Departamento de Sistemas e Computação – FURB
Curso de Ciência da Computação
Trabalho de Conclusão de Curso – 2013/2

Explorator: uma ferramenta para mineração de dados do Twitter

Acadêmico: Diego Santos Luiz
diego.santos.luiz@gmail.com

Orientador: Prof. Aurélio Hoppe
aurelio.hoppe@gmail.com

Grupo de Pesquisa em Computação Gráfica, Processamento de Imagens e Entretenimento Digital
http://www.inf.furb.br/gcg

Roteiro

- Motivação
- Trabalhos relacionados
- Trabalho proposto
- Requisitos
- Desenvolvimento
- Experimentos
- Conclusão
- Extensões
- Demonstração

Motivação

- Aumento no volume de dados, ultrapassa a capacidade humana de interpretação
- A procura por informações resumidas e relevantes
- Organizações procuram identificar a aceitação de seu produto e/ou serviço fornecido ao consumidor

Trabalhos relacionados

características / trabalhos relacionados	SANTOS L. (2010)	RAMOS E BRÄCHER (2009)	TORRES (2005)	SILVA (2010)
técnica de agrupamento	-	X	-	-
técnica de classificação	SVM	-	Naive Bayes	KnnFlex
ferramentas de auxílio	-	SAS	-	R/tm
remoção de stopwords	X	X	X	X
normalização	X	X	X	X
sinônimos	-	-	X	-

Trabalho proposto

Desenvolver uma ferramenta que fará a seleção, a classificação e a apresentação de informações extraídas do Twitter.

Objetivos:

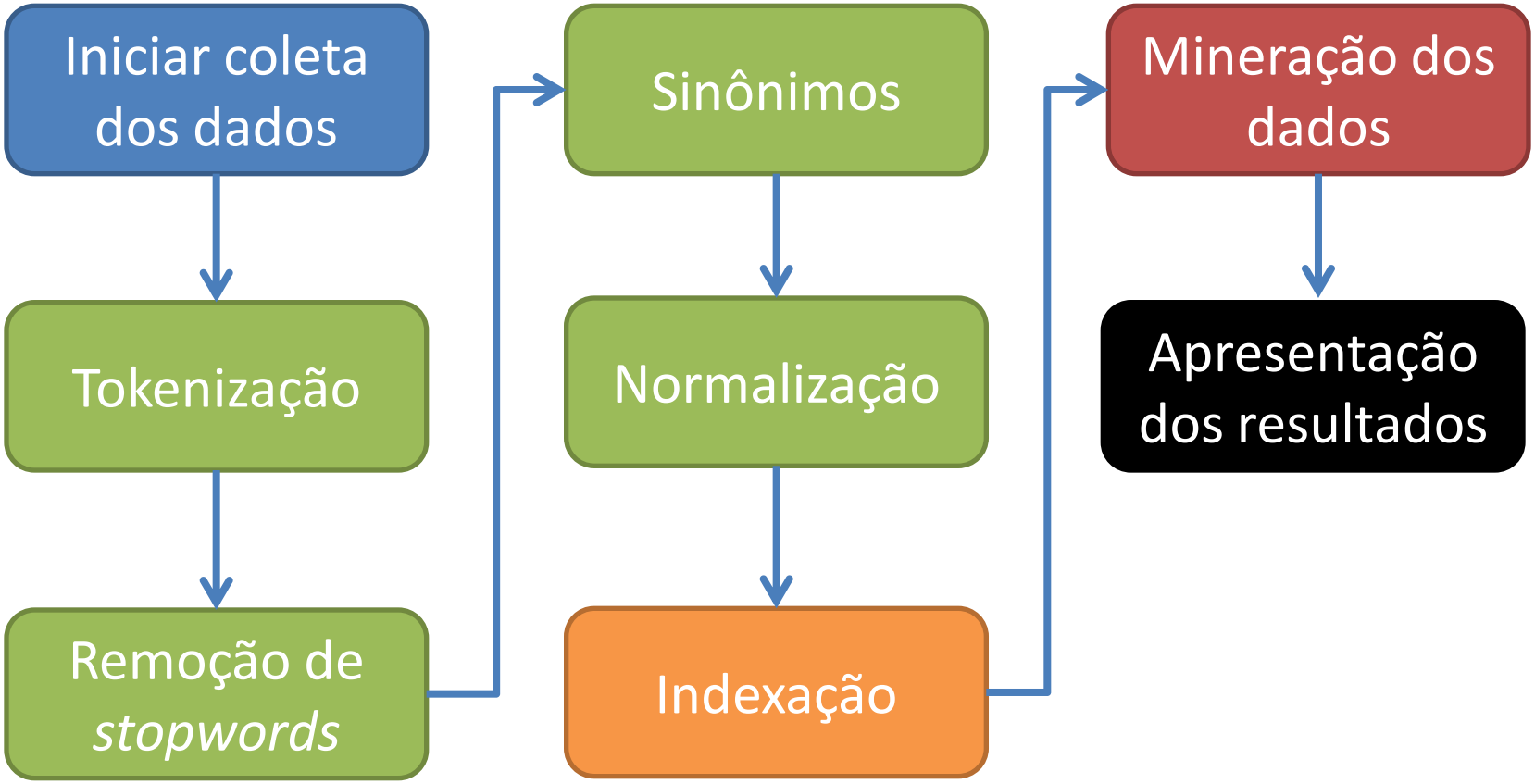
- Estabelecer conexão com a base de dados do Twitter;
- Utilizar técnicas de mineração de textos para filtrar *tweets* de acordo com os interesses do usuários;

