

Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo

Felipe Cordeiro Alves Dias

Orientador: Prof. Dr. Daniel de Angelis Cordeiro

Universidade de São Paulo

4 de Janeiro de 2018

Introdução

- Segregação urbana: dentre os mais de 12 milhões de habitantes da cidade de São Paulo, 10% estão localizados no Centro Expandido (CE) e 90% no Cinturão Periférico (CP).
 - Problemas relacionados a mobilidade urbana.
- Legislação federal e municipal sobre mobilidade urbana.
 - Lei Federal 12.587/2012: para desenvolvimento sustentável com a mitigação dos custos ambientais e socioeconômicos dos deslocamentos de pessoas.
 - Decreto 56.834: institui o *PlanMob/SP 2015* como instrumento de planejamento e gestão do Sistema Municipal de Mobilidade Urbana para os próximos 15 anos.

- *PlanMob/SP 2015*

- Criação da Central Integrada de Mobilidade Urbana (CIMU): com o objetivo de integrar as áreas de trânsito e transporte subordinadas à Secretaria Municipal de Transportes (SMT).
- A CIMU não processa conteúdo de Redes Sociais;
- não aborda melhorias dos sistemas já existentes;
- será integrada com o desfasado SIM.

- Sistema Integrado de Monitoramento e Transporte (SIM): localização e rastreamento dos ônibus, fornece informações em tempo real aos passageiros, monitora 1.353 rotas de ônibus, 10 corredores de ônibus, 28 terminais de ônibus e 19.933 mil paradas de ônibus que serviram em 2016 a aproximadamente 8 milhões de passageiros por dia.
- Apesar da importância do SIM, há inúmeras defasagens tecnológicas (que causam discrepância nas informações recebidas pelos usuários, dentre outros problemas).

- Sistemas de Transporte Inteligente (ITS — *Intelligent Transport System*).
- A lei de mobilidade urbana (12.587/2012) e o *PlanMob/SP 2015* não mencionam explicitamente ITS e TIC.
- O transporte público pode se beneficiar ao explorar ITS, e ao integrar as Redes Sociais com o planejamento, gestão e as atividades operacionais dos transportes públicos, abordando seus respectivos fatores sócio-técnicos.
 - Analisar o impacto dos eventos de exceção na operação do sistema de transporte público por ônibus na cidade de São Paulo.

Definição do problema

- Eventos de exceção: acidentes, greves, falhas na operação do metrô, manifestações, enchentes, eventos sociais, dentre outros.
- Identificação de características dos eventos de exceção.
 - Dados históricos do SIM.
 - Processamento de grandes volumes de dados;
 - qualidade dos dados comprometida.
 - Twitter.
 - Identificação dos eventos de exceção nas publicações;
 - geolocalizá-los;
 - extração e identificação de *timestamps*;
 - correlacioná-los com a base histórica.
- Uso de tais características para análise, aprendizado e simulação de como os eventos de exceção impactam o transporte público por ônibus.

Objetivos

Gerais

Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo.

Específicos

- Identificar os eventos de exceção, quando existentes, dos *tweets* coletados.
- Extrair os endereços dos eventos de exceção identificados e geolocalizá-los.
- Construir uma base de dados pública com os dados processados, disponibilizada via API (para consumo e contribuição da comunidade de software), mantendo o modelo de dados consistente.
- Criação de plataforma para exploração e visualização dos dados coletados e processados do Twitter e da SPTrans.

- É possível identificar e categorizar os eventos de exceção de acordo com os tipos de eventos encontrados pela Revisão Sistemática.
- Extração de características com o auxílio de Processamento de Linguagem Natural (NLP — *Natural Language Processing*) em conjunto com dicionários auxiliares para o contexto dos eventos de exceção mencionados.
- Extração dos endereços dos *tweets* utilizando a técnica de Expressão Regular e posterior geolocalização.

Fundamentação Teórica

Cidades Inteligentes (SC — *Smart Cities*)

São cidades sustentáveis e socialmente inclusivas, que utilizam Tecnologias da Informação e Comunicação (TICs) para gerir eficientemente seus respectivos recursos naturais.

