UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

FELIPE CORDEIRO ALVES DIAS

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

São Paulo
2017

FELIPE CORDEIRO ALVES DIAS

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

Versão original

Texto de Exame de Qualificação apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Orientador: Prof. Dr. Daniel de Angelis Cordeiro

São Paulo

2017

Resumo

DIAS, Felipe Cordeiro Alves. **Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo**. 2017. 107 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

A cidade de São Paulo é o município mais populoso do Brasil, caracterizado por uma segregação urbana responsável por inúmeros problemas relacionados a mobilidade urbana. As ações atuais para resolver os problemas de mobilidade urbana têm pouco aprofundamento em questões tecnológicas e melhorias dos sistemas computacionais existentes – como as necessárias ao desfasado Sistema Integrado de Monitoramento e Transporte (SIM), utilizado para gestão e monitoramento do transporte público por ônibus de São Paulo. Uma das possíveis melhorias é integrar o SIM às Redes Sociais. Com essa perspectiva de integração, esse trabalho tem como objetivo utilizar *tweets* e dados do SIM na caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo. Para alcançar tal objetivo, esse trabalho propõe utilizar *tweets* publicados por instituições governamentais responsáveis por reportar eventos de exceção e dados dos módulos AVL (*Automatic Vehicle Location*) do SIM, responsáveis por rastrear e localizar os ônibus do município. A hipótese é de que é possível identificar e localizar eventos de exceção nos *tweets* por meio de Processamento de Linguagem Natural e Expressão Regular, e correlacionar esses eventos com os dados históricos do SIM.

Palavras-chaves: Cidades Inteligentes. Transporte Público. Sistemas de Transporte Inteligentes. Eventos de exceção.

Sumário

1 Introdução

Neste capítulo, são apresentadas as seções referentes à motivação da proposta de pesquisa; sobre a definição do problema que pretendemos abordar; a respeito dos objetivos gerais e específicos; sobre as hipóteses a serem verificadas e sobre a organização dos capítulos desse documento.

1.1 Motivação

A cidade de São Paulo é o município mais populoso do Brasil, que passou por um rápido processo de urbanização e tem população atual estimada em 12.106.920 milhões de habitantes (com data de referência em 1º de julho de 2017)[1]. Desse total de habitantes, 10% vivem na área do Centro Expandido (CE) e 90% no Cinturão Periférico (CP) (SÁ, T. H. et al., 2017), o que caracteriza uma segregação urbana responsável por inúmeros problemas relacionados a mobilidade urbana.

Um desses problemas é conhecido como o movimento pendular, no qual longas distâncias são percorridas diariamente pelos moradores do CP para acessar os locais de emprego, educação e serviços localizados em maioria no CE. Além disso, o movimento pendular torna o CP uma região dormitória, com parte de seus respectivos moradores dependentes do Sistema de Transporte Público para acessar o CE.

Devido aos problemas de mobilidade urbana existentes no Brasil, como os da cidade de São Paulo, a Lei Federal 12.587/2012[2], relacionada ao Programa de Aceleração do Crescimento (PAC)[3], obrigou os municípios a enviarem seus respectivos planos de mobilidade urbana até o final do ano de 2015, visando promover o desenvolvimento sustentável com a mitigação dos custos ambientais e socioeconômicos dos deslocamentos de pessoas. Considerando essa lei, o Plano de Mobilidade de São Paulo (*PlanMob/SP 2015*) foi instituído pelo Decreto 56.834[4], como instrumento

[1] <https://agenciadenoticias.ibge.gov.br/media/com_mediaibge/arquivos/9bc1a0065c49fd6f81dc785b2b8d8c35.xlsx>. Acesso em Outubro, 29 de 2017.

[2] <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12587.htm>. Acesso em Outubro, 29 de 2017.

[3] <http://www.pac.gov.br>. Acesso em Outubro, 29 de 2017.

[4] <http://www.prefeitura.sp.gov.br/cidade/secretarias/transportes/planmob>. Acesso em Outubro, 29 de 2017.

de planejamento e gestão do Sistema Municipal de Mobilidade Urbana para os próximos 15 anos.

No *PlanMob/SP 2015*, a Secretaria Municipal de Transportes (SMT) propõe criar uma central de monitoramento conhecida como Central Integrada de Mobilidade Urbana (CIMU), que tem como objetivo integrar as áreas de trânsito e transporte subordinadas à SMT. Nessa proposta, observam-se os seguintes problemas que poderiam ser resolvidos em paralelo ao desenvolvimento do CIMU: (I) a CIMU não processa conteúdo de Redes Sociais, (II) não aborda melhoria dos sistemas computacionais já existentes e (III) será integrada com o desfasado Sistema Integrado de Monitoramento e Transporte (SIM), da São Paulo Transportes (SPTrans), responsável pelo monitoramento da infraestrutura de ônibus.

O SIM utiliza a tecnologia *Automatic Vehicle Location* (AVL) para localizar e rastrear os ônibus, fornecer informações em tempo real aos passageiros (RTPI — *Real Time Passenger Information*), monitorar 1.353 rotas de ônibus[5], 10 corredores de ônibus[6], 28 terminais de ônibus[7] e 19.933 mil paradas de ônibus[5] que serviram em 2016 a aproximadamente 8 milhões de passageiros por dia[8]. Apesar da importância do SIM, há inúmeras defasagens tecnológicas (que causam discrepância nas informações recebidas pelos usuários, dentre outros problemas) (CONSULO et al., 2016), que precisariam ser resolvidas antes de integrá-lo ao CIMU.

Sistemas como o SIM são classificados como Sistemas de Transporte Inteligente (ITS — *Intelligent Transport System*), e normalmente estão presentes nas Cidades Inteligentes (SC — *Smart Cities*). Por definição, ITS utilizam Tecnologias da Informação e Comunicação (TIC) para explorar dados capazes de contribuir com a melhoria da segurança, do gerenciamento, eficiência dos transportes e redução do impacto ambiental (ANTTIROIKO, 2013). Com isso, nota-se que ITS são essenciais para os objetivos mencionados na Lei Federal 12.587/2012 e no PlanMob/SP 2015.

No entanto, a lei de mobilidade urbana (12.587/2012) e o *PlanMob/SP 2015* não mencionam explicitamente ITS e TIC. O conteúdo de ambos os documentos tem um viés político-urbano, com pouco aprofundamento em questões tecnológicas e melhorias dos sistemas já existentes. Esse cenário é diferente em alguns países,

[5] <http://www.sptrans.com.br/desenvolvedores>. Acesso em Outubro, 29 de 2017.
[6] <http://www.sptrans.com.br/terminais/corredores.aspx>. Acesso em Outubro, 29 de 2017.
[7] <http://www.sptrans.com.br/terminais>. Acesso em Outubro, 29 de 2017.
[8] <http://www.sptrans.com.br/indicadores>. Acesso em Outubro, 29 de 2017.

nos quais existem planejamentos para o transporte e mobilidade urbana que estão explicitamente relacionados ao desenvolvimento e uso de novas tecnologias.

Por exemplo, os EUA têm o plano estratégico para 2015-2019 em ITS, abordando temas como veículos conectados, automação, uso de tecnologias emergentes (para apoiar decisões em tempo real), intregação de dados corporativos, interoperabilidade (comunicação entre diferentes sistemas) e entrega acelerada de projetos (United States Department of Transportation, 2017). Já a União Européia e o Japão estão centrados em padronizações de tecnologias em ITS, com o obtjetivo de serem referências nesse setor (CONSULO et al., 2016).

O contraste entre os dois parágrafos anteriores talvez seja devido ao fato de a legislação brasileira e os planos para mobilidade urbana terem sido estabelecidos como consequência do crescimento urbano acelerado e sem planejamento. Ou seja, como solução paliativa para um problema urbano, o que difere dos planos em ITS mencionados, que têm como foco otimizar o transporte e criar padrões tecnológicos.

Apesar dessas diferenças políticas e sociais, o transporte público pode se beneficiar ao explorar ITS (NELSON; MULLEY, 2013), e ao integrar as Redes Sociais com o planejamento, gestão e as atividades operacionais dos transportes públicos, abordando seus respectivos fatores sócio-técnicos (KUFLIK et al., 2017). Por exemplo, um dos benefícios possíveis é o de se conseguir analisar o impacto dos eventos de exceção na operação do sistema de transporte público por ônibus na cidade de São Paulo, usando dados do SIM (AVL) e de Redes Sociais.

1.2 Definição do problema

Eventos de exceção tais como acidentes, greves, falhas na operação do metrô, manifestações, enchentes, eventos sociais, dentre outras, podem comprometer muitos trechos do sistema de transporte público e, dependendo da proporção do impacto causado pela exceção, inúmeras pessoas podem ser afetadas. Tais eventos de exceção e seus respectivos impactos possuem características que podem ser identificadas visando melhor gestão dessas ocorrências.

Com a identificação dessas características, é possível conhecer previamente quais seriam os impactos decorrentes de um determinado evento de exceção no

funcionamento normal do transporte público. Tais características podem ser obtidas analisando o histórico do funcionamento do sistema de transportes, e utilizadas posteriormente em simulações de como o sistema responderia a determinados eventos de exceção.

Os dados históricos existentes para essa análise são os do SIM, obtidos utilizando AVL. No entanto, analisá-los envolve problemas como o (I) grande volume de dados, em virtude da frequência com que são enviados (II) e os referentes ao comprometimento da qualidade dos dados enviados, como consequência dos problemas e limitações do *hardware* responsável pela transmissão.

O uso de conteúdo de Redes Sociais pode ajudar a abordar os problemas anteriormente mencionados, o qual delimitaria o escopo da análise histórica para a identificação das características dos eventos de exceção e dos seus respectivos impactos. Usar o conteúdo de Redes Sociais envolve alguns desafios como o de (I) identificar eventos de exceção nas publicações, (II) geolocalizá-los, (III) determinar seus *timestamps* (IV) correlacioná-las com a base histórica.

1.3 Objetivos

O objetivo geral desse projeto de pesquisa é a caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo. Visando alcançar esse objetivo, serão coletados *tweets* das contas oficiais das instituições governamentais responsáveis por reportar eventos de exceção na cidade de São Paulo. Todas as contas selecionadas do *Twitter* estão listadas na tabela 1. Também, serão utilizados os dados históricos dos módulos AVL do SIM.

Além disso, temos como objetivos 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. Com isso, a necessidade de entrega dos dados a sociedade, apontada por (KUFLIK et al., 2017), será atendida.

- Criação de plataforma para exploração e visualização dos dados coletados e processados das fontes citadas na tabela 1 e da SPTrans.

Tabela 1 – Descrição e nome dos profiles selecionados do Twitter

Descrição do *profile* no *Twitter*	**Profile no *Twitter***
Comando do Corpo de Bombeiros da PMESP [a]	*@BombeirosPMESP*
Companhia de Engenharia de Tráfego de SP	*@CETSP_*
Companhia Paulista de Trens Metropolitanos	*@CPTM_oficial*
Defesa Civil do Estado de São Paulo	*@SPCEDEC*
Governo do Estado de São Paulo	*@governosp*
Metrô de São Paulo	*@metrosp_oficial*
Polícia Cívil do Estado de São Paulo	*@Policia_Civil*
Polícia Militar do Estado de São Paulo	*@PMESP*
São Paulo Agora — CCOI [b]	*@saopaulo_agora*
São Paulo Transporte	*@sptrans_*
São Paulo Turismo	*@TurismoSaoPaulo*
Secretaria de Transportes de São Paulo	*@smtsp_*

[a] Polícia Militar do Estado de São Paulo.
[b] Centro de Controle Integrado 24 Horas da Cidade de São Paulo

Fonte: Felipe Cordeiro Alves Dias

1.4 Hipóteses

Com base na Revisão Sistemática do Cap. 3, os eventos de exceção presentes nos *tweets* podem ser caracterizados, não exaustivamente, em:

1. **Acidentes**.
   a) Acidentes nas estações de transporte (ITOH et al., 2016).
   b) Incêndio (ITOH et al., 2016).
2. **Espaço-temporais**.
   a) Dia da semana (CHEN et al., 2016).
   b) Hora do dia (CHEN et al., 2016).
3. **Eventos sociais**.
   a) Feiras de rua (CHEN et al., 2016).
   b) Festivais (CHEN et al., 2016), (LECUE et al., 2014).
   c) Jogos esportivos (CHEN et al., 2016), (GAL-TZUR et al., 2014).

   d) Passeatas e maratonas (CHEN et al., 2016), (ITOH et al., 2016).

4. **Eventos urbanos**.

   a) Relacionados ao tráfego (CHEN et al., 2016); (LECUE et al., 2014).

5. **Desastres naturais**.

   a) Tempestades (ITOH et al., 2016).
   b) Terremoto (ITOH et al., 2016).
   c) Tufões (ITOH et al., 2016).

6. **Metereológicas**.

   a) Dia claro, nublado, chuvoso, nevando, com neblina (CHEN et al., 2016).
   b) Temperatura do ar (CHEN et al., 2016).

Dito isso, espera-se que seja possível identificar tais características utilizando Processamento de Linguagem Natural (NLP — *Natural Language Processing*) em conjunto com dicionários auxiliares para o contexto dos eventos de exceção mencionados.

Após a identificação dos eventos de exceção, temos como hipótese que seja possível extrair, com confiabilidade, os endereços dos *tweets* utilizando a técnica de Expressão Regular. Pois em uma análise preliminar observamos que o conteúdo das contas selecionadas, citadas na tabela 1, utilizam padrões de formatação para os endereços publicados. Com isso, podemos afirmar que esses *tweets* apresentam a característica de serem semi-estruturados, diferentemente dos *tweets* não estruturados publicados pelos usuários comuns do *Twitter*; o que consequentemente simplifica o processamento necessário para geolocalizar os eventos de exceção.

1.5 Organização do documento

Neste documento, é apresentado o Cap. 1 sobre a introdução do trabalho; o Cap. 2 a respeito da fundamentação teórica; Cap. o 3 sobre a revisão sistemática realizada; o Cap. 4 referente a proposta de pesquisa e o Cap. 5 contendo a conclusão da proposta apresentada.

2 Fundamentação Teórica

Neste capítulo, são apresentados fundamentos teóricos sobre os conceitos Cidades Inteligentes; Sistemas de Transporte Inteligentes; relacionados ao transporte público; *General Transit Feed Specification*; Redes Sociais; Processamento de Linguagem Natural; *Feature Engineering* e Aprendizado de Máquina.

2.1 Cidades Inteligentes

Embora não haja concenso, o conceito de Cidades Inteligentes (SC — *Smart Cities*) tem sido definido pela literatura principalmente como cidades sustentáveis e socialmente inclusivas (WANG; SINNOTT; NEPAL, 2016), que utilizam Tecnologias da Informação e Comunicação (TICs) para gerir eficientemente seus respectivos recursos naturais, de energia, transporte, lixo, dentre outros (AHVENNIEMI et al., 2017). As SC podem ter viés tecnológico (*TDM — Technology Driven Method; top-down*; de fornecimento), ou, humano (*HDM — Human Driven Method; bottom-up*; de demanda) (KUMMITHA; CRUTZEN, 2017).

O aspecto humano das Cidades Inteligentes começou a ser explorado recentemente, após críticas referentes aos poucos indicadores humanos existentes para SC (AHVENNIEMI et al., 2017) (FINGER; RAZAGHI, 2017). A abordagem humana das SC foca questões sociais e qualidade de vida, tais como governança participativa, segurança, cultura, lazer, sustentabilidade, desenvolvimento de capital humano, dentre outras (AHVENNIEMI et al., 2017). Na perspectiva tecnológica de SC, argumenta-se que apenas o uso de TICs seja capaz viabilizar o desenvolvimento de capital humano e de soluções para os problemas da cidade (KUMMITHA; CRUTZEN, 2017).

Independentemente dos vieses humano e tecnológico, a cidade pode ser conceituada como um complexo e dinâmico sistema sócio-técnico. Ou seja, uma cidade (região metropolitana) é composta por sistemas urbanos, com espaços físicos para a vida cotidiana e com sistemas de infraestrutura (para transporte, energia, água e tratamento de água, moradia, telecomunicações e áreas verdes). Os sistemas urbanos por natureza nunca estão em equilíbrio, possuem subsistemas imprevisíveis (FINGER; RAZAGHI, 2017).

Apesar disso, as TICs permeiam os sistemas urbanos e espaços físicos, o que tem sido acentuado com o crescente número de sensores e dispositivos conectados à Internet (*IoT — Internet of Things*), de dados voluntários enviados por pessoas via dispositivos móveis e, de conteúdo existente em Redes Sociais sobre os acontecimentos da cidade. Tais fontes heterogêneas geram grandes volumes de dados, utilizados para desenvolver serviços de Cidades Inteligentes (FINGER; RAZAGHI, 2017) (ANG et al., 2017).