Requisitos

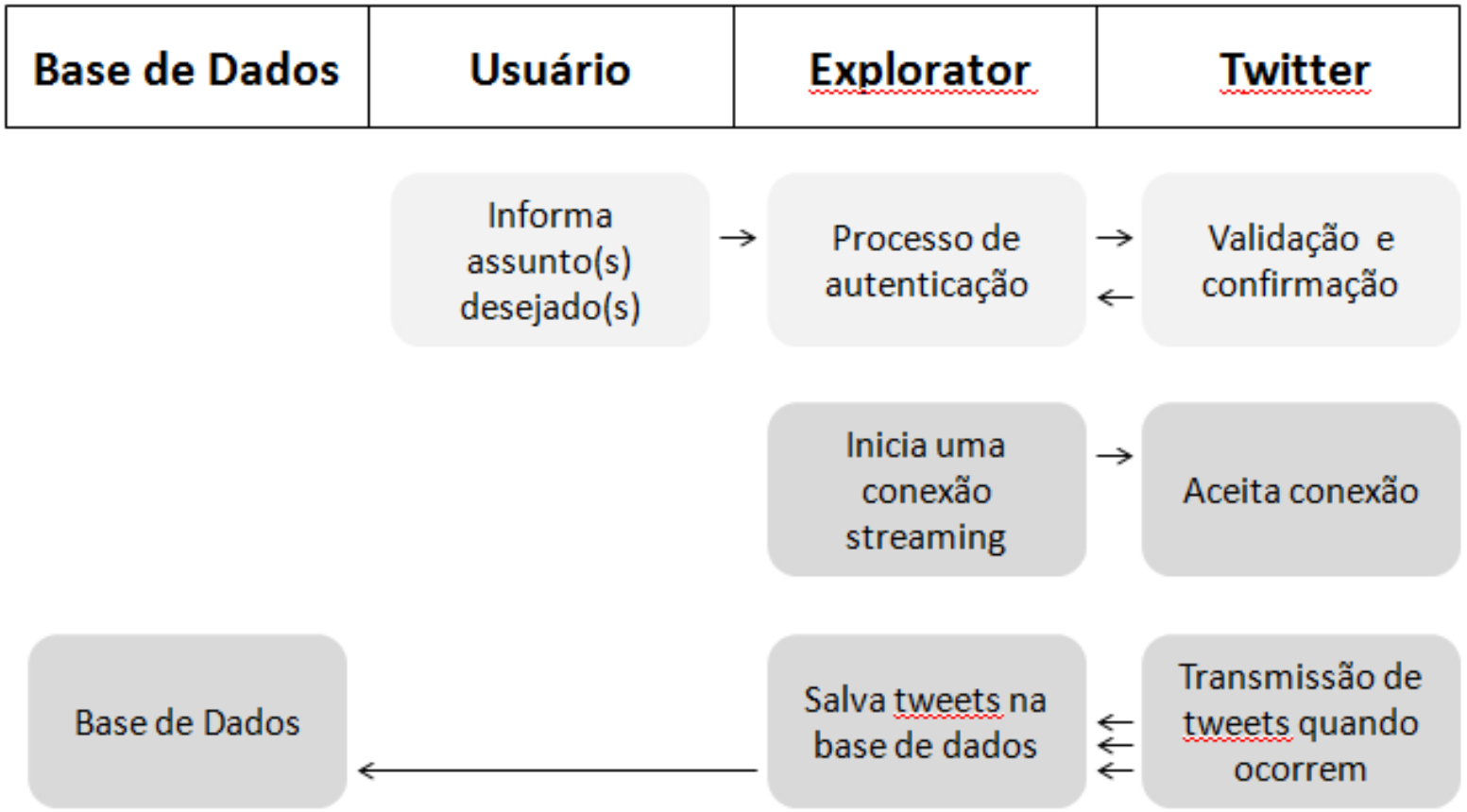
A seguir estão enumerados os requisitos funcionais do projeto:

- Implementar o processo de mineração de textos e as suas etapas, sendo elas: coleta, pré-processamento, indexação, mineração e análise (RF);
- Implementar o algoritmo de K-means (RF);

Desenvolvimento



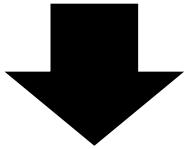
Coleta dos dados



Pré-processamento

Tokenização

O refrigerante coca-cola custa hj nos mercados R$ 3,50, mesmo sendo caro é mto boa. Mais informações no site http://www.cocacola.com.br/



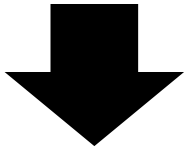
[O][refrigerante][coca][-][cola][custa][hj][nos][mercados][R$][3] [,][50] [,][mesmo][sendo][caro][é][mto][boa][.][Mais][informações][no][site] [http][:][/][/][www][.][cocacola][.][com][.][br]

Pré-processamento

Identificar abreviações

[O][refrigerante][coca][-][cola][custa][hj][nos][mercados][R$][3]

[,][50][,][mesmo][sendo][caro][é][mto][boa][.][Mais][informações][no] [site][http][:][/][/][www][.][cocacola][.][com][.][br]



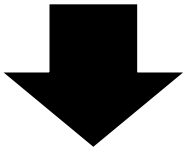
[O][refrigerante][coca][-][cola][custa][hoje][nos][mercados][R$][3][,]

[50][,][mesmo][sendo][caro][é][muito][boa][.][Mais][informações][no] [site][http][:][/][/][www][.][cocacola][.][com][.][br]

Pré-processamento

Identificação de palavras combinadas

[O][refrigerante][coca][-][cola][custa][hoje][nos][mercados][R$][3][,][50][,][mesmo][sendo][caro][é][muito][boa][.][Mais][informações][no] [site][http][:][/][/][www][.][cocacola][.][com][.][br]

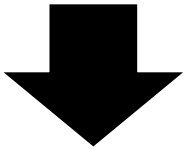


[O][refrigerante][coca-cola][custa][hoje][nos][mercados][R$][3][,][50][,][mesmo][sendo][caro][é][muito][boa][.][Mais][informações][no] [site][http][:][/][/][www][.][cocacola][.][com][.][br]

Pré-processamento

Identificação de símbolos de internet

[O][refrigerante][coca-cola][custa][hoje][nos][mercados][R$][3]
[,][50][,] [mesmo][sendo][caro][é][muito][boa][.][Mais][informações]
[no][site][http][:][/][/][www][.][cocacola][.][com][.][br]

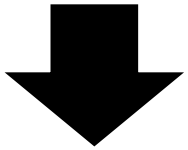


[O][refrigerante][coca-cola][custa][hoje][nos][mercados][R$][3]
[,][50][,] [mesmo][sendo][caro][é][muito][boa][.][Mais][informações]
[no][site][http://www.cocacola.com.br]

Pré-processamento

Identificação de números

[O][refrigerante][coca-cola][custa][hoje][nos][mercados][R$][3][,][50][,] [mesmo][sendo][caro][é][muito][boa][.][Mais][informações][no][site][http://www.cocacola.com.br]



