Arquitetura e Desenvolvimento de um Repositório de Tweets em Português Europeu

Gaspar Brogueira[1,2], Fernando Batista[1,2] e Joao P. Carvalho[1,3]

[1] Laboratório de Sistemas de Língua Falada - INESC-ID, Lisboa, Portugal
[2] ISCTE-IUL - Instituto Universitário de Lisboa, Portugal
[3] IST - Instituto Superior Técnico, Universidade de Lisboa, Portugal
gmrba@iscte.pt, {fernando.batista, joao.carvalho}@inesc-id.pt

Resumo. As redes sociais têm ganho bastante popularidade para a partilha de informação sobre os mais diversos tópicos desde a política, desporto ou aspectos do quotidiano. As mensagens (tweets) partilhadas no Twitter[1] são essencialmente públicas, constituindo assim uma fonte de informação, que por ser difundida em tempo real pode revelar-se vantajosa para domínios como o turismo, marketing, saúde ou segurança. Este artigo apresenta uma metodologia para a criação de um repositório de tweets em Português Europeu que, partindo do fluxo de tweets geolocalizados disponibilizados pela API[2] do Twitter, é expandido pela recolha da timeline de cada um dos utilizadores que publicam de forma geolocalizada.

Palavras-Chave: Twitter, Twitter API, MongoDB, Geolocalização, Tweets em Português

1 Introdução

A utilização do Twitter gera em média cerca de 400 milhões de novas mensagens por dia [7]. Alguma desta informação está disponível através das APIs públicas do Twitter, sem qualquer custo. A amostra do fluxo de dados cedida gratuitamente pelo Twitter, corresponde a 1% da totalidade do fluxo público de tweets em determinado momento [10]. A popularidade do Twitter como fonte de informação tem conduzido ao desenvolvimento de aplicações e investigação em diversos domínios. Informação esta que pode ser utilizada na previsão e divulgação de dados em situações de catástrofe natural, uma vez que muitas pessoas utilizam o Twitter para partilhar informação e experiências durante os momentos de crise e no rescaldo dos mesmos [13, 8]. Sakaki et al. [12] usou a informação do Twitter para prever a ocorrência de sismos e Muralidharan et al. [11] efectuou um estudo de como através de mensagens publicadas no Twitter, organizações sem fins lucrativos e os *media,* mobilizaram ajuda no esforço de apoio às vítimas do terramoto no Haiti, em 2010.

[1] https://twitter.com/
[2] https://dev.twitter.com/overview/api

Um dos tópicos de maior interesse na análise textual dos tweets, está relacionado com as opiniões formuladas pelos utilizadores do Twitter, acerca de produtos, marcas ou serviços. A análise destas opiniões ou sentimentos pode fornecer vantagens numa variedade de domínios, como por exemplo, na previsão de vendas [9], na política [16], no turismo [3], para agências governamentais [4] ou policiais Gerber [5]. Para Kontopoulos et al. [6] o conceito de análise de sentimentos é definido como um processo objectivo de determinar a polaridade de um *corpus* textual (documentos, parágrafos, frases, ...) tendendo para positivo, negativo ou neutro.

Também com informação recolhida do Twitter, nos trabalhos de Santos and Matos [14], [15] foi utilizado um conjunto de aproximadamente 2700 tweets produzidos em Portugal, para conseguir prever a taxa de incidência e disseminação do vírus influenza, na população Portuguesa. Gerber [5] refere que o Twitter é uma fonte de dados ideal para problemas de suporte à decisão e utilizando tweets marcados no espaço e no tempo tentou prever a atividade criminal na maior cidade dos Estados Unidos.

Considerando que a informação publicada no Twitter possa ser relevante para diversos tipos de estudos e o facto da mesma não ser disponibilizada integralmente de forma gratuita, neste artigo é apresentada uma metodologia para expansão de uma base de dados de tweets produzidos em Portugal e escritos em Português Europeu.