O desenvolvimento de serviços de SC envolve desafios relacionados a conectividade (infraestrutura de rede, interoperabilidade e padrões, consumo de energia e escalabilidade) e aos dados (capacidade e local de armazenamento, extração, tratamento, processamento, análise, integração e agregação dos dados) (ANG et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017). Além disso, a análise de dados pode tanger problemas referentes a correlação e inferência de dados de diferentes domínios, aprendizado de máquina, processamento em tempo real e propostas de novo uso para dados provenientes de infraestruturas já existentes (ANG et al., 2017).

Por fim, a seguir estão elencadas algumas frentes de estudo e de desenvolvimento de serviços de SC que ilustram iniciativas em Cidades Inteligentes:

- ***Smart buildings*** (TALARI et al., 2017), (MORENO et al., 2017), (ANG et al., 2017), (FINGER; RAZAGHI, 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017).
- ***Smart citizen / community / people*** (TALARI et al., 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).
- ***Smart economy*** (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017), (AHVENNIEMI et al., 2017).
- ***Smart environment*** (*electricity*, *waste*, *water*, *green space*) (SANTOS et al., 2017), (FINGER; RAZAGHI, 2017), (TALARI et al., 2017), (ANG et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).
- ***Smart governance*** (TALARI et al., 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).

- ***Smart living*** (*education*, *health*, *safety*, *cultural*) (SANTOS et al., 2017), (TALARI et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017), (AHVENNIEMI et al., 2017).
- ***Smart transportation / mobility*** (TALARI et al., 2017), (MORENO et al., 2017), (ANG et al., 2017), (FINGER; RAZAGHI, 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).

2.2 Sistemas de Transporte Inteligentes

Sistemas de Transporte Inteligentes (ITS — *Intelligent Transportation Systems*) é uma das mais antigas tecnologias presentes em Cidades Inteligentes (MENOUAR et al., 2017), que tem como fim utilizar TICs para resolver problemas relacionados ao transporte, tais como congestionamento, segurança, eficiência e conservação ambiental (FIGUEIREDO et al., 2001).

É importante notar a diferença entre o termo *Intelligent* e *Smart* de *Smart transportation / mobility*, o primeiro, respectivamente, refere-se apenas ao uso de tecnologias, enquanto que o segundo ao uso de TICs para transformar de forma significativa a vida cotidiana das pessoas (ALBINO; BERARDI; DANGELICO, 2015). A seguir, algumas das categorias de ITS estão enumeradas:

1. ***Advanced Traffic Management System* (ATMS)** — são sistemas utilizados para melhorar a qualidade do serviço de tráfego e redução de atrasos (FIGUEIREDO et al., 2001), por meio de:
    a) *Collection data team*: equipe de pessoas responsáveis por monitorar e coletar dados das condições de tráfego.
    b) *Support systems*: conjunto de câmeras, semáforos, sensores, dentre outros dispositivos auxiliares para gerenciar e controlar o tráfego em tempo real.
    c) *Real time traffic control systems*: sistemas utilizados para com base nos dados coletados controlar acesso a avenidas, semáforos, envio de mensagens para os dispositivos de monitoramento.

2. ***Advanced Travellers Information Systems* (ATIS)** — são sistemas utilizados para fornecer informação em tempo real aos viajantes (FIGUEIREDO et al., 2001).
3. ***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 (FIGUEIREDO et al., 2001), tais como:
    a) *Automatic Vehicles Identification*.
    b) *Automatic Vehicles Classification*.
    c) *Automatic Vehicles Location*.
    d) *Pedestrian Movement Detection*.
    e) *Board Computers*.
    f) *Real Time Traffic Transmissions*.
4. ***Advanced Public Transportations Systems* (APTS)** — são sistemas que utilizam ATMS e ATIS para melhorar a eficiência e operação do transporte público coletivo (FIGUEIREDO et al., 2001). É importante observar que APTS também podem utilizar CVO.
5. ***Advanced Vehicles Control Systems* (AVCS)** — são sistemas compostos por sensores, computadores e sistemas de controle para auxiliar e alertar motoristas, com o objetivo de melhorar a segurança e reduzir congestionamentos (FIGUEIREDO et al., 2001).

As categorias mencionadas anteriormente representam parte da primeira geração de tecnologias em ITS, a próxima geração tem como foco veículos autônomos e conectados, capazes de trocarem informações entre si em tempo real para melhorar a segurança dos condutores (MENOUAR et al., 2017).

2.3 Conceitos relacionados ao transporte público

Esta seção define os conceitos relacionados ao transporte público, de acordo com a perspectiva do Plano de Mobilidade Urbana do Município de São Paulo — PlanMob SP 2015[4].

2.3.1 Acessibilidade

A acessibilidade pode ser considerada como um atributo do espaço urbano, o qual é diretamente proporcional a abrangência e adequação das infraestruturas de acesso ao espaço urbano. As regiões da cidade têm diferentes padrões de infraestrutura de transporte e deslocamento, portanto, são diferenciadas no aspecto de acessibilidade. Além disso, a acessibilidade atua como instrumento de acesso as oportunidades socioeconômicas da cidade. Observa-se que a acessibilidade não é entendida como um atributo econômico relacionado ao valor das tarifas do transporte, ou, as condições de uso (como o congestionamento viário).

Uma qualidade específica do espaço urbano é a acessibilidade universal, que o caracteriza como acessível a pessoas portadoras de necessidades especiais (PNEs). A acessibilidade universal é garantida ao eliminar as barreiras físicas que impedem a participação plena e efetiva das pessoas PNEs ao espaço urbano.

2.3.2 Mobilidade

A mobilidade por ser entendida como um atributo do indivíduo, o qual está relacionado a sua capacidade de se deslocar pelo território da cidade e a sua respectiva renda (dimensão econômica); ou seja, pessoas ou famílias de maior renda tendem a ter maior número de viagens. Além disso, observa-se que a restrição da mobilidade devido a má qualidade das infraestruturas urbanas é considerada como falta de acessibilidade ao espaço e não como perda de mobilidade do indivíduo.

A condição de mobilidade pode ser calculada pelo indicador conhecido como taxa ou índice de mobilidade, determinado pelo quociente entre o total de viagens realizadas e o total da população residente em uma região. Tal indicador pode ser especializado de acordo o tipo de mobilidade, por exemplo, ao considerar apenas as viagens motorizadas, obtém-se o índice de mobilidade motorizada; e ser caracterizado como crescente ou decrescente de acordo com fatores socioeconômicos.

Além da mobilidade como atributo do indivíduo, existe a mobilidade como atributo da cidade, conhecida como mobilidade urbana. A mobilidade urbana consi-

dera um conjunto de fatores de uma aglomeração urbana que tornam a mobilidade mais qualificada e eficiente, tais como:

1. Transporte público coletivo;
2. transporte de alta capacidade;
3. acessibilidade universal nos passeios e edificações;
4. prioridade ao transporte coletivo no sistema viário;
5. terminais de transporte intermodais;
6. rede de transporte coletivo por ônibus (com acessibilidade universal);
7. rede cicloviária;
8. bicicletários e paraciclos;
9. legibilidade dos sistemas de orientação;
10. comunicação eficaz com os usuários;
11. modicidade tarifária;
12. logística eficiente no transporte de carga, dentre outros itens.

2.3.3 Viagem e modais de transporte

O conceito de viagem no setor de transportes é definido como o deslocamento de uma pessoa entre dois pontos de interesse (origem e destino), com um motivo definido e por meio de um modal de transporte. A saber, os modais de transporte considerados no *PlanMob/SP 2015* estão enumerados a seguir:

1. A pé.
    a) Independentemente do deslocamento percorrido caso o motivo seja escola ou trabalho;
    b) Superior a 500 metros de deslocamento.
2. Coletivos.
    a) Metrô;
    b) ônibus;
    c) ônibus fretado;
    d) ônibus escolar e lotação;
    e) trem.

3. Individuais.

    a) Automóveis (bicicleta, carro particular, caminhão, moto e táxi).

2.4 General Transit Feed Specification

A *GTFS — General Transit Feed Specification*[1], como o próprio nome sugere, é uma especificação de um formato comum (o que permite interoperabilidade) para troca de informações estáticas sobre transporte público. Um *feed* especificado na GTFS estática é composto por arquivos de texto (que seguem determinados requisitos semelhantes aos do formato *CSV*[1]) compactados no formato *Zip*[2], e detalhados na tabela 2. Cada arquivo modela diferentes perspectivas do transporte público, tais como paradas, trajetos, viagens e outros dados relativos a horário.

Além da GTFS estática existe a GTFS-*realtime*[1], que é uma extensão da GTFS estática, assim, para usar *feeds* em tempo real é necessário definir os arquivos estáticos da GTFS, que são utilizados na GTFS-*realtime* para obter as informações do sistema de transporte público. A GTFS-*realtime* é utilizada para transmissões em tempo real de três tipos de *feeds*[1], enumerados e detalhados a seguir:

1. Atualizações dos horários de parada.

    a) Descritor de viagem: viagem programada (de acordo ou próxima a uma programação GTFS), adicionada (não programada e adicionada, por exemplo, para atender à demanda ou substituir um veículo quebrado), desprogramada (que está sendo feita e não está associada a uma programação, por exemplo, quando não há uma programação, e os ônibus rodam em um serviço de translado), cancelada (viagem programada, mas removida), substituição (substitui uma parte da programação estática).
    b) Indefinição: especifica o erro esperado no atraso real como um número inteiro, em segundos.

2. Alertas de serviço.

    a) Intervalo de tempo: o alerta será exibido eventualmente, no intervalo de tempo especificado.

[1] <https://developers.google.com/transit>. Acesso em Outubro, 29 de 2017.
[2] <https://support.pkware.com/display/PKZIP/APPNOTE>. Acesso em Outubro, 29 de 2017.

b) Seletor de entidade: agência (afeta toda a rede de transporte público), trajeto (afeta todo o trajeto), tipo de trajeto (afeta qualquer trajeto desse tipo, por exemplo, todos os ônibus), viagem (afeta uma viagem específica) e parada (afeta uma parada específica).
c) Causa: desconhecida, outra causa (não representada por nenhuma destas opções), problema técnico, greve, manifestação, acidente, feriado, tempo, manutenção, construção, atividade policial, emergência médica.
d) Efeito: sem serviço, serviço reduzido, atrasos significativos (atrasos não significativos só devem ser fornecidos por Atualizações de viagem), desvio, serviço adicional, serviço modificado, parada deslocada, outro efeito (não representado por qualquer uma dessas opções), efeito desconhecido.

3. Posições de veículos.

a) Posição: a posição contém os dados de localização na posição do veículo, com os campos obrigatórios latitude e longitude, e com os campos opcionais rumo (direção que o veículo está seguindo), odômetro (distância que o veículo percorreu) e velocidade (velocidade no momento medida pelo veículo, em metros por segundo).
b) Nível de congestionamento: congestionamento desconhecido, fluxo estável, paradas frequentes, congestionamento e congestionamento grave.
c) Status de parada do veículo: chegando em (o veículo está prestes a chegar na parada em questão), parado em (o veículo está parado na parada em questão), em direção a (a parada em questão é a próxima parada do veículo — padrão).
d) Descritor do veículo: id único (sistema de identificação interna do veículo), etiqueta de identificação (visível ao usuário) e placa real do veículo.

No demais, os *feeds* da GTFS-*realtime* são atualizados frequentemente, serializados em *Protocol Buffers*[3] e transmitidos via protocolo HTTP[4]. A estrutura dos dados é definida em um arquivo *gtfs-realtime.proto*[1], usado para gerar o modelo de dados dos *feeds* em diferentes linguagens de programação, tais como *Java*, *C++* ou *Python*.

[3] <https://developers.google.com/protocol-buffers>. Acesso em Outubro, 29 de 2017.
[4] <https://tools.ietf.org/html/rfc2616>. Acesso em Outubro, 29 de 2017.

Tabela 2 – Detalhamento dos arquivos da GTFS

Nome do arquivo	**Condicional**	**Contéudo** [a]
agency.txt	Obrigatório	Contém uma ou mais agências de transporte público como fonte dos dados.
stops.txt	Obrigatório	Contém os locais individuais em que os veículos pegam ou deixam passageiros.
routes.txt	Obrigatório	Contém os trajetos de um grupo de viagens exibidas aos passageiros como um único serviço.
trips.txt	Obrigatório	Contém as viagens de cada trajeto. Uma viagem é uma sequência de duas ou mais paradas que ocorrem em um horário específico.
stop_times.txt	Obrigatório	Contém os horários de partida e chegada dos veículos em paradas específicas em cada viagem.
calendar.txt	Obrigatório	Contém datas para IDs de serviço que usam uma programação semanal. Especificam quando o serviço começa e termina, bem como os dias da semana em que o serviço está disponível.
calendar_dates.txt	Opcional	Contém as exceções para IDs de serviço definidos no arquivo *calendar.txt* . Se o arquivo *calendar_dates.txt* inclui *todas* as datas de serviço, ele pode ser especificado no lugar do *calendar.txt*.
fare_attributes.txt	Opcional	Contém informações sobre tarifas dos trajetos de uma empresa de transporte público.
fare_rules.txt	Opcional	Contém regras para implementação das informações de tarifa dos trajetos de uma empresa de transporte público.
shapes.txt	Opcional	Contém regras para desenhar linhas em um mapa para representar os trajetos de uma empresa de transporte público.
frequencies.txt	Opcional	Contém os intervalos entre as viagens nos trajetos.
transfers.txt	Opcional	Contém regras para conexões em pontos de baldeação entre os trajetos.
feed_info.txt	Opcional	Contém informações adicionais sobre o *feed*, incluindo editor, versão e informações sobre validade.

[a] Os campos contidos em cada arquivo da especificação GTFS estão descritos no apêndice C, nas tabela 9 - 21.

Fonte: Google Transit (adaptada)[1]

2.5 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. Tais características, por exemplo, podem ser encontradas na rede social *Facebook*[5]. O *Twitter*[6] além de possuir as características de rede social mencionadas anteriormente, pode ser caracterizado também como uma Rede de Informações. Nesse tipo de rede a interação dominante é a disseminação de informações entre os relacionamentos, com baixo baixo índice de reciprocidade (MYERS et al., 2014).

No *Twitter* as informações (*tweets*) são publicadas contendo no máximo 280 caracteres; cada publicação pode receber *retweets* (ser compartilhada por outros usuários), comentários (diretamente no *tweet* — *replies* — ou de forma privada via caixa de mensagens) e *likes* (indicador de quantos usuários gostaram da publicação). Além dessas funcionalidades, os *tweets* podem conter menções a outros usuários (@nome do *profile*) e rótulos (*#hashtag*) indicando assuntos, categorias, etc.

Devido as características citadas nos parágrafos anteriores, o *Twitter* tem sido uma rede social importante para compartilhamento de informações e acontecimentos do cotidiano. Tais acontecimentos podem ser classificados como eventos sociais, capazes de descrever desde eventos rotineiros (*shows*, jogos esportivos, etc.) a situações de crise (eventos de exceção — desastres naturais, mobilizações sociais, dentre outros) (ZHOU; CHEN, 2014), (ATEFEH; KHREICH, 2015).