- *TDM — Technology Driven Method; top-down*; de fornecimento.
- *HDM — Human Driven Method; bottom-up*; de demanda.

Cidade

Complexo e dinâmico sistema sócio-técnico. Composto por sistemas urbanos, com espaços físicos para a vida cotidiana e com sistemas de infraestrutura.

- As TICs permeiam os sistemas urbanos e espaços físicos: dados voluntários, de sensores e Redes Sociais.
- Desafios relacionados a conectividade:
 - Infraestrutura de rede.
 - interoperabilidade;
 - padrões;
 - consumo de energia;
 - escalabilidade, dentre outros.
- Desafios relacionados aos dados:
 - Capacidade e local de armazenamento;
 - extração;
 - tratamento;
 - processamento;
 - análise;
 - integração;
 - agregação de dados, dentre outros.

Sistemas de Transporte Inteligentes (ITS — *Intelligent Transportation Systems*)

Tem como fim utilizar TICs para resolver problemas relacionados ao transporte, tais como congestionamento, segurança, eficiência e conservação ambiental.

- **Commercial Vehicles Operation (CVO)** — são sistemas utilizados para a segurança de veículos comerciais e frotas, por meio de tecnologias relacionadas a gerenciamento de tráfego, controle e gerenciamento de veículos e informações aos viajantes, tais como:
 - *Automatic Vehicles Location.*

GTFS — General Transit Feed Specification

É uma especificação de um formato comum para troca de informações estáticas sobre transporte público.

- *agency.txt*: agências de transporte público como fonte de dados.
- *stops.txt*: locais de embarque e desembarque.
- *routes.txt*: trajetos de um grupo de viagens.
- *trips.txt*: viagens de cada trajeto.
- *stop_times.txt*: horários de partida e chegada em paradas.
- *calendar.txt*: início, fim e dias disponíveis dos serviços.

Conceitos relacionados ao transporte público

- Acessibilidade.
- Acessibilidade universal.
- Mobilidade.
- Mobilidade urbana.
 - Transporte público coletivo;
 - transporte de alta capacidade;
 - acessibilidade universal nos passeios e edificações;
 - prioridade ao transporte coletivo no sistema viário;
 - terminais de transporte intermodais;
 - rede de transporte coletivo por ônibus (com acessibilidade universal);
 - rede cicloviária;
 - bicicletários e paraciclos;
 - legibilidade dos sistemas de orientação;
 - comunicação eficaz com os usuários;
 - modicidade tarifária;
 - logística eficiente no transporte de carga, dentre outros itens.

Redes Sociais

As Redes Sociais podem ser definidas como redes que possuem muitos relacionamentos, com grandes componentes conectados, altos coeficientes de agrupamento e grau de reciprocidade. Ex.: Facebook.

Rede de Informações

Nesse tipo de rede a interação dominante é a disseminação de informações entre os relacionamentos, com baixo índice de reciprocidade. Ex.: Twitter.

Processamento de Linguagem Natural

Explora como computadores podem ser utilizados para entender e manipular texto ou fala em linguagem natural, o que envolve conhecimento interdisciplinar principalmente entre as áreas de ciência da computação, linguística e estatística.

- Problemas de baixo nível (comuns a NLP).
 - *Sentence boundary disambiguation*;
 - *Tokenization*;
 - *Part-of-speech tagging*;
 - dentre outros.
- Problemas de alto nível (específicos e com base nos problema de baixo nível).
 - *Spelling / grammatical error identification and recovery*;
 - *Named Entity Recognition*;
 - *Word Sense Disambiguation*;
 - dentre outros.

Feature Engineering

Processo iterativo de construção, extração e seleção de características (features), o qual depende do conhecimento de domínio e de suas respectivas métricas.

- Características (*features*) binárias, categóricas ou contínuas.
- Pré-processamento: técnicas de padronização, normalização, remoção de ruídos, redução de dimensionalidade, discretização, expansão, etc.