Este artigo encontra-se estruturado nas seguintes secções: A Secção 2 descreve a arquitectura proposta; na Secção 3 é efetuada uma análise aos tweets armazenados, comparando o volume de tweets obtidos pela Streaming e pela REST API do Twitter, de Fevereiro a Dezembro de 2014; na Secção 4 são referidas algumas aplicações da informação recolhida pela metodologia apresentada. Pos último, a Secção 5 apresenta as conclusões e menciona as propostas para continuação do trabalho desenvolvido.

2 Repositório de dados

A arquitectura desenvolvida para a criação de um repositório de tweets produzidos em Portugal e escritos em Português Europeu compreende duas etapas:

1. Armazenamento do fluxo de tweets geolocalizados disponibilizado pela API do Twitter e cuja mensagem de 140 caracteres é identificada, pelo Twitter, como estando escrita em Português;
2. Recolha da atividade recente (timeline) de todos os utilizadores que alguma vez publicaram um tweet geolocalizado em Portugal e escrito em Português Europeu, tendo sido o tweet disponibilizado pela Streaming API.

A Streaming API retorna os tweets públicos, publicados em determinado instante, que estão de acordo com um ou mais filtros. Com a Streaming API é possível pesquisar por palavras-chave, *hashtags*, *user_id* ou regiões delimitadas geograficamente. A *statuses/filter*[3] API facilita não só esta pesquisa, como per-

[3] https://dev.twitter.com/streaming/reference/post/statuses/filter

mite a recolha de um fluxo contínuo de tweets, que estejam de acordo com um critério de pesquisa indicado [7].

A identificação da localização geográfica de um tweet pode ser obtida seguindo uma das seguintes abordagens:

- Geolocalização: os utilizadores podem optar por tornar pública a informação da sua localização no momento da publicação de um novo tweet. Esta informação pode ser bastante precisa, isto se o tweet for publicado com recurso, por exemplo, a um smartphone com GPS.
- Perfil do Utilizador: a localização de determinado utilizador pode ser extraída do campo *location* do seu perfil. Esta informação é igualmente disponibilizada na API do Twitter.

Porém a Streaming API tem limitações tanto no número de parâmetros como na quantidade de resultados devolvidos. Cada pedido admite como parâmetros o máximo de 400 palavras-chave, 25 delimitações de regiões geográficas ou 5000 *user_id*.

Como resultado, são retornados todos os tweets que satisfaçam as restrições do pedido, até determinado limite, que no caso da *statuses/filter* API é de 1% do volume total de tweets produzidos no Twitter, em determinado instante [7].

2.1 Recolha de Tweets Geolocalizados

No presente trabalho, foi utilizada a primeira das abordagens enunciadas anteriormente para identificar a localização de um tweet. Tendo como objectivo recolher unicamente tweets produzidos em Portugal e escritos em Português Europeu, ao parâmetro *locations* da *statuses/filter* API foram atribuídas as coordenadas que delimitam a região de Portugal Continental e dos Arquipélagos dos Açores e da Madeira.

As APIs do Twitter podem ser acedidas apenas com pedidos autenticados. O Twitter utiliza a *Open Authentication* (OAuth) e cada pedido é autenticado com as credenciais de um utilizador do Twitter. O acesso às APIs do Twitter é limitado por um número máximo de pedidos num determinado intervalo de tempo, designado por *rate limit*. A janela de tempo correspondente ao *rate limit* é atualmente de 15 minutos e é utilizada para renovar periodicamente a quota de invocações permitidas às APIs do Twitter. Os limites são aplicáveis tanto na autenticação com contas de nível individual como na autenticação com contas de nível de aplicação, podendo ser diferentes quanto ao máximo de pedidos permitidos [7]. As contas, com as quais é permitido o acesso à API do Twitter, deverão ser criadas e devidamente registas no Twitter[4].

O acesso ao fluxo de tweets da *statuses/filter* API é permitido após a autenticação com sucesso, sendo o fluxo de tweets disponibilizado de forma contínua sem mais intervenção por parte da conta "consumidora". A resposta das APIs do Twitter é devolvida no formato JavaScript Object Notation[5] (JSON), actualmente bastante popular para a representação de objectos na Web.