2.6 Processamento de Linguagem Natural

O processamento automático de *tweets* envolve o Processamento de Linguagem Natural (NLP — *Natural Language Processing*), que explora como computadores podem ser utilizados para entender e manipular texto ou fala em linguagem natural (LIU; LI; THOMAS, 2017), o que envolve conhecimento interdisciplinar principalmente entre as áreas de ciência da computação, linguística e estatística. A

[5] <https://www.facebook.com>. Acesso em Outubro, 29 de 2017.
[6] <https://twitter.com>. Acesso em Outubro, 29 de 2017.

seguir são detalhados alguns dos problemas relacionadas a NLP, divididos em baixo e alto nível (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011):

1. Baixo nível (problemas comuns a NLP) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
    a) ***Sentence boundary disambiguation* (SBD)**: processamento para identificação do início e fim de uma sentença (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
    b) ***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 (SETIAWAN; WIDYANTORO; SURENDRO, 2017).
    c) ***Part-of-speech tagging***: processamento para identificação das classificações gramaticais (verbo, sujeito, adjetivo, etc.) das palavras em uma sentença, considerando seus respectivos significados e contexto no qual estão inseridas (ROY; MAJUMDER; NATH, 2017).
    d) ***Decomposição morfológica***: processamento para decomposição morfológica de uma determinada palavra para a sua forma inflexionada, usando *lemmatization* (identificação do lema da palavra) ou *stemming* (identificação da raiz da palavra usando heurísticas para determinar a localização de sua respectiva flexão) (SETIAWAN; WIDYANTORO; SURENDRO, 2017), (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011), (KORENIUS et al., 2004).
    e) ***Shallow parsing* (*chunking*)**: processamento para identificação de segmentos de uma sentença, tais como frases verbais, nominais, etc., com base nos *tokens* que constituem a *part-of-speech* (COLLOBERT et al., 2011), (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
2. Alto nível (aplicação de NLP a problemas específicos, com base nos problema de baixo nível) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
    a) ***Spelling / grammatical error identification and recovery***: processamento iterativo para identificação e correção de erros gramaticais e de digitação. (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

b) ***Named Entity Recognition* (NER)**: processamento para identificação e categorização de palavras ou frases específicas (entidades) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

c) ***Word Sense Disambiguation* (WSD)**: processamento para identificação do sentido de uma palavra numa sentença (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

d) ***Negation and uncertainty identification***: processamento para inferir se uma entidade está presente ou não numa sentença, assim como quantificar a quantidade de incerteza da inferência realizada (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

e) ***Extração de relacionamentos***: processamento para identificar relacionamentos entre entidades e eventos (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

f) **Extração de relacionamento / inferência temporal**: processamento para inferência de expressões e relacionamentos temporais (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

g) **Extração de informação**: processamento para extração e transformação para uma forma estruturada de informações específicas a um problema (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

2.7 Feature Engineering

Neste projeto, inicialmente pretendemos utilizar *feature extraction* e *selection*, duas das fases do processo de *feature engineering*. Dito isso, explicamos nos parágrafos seguintes o que são *features* e as fases do processo de *feature engineering*.

Um conjunto de dados pode ser representado por um número fixo de características (*features*) binárias, categóricas ou contínuas (GUYON; ELISSEEFF, 2006). O processo de construção dessas *features* é conhecido como *feature engineering*, o qual depende estritamente do conhecimento de domínio dos dados e de suas respectivas métricas (GUYON; ELISSEEFF, 2006).

O processo de *feature engineering* é iterativo entre as fases de *feature extraction*, *feature construction* e *feature selection* (MOTODA; LIU, 2002). Antes da

fase de *feature extraction* os dados podem ser pré-processados utilizando técnicas de padronização, normalização, remoção de ruídos, redução de dimensionalidade, discretização, expansão, dentre outras; observa-se que informações podem ser perdidas ao realizar essas transformações (GUYON; ELISSEEFF, 2006).

Na fase de *feature construction* é realizado um processo para descobrir informações ausentes a respeito dos relacionamentos entre as *features* e aumentar o espaço de *features* inferindo ou criando novas *features*, com o objetivo de melhorar a acurácia dos algoritmos de classificação, compreensão dos dados, obtenção de dados ocultos, etc (MOTODA; LIU, 2002). Nessa fase, a partir de um conjunto de *n features* $A_1, A_2, ..., A_n$ é possível construir *m features* adicionais $A_{n+1}, A_{n+2}, ..., A_{n+m}$, por meio de heurísticas, operadores lógicos, algoritmos (*greedy search* aplicada em árvores de decisão, genéticos, etc.), dentro outros (MOTODA; LIU, 2002).

O processo de *feature extraction*, por sua vez, utiliza uma função de mapeamento para extrair um conjunto mínimo de novas *features* com base nas *features* originais e em métricas de desempenho (o que pode ser realizado também usando *feedforward neural network*), diferentemente da análise dos relacionamentos entre as *features* realizada na fase de *feature construction* (MOTODA; LIU, 2002). Assim, com um conjunto inicial de *n features* $A_1, A_2, ..., A_n$ é possível extrair novas *features* $B_1, B_2, ..., B_m (m < n), B_i = F_i(A_1, A_2, ..., A_n)$, onde F_i é a função de mapeamento (MOTODA; LIU, 2002).

Por fim, tem-se como objetivo no processo de *feature selection* a redução ótima do espaço das *features* com base em critérios de seleção, ou seja, obter *m features* a partir de um conjunto de *n features*, onde $m \leq n$ (o que pode ser alncançado por algoritmos de busca seguindo critérios de avaliação). Dessa forma, com um subcojunto menor de *features*, é possível reduzir a dimensionalidade do espaço das *features* (evitando sobre-ajuste), otimizar algoritmos de Aprendizado de Máquina e compreender melhor seus respectivos resultados, melhorar a acurácia dos algoritmos de classificação, dentre outros benefícios (MOTODA; LIU, 2002).

2.8 Algoritmos de Aprendizado de Máquina

Os algoritmos de Aprendizado de Máquina podem ser (I) supervisionados, nos quais relações com resultados conhecidos são criadas com base nas características de entrada; (II) não-supervisionado, nos quais são conhecidas as características de entrada, mas não os resultados; (III) semi-supervisionados, nos quais podem ser definidas algumas das relações entre dados de entrada e resultados; (IV) por reforço, nos quais são estabelecidas ações com o foco em maximizar determinado ganho.

No contexto desse trabalho conhecemos como os dados de entrada podem ser classificados, devido a isso iremos utilizar aprendizado de máquina supervisionado. Os dados de entrada que pretendemos utilizar nesse trabalho são textuais. Assim, elencamos a seguir alguns dos principais algoritmos de aprendizado de máquina supervisionado para classificação textual, com base em (KHAN et al., 2010):

1. *Artificial Neural Network*;
2. *Decision Tree*;
3. *Decision Rules Classification* ;
4. *K-nearest neighbor (k-NN)*;
5. *Fuzzy correlation*;
6. *Genetic Algorithm*;
7. *Naïve Bayes Algorithm*;
8. *Rocchio's Algorithm*;
9. *Support Vector Machine*.

3 Revisão Sistemática

Este capítulo apresenta uma Revisão Sistemática (RS) com o objetivo de encontrar o estado da arte de trabalhos que visam melhorar sistemas de transporte público por meio do processamento de *tweets*. Além disso, de uma forma mais ampla, busca-se também entender como os *tweets* têm sido utilizados na caracterização de problemas urbanos. Sendo assim, o capítulo é iniciado com a seção sobre o planejamento da Revisão Sistemática; seguida das questões de pesquisa utilizadas na formulação do problema da RS; do processo de coleta dos estudos primários; da avaliação dos dados coletados; da análise e interpretação dos estudos selecionados, concluindo com as considerações finais.

3.1 Planejamento da Revisão Sistemática

A presente Revisão Sistemática utiliza a metodologia proposta por BIOLCHINI et al. (2005), composta por cinco etapas. A primeira etapa está relacionada à formulação do problema, na qual é levantada uma questão central se referindo ao tipo de evidência que deverá estar contida na revisão. Em seguida, são construídas definições que permitem estabelecer uma distinção entre os estudos relevantes e irrelevantes para o propósito específico do que se está investigando (BIOLCHINI et al., 2005).

A segunda etapa da condução está relacionada à Coleta de Dados, na qual são definidos os procedimentos que serão utilizados para encontrar a evidência relevante que foi definida na etapa anterior. Nesta fase é extremamente importante determinar as fontes que podem fornecer estudos relevantes a serem incluídos na pesquisa (BIOLCHINI et al., 2005).

Na terceira etapa a Avaliação de Dados é definida, na qual são selecionadas as fontes primárias que deverão ser incluídas na revisão. Em seguida, são aplicados os critérios de qualidade para separar estudos que podem ser considerados válidos, e determinadas as diretrizes para o tipo de informação que deve ser extraída dos relatórios de pesquisas primárias (BIOLCHINI et al., 2005).

A quarta etapa da revisão é o processo de Análise e Interpretação, na qual os dados dos estudos primários válidos são sintetizados. E, na quinta etapa são realizados os processos de Conclusão e Apresentação (BIOLCHINI et al., 2005).

3.1.1 Justificativa da Revisão Sistemática

Esta Revisão Sistemática se justifica por não terem sido encontradas revisões sistemáticas com o foco em questões urbanas e de transporte público, abordando unicamente o processamento de *tweets*. Em (CHANIOTAKIS; ANTONIOU; PEREIRA, 2016), por exemplo, foi realizado um mapeamento de forma não sistemática dos trabalhos sobre o uso das mídias sociais em problemas relacionados ao transporte público; (STEIGER; ALBUQUERQUE; ZIPF, 2015), por outro lado, desenvolveram uma revisão sistemática sobre o uso do Twitter para questões espaço-temporais; e (JUNGHERR, 2016) no contexto político.

Devido a isso, a presente revisão sistemática se diferencia por ter como objetivo encontrar o estado da arte de trabalhos que visam melhorar sistemas de transporte público por meio do processamento de *tweets*. Além disso, de uma forma mais ampla, busca-se também entender como os *tweets* têm sido utilizados na caracterização de problemas urbanos.

3.2 Questões de Pesquisa

Nesta seção, são apresentadas as questões de pesquisa utilizadas para a formulação dos problemas abordados por essa Revisão Sistemática. Por meio das quais, busca-se atender os objetivos já mencionados na seção 3.1.1.

1. Quais os tipos de problemas urbanos abordados utilizando processamentos de *tweets*?

   O propósito da QP1 é identificar quais são as contribuições do processamento de *tweets* para a mitigação de problemas urbanos. A resposta a essa questão de pesquisa ajudará especialistas das áreas multidisciplinares relacionadas ao Urbanismo (como a de Análise de Redes Sociais e Políticas Públicas) a terem

um panorama de como *tweets* podem ser utilizados para ajudar na solução de problemas urbanos.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP1): alguns dos problemas urbanos abordados estão relacionados ao transporte, mobilidade urbana, turismo e desastres naturais.

2. Como *tweets* têm sido utilizados para abordar problemas relacionados ao transporte público?

O propósito da QP2 é identificar se *tweets* têm sido utilizados para solucionar problemas relacionados ao transporte público. A resposta a essa questão de pesquisa ajudará especialistas das áreas multidisciplinares relacionadas ao Urbanismo (como a de Análise de Redes Sociais e Políticas Públicas) a terem um panorama de como *tweets* podem ser utilizados para ajudar na solução de problemas referentes a mobilidade urbana.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP2): *tweets* têm sido utilizados principalmente para questões relacionadas ao congestionamento, não tendo como foco o transporte público.

3. Quais as técnicas estatísticas utilizadas no processamento de *tweets*?

O propósito da QP3 é identificar quais as técnicas estatísticas utilizadas no processamento de *tweets*, principalmente no que se refere a análise de acurácia de classificação binária. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama de como garantir a confiabilidade ao utilizar dados oriundos de *tweets*, dentre outros aspectos relacionados a testes estatísticos.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP3): F_1 *score* é a principal técnica estatística utilizada

para análise de acurácia de classificação binária.

4. Quais os paradigmas de processamento têm sido utilizados ao lidar com *tweets*?

   O propósito da QP4 é identificar os paradigmas utilizados para processamento de *tweets*. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama das técnicas de processamento utilizadas na análise de *tweets*.

   Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP4): o principal paradigma utilizado tem sido o processamento de *tweets* em *batch* (*offline*), após um processo de armazenamento. Poucos são os estudos que constroem uma plataforma para processamento de dados em tempo real.

5. Quais são os eventos de exceção relacionados ao transporte público?

   O propósito da QP5 é identificar os eventos de exceção relacionados ao transporte público. A resposta a essa questão de pesquisa ajudará especialistas no levantamento de eventos de exceção relacionados ao transporte público, os quais podem ser utilizados em algoritmos de classificação.

   Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP5): há poucos ou nenhum estudo que, ao tratar de problemáticas relacionadas ao transporte público, realizam um levantamento dos eventos de exceção desse contexto.

6. Quais as técnicas de Aprendizado de Máquina utilizadas no processamento de *tweets*?

   O propósito da QP6 é identificar as técnicas de Aprendizado de Máquina utilizadas no processamento de *tweets*. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama das principais técnicas de Aprendi-

zado de Máquina utilizadas no processamento de *tweets*.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP6): a técnica *Support Vector Machine* tem sido utilizada na maioria dos estudos que aplicam aos *tweets* algum algoritmo de Aprendizado de Máquina.

3.3 Coleta de dados

Nesta Revisão Sistemática, os artigos foram coletados em quatro fontes de pesquisa, por meio da plataforma de indexação de trabalhos acadêmicos *Google Scholar*[1]. Constam na tabela 3 as bases pesquisadas no ano de 2017, quantidades de artigos coletados, descartados no processo de filtragem (Fig. 1, descrito na seção 3.4) e selecionados. Com base na QP1, a seguinte *string* de busca foi construída; restrita aos trabalhos publicados entre 2011 e 2016, escritos no idioma Inglês (devido ao fato das publicações relevantes, na área de Computação, estarem disponíveis nesse idioma):

***String* de busca:** twitter urban planning city (analytics OR patterns OR tweets OR social OR media) AND (public transport)

Palavras-chave: twitter, urban, planning, city, analytics, patterns, tweets, social, media e public transport.

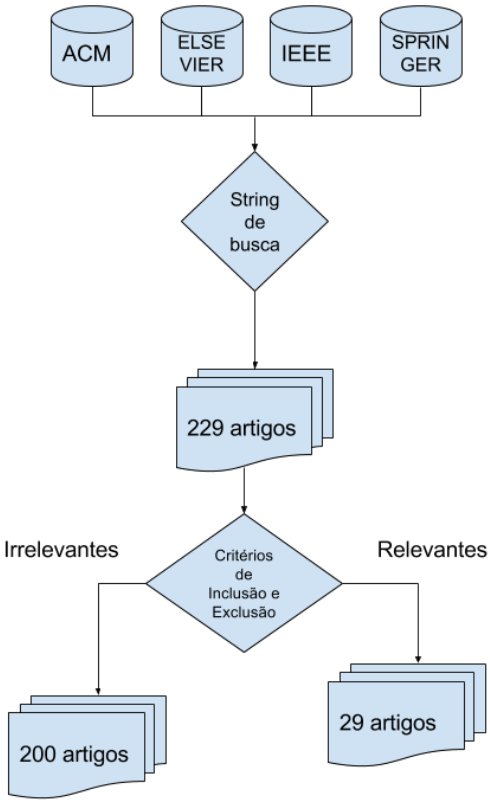
Tabela 3 – Quantidades de artigos coletados e fontes de busca

Fonte	Artigos coletados	Filtragem	Selecionados
ACM	44	34	10
IEEE	82	74	8
Elsevier	81	72	9
Springer	22	20	2
-	229	200	29

Fonte: Felipe Cordeiro Alves Dias

[1] <https://scholar.google.com>. Acesso em Outubro, 29 de 2017.

Figura 1 – Processo de Filtragem



Fonte: Felipe Cordeiro Alves Dias, 2017

3.4 Avaliação de Dados

Visando selecionar os artigos relevantes para esta Revisão Sistemática, os seguintes critérios foram utilizados no processo de filtragem:

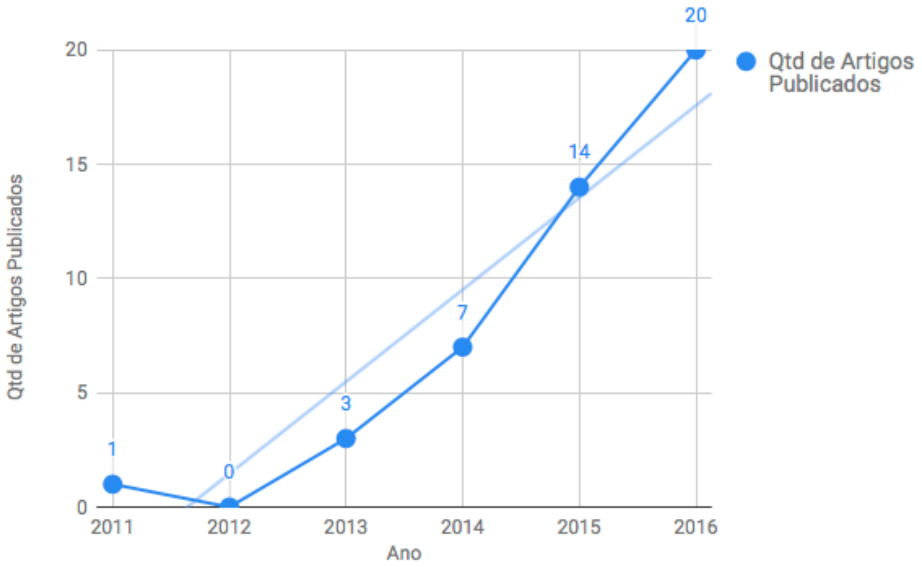
- Trabalho publicado (critério de qualidade).
- Trabalhos que utilizam *tweets* para abordar questões urbanas e de transporte público.
- Trabalhos duplicados.
- Trabalhos que estão fora do escopo da questão de pesquisa.

