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

Artur Oliveira Rodrigues

MONOGRAFIA DE PROJETO ORIENTADO EM COMPUTAÇÃO II

Como nos sentimos: visualização de sentimentos

Belo Horizonte
2012 / 1º semestre

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

Como nos sentimos: visualização de sentimentos

por

Artur Oliveira Rodrigues

Monografia de Projeto Orientado em Computação II

Apresentado como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFMG

Prof[a]. Dr[a]. Raquel Cardoso de Melo-Minardi
Orientadora

Belo Horizonte
2012 / 1º semestre

Resumo

O principal objetivo deste trabalho é apresentar um sistema de coleta, mineração, armazenamento e visualização de sentimentos expressos em Língua Portuguesa em redes sociais online. Esse sistema se baseia na visualização *We Feel Fine* (Kamvar e Harris 2011), mas difere desta por utilizar redes mais atuais e ter como a Língua Portuguesa e a sociedade brasileira como fonte e alvo.

Utilizou-se o *Twitter* e sua Interface de Programação de Aplicações como rede social para a coleta das informações. Foi feita uma busca por textos contento fragmentos que caracterizam expressão de sentimentos, como "estou" , "me sinto" e "fico". Em seguida, esses textos recuperados foram submetidos a um processamento de linguagem natural feito com base em expressões regulares para identificação de verbetes com teor sentimental. Gírias, erros de ortografia e outras características do meio virtual também foram considerados. Esses textos foram então indexados em um banco de dados, acompanhados de informações temporais e autorais. Paralelamente é realizado o *geocoding* das localizações inseridas pelos autores dos textos a fim de armazenar essa informação de forma estruturada. Quando possível também foram anexadas informações climáticas da localização do usuário no momento da criação da mensagem. Por fim, o *front-end* de visualização é responsável por exibir essas informações armazenadas através de representações e metáforas visuais. O objetivo destas visualizações é atrair o interesse do usuário na exploração dos dados apresentados e facilitar sua percepção e cognição.

O sistema descrito fica, atualmente, em execução ininterrupta e o *front-end* de visualização foi disponibilizado na internet através de um sítio, permitindo que qualquer pessoa explore essas informações em tempo real.

Palavras-chave: Mineração de Dados, Banco de Dados, Visualização de Dados, Redes Sociais.

Abstract

The main objective of this work is to present a collection, mining, storage and visualization system of feelings expressed in the Portuguese Language in online social networks. This system is based on the visualization We Feel Fine (Kamvar e Harris 2011), but differs from it by using modern social networks and having the Portuguese Language and Brazilian society as its source and target.

Twitter and its Application Programming Interface were used as the social network for the data gathering. Texts containing the fragments that characterize feelings expression, like *"estou"*, *"eu to"* and *"me sentindo"* were searched. Subsequently, these recovered texts were submited to a natural language processing based on regular expressions in order to identify entries with emotional content. Slangs, misspellings and other common social expressions were also considered. Next, the texts selected from the previous step were indexed and stored in a relational database, coupled with authorial and temporal data. At the same time the geocoding of the user location input is done in order to store this information in a structured way. When possible, weather data related to the user location, at the time of the message, was also indexed. Finally, the visualization front-end is responsible for displaying this stored information through representations and visual metaphors. The objective of these visualizations is to draw the user's attention to the exploration of the presented data, as well as making its perception and cognition smoother.

The described system is currently being run uninterruptedly and the visualization front-end was made available in the internet through a website, allowing anyone to explore the information in real time.

Keywords: Data Mining, Databases, Data Visualization, social networks.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

API	Application Programming Interface ou Interface de Programação de Aplicações
RGB	Sistema de Cores Aditivas formado pelas cores Vermelho(Red), Verde(Green) e Azul(Blue)
PLN	Processador de Linguagem Natural
GPS	Global Positioning System ou Sistema de Posicionamento Global
JSON	JavaScript Object Notation ou Notação de Objetos JavaScript
AJAX	Asynchronous JavaScript and XML ou JavaScript e XML Assíncronos

Sumário

1 INTRODUÇÃO

1.1 Visão Geral

O crescimento das redes sociais online nos últimos anos, acompanhado pelo aumento no número de textos carregados de teor sentimental, levou ao aumento do interesse na análise de sentimentos (Kamvar e Harris 2011). Na maioria das vezes esses textos são analisados a partir da aglutinação em torno de um tópico, ou até mesmo uma marca, a fim de obter dados que reflitam a opinião pública sobre esse tópico, para então traçar estratégias de produção ou marketing, por exemplo.

A quantidade de informação gerada atualmente é sem precedentes. A exploração e análise destes vastos volumes de dados se torna, portanto, cada vez mais difícil. Assim, técnicas de visualização de informações apresentam uma alternativa interessante para a compreensão desses conjuntos de dados. Além disso, a exploração visual faz com que o usuário esteja diretamente envolvido no processo de mineração (Keim 2002).

1.2 Objetivo

O objetivo deste trabalho é o desenvolvimento de um sistema de coleta, mineração, armazenamento e visualização de sentimentos expressos em Língua Portuguesa em redes sociais online. Inspirada no sistema We Feel Fine (Kamvar e Harris 2011), são coletadas expressões do tipo “me sentindo”, “eu to” e “estou”, visando a obtenção de coleções de expressões de sentimento. Estas informações são associadas a informações descritivas dos usuários tal como localização geográfica, contribuindo para a segmentação dos usuários de forma extremamente rica e possibilitando a descoberta de padrões interessantes nos dados. Os dados e padrões são então visualizados por meio de representações e metáforas visuais. O objetivo destas visualizações é atrair o interesse do usuário na exploração dos dados apresentados e facilitar sua percepção e cognição. Isso permitirá com que ele generalize o conhecimento obtido gerando respostas para suas questões e mesmo novas questões a partir de suas descobertas.

Uma rede social é um conjunto de pessoas (ou organizações ou outras entidades sociais) conectadas por um conjunto de relações sociais, como amizade, trabalho mútuo e troca de informações (Garton, Haythornthwaite e Wellman 1997). Quando essa rede vai para a Web, são criados serviços que permitem a estruturação e manutenção destas redes, como o Facebook[1] e Twitter[2], e as redes criadas nesses serviços podem, portanto, serem chamadas de redes sociais online (Ackland 2008). Nessa primeira parte do projeto foram explorados os conteúdos em língua portuguesa gerados na rede Twitter.

Essa ferramenta de visualização está inserida dentro de um sistema maior, composto por vários componentes, responsável por vasculhar redes sociais em busca de textos com teor sentimental, recuperá-los, processá-los e aplicar rótulos significativos, armazenar os textos e seus rótulos em um banco de dados e por fim exibi-los através da aplicação de visualização.

Finalmente, o sistema descrito fica, atualmente, em execução ininterrupta e o *front-end* de visualização foi disponibilizado na internet através de um sítio, permitindo que qualquer pessoa explore essas informações em tempo real.

[1] http://www.facebook.com
[2] http://www.twitter.com

2 REFERENCIAL TEÓRICO

Como explicitado na seção 1.2, para que a aplicação de visualização seja possível, foi necessário desenvolver um sistema maior composto de ferramentas que desempenham papeis de recuperação de dados da Web, indexação de contexto, processamento de linguagem natural, armazenamento em banco de dados, mineração e visualização de dados.

A parte de recuperação de dados da Web foi feita através do uso das API's da rede fonte, o que tornou a tarefa mais simples do que a coleta de dados através de Crawlers como discutido em (Thelwall 2001), (Chakrabarti 2002) e (Kleinberg 2007). A indexação do contexto foi feita, quando possível, através das APIs, seguida pelo processo conhecido como *geocoding*, descrito em (Inc. 2011). Quando possível também foi indexada a condição climática da localização do autor da mensagem, no momento da sua composição. Para essas tarefas foi utilizada a linguagem de programação Python[1].

O processamento dos textos recuperados através do passo anterior envolve conceitos de processamento de linguagem natural. Esse tipo de processamento em Língua Portuguesa está atualmente em fase inicial e a literatura existente é escassa em teoria e ferramentas. A maioria das soluções razoáveis disponíveis são proprietárias. Dessa maneira foi desenvolvida uma análise crua dos textos, utilizando expressões regulares, mas que atendeu aos propósitos e requisitos do projeto. Aqui também foi utilizada a linguagem Python. Alguns dos esforços existentes são discutidos em (Bonfante e Nunes 2002) e (Santos 2005).

Quanto ao armazenamento dos dados processados, este foi feito utilizando um modelo não relacional e uma solução aberta consolidada no mercado, MongoDB, como descrito em (Chodorow e Dirolf 2010).

Finalmente, a visualização é inspirada na aplicação We Feel Fine[2], apresentado em (Kamvar e Harris 2011). Porém, o projeto aqui proposto se diferencia ao utilizar como fontes de textos redes atuais, e a Língua Portuguesa como idioma dos textos. Além disso, a visualização descrita procura levar em consideração princípios de percepção visual e cognição

[1] http://www.python.org
[2] http://www.wefeelfine.org

(Few 2009, Tufte 1991, Tufte e Robins 1997, Cleveland 1993) na proposta de melhorias para as representações existentes e na proposta de novas técnicas. Pretendeu-se estudar e desenvolver técnicas de visualização interativas que permitam aos usuários a exploração dos dados apresentados e a descoberta de padrões interessantes. Utilizou-se as linguagens de programação Processing[3] e ProcessingJS[4] no desenvolvimento destas visualizações e posteriormente será avaliado o seu uso de forma qualitativa.

