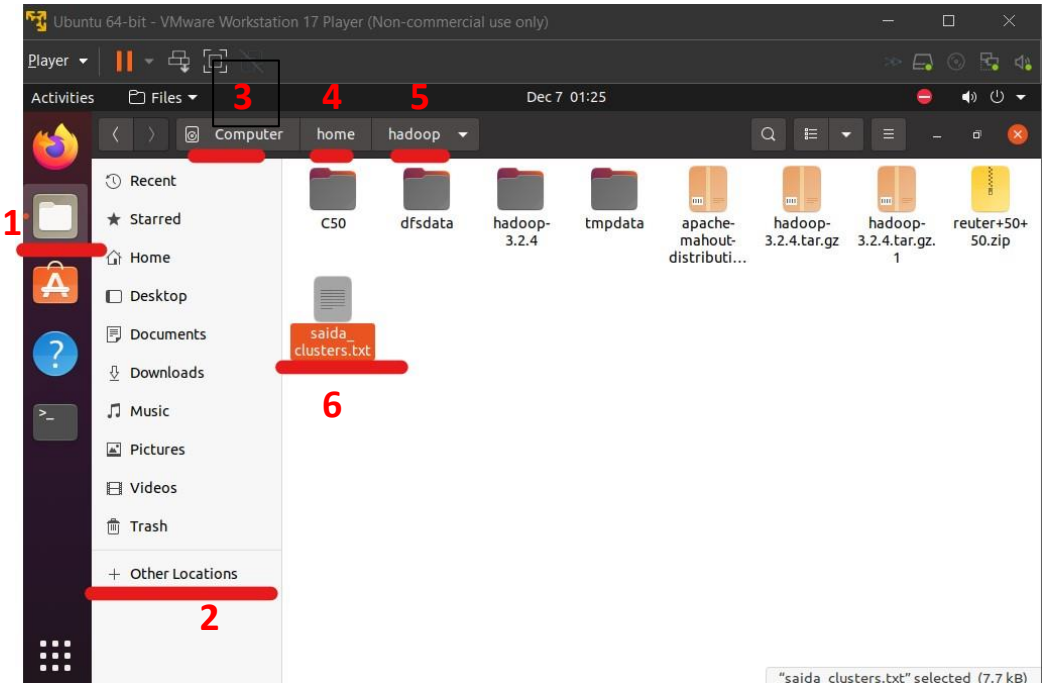


  - Vai ter essa carinha aqui:



- **A partir daí, no vídeo ele vai explicar como abrir o prompt de comando pelo Ubuntu e começar as outras configurações:**

- ✓ Baixar e Instalar OpenJDK no Ubuntu (Java)
- ✓ Baixar e Instalar OpenSSH no Ubuntu
- ✓ Instale OpenSSH no Ubuntu
- ✓ Criar usuário Hadoop
- ✓ Baixar e Instalar O Hadoop
- ✓ Definindo as variaveis de ambiente
- ✓ Atualizando bashrc
- ✓ Configurando o hadoop-env.sh
- ✓ Vai configurar alguns arquivos xml
- ✓ Vai baixar a base proposta na atividade (Baixando o Reuters C50)
- ✓ Criando diretorios / Permissões (para fazer o processamento do algoritmo)
- ✓ Configurar o Formato HDFS NameNode
- ✓ Iniciando Serviços
- ✓ Execute: jps 9vc vai abrir o navegador visualizar a interface do Hadoop)
- ✓ Instalando o mahout
- ✓ Adicionar as variaveis de ambiente
- ✓ Salve e execute
- ✓ Instalando utilitarios
- ✓ Executando o código proposto no material da Cruzeiro do Sul:
  - hadoop fs -copyFromLocal C50/ /
  - ./mahout seqdirectory -i /C50/C50train -o /seqreuters -xm sequential
  - ./mahout seq2sparse -i /seqreuters -o /train-sparse
  - ./mahout kmeans -i /train-sparse/tfidf-vectors/ -c /kmeans-train-clusters -o /train-clusters-final -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure -x 10 -k 10 -ow
  - ./mahout clusterdump -d /train-sparse/dictionary.file-0 -dt sequencefile -i /train-clusters-final/clusters-10-final -n 10 -b 100 -o ~/saida_clusters.txt -p /train-clusters-final/clustered-points
- ✓ No final será gerado um arquivo "saida_clusters.txt" com o resultado.
  - O arquivo gerado estará em home/hadoop/saida_clusters.txt
    - Siga a trilha da imagem:



####### EXECUÇÃO DO ALGORITMO KMEANS ######################

"A ideia da atividade é vocês executarem o algoritmo kmeans usando uma das pastas com os textos, e analisar qual o resultado do algoritmo. Observem os clusters gerados, e se de fato os assuntos possuem relação entre si."