O processo de condução da Revisão Sistemática foi realizado utilizando os critérios acima mencionados, e está disponível em DIAS (2017) (não incluso neste trabalho com o objetivo de não deixar o texto exaustivo), assim como seu respectivo protocolo (no qual contém o detalhamento dos critérios de inclusão e exclusão,

dentre outros artefatos da condução). Após o processo de condução, alguns dos metadados dos artigos selecionados foram sintetizados.

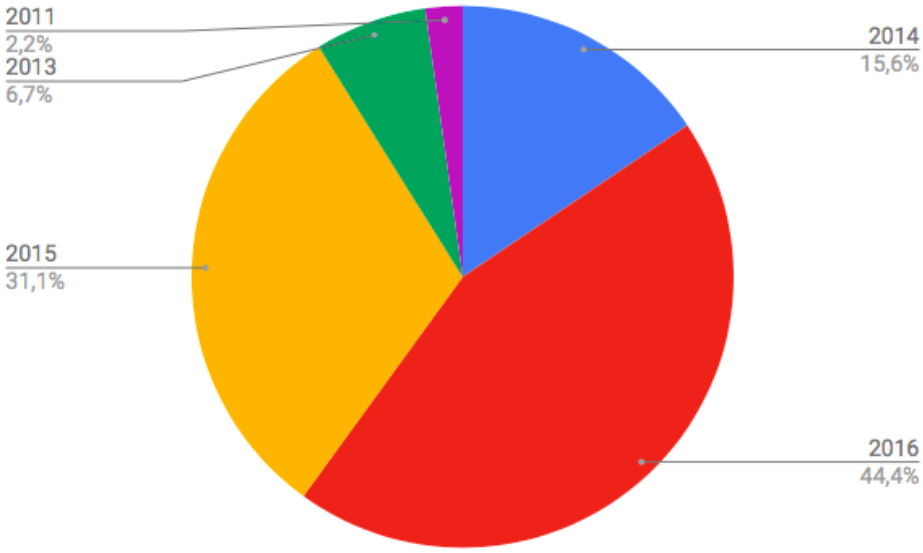
Sendo assim, a Fig. 4 apresenta uma nuvem de *tags* (gerada com a biblioteca *wordcloud* (Andreas Mueller, 2018)) sintetizando as palavras chaves dos estudos primários selecionados; e a Fig. 2 a quantidade de artigos publicados por ano, sendo possível analisar por meio dela a distribuição dos artigos entre 2011 e 2016, assim como sua respectiva porcentagem, ilustrada na Fig. 3.

Figura 2 – Quantidade de artigos publicados por ano



Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 3 – Porcentagem dos artigos publicados por ano



Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 4 – Nuvem de palavras das *keywords* dos artigos selecionados

Fonte: Felipe Cordeiro Alves Dias, 2017

3.5 Análise e Interpretação

Nesta seção é realizada a análise e interpretação dos estudos primários selecionados pela Revisão Sistemática, sendo as subseções divididas de acordo com as questões de pesquisa.

3.5.1 Tipos de problemas urbanos abordados utilizando o processamento *tweets* (QP1)

Os tipos de problemas urbanos abordados utilizando o processamento de *tweets* foram divididos nas seguintes categorias:

1. ***e-Participation*** (Interação entre cidadãos e órgãos civis) (MUKHERJEE et al., 2015), (SOOMRO; KHAN; HASHAM, 2016).
2. **Detecção de zoneamento urbano** (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).
3. **Identificação de pontos de interesse** (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (BENDLER et al., 2014), (ABBASI et al., 2015), (GKIOTSALITIS; STATHOPOULOS, 2015), (GKIOTSALITIS; STATHOPOULOS, 2016), (HASAN; UKKUSURI, 2014), (MAGHREBI et al., 2015), (DI LORENZO et al., 2013).
4. **Mobilidade** (GUTEV; NENKO, 2016), (CHEN et al., 2016), (YOUSAF et al., 2014).

5. **Padrões demográficos** (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (STEIGER et al., 2015), (GUO et al., 2016).
6. **Poluição** (ZAGAL; MATA; CLARAMUNT, 2016).
7. **Segurança Pública** (WEN; LIN; PELECHRINIS, 2016), (MATA; CLARAMUNT, 2015).
8. **Turismo** (THOMAZ et al., 2016), (ABBASI et al., 2015), (CHUA et al., 2016), (SOBOLEVSKY et al., 2015).
9. **Tráfego** (ANANTHARAM et al., 2015), (LECUE et al., 2014).

Conforme os estudos primários analisados pela Revisão Sistemática, e enumerados nessa seção, é possível interpretar que *tweets* podem ser utilizados para auxiliar na mitigação de inúmeros problemas urbanos. Apesar disso, (CHANIOTAKIS; ANTONIOU, 2015) observam que os *tweets* contendo informações sobre geolocalização são normalmente publicados em áreas relacionadas ao lazer, além de haver correlação entre regiões urbanas com maior renda *per capita* e o número de *tweets* postados. Tal evidência pode conduzir viés nas análises por representar somente algumas classes econômicas da população.

Considerando a observação anterior, um dos estudos analisados foi o realizado por (ZAGAL; MATA; CLARAMUNT, 2016), na Cidade do México. Nesse estudo, foram mapeados os pontos da cidade referenciados em publicações relacionadas a doenças respiratórias e poluição, orientando tomadas de decisão no aspecto ambiental.

Além disso, há também exemplos de trabalhos relacionados a Segurança Pública, como o estudo de caso realizado por (WEN; LIN; PELECHRINIS, 2016), no qual foi enriquecido um conjunto de dados com *tweets* geolocalizados, visando analisar o impacto dos ataques terroristas (em Paris, em novembro de 2015) nos padrões de atividades urbanas (relacionadas ao uso de transporte público, serviços, realização de compras, e atividade noturna). Em um outro caso de aplicação, estimou-se por meio de *tweets*, a probabilidade de ocorrência de crimes e ameaças nas ruas da Cidade do México, sugerindo rotas seguras aos pedestres (MATA; CLARAMUNT, 2015).

Também, foram encontrados na literatura estudos que utilizaram *tweets* para inferir padrões demográficos. Por exemplo, em (FARSEEV et al., 2015); (GKIOTSALITIS; STATHOPOULOS, 2015); (GKIOTSALITIS; STATHOPOULOS, 2016), *tweets*

foram processados para analisar a distribuição etária e de gênero da população, assim como seus respectivos pontos de interesse (HASAN; UKKUSURI, 2014) e (MAGHREBI et al., 2015) (como locais para entretenimento, residência, trabalho, recriação, compras, educação e serviços sociais).

Tais pontos de interesse, podem ser utilizados em problemas relacionados ao transporte público (GUTEV; NENKO, 2016) e também ao Turismo, como no estudo realizado por (ABBASI et al., 2015) para identificar a locomoção de visitantes e residentes em pontos turísticos de Sydney; por (CHUA et al., 2016), ao caracterizar aspectos espaciais, temporais e demográficos, dos turistas da cidade de Cilento, Itália; e por (THOMAZ et al., 2016) na cidade de Curitiba (Brasil), no contexto da Copa do Mundo de 2014.

Nesse mesmo contexto, (GUO et al., 2016) estudaram algumas questões demográficas via análise de sentimento, encontrando correlação positiva entre oportunidades de emprego e sentimentos positivos, e negativa entre felicidade e número de crianças na população da Grande Londres. Outro caso de uso, foi o desenvolvido em (STEIGER et al., 2015), no qual *tweets* foram processados para identificar diferentes tipos de atividades em Londres, correlacionando-as com informações censitárias; e em (SOBOLEVSKY et al., 2015) ao estudar a atratividade da Espanha a turistas.

Um dos problemas relacionados à identificação de pontos de interesse se refere as incertezas espaço-temporais e de determinação de tópicos, o qual foi abordado pelo trabalho realizado por (BENDLER et al., 2014). Nele, os autores contribuíram com uma técnica para minimizar o problema ao processar *tweets*; analisando a causalidade entre o tempo e local das postagens realizadas, reduzindo assim os índices de incerteza, no contexto da cidade de São Francisco, EUA. Outro problema, relaciona-se com a questão da privacidade, pois as localizações dos usuários podem ser inferidas mesmo quando não disponibilizadas. Nesse cenário, (WANG; SINNOTT; NEPAL, 2016) propõem um Sistema de Calibração de Trajetórias Privadas (PTCS), usando os mecanismos de Privacidade Diferencial e de *k-anonymity*, com isso é possível extrair informações sobre trajetórias sem exposição de informações sensíveis, testado na extração de localizações por meio de *tweets*.

Outro contexto na literatura revisada está relacionado ao processamento dos eventos que acontecem na cidade (idealmente em tempo real, como sugerem (SO-

OMRO; KHAN; HASHAM, 2016)). Um dos estudos encontrados sobre esse assunto, foi o realizado por (ANANTHARAM et al., 2015), no qual foi desenvolvida uma técnica para identificar os diferentes tipos de eventos do cotidiano urbano, rotulando-os sequencialmente, por meio da anotação de *tweets* e extração de eventos, considerando aspectos espaciais, temporais e temáticos. Para isso, utilizou conhecimentos de domínio, tais como informações sobre os locais em uma cidade e possíveis termos para os eventos, identificando assim os relacionados ao tráfego da região da Baia de São Francisco, EUA.

Sobre a mesma temática, (DI LORENZO et al., 2013) desenvolveram uma ferramenta inteligente e interativa para exploração visual da dinâmica de eventos sociais ao longo das dimensões espacial, temporal e organizacional. O tráfego também foi objeto de estudo em (CHEN et al., 2016), ao relacionar eventos do trânsito com a demanda por bicicletas; e em (LECUE et al., 2014), ao demonstrar uma plataforma para análise inteligente do tráfego (em tempo real), com base em fontes heterogêneas de dados (incluindo tweets de agências oficiais de trânsito).

Em uma abordagem mais genérica, (MUKHERJEE et al., 2015) propuseram uma plataforma para processar (em *near real time*) questões urgentes da cidade, oriundas de diversas fontes (incluindo o *Twitter*), atuando como intermediadora entre cidadãos e agências civis. No que se refere a mobilidade urbana, mas não utilizando informações sobre pontos de interesse, (YOUSAF et al., 2014) inferiram a afinidade entre usuários por meio da análise de *retweets*, possibilitando que rotas de corridas sejam compartilhadas entre pessoas com interesses em comum, tornando a viagem mais agradável.

De forma inusitada, (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014) utilizaram apenas *tweets* geolocalizados para analisar suas respectivas distribuições no espaço urbano, visando identificar a caracterização do uso da terra, considerando os zoneamentos urbanos industriais, residenciais, comerciais e de lazer. O trabalho foi realizado no contexto da cidade de Manhattan (EUA), Londres (Reino Unido) e Madrid (Espanha).

3.5.2 Casos de uso relacionados ao transporte público (QP2)

Nesta seção, são identificados os estudos primários que utilizam processamento de *tweets* tendo como foco a mitigação dos problemas relacionados ao transporte público; enumerados a seguir:

1. **Impacto de eventos no transporte público**.
    a) Impacto dos ataques terroristas em Paris no uso do transporte público (WEN; LIN; PELECHRINIS, 2016).
    b) Impacto de eventos relacionados ao tráfego na demanda por bicicletas, em Nova Iorque e Washington D.C, EUA (CHEN et al., 2016).
    c) Impacto dos pontos de interesse na demanda por transporte público (MAGHREBI et al., 2015).
    d) Impacto dos eventos anormais nas tomadas de decisão dos passageiros do Metrô de Tokyo (ITOH et al., 2016).
    e) Predição de fluxo de passageiros no Metrô de Nova Iorque (NI; HE; GAO, 2016).
2. **Planejamento e gestão do transporte público**.
    a) Análise de sentimento relacionada ao acesso ao transporte público (GUO et al., 2016).
    b) Coleta de informações relacionadas ao transporte público (GAL-TZUR et al., 2014).
    c) Identificação de locais para estações de bicicletas, em St. Petersburg, Rússia (GUTEV; NENKO, 2016).
    d) Identificação da disposição dos usuários para realizar viagens de lazer (GKIOTSALITIS; STATHOPOULOS, 2016).
    e) Plataforma para notificação de problemas relacionados ao transporte público de Bangalore, Índia (MUKHERJEE et al., 2015).

Conforme os estudos primários analisados pela Revisão Sistemática, e enumerados nessa seção, é possível interpretar que os estudos estão classificados em análise de impacto de eventos, planejamento e gestão do transporte público. Por exemplo, (WEN; LIN; PELECHRINIS, 2016) utilizaram *tweets* para analisar o im-

pacto dos ataques terroristas em Paris (2015) nos padrões de mobilidade referentes ao uso de transporte público. Semelhantemente, ITOH et al. (2016) desenvolveram uma ferramenta para analisar e explorar visualmente, com base em *tweets*, as tomadas de decisão dos passageiros do Metrô de Tokyo, ante a eventos anormais, tais como Tufões, Incêndios, Terremotos, dentre outros. Nesse mesmo contexto, (NI; HE; GAO, 2016) propuseram uma técnica de predição de fluxo de passageiros no Metrô de Nova Iorque, identificando eventos com base nas *hashtags* dos *tweets*. Enquanto que em (CHEN et al., 2016), analisaram a relação entre eventos do tráfego com a demanda por bicicletas.

No que se refere aos estudos focados no planejamento e gestão do transporte público, (MUKHERJEE et al., 2015) apresentam uma plataforma desenvolvida e utilizada pela Agência de Transporte Público de Bangalore, na Índia, a qual permite que usuários reportem questões relacionadas ao transporte público, possibilitando a melhoria do planejamento de suas respectivas operações, assim como o serviço prestado para a população. Nessa mesma linha de estudo, em (GUTEV; NENKO, 2016), *tweets* são utilizados para identificar a popularidade de determinados locais, pontos de interesse e distribuição etária, com o objetivo de determinar os melhores pontos para estações de bicicletas e incentivar assim o uso desse modal de transporte. Também relacionado aos pontos de interesse, (MAGHREBI et al., 2015) utilizaram *tweets* para identificar padrões das atividades humanas (em diferentes horários do dia) e seus respectivos impactos na demanda por transporte público.

Em (GAL-TZUR et al., 2014), por sua vez, utilizaram uma abordagem hierárquica para classificar *tweets* relacionados ao transporte. Com isso, demonstraram que é possível usar essas informações para fins de planejamento e gerenciamento do transporte. Tal técnica, foi aplicada em um estudo de caso associado a eventos esportivos, ocorridos no Reino Unido. A hierarquia é composta por três níveis, no primeiro, os *tweets* são classificados entre os que expressam a necessidade de serviços de transporte, opiniões e incidentes; o segundo, identifica a categoria do transporte; e último, relaciona *tweets* a tópicos.

Outro estudo que contribui com o planejamento do transporte público, é o realizado em (GKIOTSALITIS; STATHOPOULOS, 2015, 2016), no qual *tweets* foram processados para identificar a disposição dos usuários para realizar viagens relacionadas ao lazer (pontos de interesse), sugerindo a eles atividades com menor

tempo de percurso e probabilidade de atrasos. Além do tempo de percurso, outro ponto relevante considerado foi o de bom nível de acesso ao transporte público, o qual quando existente impacta positivamente na felicidade das pessoas e se correlaciona com sentimentos positivos, segundo a análise de sentimentos realizada por (GUO et al., 2016), utilizando *tweets* publicados na Grande Londres.

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

Nesta seção, são apresentadas as técnicas estatísticas utilizadas pelos estudos primários, no processamento de *tweets*, enumeradas a seguir:

1. **Análise de métricas relacionadas a desempenho** (erro de reconstrução relativo, qualidade dos componentes descritivos recuperados e qualidade dos componentes comuns recuperados) (WEN; LIN; PELECHRINIS, 2016).
2. ***Cosine similarity*** (YOUSAF et al., 2014), (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).
3. ***F_1 score*** (ANANTHARAM et al., 2015), (CHEN et al., 2016).
4. ***Term frequency–inverse document frequency*** (TF-IDF) (MUKHERJEE et al., 2015).
5. ***Inverse coefficient of variation*** (BENDLER et al., 2014).
6. ***Jackknife resampling*** (BENDLER et al., 2014).
7. ***Linear Regression*** (GUTEV; NENKO, 2016), (BENDLER et al., 2014), (NI; HE; GAO, 2016), (GUO et al., 2016).
8. ***Local Indicators of Spatial Association*** (LISA) (STEIGER et al., 2015).
9. ***Local Moran's*** (STEIGER et al., 2015).
10. ***Maximum likelihood estimation*** (MUKHERJEE et al., 2015).
11. ***Seasonal Autoregressive Integrated Moving Average*** (SARIMA) (NI; HE; GAO, 2016).
12. ***Optimization and Prediction with hybrid loss function*** (NI; HE; GAO, 2016).