[3]http://www.processing.org
[4]http://www.processingjs.org

3 METODOLOGIA

3.1 Tipo de Pesquisa

O projeto apresentado pode ser classificado, do ponto de vista de sua natureza, como pesquisa aplicada, pois objetiva gerar conhecimentos úteis através da aplicação prática de técnicas e teorias. Do ponto de vista da forma de abordagem o projeto é classificado como quantitativo e qualitativo pois além de procurar traduzir em números os resultados e exigir técnicas estatísticas, ele procura explicitar situações e comparar fenômenos através de um ambiente real para coleta de dados, tendo como foco o processo e seu significado. Com base em seus objetivos, o projeto apresentado é exploratório por explicitar o problema com vistas a construir hipóteses.

3.2 Procedimentos metodológicos

Após estabelecido o referencial teórico, o procedimento metodológico está descrito nas etapas abaixo. Por se tratar de um sistema complexo, o desenvolvimento de algumas das atividades infracitadas ocorreram paralelamente.

1. **Revisão Bibliográfica**. Pesquisa na literatura e portais Web por algoritmos e técnicas consolidados e eficientes passíveis de serem aplicados no desenvolvimento do trabalho.

2. **Implementação do Crawler**. Como apresentado nas seções 1.2 e 2, a recuperação de dados das redes sociais online foi realizada através da utilização das APIs de cada serviço. Os textos recuperados são colocados em uma fila para serem processados pelas partes restantes do sistema.

3. **Implementação do Processador de Linguagem Natural**. Este processador consome os textos colocados na fila citada no passo anterior, identificando os sentimentos presentes no texto através de expressões regulares. O resultado do processamento é armazenado no banco de dados.

4. **Implementação do Geocoder**. Este componente também consome a fila citada anteriormente, buscando estruturar a localização dos autores das mensagens. O resultado, novamente, é inserido no banco de dados.

5. **Implementação do Indexador Climático**. Quando possível inferir as cidades dos autores das mensagens, são recuperadas as condições climáticas para essas localizações. Essa informação também é inserida no banco de dados.

6. **Desenvolvimento do Modelo para o Banco de Dados**. Responsável por armazenar de maneira eficiente as informações processadas no passo anterior, bem como facilitar a recuperação destas pela aplicação de visualização.

7. **Desenvolvimento do *Analytics***. Módulo responsável pela mineração de dados adicional das informações presentes no banco de dados.

8. **Desenvolvimento da Aplicação de Visualização**. Plataforma responsável por exibir de maneira interativa os dados processados nas etapas anteriores.

9. **Desenvolvimento do Sítio Online**. Esse espaço foi desenvolvido para que a ferramenta de visualização fosse disponibilizada para o público em geral.

10. **Avaliação dos Resultados**. Análises quantitativas foram utilizadas através do emprego de métricas nos motores de indexação do sistema. Todavia a avaliação mais relevante do projeto será sobre o estudo qualitativo dos resultados finais observados na aplicação de visualização.

3.3 Considerações de Projeto

O objetivo do sistema e da ferramenta de visualização descrita neste documento é de possibilitar a exploração tanto estatística quanto qualitativa do sentimento das pessoas. Acredita-se que a análise estatística irá possibilitar o entendimento cognitivo de padrões emocionais.

Espera-se que por ter sido criado um meio imersivo de interagir com os dados individualmente, tenha-se afetado diretamente a maneira como os usuários finais da aplicação entendam a si próprios e outros. Através dos dados coletados, é possível, por exemplo, descobrir como as pessoas de Minas Gerais se sentiram quando o Cruzeiro Esporte Clube assumiu a liderança do Campeonato Brasileiro ou até mesmo como toda a população do Brasil se sentiu com a aprovação do casamento gay. Dessa maneira, a exploração qualitativa e estatística dos dados que são pessoais, de um grande conjunto de usuários, permite a análise de dados através de *crowdsourcing*. Os seguintes princípios foram levados em conta no projeto:

1. **Análise de textos curtos.** Quando tratamos de meios sociais online os átomos básicos são pequenos textos, sejam eles *tweets*, atualizações de status no Facebook ou *posts* em *blogs*, ao contrário de outros meios da internet, quando documentos ou páginas inteiras são relevantes, como nas máquinas de buscas. Dessa maneira, focou-se nos *tweets* no projeto em questão, como as fontes da análise.

2. **Indexação de contexto**. Através do registro temporal e dos *profiles* públicos, quando disponíveis, é possível enriquecer ainda mais a análise do que somente através dos sentimentos expressos através do texto em si. No caso do Twitter, a localização dos usuários, autores das mensagens, juntamente com a condição climática dessa localização possibilitam análises ricas.

3. **Sentimentos como princípio de organização primário**. A maioria dos trabalhos e ferramentas que exploram a análise de sentimentos o fazem buscando informações acerca de um tópico, tornando este o centro da análise. No trabalho descrito os sentimentos estão no centro da organização. A análise aqui feita vai além da qualificação positiva ou negativa.

4. **Ausência de classificação por rankings**. Sentimentos ao contrário de informações factuais que podem ser classificadas de acordo com a relevância da fonte, são difíceis de serem priorizados entre si. Assim, não foi inserida a ideia de ranking nesse projeto.

5. **Ênfase no resumo e na usabilidade**. Ao invés de priorizar os sentimentos e textos entre si, preferiu-se oferecer uma ferramenta de visualização que resume os textos com sentimentos, permitindo que seus usuários tenham uma visão geral de como a população do Brasil se sente, através de uma interface de fácil utilização.

6. **Visualizações que refletem os dados**. A interface deve ser funcional, mas por tratar de sentimentos também deve ser lúdica. Sempre que possível os dados – textos com teor sentimental – devem representar seus autores.

7. **Modularização do sistema para acesso direto aos dados**. Visando permitir a abordagem quantitativa desse documento e de futuros utilizadores do sistema aqui descrito, os dados devem estar separados da visualização em si, para que esses possam ser estudados e trabalhados livremente, independente da ferramenta construída pelo autor.

4 RESULTADOS E DISCUSSÃO

4.1 Arquitetura

Como descrito em 3.2, a arquitetura do sistema é composta por diversos componentes que tornam possível a ferramenta de visualização descrita nesse documento: o *Crawler*, o *Processador de Linguagem Natural*, o *Geocoder*, o *Indexador Climático*, o *Banco de Dados*, o *Analytics* e por fim a *Visualização* e *Sítio Online*. Uma visão geral da arquitetura é apresentada abaixo, em seguida são detalhados seus componentes.

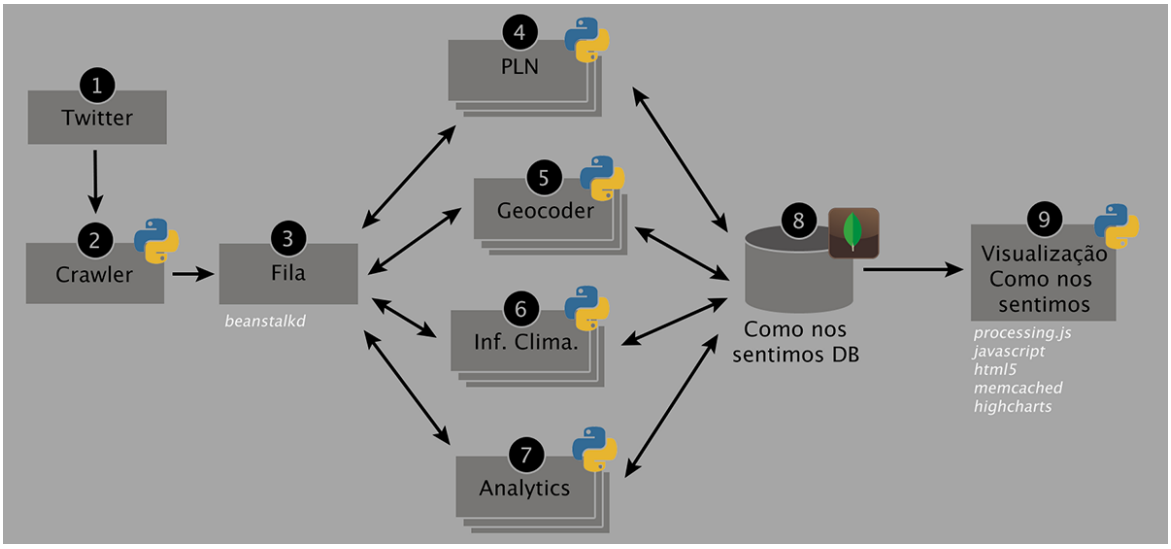


Figura 1: Visão geral da Arquitetura

4.1.1 Crawler

O Crawler Simplificado (componente 2 na figura 1) é responsável pela coleta dos textos relevantes do Twitter (componente 2 na figura 1) através de sua *Public Stream API*[1] e adicioná-los à fila (componente 3 na figura 1). Para o desenvolvimento desse Crawler foi utilizada a linguagem Python e para a comunicação com a API do Twitter utilizou-se a biblioteca Tweepy[2].

[1] https://dev.twitter.com/docs/streaming-apis/streams/public
[2] https://github.com/tweepy/tweepy/

A *Public Stream API* mencionada é uma interface que recebe como parâmetro termos a serem buscados, fornecendo continuamente os textos que possuem esses termos, a medida que são produzidos, em tempo real. Para essa coleta dos *tweets*, são passadas como parâmetros de busca as *strings* “eu to”, “me sentindo”, “me sinto”, “estou”, “fiquei”, “fico” e “ficando”. A escolha dessas *strings* foi feita após um estudo sobre as melhores opções de fragmentos que geralmente estão presentes em sentenças e textos com teor sentimental. A própria API do Twitter se encarrega de desconsiderar acentos e diferenciar maiúsculas e minúsculas.

No sistema atual, o *crawler* permanece em execução ininterruptamente, permitindo que a totalidade dos textos seja recuperada e processada. Ele também é responsável por recuperar informações dos autores dos textos, como apelido, nome e localização.