- ✓ **A explicação da execução de cada comando do algoritmo está na apostila Unid_V - Algoritmos e Big Data na Prática com Apache Mahout da disciplina Tópicos Especiais em Ciências de Dados** a **partir da página 10.**
- ✓ **A explicação do conceito da execução do Kmeans para análise de texto está na apostila Unid_III - Análise de Textos e Prática de Kmeans da disciplina Tópicos Especiais em Ciências de Dados** a partir da **página 17, mas eu recomendo fortemente ler toda apostila para entender o restante do algoritmo.**

- ✓ Comando para a cópia dos arquivos com os textos.
  - hadoop fs -copyFromLocal C50/ /

- ✓ Criar arquivos de vetores sequenciais para o processamento dos textos.

São passados parâmetros de diretório ou arquivo de entrada, o diretório de saída do comando e o parâmetro que informa para se criar o vetor sequência.

  - ./mahout seqdirectory -i /C50/C50train -o /seqreuters -xm sequential

✓ Faz efetivamente a análise de textos a partir dos vetores criados, são criados, então, os vetores contendo o TF-IDF calculado para os termos e documentos existentes.

O comando recebe como parâmetro o diretório sequencial de entrada, o diretório para saída e o parâmetro que informa a criação de vetores de análise de texto, por padrão usando o TF-IDF "seq2sparce".

  - ./mahout seq2sparse -i /seqreuters -o /train-sparse

- ✓ Executa efetivamente o algoritmo kmeans.
- ✓ O comando de execução do kmeas recebe os seguintes parâmetros:

- i: passa os vetores e valores de TF-IDF para o algoritmo;
- c: indica o diretório de clusters iniciais;
- o: indica o diretório de execução do algoritmo e onde ficarão armazenados os clusters gerados;
- dm: indica o tipo de métrica de distância a ser utilizada, nesse caso, utilizando a distância euclidiana;
- x: indica o número máximo de iterações do algoritmo, nesse caso, o valor máximo de 10 iterações;
- k: indica o número de clusters do modelo a ser gerado pelo algoritmo, nesse caso, 10 clusters;
- ow: indica que o algoritmo poderá sobrescrever os diretórios caso os indicados para saída não estejam vazios;

  - ./mahout kmeans -i /train-sparse/tfidf-vectors/ -c /kmeans-train-clusters -o /train-clusters-final -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure -x 10 -k 10 –ow