[4] https://www.apps.twitter.com/
[5] http://www.json.org/

Na arquitectura proposta, o fluxo de tweets da *statuses/filter* API é guardado numa estrutura de ficheiros comprimidos em disco, criando-se um novo ficheiro a cada hora, separando-os em directórios que contêm os ficheiros produzidos a cada dia. Posteriormente, os ficheiros são processados de modo a filtrar os tweets recolhidos considerando como "válidos" os tweets com o valor do campo *language* igual a "pt", armazenando-os numa base de dados MongoDB[6]. Os tweets cujo campo *language* difere de "pt" não serão considerados, permanecendo apenas armazenados nos ficheiros.

Tendo em consideração a dificuldade de representação exacta dos limites geográficos de Portugal, os limites indicados no parâmetro *locations* da *statuses/filter,* possibilitam a recolha de tweets geolocalizados em Espanha ou em Marrocos. Para contornar tal dificuldade, além da filtragem pelo campo *language,* é verificado igualmente se campo *place.country* contém como valor "Portugal", visto que, embora o campo *language* seja estimado de forma automática pelo Twitter por aplicação de algoritmos de detecção de linguagem[7], nem sempre um tweet cujo campo *language* tem o valor "pt", tem igualmente o valor "Portugal" no campo *place.country.*

Os tweets filtrados no passo anterior são armazenados numa base de dados MongoDB, que não seguindo o modelo relacional, permite o armazenamento de dados não estruturados. Estando os tweets disponibilizados pelas APIs do Twitter no formato JSON, podem ser armazenados directamente numa base de dados MongoDB sem quaisquer alterações, o que não seria prático se utilizada uma base de dados relacional, dado o processamento que implicaria.

No entanto, de modo a controlar os tweets armazenados assim como os autores dos mesmos, alguma informação relativa aos tweets e seus autores, é registada numa base de dados MySQL. De cada tweet "válido" é guardado o *tweet_id* que identifica univocamente cada tweet, o *user_id* que identifica o autor do tweet e a data de publicação do mesmo, que se encontra no campo *created_at.*

Para cada utilizador é guardado o seu identificador único, dado pelo campo *user.id,* e dois outros valores (*process_iteration* e *last_process_iteration_date*), cuja utilidade será explicada na Secção 2.2.

2.2 Expansão da Base de Dados

Partindo do conjunto de utilizadores obtido pelo armazenamento dos tweets do fluxo da *statuses/filter* API, procede-se à recolha da timeline de cada utilizador, ou seja, a actividade recente gerada por determinado utilizador. A timeline é disponibilizada pela REST API *statuses/user_timeline*[8], que admitindo como parâmetro o *user_id* ou o *screen_name* (nome do utilizador visível na aplicação do Twitter) retorna os últimos 3200 tweets produzidos pelo utilizador indicado.

O procedimento de recolha da timeline de todos os utilizadores, esquematizado na Fig. 1, não é um processo trivial, dadas as restrições impostas pelo

[6] https://www.mongodb.org/
[7] https://blog.twitter.com/2013/introducing-new-metadata-for-tweets
[8] https://dev.twitter.com/rest/reference/get/statuses/user_timeline

Twitter na utilização da *statuses/user_timeline*. Genericamente, a utilização da API do Twitter é dividida em períodos de 15 minutos, sendo permitido efetuar 180 pedidos à API em cada período com autenticação do nível de utilizador ou 300 pedidos no caso de autenticação com nível de aplicação [7].

A autenticação utilizada neste trabalho é do nível de utilizador. Cada invocação da *statuses/user_timeline* retorna no máximo 200 tweets, sendo portanto necessário efetuar 16 pedidos à referida API para recolher o limite máximo de 3200 tweets disponíveis. No caso extremo de todos os utilizadores terem produzido o máximo de tweets que é permitido recolher, em cada intervalo de 15 minutos conseguir-se-á recolher a timeline completa de apenas 11 utilizadores, permitindo recolher a timeline somente de cerca de 1000 utilizadores diariamente, por cada conta de autenticação utilizada no acesso à REST API *statuses/user_timeline*. No processo de recolha da timeline, foi utilizado mais do que uma conta com autenticação de nível utilizador, que em paralelo recolhem a timeline dos cerca de 76K utilizadores identificados, recolhendo cada conta em média a timeline de 3K utilizadores por dia, ou seja, considerando o número de utilizadores à data da escrita deste artigo, são necessários 4 dias para efetuar uma iteração completa por todos os utilizadores.