4.1.2 Fila

A Fila (componente 3 na figura 1) é responsável por armazenar tarefas que serão consumidas por trabalhadores, que, na figura, são os componentes 4, 5, 6 e 7. Ela possui diversas sub-filas, chamadas “tubos” que dizem respeito a cada um desses componentes. Trabalhadores ficam constantemente “escutando” um desses tubos por tarefas. Quando há uma nova tarefa, um trabalhador consome-a, primeiramente “reservando” a tarefa e, em seguida, excluindo-a.

A fila utilizada chama-se Beanstalkd[3] e para a comunicação entre ela e os trabalhadores utilizou-se a biblioteca Python Beanstalkc[4].

4.1.3 Processador de Linguagem Natural

O Processador de Linguagem Natural (componente 4 na figura 1) é um trabalhador que consome os *tweets* presentes na Fila, lá colocados pelo *Crawler*. Para cada um dos textos das mensagens, ele irá tentar corresponder uma das expressões regulares que identificam sentimentos. Essa lista de expressões regulares foi compilada manualmente e é armazenada num arquivo texto que é carregado pelo processador. A relação das expressões regulares e os sentimentos a que elas correspondem é encontrada na tabela 1.

Juntamente com a as expressões regulares e os sentimentos, é relacionado um valor RGB que será utilizado para preencher a cor das entidades na visualização.

Essas expressões regulares identificam sentimentos em somente 9% dos textos recebidos pelo *crawler*. Isso significa que a maioria dos textos retornados pelo Crawler não possuem teor sentimental identificável pelo PLN, ou, como experimentos mostraram, simplesmente não

[3] http://kr.github.com/beanstalkd/
[4] https://github.com/earl/beanstalkc

possuem teor sentimental. A frase “eu to em casa” será coletada pelo *crawler* mas não será identificada. Esse tipo de texto corresponde a grande parte dos 91% de textos não identificados.

Pode-se argumentar que parte dos textos não identificados possuem teor sentimental que não estão mapeados pelas expressões regulares. De fato isso acontece, entretanto, os sentimentos que geram maior volume de dados já estão mapeados.

Quando mais de um sentimento é identificado no texto, ambos são armazenados, mas somente um será utilizado na visualização, uma vez que múltiplos sentimentos são conflitantes com as considerações de projeto definidas em 3.3 para a ferramenta visual.

É importante ressaltar que as expressões regulares foram feitas com a intenção de identificar gírias, erros ortográficos e construções típicas do ambiente social virtual, que são igualmente relevantes para o estudo aqui realizado. Dessa maneira, a expressão regular para “feliz” identificará tanto “feliz”, quanto “felis” e até mesmo “feeeeeeellliiiis”. Além disso foram desconsiderados, quando possível, os sufixos, permitindo que tanto “morta”, “morto” e “mortinha” fossem mapeados para o sentimento “morto”.

Um exemplo de frase identificado pelo PLN é “**Estou** tão **cançaadinha** hoje”, que possui tanto erros ortográficos quanto excesso de vogais.

Sentimento	**RE**	**RGB**
abençoado	a+be+n(s+|sc|ç|c)o+a+d	24,116,41
alegre	a+l+e+gr	255,164,1
amado	a+ma+d	243,1,114
apaixonado	a*pa+i+(x+|ch)o+na+d	115,32,124
cansado	(c|k)a+n(s+|ç|sc|c)a+d	36,156,240
carente	(c|k)a+re+nt	1,72,114
chateado	(ch|x)at(i+|e)ad	61,74,123
confiante	(c|k)o+(n|m)fi+a+nt	228,146,2
confortável	(c|k)o+(n|m)fo+rta*(ála)+a*v	64,182,184
confuso	(c|k)o+(n|m)f+u+(s|z)	115,32,124
culpado	(k|c)u+l*pa+d	0,78,111
deprimido	de+pri+mi+d	40,49,82
disposto	d(e|i)+spo+st	254,153,1
doente	d(o|u)+e+nt	137,15,185
especial	e+spe+(c|s+|sc)i+a+l+	243,2,99
estranho	(e|i)+stra+nh	243,1,87
estressado	(e|i)+stre+(c|ç|s+)a+n*d	146,86,160
estúpido	(i|e)stu*(u|ú)+pi+d	2,121,117
feliz	fe+l+i+(s|z)	255,231,13
fome	(fo+me+)|(fa+mi+nt)	247,32,65
forte	fo+rt	255,178,0
fraco	(en)0,1fra+(qu|k|c)	243,1,114
frio	f+r+i+o+	127,181,183
horrível	h?o+r+i*(i|í)+ve+	228,1,10
livre	l+i+vre+	126,195,27
louco	lo+u+(c|k)	243,2,109
machucado	ma+(x+|ch)u+(c|k)a+d	35,92,147
magoado	ma+g(o|u)+a+d	35,92,147
melhor	me+lho+r	141,186,50
morto	mo+rt(a|o|i)+	200,0,39
nervoso	ne+(r|l)vo+(s|z+)	255,176,3
ofendido	o+fe+ndi+d	8,72,146
orgulhoso	o+rgu+lho+(s|z)	255,179,0
ótimo	o*(o|ó)+t(i+|e+)m	255,216,1
perdido	pe+rd+id	1,72,114
preguiçoso	pre+gu*i+(sc|ç|s|c)	78,95,193
puto	pu+t(a|o)+	89,88,132
realizado	re+a+li+(s|z)a+d	254,165,63
relaxado	re+la+(ch|x)a+d	254,215,60
revoltado	re+vo+(o|l)+t	200,0,39
saudade	sa+u+da+d	1,91,92
solitário	so+li+ta*(a+|á)+ri+	8,70,143
sono	so+no+	113,62,243
sortudo	so+rtu+d	254,215,60
sozinho	so+(z+|s)i+nh	150,173,198
tenso	te+n(c|ç|sc|s)	0,115,78
triste	tri+st	4,63,105
velho	ve+lh(a|o)+	53,44,101
vivo	vi+v(a|o)+	100,184,34

Tabela 1: Tabela das 49 expressões regulares dos sentimentos

4.1.4 Geocoder

Geocoding é o processo de conversão de endereços (como Avenida Presidente Antônio Carlos, 6627 - Pampulha - Belo Horizonte - Minas Gerais - Brasil) em coordenadas geográficas (como latitude -19.865044 e longitude -43.957453) (Inc. 2011). No caso desse projeto ela foi explorada de outra maneira. Usuários to Twitter podem colocar a informação a respeito de suas localizações livremente através de um campo de texto. Isso significa que não há nenhuma estruturação que sustente essa informação.

Um dos requisitos desse projeto é que essa informação esteja estruturada, permitindo assim a estratificação localizacional dos usuários. Assim, esse requisito foi atendido através da API pública do Google Maps (Inc. 2011). Isso permitiu que localizações como "Sampa" ou "Beaga" fossem corretamente traduzidas e estruturadas para "São Paulo - SP" e "Belo Horizonte - MG", de maneira que tanto usuários que definam "Beaga" como "BH" sejam categorizados como usuários do estado "Minas Gerais".

No que tange os resultados obtidos com essa solução, 55% dos usuários preenchem o campo localização; 71% desses usuários têm sua localização estruturada dentro do sistema a nível de estado, no mínimo, resultando em 39% do total da base de dados.

Além disso, com o fim de diminuir o número de requisições realizadas a API do Google Maps - que são limitadas - foi criada uma *cache* que guarda o texto da localização e a informação retornada pela API do Google quando é buscada uma localização ainda não existente. Essa base de dados conta atualmente com mais de 160 mil registros, capazes de reduzir significativamente o número de requisições direcionadas ao nível 2.

O Twitter permite ainda que usuários insiram a sua geolocalização através de dados provenientes de aparelhos GPS. Quando esse tipo de dado está presente a nível de estado, ela é utilizada, tornado desnecessária a utilização do Geocoder.

4.1.5 Indexador Climático

Quando se tem a localização do usuário a nível de cidade, é buscada a condição climática para essa localização, no momento da composição da mensagem. Para essa busca, é utilizada a API Climática do Google, que não é documentada.

São armazenados a condição climática, como "chuviscando", "nublado" e "trovejando", e a temperatura da cidade em questão.

4.1.6 Banco de Dados

O Banco de Dados desempenha papel fundamental no projeto, armazenando todos os dados coletados e processados que serão utilizados pela ferramenta de visualização. Foi escolhido o banco de dados MongoDB[5], que é um banco de dados não relacional que armazena informações em documentos, nos moldes descritos em (Chodorow e Dirolf 2010). Esses documentos utilizam o formato JSON, que é o mesmo formato utilizado pela API do Twitter na disponibilização dos dados, e também facilmente utilizável pela linguagem na qual a visualização é escrita, tornando a tarefa de comunicação de dados entre as diversas partes do sistema uma tarefa mais fácil e natural.

Uma única coleção – que é análoga a uma tabela no modelo relacional – armazena cada *tweet* e as informação agregadas durante os processos descritos anteriormente. Quando essa coleção é lida para a recuperação dos dados para a visualização, são buscados os documentos mais recentes.

A seguir um exemplo de documento para essa coleção, que representa um *tweet* onde foi possível determinar a localização do usuário e a condição climática de sua cidade.

```
{
    "_id" : NumberLong("217462364927311872"),
    "author" : {
        "location" : "São Paulo",
        "screen_name" : "alo_jesus",
        "name" : "Mãe Jurema das Matas"
    },
    "created_at" : ISODate("2012-06-26T03:40:42Z"),
    "created_at_local" : ISODate("2012-06-26T00:40:42Z"),
    "feelings" : [
        "chateado"
    ],
    "feelings_size" : 1,
    "location" : {
        "city" : "Sao Paulo",
        "state" : "São Paulo",
        "weather" : {
            "temp" : 14,
            "condition" : "Cloudy"
        }
    },
    "text" : "e essa hora nao passa. to ficando chateada"
}
```

As outras duas coleções existentes dizem respeito a *cache* do Geocoder e aos dados do componente Analytics. A seguir um exemplo de documento da *cache*.

