

Inteligência Competitiva – 2ª. Lista

O Exercício a seguir deve ser desenvolvido utilizando o ambiente R. Deve ser entregue o código fonte, comentado, em R. O prazo final para entrega da lista é o dia 28 de novembro, no horário da aula.

Criar um script que:

- a) Leia do Twitter mensagens que contenham a palavra Olimpíada a partir da data 17/09/2016 e que estejam em um raio de 60Km da latitude -22° 54' 10" (-22.90278) e longitude -43° 12' 27" (-43.2075). Deverá ser criado um dataframe com as mensagens lidas;
- b) A partir dos textos lidos no item a: realizar a limpeza removendo caracteres como “.-’:”, pontuação, números e URLs; converter os caracteres de todas as palavras para minúsculo; remover stopwords; remover espaços;
- c) Realizar o stemming dos textos;
- d) Criar e exibir a DTM. Apresentar os termos mais encontrados e as respectivas frequências;
- e) Apresentar um histograma (termo x frequência) para os termos que apareceram pelo menos 40 vezes;
- f) Apresentar os termos associados com o termo ouro segundo uma correlação de 60%;
- g) Apresentar um wordcloud com termos cuja frequência seja de pelo menos 20;
- h) Apresentar um wordcloud com cores;
- i) Apresentar o dendrogram do cluster hierárquico identificando os clusters;
- j) Usando o método k-means apresentar o cluster de termos;