Por cada novo utilizador, além de se registar o seu *user_id*, tal como referido anteriormente, é registado no atributo *process_iteration* um dos seguinte valores:

- -1 : Erro. Acesso bloqueado à timeline;
- 0 : Valor inicial. Novo utilizador sem timeline recolhida;
- 1 : Após a primeira recolha da timeline;
- 2 : Atualização da timeline em função do último tweet recolhido.

No atributo *last_process_iteration_date* é registado inicialmente a data em que o utilizador foi integrado na base de dados, sendo este atributo utilizado para efeitos de ordenação dos utilizadores aquando da recolha da timeline, tendo prioridade os utilizadores registados à mais tempo. Após a recolha da timeline, o valor deste atributo é atualizado para a data corrente, com o mesmo propósito de ordenação.

O atributo *process_iteration* permitirá controlar o armazenamento da timeline. Inicialmente é recolhido iterativamente a timeline dos utilizadores com o valor do atributo *process_iteration* igual a 0. Nesta fase pretende-se obter todos os tweets da timeline de cada utilizador até ao máximo permitido pela API, não tendo em consideração o histórico de tweets geolocalizados já armazenados na base de dados MongoDB, para determinado utilizador. Os tweets recolhidos da timeline, são armazenados numa *collection* MongoDB distinta da *collection* onde são guardados os tweets geolocalizados recolhidos pelo procedimento descrito na Secção 2.1. O valor do atributo *process_iteration* é incrementado para 1.

Quando o status do campo *process_iteration* de todos os utilizadores for igual a 1, o processo de recolha da timeline é reiniciado. Nesta fase, para cada utilizador são pesquisados os seus novos tweets, publicados posteriormente ao último tweet armazenado no MongoDB desse mesmo utilizador. O facto da recolha se restringir apenas aos novos tweets, permite reduzir o número de invocações à REST API *statuses/user_timeline* necessárias para guardar todos os novos tweets de

cada utilizador, permitindo o armazenamento da timeline de um maior número de utilizadores diariamente. Após esta segunda iteração, o valor do campo *process_iteration* é incrementado para 2, sendo igualmente atualizada a data de *last_process_iteration_date*.

O processo de recolha da timeline executando-se continuamente, dá prioridade à recolha da timeline dos novos utilizadores, com o valor de *process_iteration* igual a 0. A atualização da timeline dos utilizadores em que *process_iteration* assume o valor 1 ou 2, é efetuada mediante a ordenação decrescente da data em *last_process_iteration_date*. Os utilizadores que não permitem a recolha da sua timeline, por esta se encontrar bloqueada, ficarão com o atributo *process_iteration* com o valor -1, não sendo considerados no processo de recolha/atualização da timeline.