[5]http://www.mongodb.org

```
{
    "_id" : ObjectId("4fcd954095ec850a7c000000"),
    "city" : "Guarulhos",
    "state" : "São Paulo",
    "user_location" : "sp-guarulhos"
}
```

4.1.7 Analytics

Com o intuito de se obter análises e conclusões mais completas e com caráter quantitativo, foi desenvolvido um componente que armazena estatísticas acerca dos *tweets* coletados. Essas estatísticas giram em torno de três variáveis: sentimentos, estados federativos e condições climáticas. Esse componente também opera no modelo de um trabalhador, e consome os *tweets* coletados pelo Crawler e inseridos na Fila.

O primeiro armazenamento é histórico e, portanto, pretende gravar um registro temporal de como cada sentimento se comportou ao longo do tempo. A contabilização é feita para três diferentes parâmetros: horário, diário e mensal. Para cada um desses parâmetros é armazenado o número total de sentimentos identificados, a quantidade total para cada sentimento e também a quantidade para cada estado e condição climática para cada sentimento. Um exemplo reduzido de documento armazenado para as 13 horas do dia 30 de junho de 2012 é apresentado abaixo:

```
{
    "_id" : ObjectId("4fef227fbb7641f9775154b6"),
    "type" : "hourly",
    "year" : 2012
    "month" : 6,
    "day" : 30,
    "hour" : 13,
    "count" : 6058,
    "feelings" : {
        {...}
        "ótimo" : {
            "count" : 98,
            "states" : {
                "al" : 1,
                "ba" : 1,
                (...)
                "pe" : 1,
                "rj" : 4,
                "rs" : 5,
                "sp" : 13
            },
            "weather" : {
                "Clear" : 12,
                "Cloudy" : 5,
                "Mostly Cloudy" : 3,
```

```
                "Partly Cloudy" : 6
            }
        }
    }
}
```

Com esse tipo de informação é possível, por exemplo, descobrir com base em uma data, qual foi o sentimento predominante da população brasileira. É possível também fazer um acompanhamento de como um determinado sentimento se faz presente no decorrer de um ano ou mês. Perguntas do tipo “as pessoas tem se sentido mais felizes nos últimos 10 dias?” e “em qual estado as pessoas se sentiram mais tristes com a notícia da internação do ex-presidente Lula?”, podem ser facilmente respondidas.

O segundo armazenamento é absoluto, sendo portanto desconsiderada a noção cronológica dos sentimentos. Desta vez são considerados cinco parâmetros para contabilização: horário, diário, mensal, dia da semana e total. A natureza das informações armazenadas é semelhante ao do armazenamento anterior. Um exemplo reduzido de documento armazenado para os sábados é apresentado abaixo:

```
{
    "_id" : ObjectId("4fee6bafbb7641f9775154a8"),
    "count" : 107358,
    "feelings" : {
        (...),
        "cansado" : {
            "count" : 1420,
            "states" : {
                "ac" : 2,
                "al" : 4,
                "am" : 7,
                (...)
                "sp" : 134,
                "to" : 1
            },
            "weather" : {
                "Clear" : 221,
                (...)
                "Rain" : 7,
                "Thunderstorm" : 6
            }
        }
    },
    "type" : "weekday",
    "weekday" : 6
}
```

Esse tipo de armazenamento permite descobrir, por exemplo, qual o sentimento predominante em determinado dia da semana. É possível também analisar como um sentimento

se comporta no decorrer das horas do dia. Perguntas do tipo “em qual hora do dia as pessoas se sentem mais cansadas no estado de Minas Gerais?”, “as pessoas se sentem mais amadas em junho em comparação com o resto do ano?” e “qual sentimento é mais expresso no Twitter?” podem ser facilmente respondidas.

É importante ressaltar que todos esses dados poderiam ter sido obtidos através de técnicas de mineração de dados na coleção de *tweets* armazenados. Todavia, optou-se por fazer esse pré-processamento, que resume e facilita a obtenção dessas estatísticas, pois elimina-se tarefas pesadas durante a execução da visualização pelos usuários. Esse ganho de desempenho é significativo se considerarmos que são coletados 120 mil novos *tweets* todos os dias.

4.1.8 Visualização

A visualização opera como *Front-End* para dados coletados e trabalhados pelo restante do sistema. Visando sua disponibilização pública na internet, suas diversas partes foram inseridas em um sítio virtual. Para a implementação da visualização foram utilizadas as linguagens JavaScript, Processing e ProcessingJS. Maiores detalhes são dados na seção 4.2.

4.2 Interface da Visualização

A ferramenta de visualização é uma aplicação composta por diversos elementos que procura encorajar a exploração de dados no nível atômico das mensagens. A primeira parte da análise aqui feita diz respeito ao elemento principal da aplicação. Destarte, cada *tweet*, o objeto indivisível da análise desse projeto, é sempre representado por uma partícula na visualização. Essa partícula irá compor e ajudar a produzir cada uma das visualizações presentes na ferramenta, através de movimentos e cores que despertem o interesse do usuário, tornando mais fácil a cognição e interpretação dos dados que elas representam.

A visualização principal é apresentada como duas telas com dimensões fixas. Na tela maior e inferior são inseridas as partículas que representam as mensagens mais recentes coletadas do Twitter. Cada partícula possui os atributos aceleração, velocidade e posição, cada um deles representados por um vetor bidimensional. A posição é diretamente influenciada pela velocidade e esta pela aceleração. É feito um controle no sentido de impedir que as partículas abandonem a tela da ferramenta ou assumam velocidades excessivas. As partículas possuem cores que estão relacionadas aos sentimentos que representam, previamente definidas. Quando uma das partículas é clicada as demais partículas ao redor dela sofrem uma repulsão, dando o *feedback* do clique pro usuário.

Adicionalmente, a ferramenta de visualização principal é composta por quatro diferentes visualizações - chamadas Anarquia, Sentimentos, Estados e Mapa. Cada uma dessas visualizações procura oferecer um ponto de vista diferente acerca dos dados, algumas vezes dando uma vaga análise estatística.

A transição entre as visualizações é feita através de botões localizado abaixo da tela. Sendo mais específico, eles são responsáveis por alterar os estados da ferramenta, como se essa possuísse internamente uma máquina de estados.

4.2.1 Anarquia

Nessa visualização, cada partícula se movimenta livremente na tela. Do ponto de vista da implementação, isso é feito gerando valores aleatórios para o vetor aceleração. Assim, cada partícula se movimenta sem um sentido específico, daí o nome "Anarquia". A intenção dessa visualização é introduzir o usuário à ferramenta através de uma exibição visualmente interessante e lúdica, que por trás, representa exatamente a heterogeneidade dos sentimentos, das mensagens e dos autores que inundam o meio virtual com suas histórias e sentimentos.

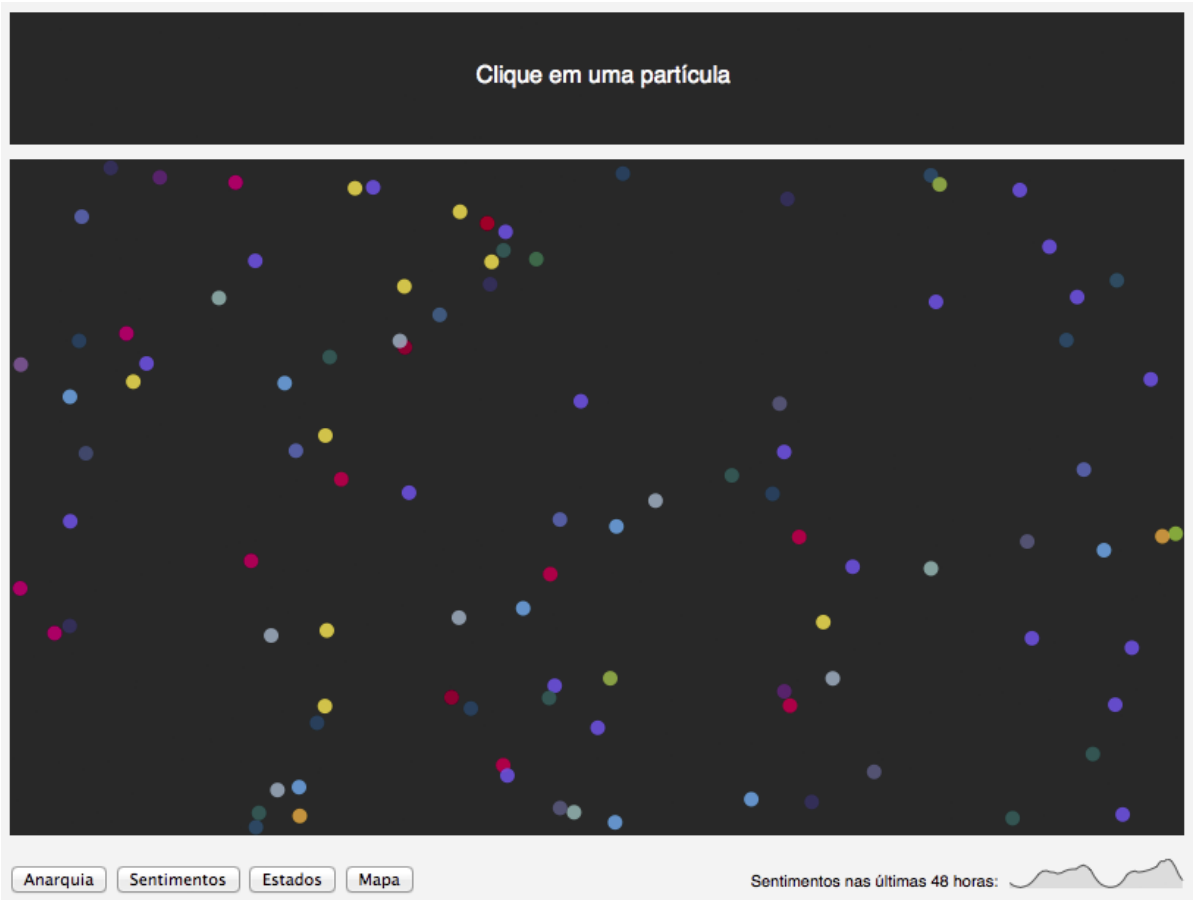


Figura 2: Visualização Anarquia

Caso o usuário passe o mouse em alguma das partículas, ela tem seu raio aumentado, e é exibido ao seu lado um texto discreto que identifica o sentimento. Sendo assim, o usuário possui um recurso simples e objetivo de navegar pelas partículas que estão sendo exibidas. Na ocasião do clique, é exibido na tela superior, o texto no qual o sentimento foi identificado, a data e horário da composição da mensagem, o apelido e nome do autor, sua localização de forma estruturada e se presente, a condição climática. Caso sua localização não tenha sido identificada no processo de *Geocoding*, é exibido "em algum lugar". Essa funcionalidade é exibida na figura 3.

