



### INF-0613 – MINERAÇÃO DE DADOS

PROF<sup>A</sup>. SANDRA AVILA

sandra@ic.unicamp.br

### Projeto Prático

**Data de entrega:** 11 de maio de 2018 até 23h59.

O projeto deve ser desenvolvido **em dupla**.

## Objetivo

Efetuar mineração de chamadas de notícias. A partir do conjunto de dados fornecido, você deve utilizar métodos de redução de dimensionalidade e agrupamentos para encontrar grupos de notícias que possuam um tema comum.

## Descrição do conjunto de dados

As chamadas de notícia foram coletadas do site de notícias da *Australian Broadcasting Corporation* (ABC) e representam notícias publicadas do período de janeiro de 2012 a dezembro de 2017. Você terá acesso a dois arquivos no formato CSV com os dados: **headlines.csv** e **features.csv**.

O arquivo **headlines.csv** contém duas colunas:

- **publish\_date**: números inteiros representando a data de publicação da notícia, no formato **aaaammdd**.  
Exemplo: o número 20160113 representa o dia 13 de janeiro de 2016.
- **headline\_text**: o texto da chamada da notícia (somente caracteres ASCII) em minúsculo.  
Exemplo: `claims north korea faked missile test footage`

O arquivo **features.csv** contém diversas colunas. As colunas neste arquivo representam as *features* (ou características) extraídas dos textos das chamadas utilizando o modelo *bag-of-words* baseado em frequência de termos. As linhas nos dois arquivos são organizadas de tal forma que uma linha no arquivo **headlines.csv** e a mesma linha no arquivo **features.csv** representam a mesma notícia.

## Atividades

1. Carregue o arquivo **features.csv** e observe o número de dimensões dos dados. Considere o uso de PCA para redução de dimensionalidade. Com quantas componentes principais conseguimos preservar 85% da variância dos dados? E 90%? Para os itens seguintes, escolha entre preservar 85% e 90% da variância e utilize os dados apenas com tais componentes principais.
2. Efetue o agrupamento dos dados com o k-means e determine o número de *clusters* adequado.
  - (a) Faça isso comparando tanto o coeficiente de silhueta quanto o valor do erro quadrático do resultado obtido variando o *k* de 4 a 20.
  - (b) Como o uso de normalização (parâmetro **scale** do **prcomp**) antes de efetuar o PCA afeta os resultados?
  - (c) Explore duas variações do k-means. Por exemplo, k-medians, k-medoids e fuzzy c-means.
3. Analise os *clusters* calculando os bigramas<sup>1</sup> mais frequentes de cada *cluster*.

---

<sup>1</sup>Utilize a função **ngrams** do pacote **NLP** para isso.

- (a) Quais são os 3 bigramas mais frequentes de cada um?
  - (b) O que eles dizem sobre o tema das notícias dos seus *clusters*?
4. Utilizando os dados com a dimensionalidade reduzida, efetue a mesma análise do item 3 apenas para notícias de 2016.
- (a) O número de clusters utilizado é o mais adequado?
  - (b) Existem temas recorrentes que surgem tanto na análise com os dados completos quanto na análise deste ano isoladamente?

## Avaliação

Os projetos serão avaliados de acordo com o cumprimento das atividades aqui descritas. Para tal, cada dupla deve elaborar um relatório de 4 páginas descrevendo as atividades desenvolvidas e mostrando os resultados obtidos. Além disso, todo o código escrito durante o desenvolvimento do projeto deve ser entregue. Tanto o relatório quanto o código devem ser submetidos via *Moodle*. Não serão aceitas entregas por outros meios (Slack, e-mail, etc.).