[O][refrigerante][coca-cola][custa][hoje][nos][mercados][R$3,50][,][mesmo][sendo][caro][é][muito][boa][.][Mais][informações][no][site][http://www.cocacola.com.br]

Pré-processamento

Remoção de *stopwords*

Lista de stopwords	
artigos	o, os, a, as, um, uns, uma, umas
pronomes	eu, tu, eles, ele, elas, ela, nós, vós, meus, meu, minhas, minha, teu, teu, tuas, tua, seus, seu, suas, sua, nossos, nosso, nossas, nossa, vossos, vosso, vossas, vossa, deles, dele, delas, dela, isto, isso, aquilo, aquelas, aquela, aqueles, aquele, essas, essa, estas, esta, naquelas, naquela, algumas, alguma, alguns, algum, nenhum, nenhuma, todos, todo, todas, toda, muitos, muito, muitas, muita, poucos, pouco, poucas, pouca, tantos, tanto, tantas, tanta, certos, certo, certas, certa, vários, vário, várias, vária, outros, outro, outras, outra, quantos, quanto, quantas, quanta, tais, tal, quais, qual, quaisquer, qualquer, que, onde, quem, cujo, cuja, me, te, lhes, lhe, se, mim, ti, si

Pré-processamento

Sinônimos

A ferramenta possui um cadastro de sinônimos

Usuário informar sua coleção de palavras positivas e negativas

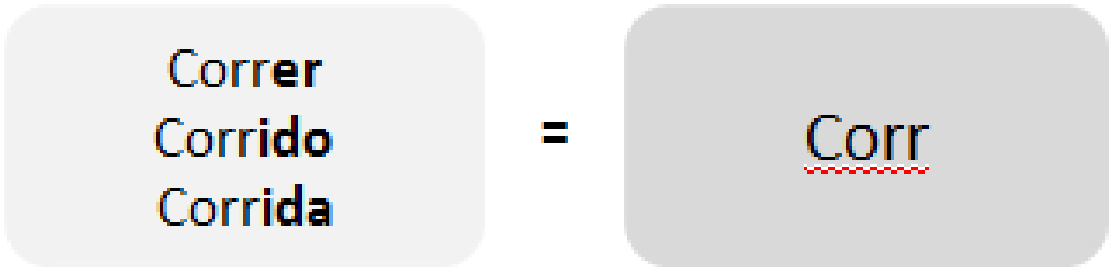
Verificado se as mesmas existem, caso existam é concatenado a coleção informado pelo usuário os sinônimos das palavras

Pré-processamento

Normalização

Utilizado o método de Porter

Busca o radical das palavras



Indexação

- A indexação tem o objetivo de identificar os tweets que precisam ser classificados
- Os tweets são agrupados em dois grupos: relevantes e não relevantes
- Por padrão os tweets são considerados como não relevantes
- Análisado cada tweet procurando identificar se o mesmo contém uma palavra positiva ou negativa;
- Encontrado uma ou mais ocorrências, o tweet é classificado como relevante

Mineração dos dados

- Na etapa de MD será realizado a classificação dos *tweets* em positivos, negativos e indefinidos
- Aplicado apenas para os tweets relevantes
- O processo de classificação soma as ocorrências das palavras positivas e negativas de cada *tweet*
- Se a quantidade for igual, o *tweet* é classificado como indefinido. Senão, o *tweet* é classificado levando em consideração a classe com o maior número de ocorrências

Demonstração da mineração de dados

- Tweets a serem classificados:
  - “Primeiros PlayStation 4 vendidos apresentam problemas no HDMI #gamegen”;
  - “PlayStation 4 apresenta gráficos lindos e perfeitos”;
  - “XBOX ONE teve seu lançamento e diversos problemas como PlayStation 4”;
  - “Veja o comando de voz do Playstation 4 em ação, ele é muito rápido”;
  - “PlayStation 4 apresenta o problema da tela azul”;
  - “Encontrada solução para problema no HDMI do PlayStation 4 - http://tinyurl.com/l4m9cjd #gamegen”;
- Palavras positivas: lindo, rápido, perfeito, solução;
- Palavras negativas: problema;