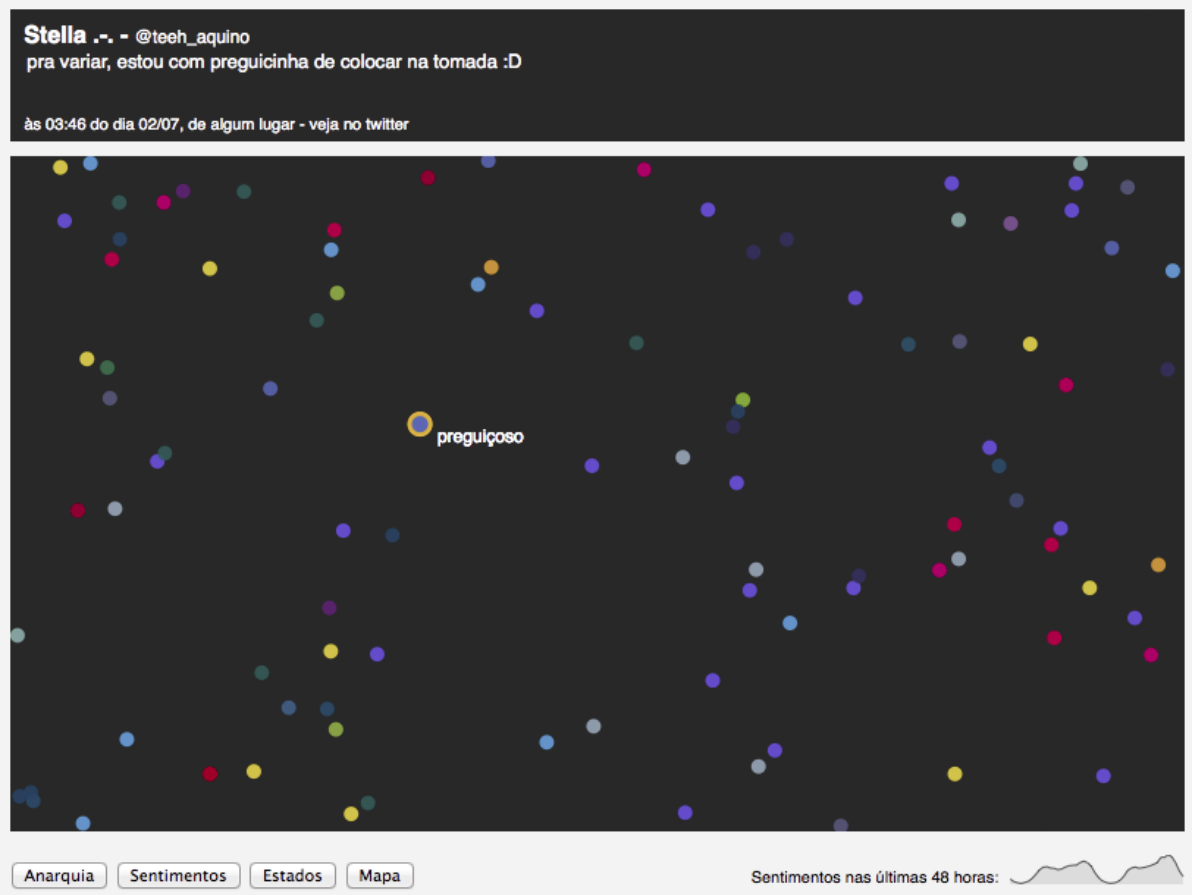


Figura 3: Visualização Anarquia - Exibição do texto

4.2.2 Sentimentos

Essa visualização procura inserir uma noção estatística à visualização. É apresentado um histograma dos sentimentos, informando assim a frequência de cada um deles dentre a totalidade das partículas exibidas. Assim, as partículas se movimentam no sentido de formar os retângulos do histograma.

Cada um dos sentimentos é identificado por um texto, seguido do número absoluto. Na ocasião de existirem mais partículas para aquele sentimento do que o compartimento de uma

sequência, é montada uma nova linha. Da mesma maneira, caso existam mais sentimentos do que o comporta somente uma coluna, é montada uma nova coluna com os sentimentos menos frequentes como visto na figura 4.

Figura 4: Visualização Sentimentos

4.2.3 Estados

A Visualização Estados é semelhante à visualização Sentimentos, no sentido que é apresentado um histograma, entretanto, dessa vez explicitando a frequência de cada uma das unidades federativas do Brasil. Através dessa visualização, o usuário tem uma ideia clara de qual estado produz maior número de textos. Partículas cujos autores que não tiveram sua localização identificada se direcionam para o canto inferior direito formando um ponto de interrogação. Novamente, a unidade indivisível da ferramenta, os textos, se combinam com o intuito de transmitir uma mensagem visual.

Figura 5: Visualização Estados

4.2.4 Mapa

Nesta visualização é apresentado um mapa do Brasil, e as partículas se dirigem para seu estado de origem novamente explicitando a frequência de cada uma das unidades federativas, dando uma nova visualização para os mesmos dados da visualização Estados. Diferentemente da última, essa possibilita a visão geral de como a internet é utilizada pelo brasileiro: quais são as regiões do país que mais utilizam o serviço Twitter, qual estado desponta em sua região e assim por diante. Através do mapa essa informação sai de forma muito mais natural como visto na figura 6.

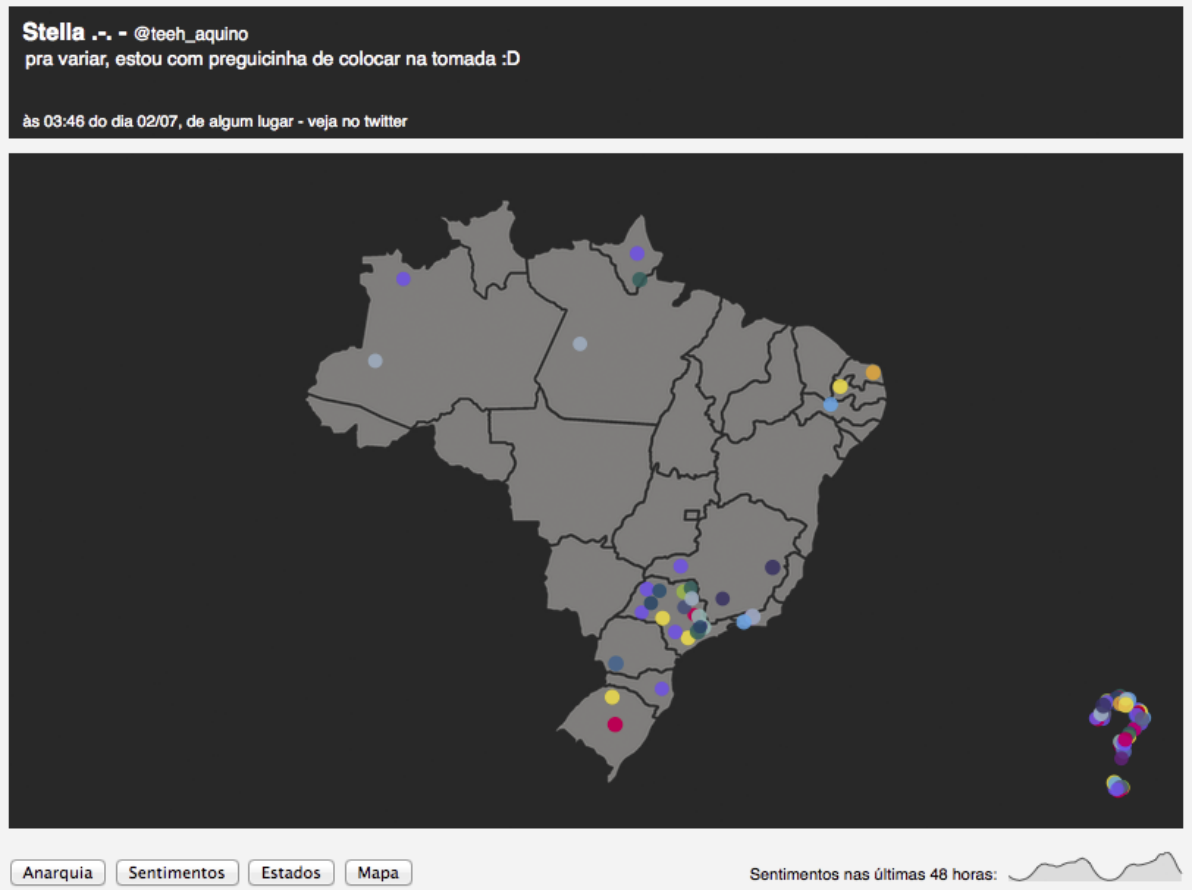


Figura 6: Visualização Mapa

Para a realização dessa visualização foi montado um mapa do Brasil onde cada estado foi colorido com uma cor da palheta RGB, como na figura 7. Posteriormente, já no algoritmo da visualização, é feita uma varredura dessa imagem, onde cada estado tem sua cor RGB pré-definida, e quando um pixel da imagem é identificado com determinada cor, esse pixel é adicionado a um arranjo de pixels de seu estado. Dessa maneira, para cada partícula é dada aleatoriamente um dos pixels do *array* do seu estado. Pode-se concluir, portanto, que não é feita a localização a nível de cidades no mapa.