- Supervisionados;
 - *Artificial Neural Network*;
 - *Decision Tree*;
 - *Decision Rules Classification*;
 - *K-nearest neighbor (k-NN)*;
 - *Fuzzy correlation*;
 - *Genetic Algorithm*;
 - *Naïve Bayes Algorithm*;
 - *Rocchio's Algorithm*;
 - *Support Vector Machine*.
- não-supervisionados;
- semi-supervisionados;
- por reforço.

Revisão Sistemática

Quais os tipos de problemas urbanos abordados utilizando processamentos de *tweets*? (QP1)

- *e-Participation*.
- Detecção de zoneamento urbano.
- Identificação de pontos de interesse.
- Mobilidade.
- Padrões demográficos.
- Poluição.
- Segurança Pública.
- Turismo.
- Tráfego.

Casos de uso relacionados ao transporte público (QP2)

- Impacto de eventos no transporte público.
 - Impacto dos ataques terroristas em Paris no uso do transporte público.
 - Impacto de eventos relacionados ao tráfego na demanda por bicicletas, em Nova Iorque e Washington D.C, EUA.
 - Impacto dos pontos de interesse na demanda por transporte público.
 - Impacto dos eventos anormais nas tomadas de decisão dos passageiros do Metrô de Tokyo.
 - Predição de fluxo de passageiros no Metrô de Nova Iorque.
- Planejamento e gestão do transporte público.
 - Análise de sentimento relacionada ao acesso ao transporte público.
 - Coleta de informações relacionadas ao transporte público.
 - Identificação de locais para estações de bicicletas, em St. Petersburg, Rússia.
 - Identificação da disposição dos usuários para realizar viagens de lazer.
 - Plataforma para notificação de problemas relacionados ao transporte público de Bangalore, Índia.

Técnicas estatísticas utilizadas no processamento de *tweets* (QP3)

- Análise de métricas relacionadas a desempenho.
- *Cosine similarity*.
- F_1 score.
- *Term frequency–inverse document frequency* (TF-IDF).
- *Inverse coefficient of variation*.
- *Jackknife resampling*.
- *Linear Regression*.
- *Local Indicators of Spatial Association* (LISA).
- *Local Moran's*.
- *Maximum likelihood estimation*.
- *Seasonal Autoregressive Integrated Moving Average* (SARIMA).
- *Optimization and Prediction with hybrid loss function*.

Paradigmas de processamento (QP4)

- *Batch processing* (offline).
- *Near Real Time*.
- *Real Time*.

Eventos de exceção relacionados ao transporte público (QP5)

- Acidentes.
 - Acidentes nas estações transporte.
 - Incêndio.
- Espaço-temporais.
 - Dia da semana.
 - Hora do dia.

Eventos de exceção relacionados ao transporte público (QP5)

- Eventos sociais.
 - Feiras de rua.
 - Festivais.
 - Jogos esportivos.
 - Passeatas e maratonas.
- Eventos urbanos.
 - Relacionados ao tráfego.
- Desastres naturais.
 - Tempestades.
 - Terremoto.
 - Tufões.
- Meteorológicas.
 - Dia claro, nublado, chuvoso, nevando, com neblina.
 - Temperatura do ar.

Técnicas de Aprendizado de Máquina utilizadas no processamento de *tweets* (QP6)

- *Bayes classification.*
- *C5.0 algorithm.*
- *Conditional Random Field (CRF) with Logistic Regression.*
- *Event extraction based on tweet hashtags.*
- *Latent Dirichlet Allocation (LDA).*
- *Monte Carlo simulation.*
- *PairFac (técnica inovadora que utiliza Tensor Factorization).*
- *Random Forest classification.*
- *Support Vector Machine.*
- *Self-Organizing Maps.*