Em (GUTEV; NENKO, 2016), os autores utilizaram Regressão Linear (RL) para analisar a demanda por bicicletas de acordo com as localizações extraídas dos *tweets*. Enquanto que (BENDLER et al., 2014) usaram RL para fornecer evidências de que

as categorias dos pontos de interesse se relacionam com as variáveis referentes ao espectro espaço-temporal; e (GUO et al., 2016) para analisar a correlação entre sentimentos positivos com as oportunidades de trabalho, com a quantidade de crianças, e com o acesso a transporte. (NI; HE; GAO, 2016), por outro lado, uniram Regressão Linear com a técnica *Seasonal Autoregressive Integrated Moving Average*, propondo uma abordagem baseada em otimização paramétrica e convexa, chamada *Optimization and Prediction with hybrid loss function* e adequada para modelagem utilizando séries temporais.

Devido aos problemas relacionados a ambiguidade e identificação de contextos, (ANANTHARAM et al., 2015); (CHEN et al., 2016) e (GAL-TZUR et al., 2014) aplicaram a técnica F_1 *score* para analisar a acurácia do processamento de *tweets*. Por outro lado, em (MUKHERJEE et al., 2015), utilizaram a técnica *Maximum likelihood estimation* para determinar a probabilidade de ocorrência de um evento, assim como a confiabilidade da informação.

No que se refere a agrupamento, (YOUSAF et al., 2014) agruparam usuários de acordo com a *Cosine similarity*, unindo pessoas com interesses em comum nos mesmos grupos. (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014), por outro lado, usou a mesma técnica para agrupar *tweets* de acordo com suas semelhanças quanto aos tipos de zoneamento urbano.

De forma isolada, no trabalho realizado por (MUKHERJEE et al., 2015), utilizaram a técnica TF-IDF na fase de classificação para o definir o *score* de categorias de eventos, escolhendo a mais relevante a ser buscada em um dicionário de categorias. Também isoladamente, (STEIGER et al., 2015) usaram a técnica LISA na identificação de *clusters* espaciais e valores esporádicos espaciais, obtendo assim os locais com atividades sociais. Além disso, também utilizaram a técnica *Local Moran's* para detectar diferentes padrões de atividade de acordo com o espaço geográfico.

Por último, (BENDLER et al., 2014) inovaram ao utilizar a técnica *Jackknife resampling* como inspiração para o desenvolvimento de uma abordagem que visa analisar a estabilidade estatística de um conjunto de categorias. Além disso, usaram também a análise *Iverse Coefficient of variation* para verificar a dispersão negativa da distribuição de um conjunto de variáveis.

3.5.4 Paradigmas de processamento (QP4)

Nesta seção, encontram-se a seguir apenas os paradigmas de processamento extraídos dos estudos primários analisados:

1. ***Batch processing*** (offline) (ANANTHARAM et al., 2015), (WEN; LIN; PELECHRINIS, 2016), (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (MATA; CLARAMUNT, 2015), (CHEN et al., 2016), (ABBASI et al., 2015), (BENDLER et al., 2014), (BENDLER et al., 2014), (YOUSAF et al., 2014), (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014), (STEIGER et al., 2015), (GAL-TZUR et al., 2014), (GKIOTSALITIS; STATHOPOULOS, 2016), (DI LORENZO et al., 2013), (ITOH et al., 2016), (CHANIOTAKIS; ANTONIOU, 2015).
2. ***Near Real Time*** (MUKHERJEE et al., 2015).
3. ***Real Time*** (SOOMRO; KHAN; HASHAM, 2016), (LECUE et al., 2014).

3.5.5 Eventos de exceção relacionados ao transporte público (QP5)

Nesta seção, encontram-se a seguir os eventos de exceção relacionados ao transporte público, extraídos dos estudos primários:

1. **Acidentes**.
    a) Acidentes nas estações transporte (ITOH et al., 2016).
    b) Incêndio (ITOH et al., 2016).
2. **Espaço-temporais**.
    a) Dia da semana (CHEN et al., 2016).
    b) Hora do dia (CHEN et al., 2016).
3. **Eventos sociais**.
    a) Feiras de rua (CHEN et al., 2016).
    b) Festivais (CHEN et al., 2016), (LECUE et al., 2014).
    c) Jogos esportivos (CHEN et al., 2016), (GAL-TZUR et al., 2014).
    d) Passeatas e maratonas (CHEN et al., 2016), (ITOH et al., 2016).
4. **Eventos urbanos**.

a) Relacionados ao tráfego (CHEN et al., 2016), (LECUE et al., 2014).

5. **Desastres naturais**.

   a) Tempestades (ITOH et al., 2016).
   b) Terremoto (ITOH et al., 2016).
   c) Tufões (ITOH et al., 2016).

6. **Metereológicas**.

   a) Dia claro, nublado, chuvoso, nevando, com neblina (CHEN et al., 2016).
   b) Temperatura do ar (CHEN et al., 2016).

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

Nesta seção, são apresentadas as técnicas de Aprendizado de Máquina utilizadas para processamento de *tweets*, extraídas dos estudos primários e enumeradas a seguir:

1. ***Bayes classification*** (MATA; CLARAMUNT, 2015).
2. ***C5.0 algorithm*** (ZAGAL; MATA; CLARAMUNT, 2016).
3. ***Conditional Random Field (CRF) with Logistic Regression*** (ANANTHARAM et al., 2015).
4. ***Event extraction based on tweet hashtags*** (NI; HE; GAO, 2016).
5. ***Latent Dirichlet Allocation* (LDA)** (FARSEEV et al., 2015), (ABBASI et al., 2015), (HASAN; UKKUSURI, 2014), (DI LORENZO et al., 2013).
6. ***Monte Carlo simulation*** (CHEN et al., 2016).
7. ***PairFac*** (técnica inovadora que utiliza *Tensor Factorization*) (WEN; LIN; PELECHRINIS, 2016).
8. ***Random Forest classification*** (FARSEEV et al., 2015).
9. ***Support Vector Machine*** (MUKHERJEE et al., 2015), (GAL-TZUR et al., 2014).
10. ***Self-Organizing Maps*** (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).

No contexto urbano, inúmeros eventos podem acontecer e impactar a população. O trabalho realizado por (WEN; LIN; PELECHRINIS, 2016), desenvolveu uma

técnica que utiliza a análise de tensores discriminantes para aprender e de forma automatizada descobrir os impactos de um determinado evento no cotidiano da cidade. Numa abordagem mais simples, (CHEN et al., 2016) utilizou *Monte Carlo simulation* para treinar um modelo para predição de demanda por bicicletas, devido a dificuldade de encontrar exemplos suficientes para usar outras abordagens de treinamento.

Especificamente sobre as técnicas de classificação, (MUKHERJEE et al., 2015) utilizaram *Support Vector Machine* para classificar os eventos recebidos de diversas fontes. Referente a essa abordagem, (GAL-TZUR et al., 2014) analisaram inúmeras técnicas de Aprendizado de Máquina, obtendo a melhor performance com o SVM, além disso, observaram como principal vantagem a sua capacidade de adaptação ao gênero e tarefas subjacentes.

Apesar disso, (GUO et al., 2016) utilizaram Processamento de Linguagem Natural (baseado em palavras chaves) para rotular sentimentos de *tweets*, devido a facilidade de escalar essa técnica (para processamento de milhões de *tweets*), em comparação a SVM. Outro caso de divergência é o do estudo realizado por (FARSEEV et al., 2015), no qual foi escolhido o modelo de classificação *Random Forest*, devido ao fato de ser mais adequado para classificação em espaço dimensional elevado, em vez das técnicas SVM e *Naive Bayes*, no que se refere a predição de idade e gênero usando *tweets*.

MATA; CLARAMUNT (2015), por sua vez, aplicou-se a técnica *Bayes Classification* em *tweets*, visando obter probabilidades relacionadas a crimes e ameaças em uma determinada localização. Por outro lado, (ZAGAL; MATA; CLARAMUNT, 2016) usaram o *C5.0 algorithm* devido a melhor desempenho em relação a *Bayes*, dependendo do tópico que está sendo classificado.

Para anotação de eventos, (ANANTHARAM et al., 2015) treinaram um modelo CRF (usando anotações baseadas em dicionários) para determinar os locais da cidade e os termos relacionas aos eventos expressos em *tweets*. E, isoladamente (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014) utilizaram a técnica *Self-Organizing Maps*, tendo como entrada os valores de latitude e longitude de *tweets*. Com isso, construíram um mapa segmentado em áreas urbanas, baseando-se nas regiões com diferentes concentrações de *tweets*.

Segundo (FARSEEV et al., 2015), a técnica LDA tem sido muito utilizada para identificação de pontos de interesses mencionados em *tweets*, sendo adequada para grandes bases de dados e agrupamento de *tweets* com tópicos similares, de acordo com (STEIGER et al., 2015). (ABBASI et al., 2015) exemplificou isso ao aplicar LDA para identificação de *tweets* relacionados ao Turismo; (HASAN; UKKUSURI, 2014), para identificação de padrões de atividades humanas; e (DI LORENZO et al., 2013), para identificação de eventos sociais.

No entanto, (NI; HE; GAO, 2016) em vez de usarem LDA, extraíram *hashtags* de *tweets* para um vetor, utilizando-o para medir as atividades sociais e identificar seus respectivos contextos. Segundo (NI; HE; GAO, 2016), isso se justifica devido ao fato de que há uma grande chance do alto volume de *tweets* não indicar necessariamente eventos e atendimentos a eles. Além disso, afirmam que o método baseado em *hashtag* é capaz de indicar sobre o que é o evento, mesmo não utilizando o Inglês formal.

3.6 Considerações finais sobre a revisão sistemática

Em uma análise quantitativa dos estudos primários selecionados, podemos concluir que a quantidade de artigos publicados sobre o uso de *tweets* na caracterização de problemas urbanos e relacionados ao transporte público tem crescido consideravelmente, entre 2011 e 2016. Provavelmente, devido ao fato da popularização das Redes Sociais e grande quantidade de dados disponíveis para processamento.

Tais estudos estão concentrados em maioria na identificação de pontos de interesse, utilizando-os em diferentes contextos, tais como o de turismo,mobilidade. Além disso, abordam também problemas relacionados ao transporte e desastres naturais, confirmando a primeira hipótese (HP1) dessa Revisão Sistemática. As temáticas não abordadas pela HP1 foram as relacionadas a *e-Participation*, detecção de zoneamento urbano, padrões demográficos e segurança pública, demonstrando a variedade de problemas urbanos explorados com o processamento de *tweets*.

Referente a segunda hipótese, os estudos exploraram principalmente o impacto de eventos no transporte público, confirmando-a parcialmente. Isso, devido ao fato de um dos trabalhos explorar como os eventos relacionados ao tráfego impactam na demanda por bicicletas; não havendo nenhum outro sobre processamento

de *tweets* para mitigação dos problemas envolvendo Tráfego. Outra temática não mencionada pela HP2 e sobre a qual há uma quantidade considerável de estudos, foi a do uso de *tweets* para o planejamento e gerenciamento do transporte público.

Independentemente dos problemas abordados por meio do processamento de *tweets*, dentre as 12 técnicas estatísticas elencadas, F_1 *score* foi a única referenciada como ferramenta para análise de acurácia de classificação binária, confirmando a terceira hipótese (HP3). Apesar disso, a HP3 não considerou outras técnicas importantes (com propósitos distintos), como a *Linear Regression*, amplamente utilizada nos estudos analisados. Referente as técnicas de Aprendizado de Máquina, a mais utilizada foi a *Latent Dirichlet Allocation* (LDA), seguida da *Support Vector Machine* (SVM), confirmando parcialmente a sexta hipótese (HP6).

Por fim, apenas quatro dos vinte e nove estudos analisados, cerca de 14%, mencionaram *features* relacionadas ao transporte público, confirmando assim a quinta hipótese (HP5). Assim como a quantidade de trabalhos que realizam processamento de *tweets* em tempo real, sendo apenas dois do total analisado, cerca de 6%, que utilizam esse paradigma de processamento, o que confirma a quarta hipótese (HP4). É importante ainda observar que, outros estudos que mencionaram processamento em tempo real, realizaram na verdade coleta de *tweets* em tempo real, para análises a posteriori via processamento em *batch* (offline), categoria na qual a maioria dos estudos foram enquadrados.

4 Proposta de pesquisa

Neste capítulo, são apresentadas as seções referentes a proposta de pesquisa para a dissertação. Assim, abordaremos a formalização do problema; solução proposta; construção do conjunto de dados; exploração e visualização do conjunto de dados; identificação dos eventos de exceção; correlação dos eventos de exceção com os dados AVL da SPTrans e, por fim, o plano de trabalho.

4.1 Formalização do problema

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

Posto que os conjuntos *E* e *I* existem, 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. Tal correlação pode ser definida por uma função sobrejetora, representada formalmente em lógica de primeira ordem pela expressão:

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

Dessa forma, para todo impacto ι pertencente ao conjunto I existe um evento de exceção ε_i pertencente ao conjunto *E*.

4.2 Solução proposta

A solução proposta pretende coletar *tweets* dos *profiles* contidos na tabela 1, pré-processá-los, extrair e selecionar *features* para serem utilizadas em algo-

ritmos de classificação, obtendo dessa forma os eventos de exceção. Com esses eventos de exceção pretendemos analisar a base histórica da SPTrans de dados AVL (transmitidos utilizando AVL), e identificar os possíveis impactos de cada evento de exceção. O processo de identificação dos impactos contidos na base histórica da SPTrans pode ser definido com base nas localizações extraídas dos *tweets* coletados e posteriormente geolocalizadas.

As localizações dos *tweets* podem ser extraídas usando a seguinte fórmula em expressão regular:

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

Tal expressão regular é dividida em dois conjuntos, no primeiro ($\{L_1|S_1|L_2|S_2|...|L_n|S_n\}$), os logradouros (L) e siglas (S) contidas na tabela 8 (no apêndice B) são concatenadas, especificando um filtro para identificar cadeias de caracteres iniciadas com um logradouro ou sigla. No segundo conjunto ($\{[a - zA - Z\backslash s]+\}$), o filtro especificado identifica as palavras seguintes aos logradouros e siglas.

Em resumo, propomos solucionar o problema de caracterização dos eventos de exceção e de seus respectivos impactos seguindo os seguintes passos: (I) coleta de *tweets* de órgãos responsáveis por notificar eventos de exceção; (II) identificação dos *tweets* relacionados a eventos de exceção; (III) extração e geolocalização dos endereços dos eventos de exceção e (IV) análise e correlação com os dados AVL.

4.3 Construção do conjunto de dados

Nesta seção, são apresentados os conjuntos de dados referentes a proposta de pesquisa.

4.3.1 *Corpus Twitter*

A Rede Social *Twitter*, foi escolhida como fonte de dados para a construção do conjunto de dados relacionados aos eventos de exceção. Isso devido ao fato de cada publicação ser limitada em 280 caracteres, o que reduz a complexidade de processamento do conteúdo publicado, e devido aos estudos existentes abordando

problemas urbanos e de mobilidade urbana, conforme os analisados na revisão sistemática do Cap. 3.

Assim, o conjunto de dados utilizado para a identificação dos eventos de exceção é composto por *tweets*, em português brasileiro, dos *profiles* contidos na tabela 1. É importante observar que, para esse projeto de pesquisa, apenas os *tweets* publicados pelas contas selecionadas serão considerados, descartando os relacionados às interações entre diferentes *profiles* (*retweets* e *replies*). Ou seja, os dados utilizados estão relacionados ao canal unidirecional de comunicação (no contexto de *e-participation*). Com essa restrição, evitamos problemas referentes a confiabilidade dos dados, o que nos permite focarmos na caracterização dos eventos de exceção e de seus respectivos impactos.

Sobre a seleção dos *profiles*, todos foram selecionados manualmente de acordo com os órgãos responsáveis por notificar eventos de exceção. Tais *profiles* são de caráter público, ou seja, o acesso aos *tweets* não envolve questões de privacidade. Apesar do acesso facilitado aos *tweets*, a API do *Twitter* limita a quantidade e frequência de requisições aos *endpoints*. Por exemplo, no protótipo desenvolvido (na linguagem de programação Java), há um artefato que coleta (utilizando o *plugin Twitter4J*[1]) os 3.200 *tweets* mais recentes (se disponíveis) de cada conta, através do *endpoint statuses/user_timeline*; o qual permite no máximo 180 requisições, em um intervalo de 15 minutos, com autenticação via conta de usuário[2].