Figura 7: Mapa do Brasil com suas unidades diferenciadas por cor RGB

Técnica semelhante é utilizada para o posicionamento e formação do ponto de interrogação presente nas visualizações Estados e Mapa.

4.2.5 Recurso de Filtragem

Foram criados recursos de filtragem que permitem que o usuário escolha quais sentimentos e estados exibir na ferramenta, e também que somente *tweets* com localização definida sejam carregados. Quando a visualização é iniciada pela primeira vez, todos os sentimentos e estados são exibidos, incluindo textos sem localização. Desse ponto, o usuário pode escrever ou selecionar os sentimentos e estados que deseja. Quando essa seleção é carregada, a página é recarregada com os novos dados e a visualização reiniciada.

Figura 8: Recurso de filtragem de sentimento

4.2.6 Estatísticas Adicionais

Componentes visuais adicionais que são partes integrantes do *Front-End* de visualização e do sítio online e que complementam as informações da visualização principal também foram criados. Eles procuram levar ao usuário, através de gráficos impossíveis com as partículas da visualização, informações resumidas e conclusões da base de dados, em especial aquelas possibilitadas pelo Analytics discutido na seção 4.1.7.

O primeiro desses componentes é exibido quando nenhum filtro é utilizado e mostra como os cinco sentimentos mais frequentes nas últimas 24 se comportaram, em termos de participação na totalidade de sentimentos coletados. Esse componente é exibido na figura 9.

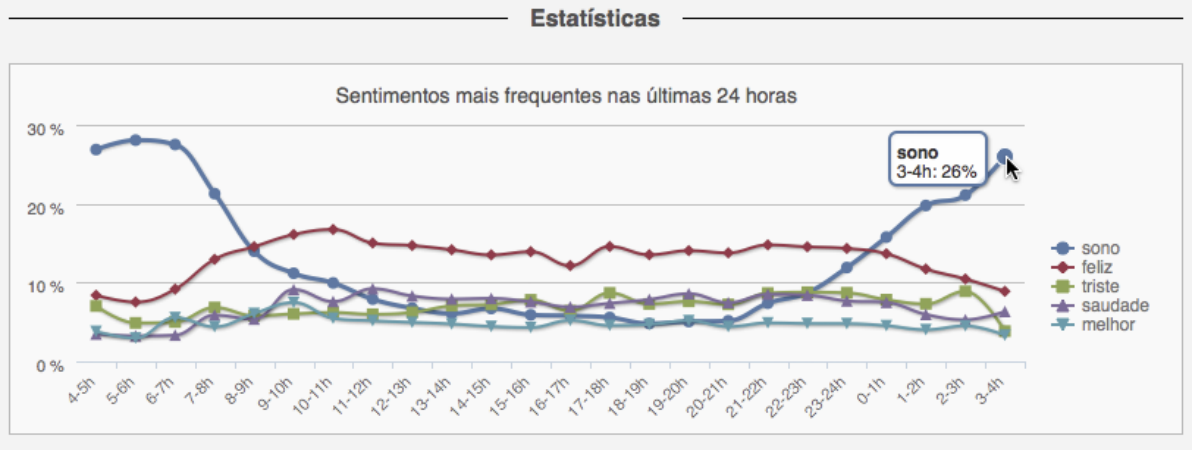


Figura 9: Sentimentos mais frequentes nas últimas 24 horas

Esse mesmo gráfico é atualizado quando é feita uma filtragem por sentimentos, com os sentimentos escolhidos. Assim, esse componente responde dinâmicamente da mesma forma que a visualização principal. Um exemplo para a filtragem para os sentimentos “fome” e “cansado” é exibido na figura 10.

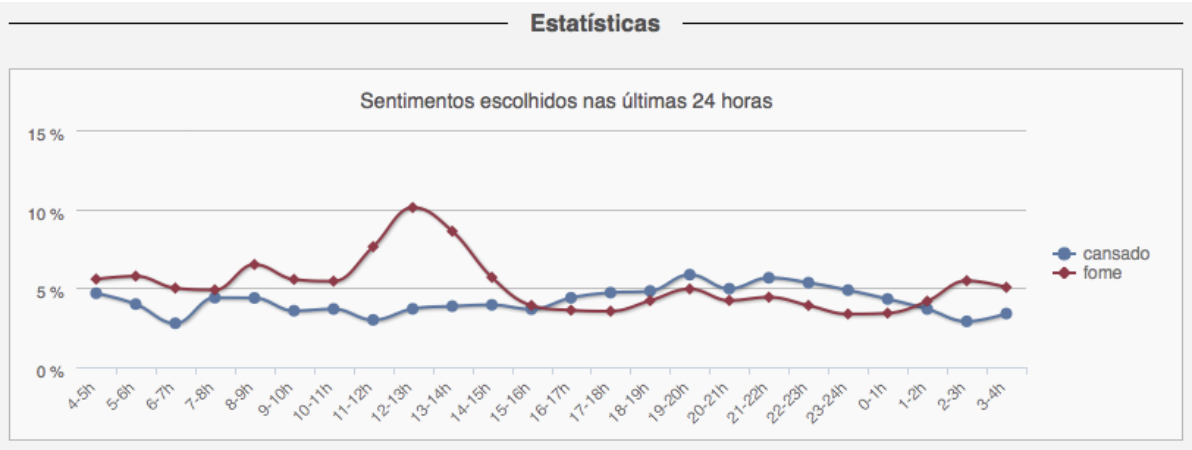


Figura 10: Sentimentos filtrados nas últimas 24 horas

Finalmente, quando estados federativos são filtrados, são exibidos gráficos de coluna com os 10 sentimentos mais frequentes, em toda a base de dados, para cada um dos estados. Novamente, por frequência entende-se participação na totalidade dos sentimentos para aquele estado. Um exemplo para a filtragem por “Rio Grande do Norte” e “Minas Gerais” é exibido na figura 11.

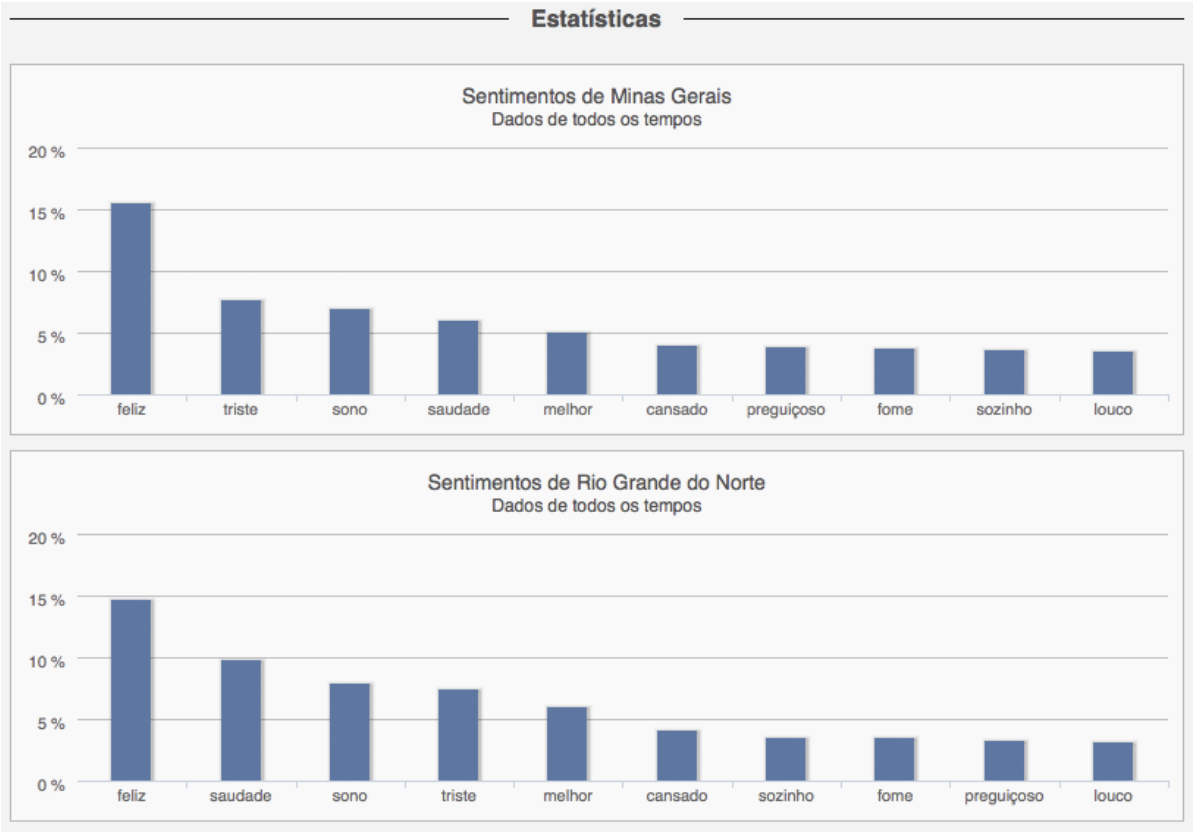


Figura 11: Dez sentimentos mais frequentes para os estados filtrados

É importante ressaltar que, como mostrado na figura 9, é possível que os valores exibidos sejam aferidos passando-se o cursor do mouse por cima dos elementos das séries, quando então é exibida uma pequena caixa com informações sobre o ponto em questão.