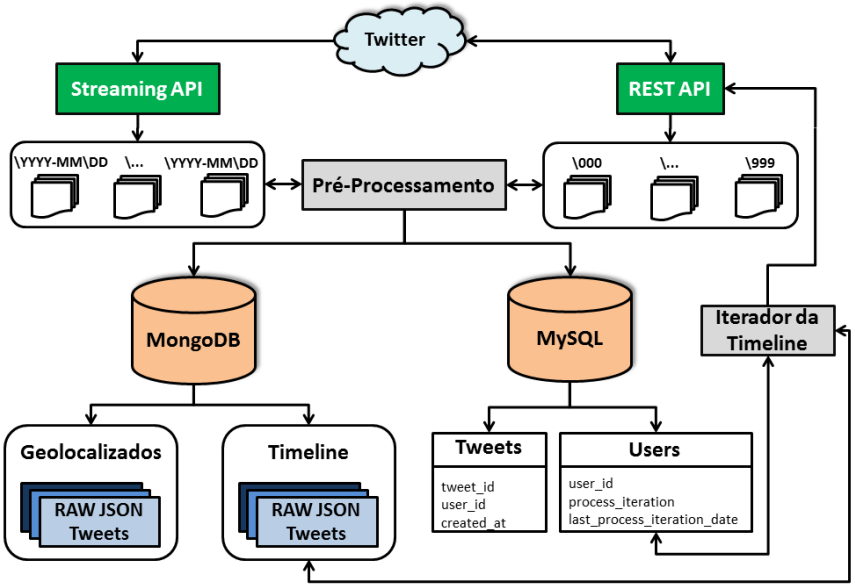

Fig. 1. Arquitectura para expansão de base de dados de tweets.

3 Estatísticas sobre os dados recolhidos

Os dados em análise neste artigo foram recolhidos no período de 20 de Fevereiro a 31 de Dezembro de 2014, tendo sido armazenados um total de aproximadamente 18.4M de tweets geolocalizados, com uma média diária de cerca de 61.8K tweets, produzidos por 75949 utilizadores. Na Fig. 2 é apresentada a distribuição da recolha dos tweets, no período de Fevereiro a Dezembro de 2014, observando-se que o número de tweets geolocalizados recolhidos aumentou ao longo do ano. Os dias com cor branca deveram-se a problemas pontuais na recolha dos tweets.

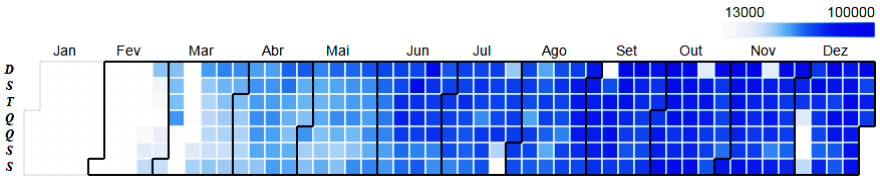

Fig. 2. Distribuição da recolha dos tweets ao longo de 2014.

Pelo procedimento descrito na Secção 2.1, são integrados em média 337 novos utilizadores por dia. Analisando o gráfico da Fig. 3 verifica-se que o número de novos utilizadores por dia é constante, perspectivando-se que a execução contínua deste procedimento, tenderá para uma cobertura praticamente total de todos os utilizadores portugueses que publicam tweets de forma geolocalizada. Deste modo, será igualmente possível a recolha de praticamente todos os novos tweets produzidos pela comunidade portuguesa do Twitter.

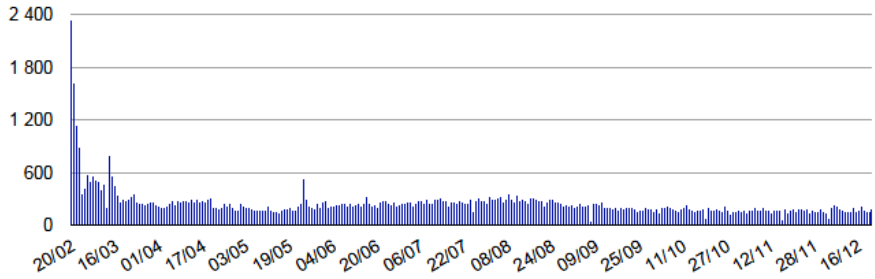

Fig. 3. Novos utilizadores por dia.

No gráfico da Fig. 4 verifica-se uma maior atividade de publicação de tweets nos primeiros dias da semana, nomeadamente ao Domingo e Segunda-Feira, decrescendo a atividade com o decorrer da semana, sendo a Sexta-Feira o dia com menor produção de tweets. Este comportamento poderá justificar-se, pelo facto da comunidade Portuguesa presente no Twitter, ser essencialmente composta por adolescentes ou jovens adultos, que aproveitam as Sextas-Feiras e Sábados para saírem com os amigos, diminuindo a sua atividade no Twitter [1].

No gráfico da Fig. 5 analisa-se a publicação de tweets ao longo do dia, verificando-se dois períodos em que a atividade é mais intensa, o primeiro situa-se por volta da hora de almoço, crescendo progressivamente a partir do final da tarde, atingindo-se o pico de atividade entre as 21h e as 22h.