Durante a coleta dos *tweets*, eles são mapeados para a seguinte classe do modelo da aplicação: *TweetInfo*, que contém as informações respectivas ao *id*, texto da publicação, *timestamp*, endereço extraído, latitude e longitude. Em seguida, o modelo é persistido no banco de dados não relacional *MongoDB*[3] e também no banco de dados de séries temporais *Druid*[4] para exploração e visualização dos dados, precesso explicado na seção 4.4. Os detalhes sobre o intervalo de tempo e o número de *tweets* coletados constam na tabela 4.

Além dos *tweets* coletados, foram extraídos 625 endereços e seus respectivos dados de geolocalização. No entanto, por meio de uma análise manual percebemos dois problemas: (I) alguns endereços não foram extraídos; (II) apesar de o endereço

[1] <twitter4j.org>. Acesso em Outubro, 29 de 2017.
[2] <https://dev.twitter.com>. Acesso em Outubro, 29 de 2017.
[3] <https://www.mongodb.com>. Acesso em Outubro, 29 de 2017.
[4] <http://druid.io>. Acesso em Outubro, 29 de 2017.

Tabela 4 – Intervalo de tempo e número de *tweets* coletados

Profile no *Twitter*	**# tweets [a]**	***Timestamp 1* [b]**	***Timestamp 2* [c]**
BombeirosPMESP	5.750	2017-05-21 02:10:39	2017-10-29 23:07:08
CETSP_	5.042	2017-02-20 14:07:04	2017-10-29 21:45:54
CPTM_oficial	5.435	2017-04-24 13:00:17	2017-10-29 10:00:40
governosp	5.450	2017-05-10 17:00:05	2017-10-29 22:00:03
metrosp_oficial	7.296	2017-06-07 17:23:34	2017-10-29 17:48:12
Policia_Civil	3.360	2015-04-15 17:44:44	2017-10-27 10:01:53
PMESP	3.956	2016-06-02 17:21:32	2017-10-29 20:25:37
saopaulo_agora	3.671	2016-11-18 07:36:12	2017-10-29 20:56:28
smtsp_	1.128	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_	8.447	2017-06-13 15:19:56	2017-10-29 22:01:44
TurismoSaoPaulo	3.308	2012-06-12 22:00:38	2017-10-27 17:46:59
Total	53.788	-	-

[a] Número de *tweets* coletados.
[b] *Timestamp* mais antigo.
[c] *Timestamp* mais recente.

Fonte: Felipe Cordeiro Alves Dias

ser extraído corretamente, encontramos geolocalizações fora do estado de São Paulo e do país. Assim, pretendemos melhorar o processo de extração dos endereços dos *tweets* e o restringir a geolocalização para a região de São Paulo.

4.3.2 *Corpus* SPTrans

O corpus SPTrans é composto por dados obtidos do SIM, transferidos via AVL, e por dados fornecidos pela SPTrans especificados em GTFS, detalhados na tabela 5. Os dados de ambas as fontes não são triviais de serem processados (grande volume de dados, dados sem tipo explicitamente definido – não tratados, dados separados em lotes de dados – um arquivo para cada hora de movimentação dos ônibus, dados fora do formato convencional – por exemplo, 24h em vez de 0h), devido a isso foram desenvolvidos *scripts* para um processo de ETL (*Extract*, *Tranform* and *Load*).

No caso dos dados especificados em GTFS, convertemos os dados originais de *string* para os seus respectivos tipos (*long*, *double*, *int* ou *string*) e padronizamos os valores referentes a hora para *POSIX timestamp*, e os referentes a latitude e longitude para *legacy coordinate pairs*[5]. Além disso, visando viabilizar *geospatial*

[5] <https://docs.mongodb.com/manual/geospatial-queries>. Acesso em Outubro, 29 de 2017.

queries, foram criados *geospatial indexes*[5] nas *collections* contendo informações geolocalizadas, logo após serem criadas no *MongoDB*. Dessa forma, conseguimos usar *geospatial queries* para identificar as linhas afetadas por um determinado evento de exceção.

Tabela 5 – Conjuntos e quantidades de dados especificados em GTFS pela SPTrans

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

Fonte: Felipe Cordeiro Alves Dias

Por sua vez, os dados transmitidos por meio de AVL (ou simplesmente dados AVL) são em grande volume, devido a isso demandam processamento distribuído para que seja possível consultá-los. Visando abordar esse problema, utilizamos o *Druid* como banco de dados de séries temporais, por possuir uma arquitetura simples de escalar, ser distribuído e específico para a análises temporais.

Os dados AVL utilizados nesta análise (fornecidos pela empresa *Scipopulis*[6] — financiada pela FAPESP[7]) são referentes as linhas de ônibus 917H-10 (Terminal Pirituba — sentido 1 / Metrô Vila Mariana — sentido 0) e 477P (Ipiranga — sentido 1 / Rio Pequeno — sentido 0). Especificamente para a linha 477P dispomos dos dados de ambos os sentidos, e apenas do sentido 0 para a linha 917H, a quantidade de dados para cada *trip_id* (id da viagem — popularmente conhecido como linha do ônibus) é detalhada na tabela 6. No demais, os campos existentes nesse conjunto de dados são: *bus_id* (id do ônibus), *trip_id* (id da viagem), *hr* (GMT-03:00:00), *timestamp* (GMT), *lat* (latitude) e *lng* (longitude).

[6] <https://www.scipopulis.com/>. Acesso em Outubro, 29 de 2017.

[7] Fundação de Amparo à pesquisa do Estado de São Paulo: <http://www.fapesp.br/>. Acesso em Outubro, 29 de 2017.

Tabela 6 – 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.

Fonte: Felipe Cordeiro Alves Dias

Conforme pode ser observado, a base de dados AVL está limitada as linhas 4779-10 e 917H-10 (apenas um dos sentidos). Ainda é importante mencionar, que em 2017 foi noticiado[8] que a linha 917H-10 teve 96 reclamações e a 477P-10 75, classificado-as como as linhas com os maiores números de queixas.

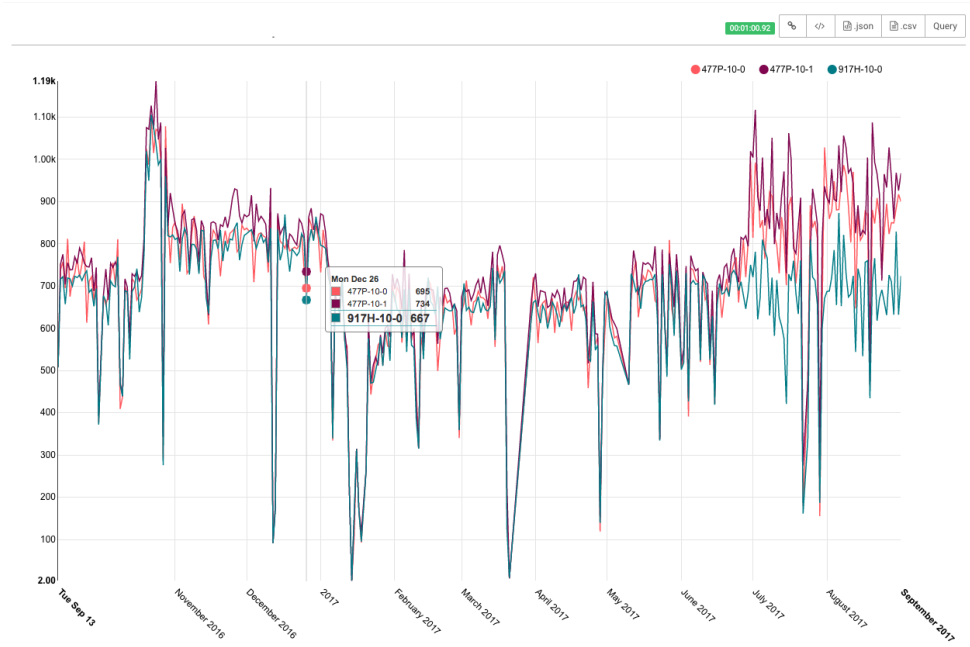
4.4 Exploração e visualização do conjunto de dados

O grande volume de dados que será analisado pode conter padrões complexos e difíceis de serem identificados. Devido a isso pretendemos construir visualizações auxiliares ao processo de análise dos dados. Com esse objetivo, utilizaremos o *Apache Superset*[9], que é integrado nativamente com o *Druid*, para exploração e visualização dos corpus desse projeto de pesquisa. As Fig. 5, 6 e 7 são exemplos de algumas visualizações construídas para os dados das linhas 477P-10 e 917H-10.

A Fig. 5 exibe séries temporais referentes a quantidade de dados enviados pelas viagens 477P-10-0, 477P-10 e 971H-10-0. Com essa visualização é possível observar, por exemplo, a oscilação da quantidade de dados enviados, assim como os picos de maior e menor volume de envio de dados, os quais podem indicar inúmeros problemas relacionados a essas viagens. O mesmo pode ser afirmado para a Fig. 6, que representa séries temporais sobre dados enviados por mês dos ônibus pertencentes as mesmas viagens já mencionadas. Por fim, o mapa exibido pela Fig. 7 ajuda a identificar a localização de onde os dados estão sendo enviados, o que permite identificar possíveis pontos de falha durante a transmissão desses dados.

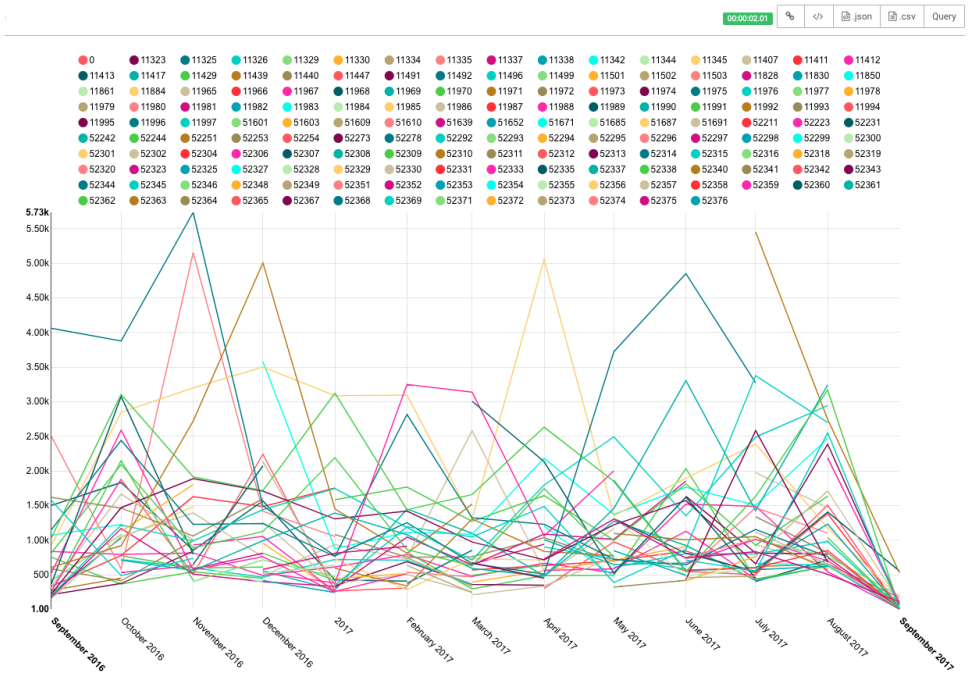
[8] <https://vejasp.abril.com.br/cidades/como-e-andar-na-pior-linha-de-onibus-da-capital>. Acesso em Outubro, 29 de 2017.
[9] <https://superset.incubator.apache.org>. Acesso em Outubro, 29 de 2017.

Figura 5 – Quantidade de dados enviados: viagens 477P-10-0, 477P-10-1 e 971H-10-0



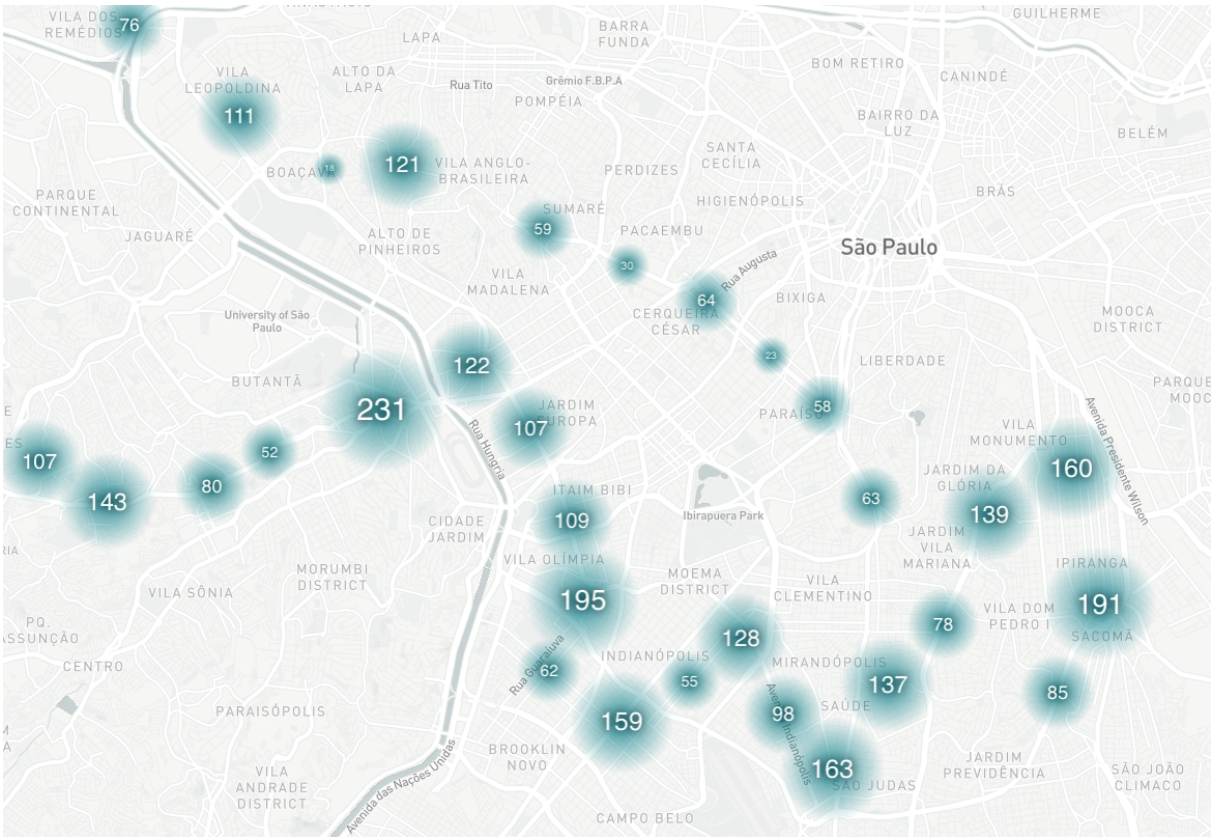
Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 6 – Quantidade de dados enviados por ônibus / mês: viagens 477P-10-0, 477P-10-1 e 971H-10-0



Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 7 – Localizações de envio dos dados: viagens 477P-10-0, 477P-10-1 e 971H-10-0



Fonte: Felipe Cordeiro Alves Dias, 2017

4.5 Identificação dos eventos de exceção

Nesta seção são apresentadas as atividades planejadas para realizar a identificação dos eventos de exceção com base nos dados do Corpus *Twitter*.

4.5.1 Pré-processamento

Numa pré-análise do *Corpus Twitter*, podemos afirmar que os *tweets* publicados pelos *profiles* selecionados evitam o uso de gírias, abreviações, erros de digitação; conforme consta nos *tweets* de exemplo contidos no trecho de código em *json*, no apêndice A. Isso diferencia tais *tweets* dos *tweets* publicados por usuários comuns do *Twitter*, que contém erros gramaticais, de sintaxe e que normalmente dependem de análise contextual para que possam ser interpretados.

Apesar disso, com base na literatura analisada ((STEIGER et al., 2015), (MIDDLETON; MIDDLETON; MODAFFERI, 2014), (KOBDANI; SCHÜTZE; BURKOVSKI, 2010), (SETIAWAN; WIDYANTORO; SURENDRO, 2017), (ZAGAL; MATA; CLARAMUNT, 2016)), as seguintes etapas de pré-processamento serão realizadas:

- ***Case folding***: processamento de normalização de todas as letras do texto (de a-z) para minúsculas (SETIAWAN; WIDYANTORO; SURENDRO, 2017).
- ***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 (SETIAWAN; WIDYANTORO; SURENDRO, 2017).
- **Remoção de *stopwords***: processamento para remoção do conjunto de *tokens* de palavras sem significado ou importância (SETIAWAN; WIDYANTORO; SURENDRO, 2017), o que reduz a quantidade de ruído do conteúdo *tweet* (STEIGER et al., 2015).
- ***Stemming***: processamento para encontrar a raiz de uma palavra, removendo sufixos e prefixos (no caso do Português Brasileiro) das palavras derivadas (SETIAWAN; WIDYANTORO; SURENDRO, 2017).