4.3 Base de Dados

A arquitetura apresentada na seção 4.1 permite que o sistema colete dados ininterruptamente, 24 horas por dia, 7 dias por semana. Essa base de dados crescente permite que sejam feitas várias análises e discussões que vão além daquelas mostradas pela ferramenta de visualização. Na realidade, como será discutido no seção 5, a criação de uma API para esses dados coletados permitirá que novas aplicações e novos estudos de terceiros surjam e respondam a novas perguntas.

Até a data deste documento o sistema havia rodado por seis dias, permitindo somente uma análise diária de alguns comportamentos dos usuários brasileiros na rede Twitter no tocante a sentimentos. Sendo assim, é possível fazer uma análise adicional da base de dados, com caráter quantitativo.

Inicialmente, ressalta-se que são indexados em média 100 mil novos *tweets* por dia. O gráfico de barras na figura 12 nos mostra quais são os sentimentos que os usuários mais expressam no meio virtual. Podemos destacar "feliz" presente em aproximadamente 14% dos textos, seguido por "sono" com 9%. Outros sentimentos também expressivos como "saudade", "triste" e "fome" confirmam a crença popular de que pessoas tendem a expressar seus sentimentos quando estes são negativos e refletem algum incômodo.

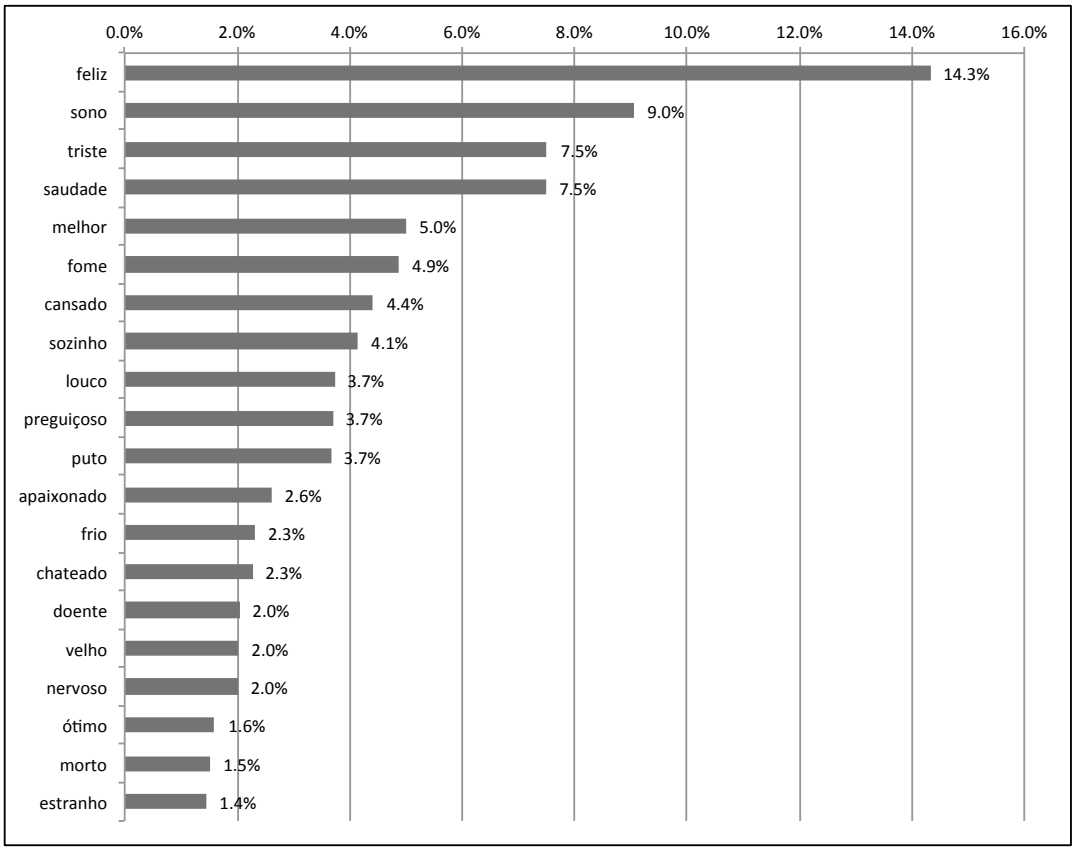


Figura 12: Participação de cada um dos sentimentos

Por outro lado, o gráfico da figura 13 nos mostra quais são as unidades federativas que possuem o maior número de usuários na rede Twitter. Com quase três vezes mais usuários que o segundo lugar, o estado de São Paulo lidera, refletindo tanto sua grande população, quanto seu desenvolvimento, nesse caso tecnológico. Contudo, alguns estados com números também expressivos, como "Rio Grande do Norte" na sétima posição, refletem o recente desenvolvimento social experimentado por suas populações. Todavia, as regiões Sul e Sudeste, preenchendo completamente as primeiras posições, evidenciam a ainda grande desigualdade social no país.

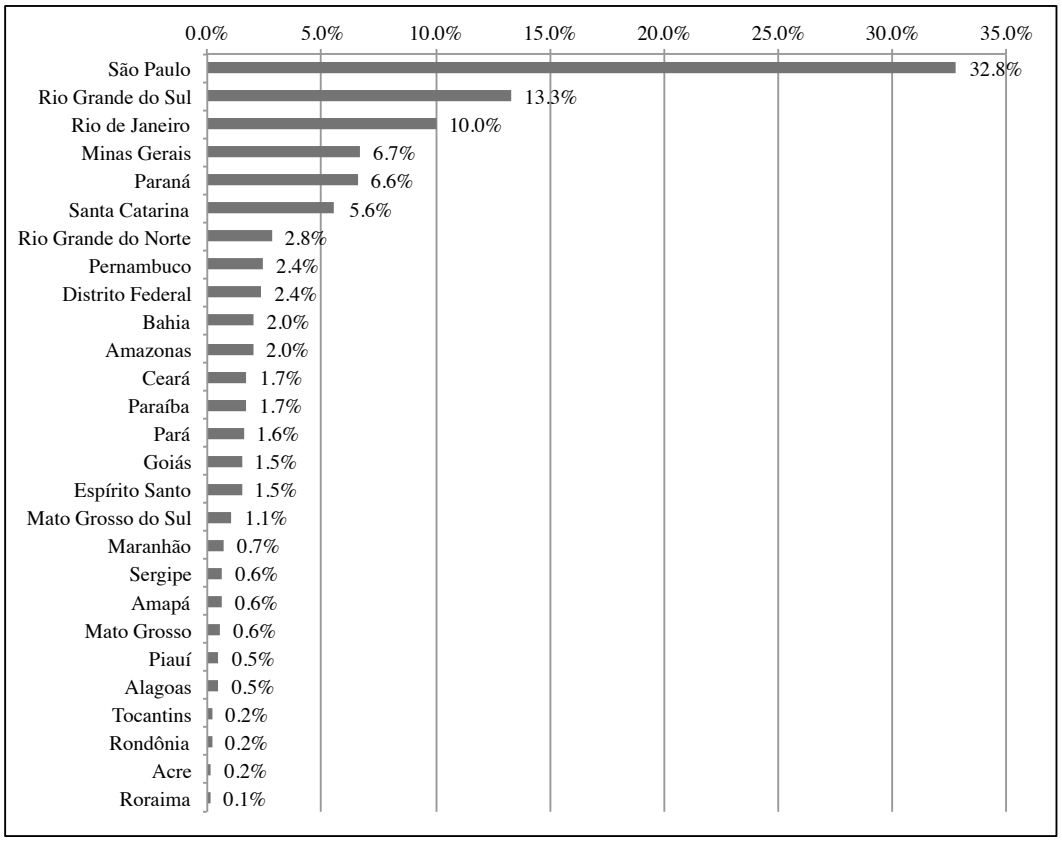


Figura 13: Participação de cada uma das unidades federativas

O gráfico de radar exibido na figura 14 nos dá a ideia de quais são os horários nos quais a rede é mais utilizada. É fácil notar que os brasileiros utilizam o Twitter, e consequentemente a internet para fins similares, mais intensamente na segunda metade do dia, principalmente entre as 16h e 0h.

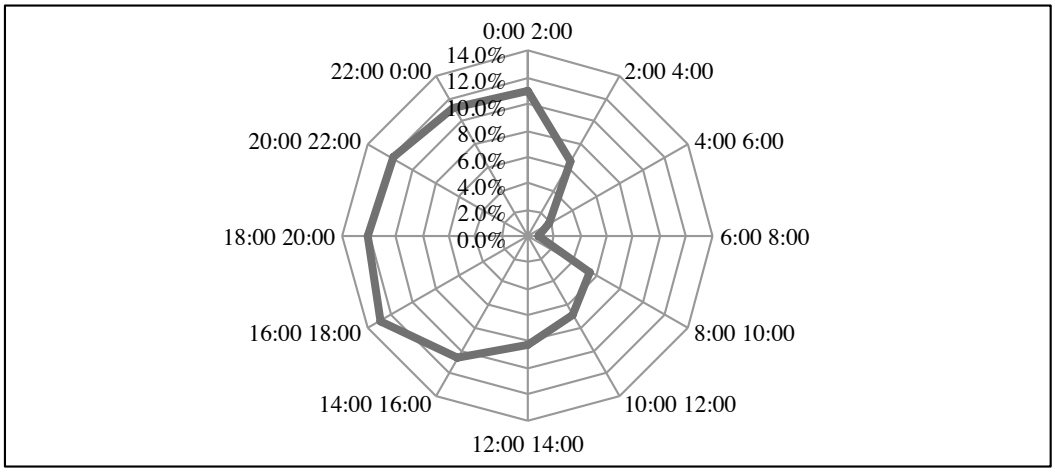
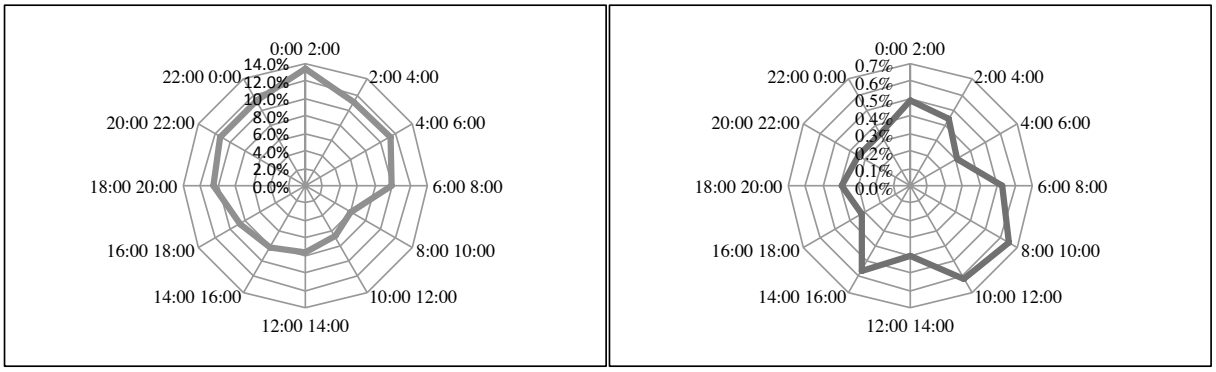