Da execução do procedimento descrito na Secção 2.2, resultou a recolha de cerca de 158.8M de tweets. De entre os tweets recolhidos pela REST API, 134.2M correspondem ao intervalo de 20 de Fevereiro a 31 de Dezembro de 2014, o que

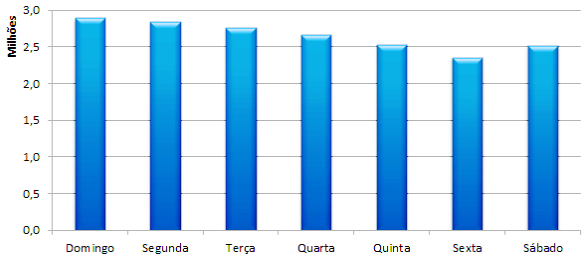

Fig. 4. Distribuição dos tweets recolhidos por dia da semana.

representa um acréscimo em média de 7 vezes no número de tweets recolhidos, por comparação com os tweets disponibilizados no fluxo da Streaming API.

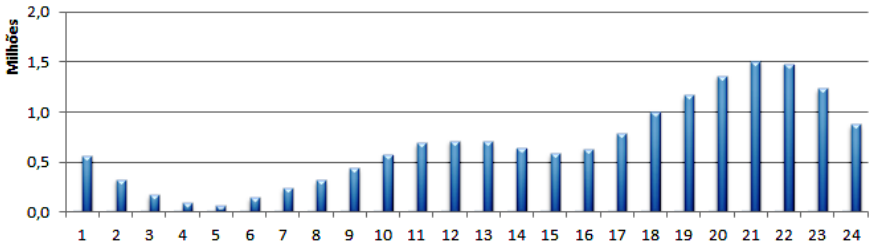

Fig. 5. Distribuição dos tweets recolhidos por hora ao longo de um dia.

Na Fig. 6 é notório o acréscimo de tweets recolhidos pela metodologia apresentada neste artigo, por comparação com volume de tweets disponibilizados pela Streaming API. A linha identificada como "Streaming API" ("paralela" ao eixo das abcissas) representa o número de tweets recolhidos da Streaming API, no período em análise neste artigo. As linhas A a E representam a evolução do volume total de tweets recolhidos da timeline de todos os utilizadores em determinado instante, indicado pela data de cada uma das linhas na legenda da Fig. 6. Cada uma das linhas é resultante de uma nova iteração de recolha/atualização da timeline de todos os utilizadores.

Ainda relativamente à Fig. 6 verifica-se que a oscilação das linhas C a E, na parte direita do gráfico, corresponde nos picos superiores, ao início das semanas, nomeadamente aos dias de Domingo e Segunda-Feira e os picos inferiores, aos dias do final da semana, mais concretamente a Sexta-Feira.

O dia 14 de Dezembro de 2014, foi o dia para o qual se recolheram mais tweets pelo método apresentado, no total de 921581 tweets, enquanto que para o mesmo dia foram recolhidos somente 93167 tweets da Streaming API, o que representa um acréscimo de sensivelmente 10 vezes em relação aos tweets disponibilizados no fluxo da Streaming API.

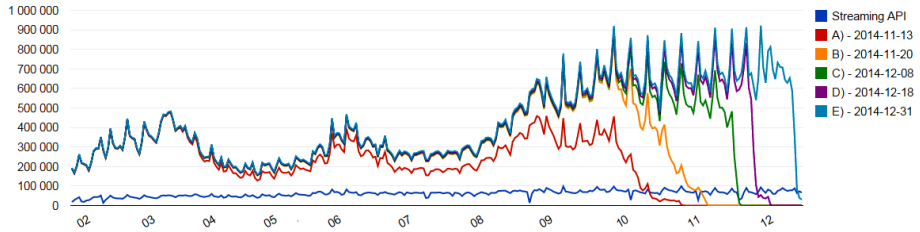

Fig. 6. Comparação do volume de tweets obtidos da Streaming API e da REST API.