É importante observar que a extração dos endereços será feita após a fase de *case folding*, pois as demais podem comprometê-la ao remover trechos do texto necessários para a expressão regular definida para identificação dos padrões de logradouro. Também não está planejado para o pré-processamento proposto uma etapa de transformação do conteúdo dos *tweets*, embora seja utilizada em trabalhos como os relacionados a identificação de sentimentos, para transformar *emoticons* nos sentimentos que eles representam (ZAGAL; MATA; CLARAMUNT, 2016).

4.5.2 *Feature extraction*

A fase de *feature extraction* depende do conhecimento do domínio do objeto de estudo, além de normalmente envolver inúmeras iterações para obter um conjunto plausível de *features* (ZHU et al., 2013). Assim, inicialmente exploramos o domínio dos eventos de exceção relacionados ao transporte público (detalhados em 3.5.5) com o auxílio da questão QP5 (em 3.2) da revisão sistemática do Cap. 3.

Após a exploração do conhecimento do domínio (já realizada), pretendemos na primeira iteração para extração de *features* 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 entrei si e derivar novas *features*, de acordo com o conhecimento do domínio.

4.5.3 *Feature selection*

Pretendemos na fase de *feature selection* encontrar as *features* mais relevantes para a classificação dos eventos de exceção, pois com um conjunto relevante de *features* evitamos um modelo de classificação com sobre-ajuste (*overfitting*) e de alto custo computacional (ZHU et al., 2013). Assim, 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*.

4.5.4 Algoritmos de Aprendizado de Máquina

Após a extração e seleção de *features*, planejamos classificar manualmente 30% dos *tweets* com base em suas respectivas *features*, utilizando-os como conjunto de teste. Posteriormente, pretendemos analisar os algoritmos de aprendizado de máquina elencados pela revisão sistemática em 3.5.6 para escolhermos dentre eles o com maior acurácia para identificar eventos de exceção, por meio de classificação.

4.6 Correlação dos eventos de exceção com os dados AVL da SPTrans

Dado que os eventos de exceção podem ser identificados utilizando *tweets* dos *profiles* contidos na tabela 1, há também a possibilidade de caracterizarmos seus respectivos impactos analisando a base histórica dos dados AVL da SPTrans, especificamente os dados referentes a *timestamp*, latitude, longitude, *bus_id* e *trip_id*. Dito isso, inicialmente pretendemos caracterizar os impactos em:

- 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.

4.7 Plano de trabalho

Nesta seção, são apresentados o cronograma (tabela 7) e as atividades realizadas e planejadas para o desenvolvimento do projeto referente a proposta de pesquisa, enumeradas a seguir:

1. Revisão Bibliográfica.
    a) Revisão Sistemática sobre estudos de caso utilizando *tweets* na caracterização de problemas urbanos e relacionados ao transporte público.
2. Desenvolvimento de protótipo.
    a) Serviço para coleta, processamento e armazenamento de *tweets* dos *profiles* selecionados contidos na tabela 1.
    b) Extração e geolocalização dos endereços contidos nos *tweets*.
    c) Implementação de *scripts* para extração, transformação e armazenamento dos dados AVL da SPTrans.
    d) Implementação de *scripts* para extração, transformação e armazenamento dos dados da GTFS da SPTrans.
    e) Criação de especificações de ingestão de dados para os dados AVL da SPTrans e dos *tweets* dos *profiles* selecionados contidos na tabela 1 para o banco de dados de séries temporais *Druid*[10].

[10] <http://druid.io>. Acesso em Outubro, 29 de 2017.

f) Integração da ferramenta *Superset*[11] com o *Druid*[10] para exploração e visualização dos dados AVL da SPTrans e dos *tweets* dos *profiles* selecionados contidos na tabela 1.

g) Implementação de *scripts* para automação dos processos de ingestão de dados e *deploy* e monitoramento dos serviços do *Druid*[10] e Superset[11].

h) Configuração do ambiente em nuvem para execução do *Druid*[10] e *Superset*[11].

3. Construção do conjunto de dados.

   a) Obtenção dos dados AVL de janeiro a dezembro de 2016 e de janeiro a setembro de 2017, de todas as linhas de ônibus de São Paulo.

   b) Obtenção da base de dados dos últimos cinco anos das reclamações relacionadas a SPTrans, realizadas na Central de Atendimento ao Cidadão (156) da Prefeitura de São Paulo.

   c) Obtenção de um conjunto de *tweets* dos *profiles* selecionados contidos na tabela 1.

4. Implementação da solução proposta.

   a) Identificação dos eventos de exceção (pré-processamento, *feature extraction* e *feature selection* dos *tweets* coletados).

   b) Estudo dos algoritmos de classificação e implementação de um artefato de *software* para classificação dos *tweets* de acordo com seus respectivos eventos de exceção.

   c) Correlação dos eventos de exceção com os dados AVL da SPTrans.

5. Avaliação dos resultados parciais obtidos durante e após o desenvolvimento da solução proposta.
6. Escrita de artigo para submissão em periódicos ou eventos da área.
7. Escrita da dissertação.

[11] <http://superset.apache.org>. Acesso em Outubro, 29 de 2017.

Tabela 7 – Cronograma de atividades

Atividade		2017												2018					
Número	Descrição	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							
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

Fonte: Felipe Cordeiro Alves Dias, 2017

5 Considerações finais

Neste capítulo, são apresentadas as contribuições e resultados esperados com o projeto de pesquisa, as limitações a ameaças à validade do estudo.

5.1 Contribuições esperadas

A principal contribuição deste projeto é propor 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. Além disso, a solução proposta visa disponibilizar os conjuntos de dados que foram construídos e uma plataforma para que esses dados possam ser visualizados e explorados, de forma a contribuir com projetos e pesquisas futuras correlatas.

Em relação a publicações científicas, serão submetidos 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.

5.2 Limitações e riscos à validade do estudo

As principais limitações deste projeto estão relacionadas ao processamento de *tweets* em português brasileiro e oriundos das contas selecionadas e referenciadas na tabela 1, o que pode tornar a solução não generalista. Dentre os riscos, apesar das análises preliminares realizadas para extração de endereços dos conteúdos dos *tweets* por meio de Expressão Regular, é possível que sejam encontrados novos desafios que inviabilizem o uso dessa técnica.

Referências

ABBASI, A. et al. Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play. *Proc. 8th ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN'15*, p. 1–9, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2830657.2830660>. Citado 6 vezes nas páginas 33, 34, 35, 41, 42 e 44.

AHVENNIEMI, H. et al. What are the differences between sustainable and smart cities? *Cities*, Elsevier B.V., v. 60, p. 234–245, 2017. ISSN 02642751. Disponível em: <http://dx.doi.org/10.1016/j.cities.2016.09.009>. Citado 3 vezes nas páginas 12, 13 e 14.

ALBINO, V.; BERARDI, U.; DANGELICO, R. M. Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, Taylor & Francis, v. 22, n. 1, p. 3–21, 2015. Citado na página 14.

ANANTHARAM, P. et al. Extracting City Traffic Events from Social Streams. *ACM Trans. Intell. Syst. Technol.*, v. 6, n. 4, p. 1–27, 2015. ISSN 21576904. Disponível em: <http://dl.acm.org/citation.cfm?doid=2801030.2717317>. Citado 7 vezes nas páginas 34, 36, 39, 40, 41, 42 e 43.

Andreas Mueller. 2018. <https://pypi.python.org/pypi/wordcloud>. Acesso em Fevereiro, 13 de 2018. Citado na página 32.

ANG, L.-M. et al. Big Sensor Data Systems for Smart Cities. *IEEE Internet Things J.*, v. 4, n. 5, p. 1–1, 2017. ISSN 2327-4662. Disponível em: <http://ieeexplore.ieee.org/document/7903653/>. Citado 2 vezes nas páginas 13 e 14.

ANTTIROIKO, A. V. U-cities reshaping our future: Reflections on ubiquitous infrastructure as an enabler of smart urban development. *AI Soc.*, v. 28, n. 4, p. 491–507, 2013. ISSN 09515666. Citado na página 7.

ATEFEH, F.; KHREICH, W. A survey of techniques for event detection in twitter. *Computational Intelligence*, Wiley Online Library, v. 31, n. 1, p. 132–164, 2015. Citado na página 21.

BARTH, J. et al. Informational urbanism . A conceptual framework of smart cities. *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, p. 2814–2823, 2017. Citado 2 vezes nas páginas 13 e 14.

BENDLER, J. et al. Taming Uncertainty in Big Data. *Bus. Inf. Syst. Eng.*, v. 6, n. 5, p. 279–288, 2014. ISSN 1867-0202. Disponível em: <http://link.springer.com/10.1007/s12599-014-0342-4>. Citado 5 vezes nas páginas 33, 35, 39, 40 e 41.

BIOLCHINI, J. et al. Techincal report rt-es 679/05: Systematic review in software engineering. *COPPE/UFRJ, 2005Rio de Janeiro*, 2005. Citado 2 vezes nas páginas 26 e 27.

CHANIOTAKIS, E.; ANTONIOU, C. Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter. *IEEE Conf. Intell. Transp.*

Syst. Proceedings, ITSC, v. 2015-Octob, n. 1, p. 214–219, 2015. Citado 2 vezes nas páginas 34 e 41.

CHANIOTAKIS, E.; ANTONIOU, C.; PEREIRA, F. Mapping Social media for transportation studies. *IEEE Intell. Syst.*, v. 31, n. 6, p. 64–70, 2016. ISSN 15411672. Citado na página 27.

CHEN, L. et al. Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. *UBICOMP*, p. 841–852, 2016. Citado 11 vezes nas páginas 10, 11, 33, 36, 37, 38, 39, 40, 41, 42 e 43.

CHUA, A. et al. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tour. Manag.*, Elsevier Ltd, v. 57, p. 295–310, 2016. ISSN 02615177. Disponível em: <http://dx.doi.org/10.1016/j.tourman.2016.06.013>. Citado 2 vezes nas páginas 34 e 35.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, v. 12, n. Aug, p. 2493–2537, 2011. Citado na página 22.

CONSULO, M. et al. An evaluation of the proposed its system for the city of são paulo based on the 2015 tender. In: EDP SCIENCES. *MATEC Web of Conferences*. [S.l.], 2016. v. 76, p. 03004. Citado 2 vezes nas páginas 7 e 8.

DI LORENZO, G. et al. EXSED: An intelligent tool for exploration of social events dynamics from augmented trajectories. *Proc. - IEEE Int. Conf. Mob. Data Manag.*, v. 1, p. 323–330, 2013. ISSN 15516245. Citado 5 vezes nas páginas 33, 36, 41, 42 e 44.

DIAS, F. *Repositório contendo os artefatos da Revisão Sistemática*. 2017. Disponível em: <https://github.com/fcas/dissertacao>. Citado na página 31.

FARSEEV, A. et al. Harvesting Multiple Sources for User Profile Learning. *Proc. 5th ACM Int. Conf. Multimed. Retr. - ICMR '15*, p. 235–242, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2671188.2749381>. Citado 6 vezes nas páginas 33, 34, 41, 42, 43 e 44.

FIGUEIREDO, L. et al. Towards the development of intelligent transportation systems. In: IEEE. *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*. [S.l.], 2001. p. 1206–1211. Citado 2 vezes nas páginas 14 e 15.

FINGER, M.; RAZAGHI, M. Conceptualizing "Smart Cities". *Informatik-Spektrum*, v. 40, n. 1, p. 6–13, 2017. ISSN 1432122X. Citado 3 vezes nas páginas 12, 13 e 14.

FRIAS-MARTINEZ, V.; FRIAS-MARTINEZ, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.*, Elsevier, v. 35, p. 237–245, 2014. ISSN 09521976. Disponível em: <http://dx.doi.org/10.1016/j.engappai.2014.06.019>. Citado 7 vezes nas páginas 33, 36, 39, 40, 41, 42 e 43.

GAL-TZUR, A. et al. The potential of social media in delivering transport policy goals. *Transp. Policy*, v. 32, p. 115–123, 2014. ISSN 0967070X. Citado 7 vezes nas páginas 10, 37, 38, 40, 41, 42 e 43.

GKIOTSALITIS, K.; STATHOPOULOS, A. A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data. *Transp. Res. Part C Emerg. Technol.*, Elsevier Ltd, v. 58, p. 265–277, 2015. ISSN 0968090X. Disponível em: <http://dx.doi.org/10.1016/j.trc.2014.12.006>. Citado 3 vezes nas páginas 33, 34 e 38.

GKIOTSALITIS, K.; STATHOPOULOS, A. Joint leisure travel optimization with user-generated data via perceived utility maximization. *Transp. Res. Part C Emerg. Technol.*, Elsevier Ltd, v. 68, p. 532–548, 2016. ISSN 0968090X. Disponível em: <http://dx.doi.org/10.1016/j.trc.2016.05.009>. Citado 5 vezes nas páginas 33, 34, 37, 38 e 41.

GUO, W. et al. Understanding happiness in cities using twitter: Jobs, children, and transport. *IEEE 2nd Int. Smart Cities Conf. Improv. Citizens Qual. Life, ISC2 2016 - Proc.*, 2016. Citado 6 vezes nas páginas 34, 35, 37, 39, 40 e 43.

GUTEV, A.; NENKO, A. Better Cycling - Better Life: Social Media Based Parametric Modeling Advancing Governance of Public Transportation System in St. Petersburg. *Proc. Int. Conf. Electron. Gov. Open Soc. Challenges Eurasia*, p. 242–247, 2016. Disponível em: <http://doi.acm.org/10.1145/3014087.3014123>. Citado 7 vezes nas páginas 33, 34, 35, 37, 38, 39 e 41.

GUYON, I.; ELISSEEFF, A. An introduction to feature extraction. *Feature extraction*, Springer, p. 1–25, 2006. Citado 2 vezes nas páginas 23 e 24.

HASAN, S.; UKKUSURI, S. V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C Emerg. Technol.*, Elsevier Ltd, v. 44, p. 363–381, 2014. ISSN 0968090X. Disponível em: <http://dx.doi.org/10.1016/j.trc.2014.04.003>. Citado 4 vezes nas páginas 33, 35, 42 e 44.

ITOH, M. et al. Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network. *IEEE Trans. Big Data*, v. 2, n. 1, p. 85–99, 2016. ISSN 2332-7790. Disponível em: <http://ieeexplore.ieee.org/document/7445832/>. Citado 6 vezes nas páginas 10, 11, 37, 38, 41 e 42.

JUNGHERR, A. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, Taylor & Francis, v. 13, n. 1, p. 72–91, 2016. Citado na página 27.

KHAN, A. et al. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, Academy Publisher, PO Box 40 Oulu 90571 Finland, v. 1, n. 1, p. 4–20, 2010. Citado na página 25.

KOBDANI, H.; SCHÜTZE, H.; BURKOVSKI, A. Relational feature engineering of natural language processing. *Proc. 19th . . .*, n. ii, p. 1705–1708, 2010. Disponível em: <http://dl.acm.org/citation.cfm?id=1871709>. Citado na página 54.

KORENIUS, T. et al. Stemming and lemmatization in the clustering of finnish text documents. In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2004. (CIKM '04), p. 625–633. ISBN 1-58113-874-1. Disponível em: <http://doi.acm.org/10.1145/1031171.1031285>. Citado na página 22.

KUFLIK, T. et al. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 77, p. 275–291, 2017. Citado 2 vezes nas páginas 8 e 9.

KUMMITHA, R. K. R.; CRUTZEN, N. How do we understand smart cities? An evolutionary perspective. *Cities*, Elsevier, v. 67, n. July 2016, p. 43–52, 2017. ISSN 02642751. Disponível em: <http://dx.doi.org/10.1016/j.cities.2017.04.010>. Citado 3 vezes nas páginas 12, 13 e 14.

LECUE, F. et al. Smart traffic analytics in the semantic web with STAR-CITY: Scenarios, system and lessons learned in Dublin City. *J. Web Semant.*, Elsevier B.V., v. 27, p. 26–33, 2014. ISSN 15708268. Disponível em: <http://dx.doi.org/10.1016/j.websem.2014.07.002>. Citado 6 vezes nas páginas 10, 11, 34, 36, 41 e 42.

LIU, D.; LI, Y.; THOMAS, M. A. A roadmap for natural language processing research in information systems. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. [S.l.: s.n.], 2017. Citado na página 21.