Figura 14: Utilização da rede durante as horas do dia

É possível também fazer uma análise quantitativa de como cada sentimento é extravasado por seus autores no decorrer do dia, permitindo que análises sofisticadas sejam elaboradas.

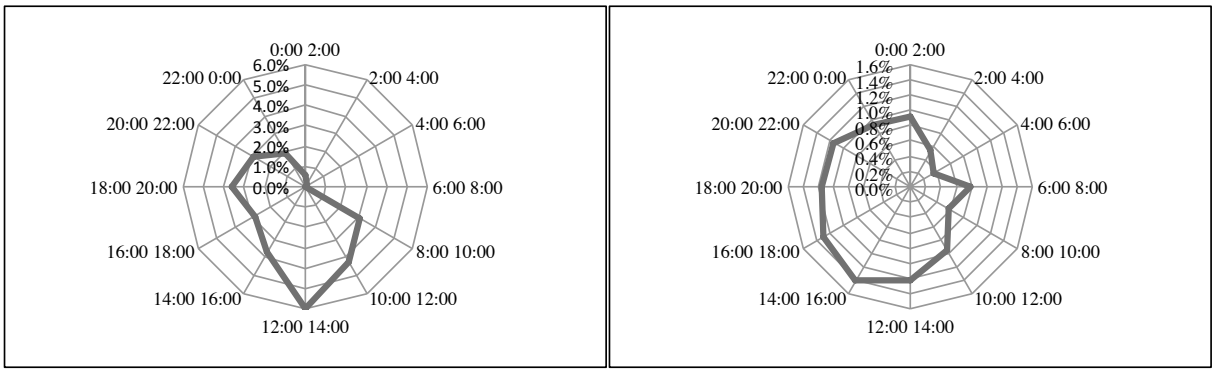
Observando a dispersão dos sentimentos “cansado” e “disposto” no decorrer do dia na figura 15, que mostra a porcentagem dos textos rotulados com os respectivos sentimentos no decorrer de um dia, podemos concluir que as pessoas se sentem mais cansadas durante as noites e madrugadas, enquanto elas aparentam estar mais dispostas pelas manhãs, como era de se imaginar.



(a) Cansado (b) Disposto

Figura 15: Disperção dos sentimento Cansado e Disposto durante o dia

Realizando a mesma análise para o sentimento “fome” como visto na figura 16a, vemos que os “internautas” sentem e tendem a expressar sua fome em torno da metade do dia, não coincidentemente no horário do almoço. Finalmente para o sentimento “estressado” é possível notar que a população está mais estressada na segunda metade do dia.



(a) Fome (b) Estressado

Figura 16: Disperção dos sentimento Fome e Estressado durante o dia

5 CONCLUSÕES E TRABALHOS FUTUROS

O projeto apresentado cumpriu com sua proposta de criação de um sistema capaz de coletar, trabalhar e exibir dados carregados de sentimentos de redes sociais online. Com o desenvolvimento desse sistema, composto por diversos componentes, o autor desse documento exercitou diversas técnicas e conceitos da Ciência da Computação, como coleta e mineração de dados, processamento de linguagem natural e visualização de dados. Apesar de ter atingido o objetivo principal que era a construção de uma ferramenta de visualização, ao mesmo tempo eficaz e que atrai o interesse do público, os produtos desse trabalho podem ser ainda mais explorados.

Os dados coletados têm grande potencial de utilização e estudo por outros autores. As técnicas desenvolvidas tanto de coleta como de análise dos textos podem ser estendidas para novas abordagens que não somente a de sentimentos. A ferramenta de visualização em si pode ser adaptada para ser uma *framework* para visualização de dados estratificados.

A visualização foi disponibilizada em um sítio online e pode ser livremente acessada por qualquer pessoa com acesso à internet. Quando a ferramenta foi apresentada para um grupo de pessoas convidado, foi relatado que a interface simples e amigável tornou transparente a apreensão do contexto, comprovando a proposta de facilitar a percepção dos dados representados.

As perspectivas futuras para a ferramenta são muito promissoras. O componente Analytics descrito na seção 4.1.7 permite que as estatísticas adicionais discutidas em 4.2.6 sejam amplamanete estendidas. A inserção de um filtro de data e hora permitiria que os usuários navegassem no histórico da visualização.

Outra melhoria seria a de explorar melhor, na visualização principal, as informações climáticas, através de um histograma adicional ou de novas metáforas visuais que representem as características desse tipo de informação.

Uma proposta interessante e que estava presente no início do projeto é a de ter o Facebook como fonte de textos. Atualmente, ela é a rede que mais cresce entre os usuários brasileiros e, portanto, sua inserção enriqueceria a análise e a qualidade das conclusões retiradas pelo sis-

tema. Infelizmente, a disponibilidade pública dos textos produzidos por seus usuários é muito mais restrita que no Twitter, tornando a sua integração mais complexa.

Uma avaliação da ferramenta com usuários também teria sido muito enriquecedora, mas devido às restrições temporais do projeto, não foi possível receber esse *feedback*. Todavia, a disponibilização online da ferramenta permitirá que o autor receba, posteriormente, as avaliações e opiniões do público alvo da visualização, os usuário da internet no Brasil.

Finalmente, seria enriquecedor que fosse criado um componente baseado nas tecnologias AJAX que atuasse como uma API para a leitura dos dados presentes no banco de dados. Através de tal componente, seria possível tornar pública e acessível a base de dados do sistema, permitindo o surgimento de novos estudos e análises sobre os sentimentos dos usuários de internet no Brasil, antes não imaginados pelo criador do sistema.

Referências

Ackland 2008 ACKLAND, R. Using Facebook as a Data Source and Platform for e-Researching Social Networks. In: *refereed paper presented at the Fourth International Conference on e-Social Science*. [S.l.: s.n.], 2008. p. 18–20.

Bonfante e Nunes 2002 BONFANTE, A.; NUNES, M. Parsing probabilístico para o Português do Brasil. In: SN. *Brazilian Symposium on Artificial Intelligence, 16*. [S.l.], 2002.

Chakrabarti 2002 CHAKRABARTI, S. Mining the Web: Analysis of hypertext and semi structured data. Morgan Kaufmann, 2002.

Chodorow e Dirolf 2010 CHODOROW, K.; DIROLF, M. *MongoDB - The Definitive Guide: Powerful and Scalable Data Storage.* [S.l.]: O'Reilly, 2010. I-XVII, 1-193 p. ISBN 978-1-449-38156-1.

Cleveland 1993 CLEVELAND, W. *Visualizing data*. [S.l.]: Hobart Press, 1993. ISBN 0963488406.

Few 2009 FEW, S. *Now you see it: simple visualization techniques for quantitative analysis*. [S.l.]: Analytics Press, USA, 2009. ISBN 0970601980.

Garton, Haythornthwaite e Wellman 1997 GARTON, L.; HAYTHORNTHWAITE, C.; WELLMAN, B. Studying online social networks. *Journal of Computer-Mediated Communication*, Wiley Online Library, v. 3, n. 1, p. 0–0, 1997. ISSN 1083-6101.

Inc. 2011 INC., G. *The Google Geocoding API*. 2011. [Online; accessed 25-junho-2011]. Disponível em: <http://code.google.com/apis/maps/documentation/geocoding/>.

Kamvar e Harris 2011 KAMVAR, S.; HARRIS, J. We feel fine and searching the emotional web. In: ACM. *Proceedings of the fourth ACM international conference on Web search and data mining*. [S.l.], 2011. p. 117–126.

Keim 2002 KEIM, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 8, n. 1, p. 1–8, jan. 2002. ISSN 1077-2626. Disponível em: <http://dx.doi.org/10.1109/2945.981847>.

Kleinberg 2007 KLEINBERG, J. Challenges in mining social network data: processes, privacy, and paradoxes. In: ACM. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2007. p. 4–5.

Santos 2005 SANTOS, C. dos. Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro. 2005.

Thelwall 2001 THELWALL, M. A web crawler design for data mining. *Journal of Information Science*, Sage Publications, v. 27, n. 5, p. 319, 2001. ISSN 0165-5515.

Tufte 1991 TUFTE, E. Envisioning information. *Optometry & Vision Science*, v. 68, n. 4, p. 322, 1991. ISSN 1040-5488.

Tufte e Robins 1997 TUFTE, E.; ROBINS, D. Visual explanations. *Cataloging and Classification Quarterly*, New York, NY: Haworth Press,[c1981-, v. 24, n. 3, p. 152–155, 1997. ISSN 0163-9374.