4 Aplicações

A informação armazenada tem sido solicitada para análise sobre um conjunto alargado de temáticas, como por exemplo, estudos sociológicos relacionados com os Meets no Centro Comercial do Vasco da Gama ocorridos no final de Agosto de 2014 ou na análise das reações publicadas no Twitter relativamente à detenção do ex Primeiro Ministro José Sócrates.

A caracterização da comunidade portuguesa do Twitter [2], a detecção de género e faixa etária através da análise dos tweets ou a caracterização dos distritos de Portugal pela análise da quantidade de tweets recolhidos, são estudos resultantes da análise dos dados recolhidos, pela metodologia apresentada neste artigo.

5 Conclusões e Trabalho Futuro

Apresenta-se um método para o desenvolvimento de um repositório de tweets em Português Europeu que permite a constituição de uma base de dados de tweets. Este método tenderá a armazenar grande parte dos tweets produzidos pela comunidade Portuguesa do Twitter, dado que se adapta aos novos utilizadores integrados na base de dados, assim como continuará a recolher os novos tweets produzidos pelos restantes utilizadores que em algum momento foram identificados como autores de mensagens no Twitter, escritas em Portugês Europeu. A arquitectura desenvolvida permite a recolha de cerca de 7 vezes mais informação do que a obtida apenas usando a *Streaming API*.

Na continuação do trabalho desenvolvido, pretende-se que o método esteja suportado somente na base de dados MongoDB, dispensando a utilização do MySQL. A disponibilização dos tweets armazenados, baseada numa REST API, será outro dos objetivos a desenvolver.

Agradecimentos Este trabalho foi financiado com fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, pelos projetos PTDC/IVC-ESCT/4919/2012 (MISNIS) e PEst-OE/EEI/LA0021/2013.

Bibliography

[1] Brogueira, G., Batista, F., Carvalho, J. P., and Moniz, H. (2014a). Expanding a Database of Portuguese Tweets. In *3rd Symposium on Languages, Applications and Technologies*, volume 38, pages 275–282.

[2] Brogueira, G., Batista, F., Carvalho, J. P., and Moniz, H. (2014b). Portuguese geolocated tweets: An overview. ISDOC '14, pages 178–179. ACM.

[3] Claster, W., Cooper, M., and Sallis, P. (2010). Thailand – tourism and conflict: Modeling sentiment from twitter tweets using naïve bayes and unsupervised artificial neural nets. In *CIMSiM'2010*, pages 89–94.

[4] Dehkharghani, R., Mercan, H., Javeed, A., and Saygin, Y. (2014). Sentimental causal rule discovery from twitter. *Expert Systems with Applications*, 41(10):4950 – 4958.

[5] Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61(0):115 – 125.

[6] Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10):4065 – 4074.

[7] Kumar, S., Morstatter, F., and Liu, H. (2014). *Twitter Data Analytics*.

[8] Lachlan, K. A., Spence, P. R., and Lin, X. (2014). Expressions of risk awareness and concern through twitter: On the utility of using the medium as an indication of audience needs. *Computers in Human Behavior*, 35(0):554 – 559.

[9] Liu, Y., Huang, X., An, A., and Yu, X. (2007). Arsa: A sentiment-aware model for predicting sales performance using blogs.

[10] Magalhaes, T. and Nunes, S. (2013). Construção de amostras de dados do twitter. In *INForum 2013*, Évora, Portugal.

[11] Muralidharan, S., Rasmussen, L., Patterson, D., and Shin, J.-H. (2011). Hope for haiti: An analysis of facebook and twitter usage during the earthquake relief efforts. *Public Relations Review*, 37(2):175 – 177.

[12] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM.

[13] Saleem, H. M., Xu, Y., and Ruths, D. (2014). Effects of disaster characteristics on twitter event signature. *Procedia Engineering*, 78(0):165 – 172. Humanitarian Technology: Science, Systems and Global Impact 2014, HumTech2014.

[14] Santos, J. C. and Matos, S. (2013). Predicting flu incidence from portuguese tweets. In *Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2013*, pages 11–18, Granada, Spain.

[15] Santos, J. C. and Matos, S. (2014). Analysing twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, 11(1).

[16] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185.