



**Universidade Federal de Pernambuco  
Centro de Informática  
Graduação em Ciência da Computação**

# **Uma ferramenta para descoberta de notícias utilizando algoritmo K-Médias**

Proposta de Trabalho de Graduação

Aluno: **Lucas Vicente Tenório**

Orientador: **Flávia de Almeida Barros**

Recife  
Setembro de 2016

## **Sumário**

- [1. Contexto](#)
- [2. Objetivo](#)
- [3. Cronograma](#)
- [4. Referências Bibliográficas](#)
- [5. Assinaturas](#)

## 1. Contexto

Com o busca por notícias movendo cada vez mais para a web[1], o processo de descoberta de novas fontes de informação jornalística e pontos de vista é essencial.

O foco deste trabalho é em construir um serviço de leitura de notícias na web que colete e analise as notícias do dia para assim determinar quais os assuntos mais comentados, estimulando a descoberta de novos interesses.

## 2. Objetivo

Sendo o objetivo principal a detecção dos tópicos de interesse mais comentados, usaremos algoritmos de agrupamento para determinar quais artigos têm o mesmo tema geral.

Para alcançar este objetivo, serão necessários quatro módulos fundamentais para o projeto:

Um **coletor** que irá obter notícias de fontes pré estabelecidas assim que forem publicadas. A coleta será feita em sua totalidade utilizando o RSS<sup>1</sup> das fontes de notícia. Este módulo irá coletar no mínimo mil artigos por dia para garantir um bom resultado dos módulos seguintes. Este módulo já foi implementado e já conta com mais de 30.000 artigos coletados.

Um **extrator** que irá analisar os sites e extrair informações necessárias para o agrupamento, como título, autor, data e texto principal livre de HTML e ruídos. Alguns dados extras como vídeos e imagens relacionadas serão guardados para uma visualização mais rica. As funções principais deste módulo já foram implementadas porém novos ajustes serão necessários na medida em que novas fontes de artigos sejam adicionadas.

Um **selecionador** que irá separar as notícias em grupos por similaridade, assim encontrando as tendências do dia. Será utilizada uma versão modificada do algoritmo K-Médias[2] mas outros algoritmos também serão testados. As modificações do algoritmo de agrupamento serão feitas para eliminar a necessidade de um número fixo de grupos de artigos, dado que a quantidade de grupos ótima pode mudar dependendo do dia.

Um **visualizador** que irá expor os grupos de notícias mais relevantes. Este será o módulo mais simples do projeto, sua única função é mostrar para o usuário os melhores grupos de artigos que surgiram no dia. O design e implementação de um ambiente de leitura ótimo foge do escopo deste projeto.

---

<sup>1</sup> <https://cyber.harvard.edu/rss/rss.html>

### 3. Cronograma

| Atividade                                                       | Agosto |  | Setembro |  | Outubro |  | Novembro |  | Dezembro |  |
|-----------------------------------------------------------------|--------|--|----------|--|---------|--|----------|--|----------|--|
| Implementação do coletor de notícias.                           |        |  |          |  |         |  |          |  |          |  |
| Elaboração da proposta.                                         |        |  |          |  |         |  |          |  |          |  |
| Implementação e teste do algoritmo de agrupamento das notícias. |        |  |          |  |         |  |          |  |          |  |
| Implementação da ferramenta de visualização de notícias.        |        |  |          |  |         |  |          |  |          |  |
| Teste de sistema.                                               |        |  |          |  |         |  |          |  |          |  |
| Escrita do relatório final.                                     |        |  |          |  |         |  |          |  |          |  |
| Apresentação.                                                   |        |  |          |  |         |  |          |  |          |  |

#### 4. Referências Bibliográficas

- [1] Lenhart, Amanda; Purcell, Kristen; Smith, Aaron; Zickuhr, Kathryn (2010). Social Media & Mobile Internet Use among Teens and Young Adults. Millennials; Pew Internet & American Life Project
- [2] Liping Jing ; Michael K. Ng ; Joshua Zhexue Huang. An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data. IEEE Transactions on Knowledge and Data Engineering ( Volume: 19, Issue: 8, Aug. 2007 )

## 5. Assinaturas

---

Aluno: **Lucas Vicente Tenório**

---

Orientador: **Flávia de Almeida Barros**