










binários que devem ser carregados antes de qualquer consulta. Esse carregamento é feito quando o AC é inicializado.

A Figura 9 mostra os documentos e diretórios do MPC. O diretório do MPC é organizado da seguinte forma:

- **“DOCUMENTOS”**: guarda todos os documentos que serão processados. Para fins de organização, estes documentos podem estar contidos em subpastas;
- **“COLECAO VETORES BINARIOS”**: é o diretório onde os vetores binários dos documentos são escritos;
- **“Vetores binarios.txt”**: é o documento em que todos os vetores binários são agrupados;
- **“VISA LOGICA”**: armazena os documentos de visão lógica de cada documento;
- **“Vocabulario.txt”**: contém o vocabulário de toda a coleção.

Figura 9 - Diretório do Modelo Probabilístico

 COLECAO BM25	06/04/2016 10:44	Pasta de arquivos
 COLECAO VETORES BINARIOS	14/03/2016 02:23	Pasta de arquivos
 DOCUMENTOS	11/03/2016 00:53	Pasta de arquivos
 VISA LOGICA	11/03/2016 01:12	Pasta de arquivos
 Tamanho dos documentos.txt	21/03/2016 02:51	Documento de Texto
 Tamanho medio dos documentos.txt	21/03/2016 02:51	Documento de Texto
 Vetores binarios.txt	21/03/2016 23:34	Documento de Texto
 Vetores BM25.txt	06/04/2016 11:31	Documento de Texto
 Vocabulario.txt	11/03/2016 01:24	Documento de Texto

Fonte: Elaborado pelo autor

A coleção de documentos pode ser muito grande, diante disso é possível presumir que o tempo necessário para processamento e gravação em memória secundária vai ser diretamente proporcional. Para tratar, verificar e escrever os 51830 documentos que formam a base inicial de documentos foram necessárias 20 horas. O tempo para trabalhar todos os documentos não precisa ser contínuo, o AC permite que o processamento seja interrompido e retomado posteriormente. Esta interrupção é possível porque cada vetor binário é escrito na memória secundária logo após ser terminado.

Quando uma consulta é enviada para o MPC, uma lista de documentos ordenados pela relevância de conteúdo é feita aplicando-se a equação de similaridade do modelo probabilístico com os valores obtidos dos vetores binários.