Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Trabalho Prático 2

Tendo em vista o bom trabalho feito para o TCU, sua competência quanto ao estudo aplicado de Mineração de Dados foi difundida. Os boatos chegaram aos ouvidos do Ricardo Teixeira, que resolveu procurar lhe para realizar uma interessante tarefa para a CBF. Preocupada com a imagem da seleção brasileira, a CBF gostaria de saber o que os torcedores brasileiros estão falando sobre nossa seleção na WEB. De forma a compor um cenário mais controlado, você sugeriu que tais análises começassem pelo *Twitter*. Além disso, a CBF ficou sabendo de um interessante aplicativo existente no Observatório da Web[1] no contexto de eleições e gostaria de começar as análises de forma análoga. Ou seja, sua primeira atividade consiste em identificar em um dado jogo da seleção brasileira o que os usuários postam sobre a seleção no decorrer do jogo. Para tanto, a CBF disponibilizou para você, como dados de teste, um conjunto de tweets que eles coletaram durante o jogo entre Brasil e Holanda nas quartas de final da última copa do Mundo de 2010. Dessa forma, pede-se:

- Proponha um processo de tratamento e manipulação de dados que identifique quais são os termos e/ou frases coerentes mais freqüentes em tweets postados durante o jogo;
- Implemente tal processo e gere como saída, em um arquivo texto, os padrões encontrados bem como as correspondentes freqüências de ocorrência;
- A granularidade temporal de análise dos padrões será de 5 minutos. Ou seja, você deve primeiramente agrupar os tweets em intervalos de 5 minutos e identificar os **50** termos e/ou frases mais freqüentes em cada grupo. Sua saída deve estar ordenada cronologicamente.

Deve ser entregue uma documentação com todas as análises e decisões tomadas. Os critérios de correlação do presente trabalho serão todos embasados na qualidade das discussões realizadas, bem como nas justificativas para as decisões tomadas e saídas geradas. Vocês estão livres para utilizar

[1] http://www.observatorio.inweb.org.br/eleicoes2010/evento

qualquer tipo de algoritmo de identificação de padrões freqüentes, bem como utilizar implementações disponíveis na WEB, ou bibliotecas de ferramentas tais como *Weka*, *R*, *RapidMiner*, *Pentaho* dentre outras. A documentação deve conter no máximo 10 páginas, com espaçamento simples entre linhas, tamanho de fonte 12 e fonte *Times New Roman*.

Deve-se também entregar o conjunto de código ou scripts utilizados para achar os padrões freqüentes e a saída encontrada para o conjunto de entrada disponibilizado.

Dicas:

- É necessário fazer tratamentos no texto de forma a reduzir variações textuais e, conseqüentemente, aumentar as freqüências observadas. Não é exigido nenhum tratamento específico mas, as avaliações levarão em conta quais tipos de tratamento foram realizados;
- Não é necessário prover como saída frases coesas e sintaticamente corretas. Basta proverem os padrões encontrados mesmo que não formem um frase com preposições, pronomes ou outros elementos gramaticais de ligação. Entretanto, é necessário que os padrões gerados sejam coerentes;
- Embora em termos práticos haja um requisito forte por tempo, uma vez que tal processo é feito em tempo-real durante os jogos, no trabalho tal requisito foi removido. Ou seja, não se preocupem em realizar execuções efetivamente eficientes.

Data de entrega: 14 de Setembro