UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
DISCIPLINA: TÓPICOS AVANÇADOS EM INTELIGÊNCIA ARTIFICIAL
PROFESSOR: DOUGLAS VITÓRIO
ALUNO: Ivo Ireneu de Souza Júnior

1ª VA

01. Pesquise uma aplicação/uso das seguintes tarefas de Mineração de Texto:

a) Classificação de texto

- Classificar Twits quanto ao sentimento de mercado em relação a uma ação na bolça de valores. Já existe uma base de dados predefinida (usada para aprendizado supervisionado) segundo sua classe e/ou categoria, com a qual é possível fazer a classificação.

b) Clusterização de texto

- Nesta operação, não existe uma base de conhecimento pra categorização ou categorias predefinidas, o que existe um aprendizado não supervisionado, para agrupamento dos twits que apresentam similaridade em seus textos.

c) Sumarização

- Ele vai criar uma versão simplificada, ou tipificada de cada texto, que contenham os principais temas. A exemplo, um twit de Elon Musk falando sobre algum cripto ativo, nesse texto os tópicos serão o ativo citado, o sentimento dele sobre esse ativo no mercado e a projeção do mesmo para o futuro.

d) Extração de informação

- A ideia aqui, é encontrar as idéias chave do texto através de sequencias predefinidas. Para o nosso exemplo, sequencias predefinidas como: [EUR vai valorizar], [O bitcoin vai cair], entre outras.

e) Recuperação de informação

- Para obtenção de documentos sobre balanços patrimoniais de certas empresas, para observar a possibilidade de valorização ou não da mesma, para isso existem documentos modelo, os quais serviram para fundamentar a busca e assim retornar documentos de mesmo cunho.

02. O que você entendeu por **Mineração de Opinião**?

- A opinião é um posicionamento de um indivíduo sobre algo, ou alguém. Minerar opinião, está relacionado a busca por opiniões especificas, no meio de muitas e aleatórias encontrar um certo grupo de seleto de opiniões de acordo com alguma métrica.

03. Explique as 4 etapas do processo de mineração de opinião.

- Primeiro será executado a coleta de informações, cuja dará respaldo para o processo de mineração, nele será composto o corpus do da base de dado que é um conjunto de documentos de texto que contem opinião;
- Os dados precisam ser pré-processados para que possam ser legíveis ao classificador.
- Processar as informações e classificar com algoritmos de aprendizado de máquina.
- Por fim, avaliar os dados retornados e testa-los, para que possa validar os resultados da mineração.

04. Por que o tipo de texto oriundo do Twitter (informal e curto) é mais difícil de ser processado?

- Por que pode apresentar muitos ruídos de informação, que podem atrapalhar ou levar a uma avaliação equivoca. O melhor caso para dados considerados muito bons, seriam os dados que mais se encontra na linguagem formal, pois facilita a associação da máquina, dessa forma os dados podem estar na “mesma linguagem”, de outra forma, quando os dados estão em linguagem informal, coloquial e de forma abreviada ou de forma cultural, dificulta a associação de conhecimento por parte do algoritmo utilizado, já que esse tipo de informação normalmente não existe um padrão.

05. O que é um corpus desbalanceado e por que é melhor se ter um corpus balanceado?

- Corpus desbalanceado se refere a quantidade diferente de documentos em cada classe da base de dados, o corpus balanceado é a melhor opção, por se tratar de uma ferramenta utilizada para treinamento de um algoritmo, assim se as bases estiverem desbalanceadas, ocorre um grande risco dessas classes estarem mal treinadas, logo o algoritmo não desempenhará sua capacidade máxima, e poderá ter um baixo desempenho nos resultados de classificação. O balanceamento das classes existe, na tentativa de evitar tal problema.

06. Por que é necessário pré-processar os textos?

- A linguagem legível pelos classificadores ou algoritmos de aprendizado, necessitam de uma padronização da informação que é dada como entrada, para que possam apresentar uma saída mais eficiente. Esse é o papel do pré-processamento dos dados, fazer a estruturação dos dados para que possam ser legíveis pelos classificadores.

07. O que tokenização? Como ficaria a seguinte string ao passar pelo TweetTokenizer?

“A @vivobr pode fazer o projeto que for como propaganda mas a loja da empresa vende e não entrega”

- A tokenização cria uma lista de sub strings a partir de uma sentença principal. Segue a frase:
  - “A”, “@vivobr”, “pode”, “fazer”, “o”, “projeto”, “que”, “for”, “como”, “propaganda”, “mas”, “a”, “loja”, “da”, “empresa”, “vende”, “e”, “não”, ,“entrega”.

08. O que são stopwords? Por que é interessante removê-las?

- São palavras comuns a língua portuguesa, e sua remoção não irá interferir no sentido do texto. É interessante a sua remoção, por que sua presença só estará ocupando espaço.

09. O que é o processo de stemming? Utilize um stemmizador na string da questão 07 e coloque aqui o resultado.

- Converte palavras flexionadas para seu Stem correspondente. O exemplo:
  - [‘a’, ‘@vivobr’, ‘pod’, ‘faz’, ‘o’, ‘projet’, ‘que’, ‘for’, ‘como’, ‘propaga’, ‘mas’, ‘a’, ‘loj’, ‘da’, ‘empres’, ‘vend’, ‘e’, ‘não’, ‘entreg’].

10. Quais são as formas mais usadas de se construir um VSM?

- Valor Binário:
  - Cruza um documento como primeira classe, com tokens que fazem parte de cada documento, com cada célula indicando se possui(1) ou não(2) cada tokem.
- Frequência:
  - Muito parecido ao anterior, no entanto indica quantas vezes cada tokem aparece em cada documento.
- TF-IDF:
  - Aplica uma formula em cada célula de cada documento, calculando o produto no número de ocorrência pelo log da divisão do número total de documentos pela número de documentos que contém ocorrências.