Proposta de pesquisa

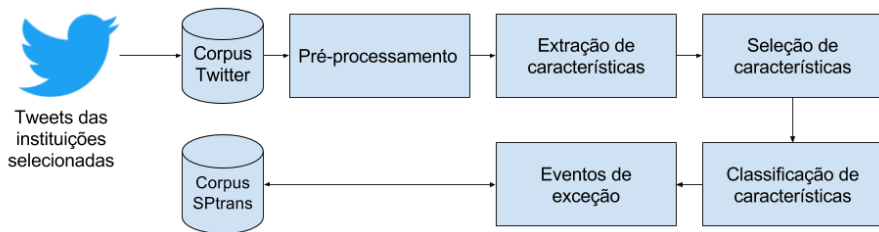
Formalização do problema

O problema de caracterização de eventos de exceção e de seus respectivos impactos envolve a fase conhecida como *feature extraction*. *Feature extraction* consiste na extração de um conjunto de características $\alpha = \{\chi_1, \chi_2, \dots, \chi_n\}$ a partir de um dado de entrada χ . Sendo assim, nessa proposta de pesquisa pretendemos extrair o conjunto de características $E = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$, referente a cada evento de exceção, e o conjunto $I_{\varepsilon_i} = \{\iota_{1\varepsilon_i}, \iota_{2\varepsilon_i}, \dots, \iota_{n\varepsilon_i}\}$, contendo as características de cada impacto decorrente de um determinado evento de exceção $\varepsilon_i \in E$.

Tem-se também como problema a correlação de cada evento de exceção com o seu respectivo impacto, permitindo assim uma análise histórica para identificação de padrões de causa e consequência.

$$\forall \iota \in I, \exists \varepsilon \in E (\iota = f(\varepsilon))$$

Proposta de pesquisa



Expressão regular para extração de endereços:

$$ER = \{L_1|S_1|L_2|S_2|...|L_n|S_n\}\{[a-zA-Z\setminus s]^+\} \quad (1)$$

Geolocalização dos endereços usando a API do Google Geocoding.

Tabela: Intervalo de tempo e número de *tweets* coletados

Profile no Twitter	# tweets ^a	Timestamp 1 ^b	Timestamp 2 ^c
BombeirosPMESP	5750	2017-05-21 02:10:39	2017-10-29 23:07:08
CETSP_	5042	2017-02-20 14:07:04	2017-10-29 21:45:54
CPTM_oficial	5435	2017-04-24 13:00:17	2017-10-29 10:00:40
governosp	5450	2017-05-10 17:00:05	2017-10-29 22:00:03
metrosp_oficial	7296	2017-06-07 17:23:34	2017-10-29 17:48:12
Policia_Civil	3360	2015-04-15 17:44:44	2017-10-27 10:01:53
PMESP	3956	2016-06-02 17:21:32	2017-10-29 20:25:37
saopaulo_agora	3671	2016-11-18 07:36:12	2017-10-29 20:56:28
smtsp_	1128	2017-04-26 10:44:26	2017-10-29 23:00:11
SPCEDEC	945	2015-06-09 10:50:23	2017-10-29 23:38:36
sptrans_	8447	2017-06-13 15:19:56	2017-10-29 22:01:44
TurismoSaoPaulo	3308	2012-06-12 22:00:38	2017-10-27 17:46:59
Total	53788	-	-

^a Número de *tweets* coletados.

^b *Timestamp* mais antigo.

^c *Timestamp* mais recente.

Tabela: Conjuntos e quantidades de dados especificados em GTFS pela SPTrans

Conjunto de dados	Quantidade de dados
<i>agency.txt</i>	1
<i>calendar.txt</i>	6
<i>fare_attributes.txt</i>	6
<i>fare_rules.txt</i>	5.400
<i>frequencies.txt</i>	39.625
<i>routes.txt</i>	291.634
<i>shapes.txt</i>	800.767
<i>stop_times.txt</i>	95.134
<i>stops.txt</i>	19.933
<i>trips.txt</i>	2.273
Total	1.254.779

Tabela: Quantidade de dados enviados pelos módulos AVL, por *id* de viagem

trip_id	Qtd. de dados ^a	Timestamp 1 ^b	Timestamp 2 ^c
4779-10-0	259.382	2016-09-13 08:24:57.936Z	2017-09-02 02:11:42.274Z
4779-10-1	271.671	2016-09-13 08:24:57.937Z	2017-09-02 02:11:42.285Z
917H-10-0	256.648	2016-09-13 08:25:59.943Z	2017-09-02 02:11:42.250Z
Total	787.701	-	-

^a Quantidade de dados.