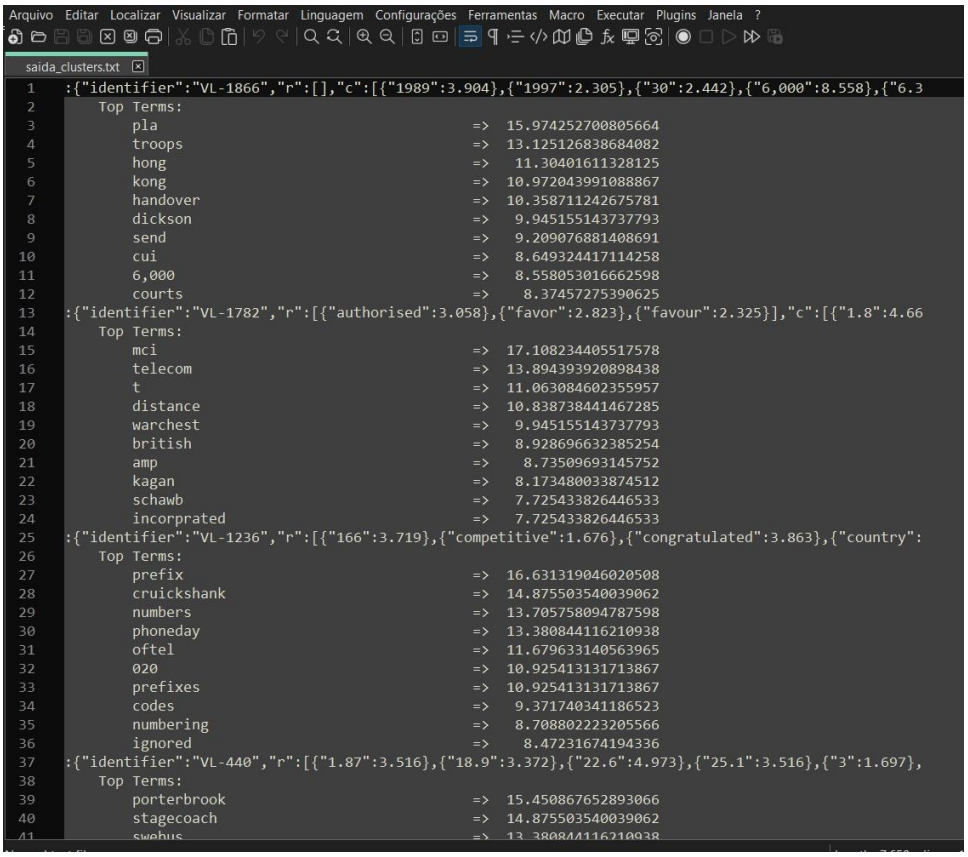
- ✓ A saída do algoritmo será no formato de vetores e será necessário converter os vetores de saída para texto plano.
- ✓ O Mahout disponibiliza a ferramenta "clusterdump" que permite a criação de um arquivo de saída com texto plano e no diretório de trabalho fora do HDFS.
- ✓ Os parâmetros passados para o comando "clusterdump" são:

-d: indica o arquivo de dicionário, nesse caso, o dicionário inicialmente criado pelo comando de criação dos valores de TF-IDF;
-dt: indica o tipo de dicionário que foi passado como parâmetro, nesse caso, o tipo era arquivos/vetores sequenciais;
-i: indica o diretório de entrada para o comando, nesse caso, contendo os clusters criados pelo algoritmo kmeans;
-n: indica o número de top termos existentes no cluster, nesse caso, o arquivo de saída conterá os 10 principais termos que caracterizam o cluster;
-b: indica o texto que caracterizará o centroide, nesse caso, o parâmetro 100 indica os 100 caracteres inicias que caracterizam o centroide;
-o: indica o arquivo de saída do comando, nesse caso, fora do HDFS;

-p: indica os pontos que foram passados como parâmetros para a criação dos vetores e clusters.

- ./mahout clusterdump -d /train-sparse/dictionary.file-0 -dt sequencefile -i /train-clusters-final/clusters-10-final -n 10 -b 100 -o ~/saida_clusters.txt -p /train-clusters-final/clustered-points

Agora vamos analisar o arquivo saida_cluster.txt e montar nosso documento de entrega da atividade.



####### ANÁLISE DOS CLUSTERS E DOCUMENTAÇÃO ######################

Para analisar a saída do algoritmo Kmeans dentro da proposta da atividade "*Observem os clusters gerados, e se de fato os assuntos possuem relação entre si.”* **gerei 2 questionamentos no ChatGPT e montei meu documento:**

1. Utilizando uma das pastas com textos da Base Reuters C50train, executei algoritmo kmeans original. Com objetivo de analisar o resultado do algoritmo, observando os clusters gerados, como eu descreveria, para cada cluster, a análise para verificar sede fato os assuntos possuem relação entre si?
   A partir da linha de comando './mahout kmeans -i /train-sparse/tfidf-vectors/ -c /kmeans-train-clusters -o /train-clusters-final -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure -x 10 -k 10 -ow', segue a saída de clusters gerados no haddop mahout : {aqui copiei e colei o resultado do arquivo "saida_clusters.txt"}

2. Após a execução do algoritmo Kmeans, o que podemos inferir e analisar em cada um dos clusters no que diz respeito aos percentuais apresentados dos principais termos?