Tabela de demonstração

DEMONSTRATIVO DA CLASSICAÇÃO DOS *TWEETS*	
Tweets	**Classificação**
Primeiros PlayStation 4 vendidos apresentam problemas no HDMI #gamegen	Negativo
PlayStation 4 apresenta gráficos lindos e perfeitos	Positivo
XBOX ONE teve seu lançamento e diversos problemas como PlayStation 4	Negativo
Veja o comando de voz do Playstation 4 em ação, ele é muito rápido	Positivo
PlayStation 4 apresenta o problema da tela azul	Negativo
Encontrada solução para problema no HDMI do PlayStation 4 - http://tinyurl.com/l4m9cjd #gamegen	Indefinido

Experimentos

- Realizado três experimentos nas seguintes bases: PlayStation 4, Submarino e Thor
- Resultado da ferramenta foram comparados com os resultados alcançados/realizados de forma manual
- Analisou-se o texto do tweet para identificar se o autor expressava um opinião positiva ou negativa

Experimento 1 - Thor

Testes	Classificados manualmente			Explorator		
	Positivo	Negativo	Indef.	Positivo	Negativo	Indef.

Palavras positivas: ver, lindo, primeiro, assistir, obrigado, gostei, melhor, lidera, delicia, fofinho, massa, 1º lugar, largar, perfeito, boa, topo, preferidos, força, legal, foda e favoritos. **Palavras negativas:** causa, sofrendo, dramalhão, bicha, odeio, idiota, não iamos. **Buscar interior das palavras**: marcado

Teste 1	45	5	-	53	3	2

Utilizado as mesmas palavras. **Buscar interior das palavras**: desmarcado

Teste 2	45	5	-	40	4	1

Palavras positivas: assistimos, assistindo, assisti, liderança, lindinho, primeira, gosto e obrigada

Teste 3	45	5	-	50	3	2

Experimento 2 - Submarino

Testes	Classificados manualmente			Explorator		
	Positivo	Negativo	Indef.	Positivo	Negativo	Indef.

Palavras positivas: uhul, por apenas, felicidade, fiel, demais, baratinho, quero, promoção, fantástica, desconto, menor preço, interessante, oferta.
Palavras negativas: afundar, chorando, aguentar, não vão, somente, não vende, sofre, palhaçada, não mesmo, puta, não tem, amassada, quebrada, problema, desonestas. **Buscar interior das palavras**: marcado

Teste 1	38	15	-	38	12	4

Utilizado as mesmas palavras. **Buscar interior das palavras**: desmarcado

Teste 2	38	15	-	31	11	3

Palavras positivas: promoções, descontos, ofertas e queria

Teste 3	38	15	-	35	11	3

Experimento 3 - PlayStation 4

Testes	Classificados manualmente			Explorator		
	Positivo	Negativo	Indef.	Positivo	Negativo	Indef.

Palavras positivas: vende mais, gratuito, solução, impressiona, rápido, legal. **Palavras negativas:** não vai, mais caro, não descerá, reclamam, falha, preço absurdo, não deve, defeito, não funcionam, instabilidade, problemas, defeito, uma bosta, não terá, polêmica desnecessária, vish. **Buscar interior das palavras**: marcado

Teste 1	6	22	-	6	22	1

Utilizado as mesmas palavras. **Buscar interior das palavras**: desmarcado

Teste 2	6	22	-	7	20	-

Palavras negativas: problema, reclamando e não funcionam

Teste 3	6	22	-	6	22	1

Conclusão

- Resultados satisfatórios, com uma margem de acerto acima de 80%
- A classificação depende do conhecimento da base de dados

Extensões

- disponibilizar a coleta dos dados para outras redes sócias, exemplo: Facebook, Linkedin;
- adicionar novos recursos ao pré-processamento como, por exemplo: correção ortográfica, redução do léxico, detecção automática de sinônimos, nomes truncados e *parsing* (Análise sintática);
- utilizar outras técnicas de normalização, por exemplo: método de *Stemmer* S, método de *Lovins* e *Lemmatization;*

Extensões

- análise da base de dados antes de iniciar processo de MT, para identificar novas *stopwords*
- criar serviços que disponibilizam a etapa de pré-processamento, indexação e classificação. Assim o usuário pode utilizar somente a etapa que melhor lhe atende
- criar novas interfaces de acesso, por exemplo: Mobile, Desktop

Demonstração

Obrigado!