^b *Timestamp* mais antigo.

^c *Timestamp* mais recente.

- **Case folding:** processamento de normalização de todas as letras do texto (de a-z) para minúsculas.
- **Tokenization:** processamento realizado para obtenção das palavras (*tokens*) que compõem uma sentença, inclui a remoção de números, pontuações e caracteres que não pertencem ao alfabeto.
- **Remoção de stopwords:** processamento para remoção do conjunto de *tokens* de palavras sem significado ou importância, o que reduz a quantidade de ruído do conteúdo *tweet*.
- **Stemming:** processamento para encontrar a raiz de uma palavra, removendo sufixos e prefixos (no caso do Português Brasileiro) das palavras derivadas.

Extração, seleção e classificação de características

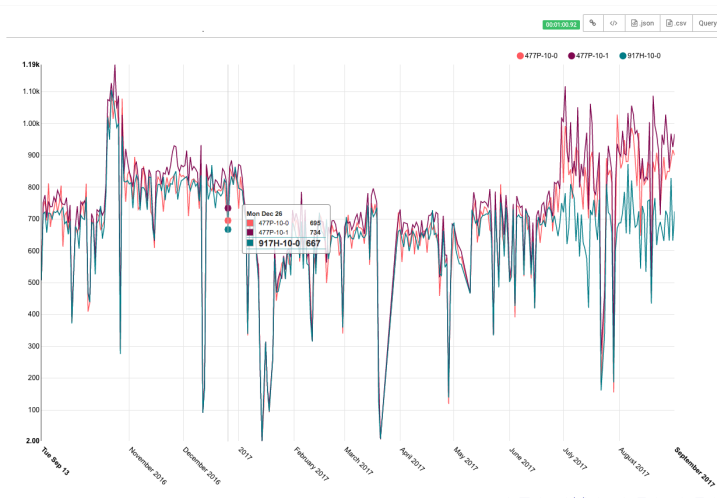
- Extração de características: pretendemos na primeira iteração para extração de características usar os *tokens* (*features*) obtidos no pré-processamento para selecionarmos as palavras mais frequentes (*features*) para cada conjunto de dados do *Corpus Twitter*. Nas iterações seguintes, planejamos analisar as *features* selecionadas, combiná-las entre si e derivar novas *features*, de acordo com o conhecimento do domínio.
- Seleção de características: pretendemos selecionar as *features* mais relevantes utilizando a medida estatística *tf-idf* (*term frequency-inverse document frequency*) para obtermos os termos mais frequentes de cada conjunto de dados do *Corpus Twitter*.
- Classificação de características: classificação manual de 30% do *Corpus Twitter* e análise dos algoritmos de aprendizado de máquina elencados pela revisão sistemática para classificação automatizada dos 70% restantes.

Correlação dos eventos de exceção com os dados AVL da SPTrans

- Atraso médio induzido nas viagens.
- Ônibus frequentemente afetados por eventos de exceção.
- Ônibus frequentemente afetados por determinado evento de exceção.
- Padrão de ocorrência dos eventos de exceção no espaço-tempo (localizações e *timestamps*).
- Quantidade e viagens afetadas.
- Quantidade e regiões da cidade de São Paulo afetadas.
- Viagens frequentemente afetadas por eventos de exceção.
- Viagens frequentemente afetadas por determinado evento de exceção.

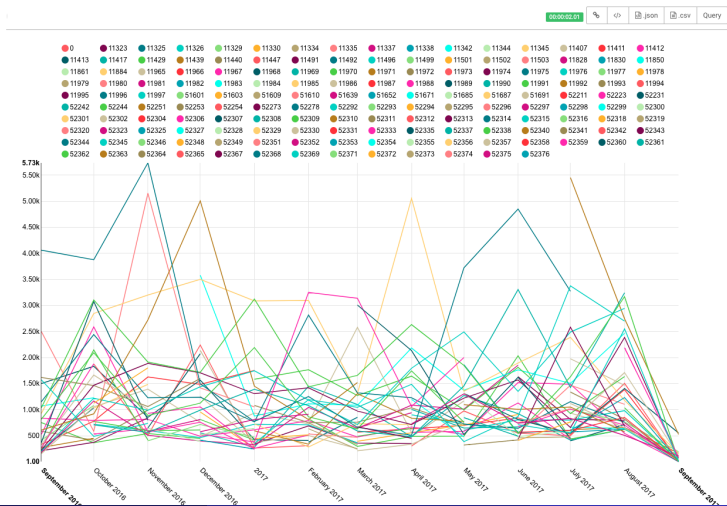
Exploração e visualização do conjunto de dados

Figura: Quantidade de dados enviados: viagens 477P-10-0, 477P-10-1 e 971H-10-0



Exploração e visualização do conjunto de dados

Figura: Quantidade de dados enviados por ônibus / mês: viagens 477P-10-0, 477P-10-1 e 971H-10-0



Exploração e visualização do conjunto de dados

Figura: Localizações de envio dos dados: viagens 477P-10-0, 477P-10-1 e 971H-10-0

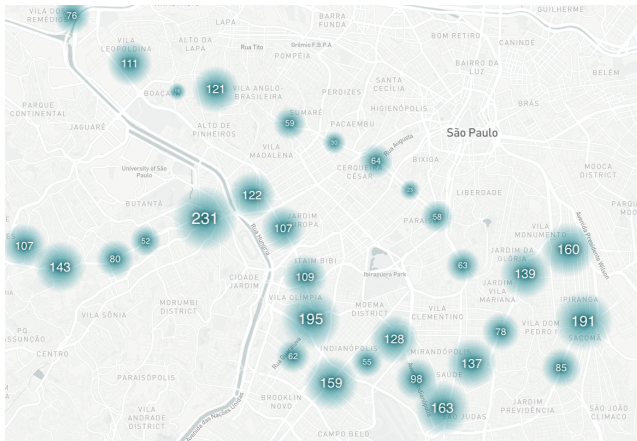


Tabela: Cronograma de atividades

Número	Atividade Descrição	2017												2018					
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
1	Revisão bibliográfica	X	X	X	X	X	X												
2	Desenvolvimento de protótipo			X	X	X	X												
3	Construção do conjunto de dados			X	X	X	X	X	X	X	X	X							
4	Implementação da solução proposta												X	X	X	X	X	X	
5	Avaliação dos resultados													X		X		X	
6	Escrita de artigo														X	X			
7	Escrita da dissertação			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

- Implementação da solução proposta.
 - Identificação dos eventos de exceção (pré-processamento, *feature extraction* e *feature selection* dos *tweets* coletados).
 - Estudo dos algoritmos de classificação e implementação de um artefato de *software* para classificação dos *tweets* de acordo com seus respectivos eventos de exceção.
 - Correlação dos eventos de exceção com os dados AVL da SPTrans.

Contribuições esperadas

- Uma solução para o problema de caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo, por meio de *tweets* e de dados históricos dos módulos AVL do SIM;
- disponibilizar os conjuntos de dados que foram construídos;
- uma plataforma para que esses dados possam ser visualizados e explorados, de forma a contribuir com projetos e pesquisas futuras correlatas;
- submissão de artigos com os resultados obtidos para veículos de disseminação de conhecimento científico nas áreas de: Análise de Redes Sociais, Sistemas de Transporte Inteligentes, Cidades Inteligentes.

Limitações e riscos à validade do estudo

- Processamento de *tweets* em português brasileiro e oriundos das contas selecionadas, o que pode tornar a solução não generalista.
- É possível que sejam encontrados novos desafios que inviabilizem o uso de Expressão Regular para extração dos endereços dos *tweets*.