MAGHREBI, M. et al. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, v. 2015-Octob, p. 208–213, 2015. Citado 4 vezes nas páginas 33, 35, 37 e 38.

MATA, F.; CLARAMUNT, C. A Mobile Trusted Path System Based on Social Network Data. *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, p. 101:1—-101:4, 2015. Disponível em: <http://doi.acm.org/10.1145/2820783.2820799>. Citado 4 vezes nas páginas 34, 41, 42 e 43.

MENOUAR, H. et al. Uav-enabled intelligent transportation systems for the smart city: Applications and challenges. *IEEE Communications Magazine*, IEEE, v. 55, n. 3, p. 22–28, 2017. Citado 2 vezes nas páginas 14 e 15.

MIDDLETON, S. E.; MIDDLETON, L.; MODAFFERI, S. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, v. 29, n. 2, p. 9–17, 2014. ISSN 15411672. Citado na página 54.

MORENO, M. V. et al. Applicability of Big Data Techniques to Smart Cities Deployments. *IEEE Trans. Ind. Informatics*, v. 13, n. 2, p. 800–809, 2017. ISSN 15513203. Citado 2 vezes nas páginas 13 e 14.

MOTODA, H.; LIU, H. Feature selection, extraction and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol*, v. 5, p. 67–72, 2002. Citado 2 vezes nas páginas 23 e 24.

MUKHERJEE, T. et al. Janayuja: A People-centric Platform to Generate Reliable and Actionable Insights for Civic Agencies. *Acm Dev 2015*, p. 137–145, 2015. Citado 9 vezes nas páginas 33, 36, 37, 38, 39, 40, 41, 42 e 43.

MYERS, S. A. et al. Information network or social network?: the structure of the twitter follow graph. In: ACM. *Proceedings of the 23rd International Conference on World Wide Web*. [S.l.], 2014. p. 493–498. Citado na página 21.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011. Citado 2 vezes nas páginas 22 e 23.

NELSON, J. D.; MULLEY, C. The impact of the application of new technology on public transport service provision and the passenger experience: A focus on implementation in Australia. *Res. Transp. Econ.*, Elsevier Ltd, v. 39, n. 1, p. 300–308, 2013. ISSN 07398859. Disponível em: <http://dx.doi.org/10.1016/j.retrec.2012.06.028>. Citado na página 8.

NI, M.; HE, Q.; GAO, J. Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Trans. Intell. Transp. Syst.*, v. 18, n. 6, p. 1623–1632, 2016. ISSN 15249050. Citado 6 vezes nas páginas 37, 38, 39, 40, 42 e 44.

ROY, A.; MAJUMDER, A. G.; NATH, A. Understanding natural language processing and its primary aspects. *International Journal*, v. 5, n. 8, 2017. Citado na página 22.

SANTOS, H. et al. Contextual data collection for smart cities. *CoRR*, abs/1704.01802, 2017. Disponível em: <http://arxiv.org/abs/1704.01802>. Citado 2 vezes nas páginas 13 e 14.

SETIAWAN, E. B.; WIDYANTORO, D. H.; SURENDRO, K. Feature expansion using word embedding for tweet topic classification. *Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol.*, n. 2011, 2017. Citado 2 vezes nas páginas 22 e 54.

SOBOLEVSKY, S. et al. Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity. *Proc. - 2015 IEEE Int. Congr. Big Data, BigData Congr. 2015*, p. 600–607, 2015. ISSN 2379-7703. Citado 2 vezes nas páginas 34 e 35.

SOOMRO, K.; KHAN, Z.; HASHAM, K. Towards Provisioning of Real-time Smart City Services Using Clouds. *ACM 9th Int. Conf. Util. Cloud Comput. Towar.*, v. 1691, p. 50–59, 2016. ISSN 16130073. Citado 3 vezes nas páginas 33, 36 e 41.

STEIGER, E.; ALBUQUERQUE, J. P.; ZIPF, A. An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, Wiley Online Library, v. 19, n. 6, p. 809–834, 2015. Citado na página 27.

STEIGER, E. et al. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Comput. Environ. Urban Syst.*, Elsevier Ltd, v. 54, p. 255–265, 2015. ISSN 01989715. Disponível em: <http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.007>. Citado 7 vezes nas páginas 34, 35, 39, 40, 41, 44 e 54.

SÁ, T. H. et al. Health impact modelling of different travel patterns on physical activity, air pollution and road injuries for são paulo, brazil. *Environment International*, v. 108, n. Supplement C, p. 22 – 31, 2017. ISSN 0160-4120. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0160412017305974>. Citado na página 6.

TALARI, S. et al. A Review of Smart Cities Based on the Internet of Things Concept. *Energies*, v. 10, n. 4, p. 421, 2017. ISSN 1996-1073. Disponível em: <http://www.mdpi.com/1996-1073/10/4/421>. Citado 2 vezes nas páginas 13 e 14.

THOMAZ, G. M. et al. Content mining framework in social media: A FIFA world cup 2014 case analysis. *Inf. Manag.*, Elsevier B.V., 2016. ISSN 03787206. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0378720616303354>. Citado 2 vezes nas páginas 34 e 35.

United States Department of Transportation. *ITS Strategic Plan 2015-2019*. 2017. <https://www.its.dot.gov/strategicplan.pdf>. Acesso em Setembro, 17 de 2017. Citado na página 8.

WANG, S.; SINNOTT, R.; NEPAL, S. Privacy-protected social media user trajectories calibration. *Proc. 2016 IEEE 12th Int. Conf. e-Science, e-Science 2016*, p. 293–302, 2016. Citado 2 vezes nas páginas 12 e 35.

WEN, X.; LIN, Y.-R.; PELECHRINIS, K. PairFac: Event Analytics through Discriminant Tensor Factorization. *Cikm*, p. 519–528, 2016. Citado 5 vezes nas páginas 34, 37, 39, 41 e 42.

XIAO, Z.; LIM, H. B.; PONNAMBALAM, L. Participatory Sensing for Smart Cities: A Case Study on Transport Trip Quality Measurement. *IEEE Trans. Ind. Informatics*, v. 13, n. 2, p. 759–770, 2017. ISSN 1551-3203. Citado 2 vezes nas páginas 13 e 14.

YOUSAF, J. et al. Generalized multipath planning model for ride-sharing systems. *Front. Comput. Sci.*, v. 8, n. 1, p. 100–118, 2014. ISSN 20952228. Citado 5 vezes nas páginas 33, 36, 39, 40 e 41.

ZAGAL, R.; MATA, F.; CLARAMUNT, C. Geographical Knowledge Discovery applied to the Social Perception of Pollution in the City of Mexico. *LBSN*, 2016. Citado 4 vezes nas páginas 34, 42, 43 e 54.

ZHOU, X.; CHEN, L. Event detection over twitter social media streams. *The VLDB journal*, Springer, v. 23, n. 3, p. 381–400, 2014. Citado na página 21.

ZHU, Y. et al. Feature engineering for semantic place prediction. *Pervasive Mob. Comput.*, Elsevier B.V., v. 9, n. 6, p. 772–783, 2013. ISSN 15741192. Disponível em: <http://dx.doi.org/10.1016/j.pmcj.2013.07.004>. Citado 2 vezes nas páginas 54 e 55.

Apêndices

Apêndice A – Exemplos de *tweets*

Exemplos de *tweets* dos *profiles* selecionados citados na tabela 1


```
{
    "tweet_id" : 895060642952077314,
    "tweet_account": "BombeirosPMESP",
    "text" : "19h58 Colisão de Carro x Caminhão, Estrada Sta Isabel,
       5950 Itaquaquecetuba. 2 Vítimas, 1 Vtr. Aguardando maiores
       informes"
}
{
    "tweet_id" : 894707930217447427,
    "tweet_account": "CETSP_",
    "text" : "Referente manifestação Rua Augusta, pista liberada.#ZC"
}
{
    "tweet_id" : 894147793060716544,
    "tweet_account": "CPTM_oficial",
    "text" : "#L11 Hoje, das 8h à meia-noite, circulação interrompida
       entre Luz e Brás. P/ seguir viagem, use a L7-Rubi q prestará
       serviço até a Est. Brás"
}
{
    "tweet_id" : 895054721026838530,
    "tweet_account": "governosp",
    "text" : "@SANROGE Lamentamos o ocorrido, Rogerio. Estamos
       trabalhando continuamente para melhorar a segurança na região.
       Entre maio e junho, [+] [1]"
}
{
    "tweet_id" : 895000711284621312,
    "tweet_account": "metrosp_oficial",
```

```
    "text" : "08/08/2017 16:16: #metrosp : Linha 5-Lilás: Velocidade
       Reduzida. Mais informações em https://t.co/CaeqD26iJR"
}
{
    "tweet_id" : 884039273493803008,
    "tweet_account": "PMESP",
    "text" : "AGORA: Desfile Cívico-Militar de 9 de Julho no Obelisco
       - Ibirapuera SP, transmissão ao vivo na página oficial Facebook
        da Polícia Militar.",
    "dateTime" : "2017-07-09 10:19:22"
}
{
    "tweet_id" : 887315002117500932,
    "tweet_account": "Policia_Civil",
    "text" : "Polícia Civil realiza operação para combater a prática
       do Jogo conhecido como "Baleia Azul"... https://t.co/kh2HW6UZvT
       ",
}
{
    "tweet_id" : 895004079910518788,
    "tweet_account": "saopaulo_agora",
    "text" : "#ItaimPaulista Incêndio na Rua Mateus Barbosa de Resende
        n° 235. Defesa Civil Regional acionada para o local. (CCOI) #
       spagora"
}
{
    "tweet_id" : 894694704989732864,
    "tweet_account": "smtpsp_",
    "text" : "A @sptrans_ irá modificar 14 linhas na Zona Leste para
       obras no Monotrilho Saiba mais: https://t.co/fCA0T7WCSY"
}
{
```

```
    "tweet_id" : 902953598857949184,
    "tweet_account": "SPCEDEC",
    "text" : "30-08-2017 - Acidente com produto perigoso em  com 36 ,
       deixa 21 vítimas feridas e 02 ."
}
{
    "tweet_id" : 895065137484320769,
    "tweet_account": "sptrans_",
    "text" : "Obras do Monotrilho desviam itinerários de 14 linhas que
        atendem a Av. Sapopemba entre 5 e 11/08, das 23h às 5h: https:
       //t.co/jH4LFgrSKZ"
}
{
    "tweet_id : 895042604068458497,
    "tweet_account": "TurismoSaoPaulo",
    "text" : "Veganos, vegetarianos e simpatizantes: vem aí o Vegan
       Club, em 12/08, no Centro de SP! #crueltyfree #veganfood...
       https://t.co/7f7ggr4vn4"
}
```

Apêndice B – Logradouros utilizados

Tabela 8 – Tabela de logradouros com abreviaturas

Abreviatura	Logradouro
ACAMP	Acampamento
AC	Acesso
AD	Adro
ERA	Aeroporto
AL	Alameda
AT	Alto
A	Area
AE	Area especial
ART	Arteria
ATL	Atalho
AV	Avenida
AV-CONT	Avenida contorno
BX	Baixa
BLO	Balao
BAL	Balneario
BC	Beco
BELV	Belvedere
BL	Bloco
BSQ	Bosque
BVD	Boulevard
BCO	Buraco
C	Cais
CALC	Calcada
CAM	Caminho
CPO	Campo
CAN	Canal
CHAP	Chacara

Continua na próxima página

Tabela 8 – continuação da página anterior

Abreviatura	Logradouro
CHAP	Chapadao
CIRC	Circular
COL	Colonia
CMP-VR	Complexo viario
COND	Condominio
CJ	Conjunto
COR	Corredor
CRG	Corrego
DSC	Descida
DSV	Desvio
DT	Distrito
EVD	Elevada
ENT-PART	Entrada particular
EQ	Entre quadra
ESC	Escada
ESP	Esplanada
ETC	Estacao
ESTC	Estacionamento
ETD	Estadio
ETN	Estancia
EST	Estrada
EST-MUN	Estrada municipal
FAV	Favela
FAZ	Fazenda
FRA	Feira
FER	Ferrovia
FNT	Fonte
FTE	Forte
GAL	Galeria

Continua na próxima página

Tabela 8 – continuação da página anterior

Abreviatura	Logradouro
GJA	Granja
HAB	Habitacional
IA	Ilha
JD	Jardim
JDE	Jardinete
LD	Ladeira
LG	Lago
LGA	Lagoa
LRG	Largo
LOT	Loteamento
MNA	Marina
MOD	Modulo
TEM	Monte
MRO	Morro
NUC	Nucleo
PDA	Parada
PDO	Paradouro
PAR	Paralela
PRQ	Parque
PSG	Passagem
PSC-SUB	Passagem subterranea
PSA	Passarela
PAS	Passeio
PAT	Patio
PNT	Ponta
PTE	Ponte
PTO	Porto
PC	Praca
PC-ESP	Praça de esportes

Continua na próxima página

Tabela 8 – continuação da página anterior

Abreviatura	Logradouro
PR	Praia
PRL	Prolongamento
Q	Quadra
QTA	Quinta
QTAS	Quinta
RAM	Rama
RMP	Rampa
REC	Recanto
RES	Residencial
RET	Reta
RER	Retiro
RTN	Retorno
ROD-AN	RodoAnel
ROD	Rodovia
RTT	Rotatoria
ROT	Rotula
R	Rua
R-LIG	Rua de ligação
R-PED	Rua de pedrestre
SRV	Servidao
ST	Setor
SIT	Sitio
SUB	Subida
TER	Terminal
TV	Travessa
TV-PART	Travessa particular
TRV	Trecho
TRV	Trevo
TCH	Trincheira

Continua na próxima página

Tabela 8 – continuação da página anterior

Abreviatura	Logradouro
TUN	Tunel
UNID	Unidade
VAL	Vala
VLE	Vale
VRTE	Variante
VER	Vereda
V	Via
V-AC	Via de acesso
V-PED	Via de pedestre
V-EVD	Via elevado
V-EXP	Via expressa
VD	Viaduto
VLA	Viela
VL	Vila
ZIG-ZAG	Zigue-zague

Fonte: MS/SAS/DRAC/CGSI - Coordenação Geral dos Sistemas de Informação (adaptada)[1]

[1] <http://www.pmf.sc.gov.br/arquivos/arquivos/pdf/04_01_2010_10.27.25.2b615e6755138defe1bdb00f1c86031f.PDF>. Acesso em Outubro, 29 de 2017.

Apêndice C – Detalhamento dos campos da GTFS

Tabela 9 – Detalhamento dos campos do arquivo *agency.txt* da GTFS

Nome do campo	Condicional	Descrição
agency_id	Opcional	Identifica uma agência de transporte público. Um *feed* de transporte público pode representar dados de mais de uma agência. Este campo é opcional para *feeds* de transporte público que contenham somente dados de uma única agência.
agency_name	Obrigatório	Contém o nome completo da agência de transporte público.
agency_url	Obrigatório	Contém o *URL* da agência de transporte público.
agency_timezone	Obrigatório	Contém o fuso horário de onde a agência de transporte público está localizada.
agency_lang	Opcional	Contém um código *ISO 639-1* de duas letras para o idioma principal usado por essa agência de transporte público.
agency_phone	Opcional	Contém um único número de telefone da agência especificada.
agency_fare_url	Opcional	Especifica o *URL* de uma página da *Web* que permite que um passageiro compre passagens ou outros instrumentos de tarifas dessa agência *on-line*.

Fonte: Google Transit (adaptada)[1]

[1] <https://developers.google.com/transit>. Acesso em Outubro, 29 de 2017.

Tabela 10 – Detalhamento dos campos do arquivo *stops.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
stop_id	Obrigatório	Contém um ID que identifica uma parada ou uma estação. Diversos trajetos podem usar a mesma parada.
stop_code	Opcional	Contém um pequeno texto ou um número que identifica a parada para os passageiros. Os códigos das paradas são usados muitas vezes em sistemas de informações sobre transporte público por telefone ou impressos em sinalizações nas paradas para que os passageiros possam obter informações sobre o horário das paradas com mais facilidade ou sobre chegadas de uma parada específica em tempo real. O campo *stop_code* só deve ser usado para códigos de parada exibidos aos passageiros. Para os códigos internos, use *stop_id*. Este campo deve ser deixado em branco para as paradas que não têm um código.
stop_name	Obrigatório	Contém o nome de uma parada ou estação. Use um nome compreensível para as pessoas locais e linguagem turística.
stop_desc	Opcional	Contém uma descrição de uma parada. Forneça informações úteis e de qualidade. Não basta repetir o nome da parada.
stop_lat	Obrigatório	Contém a latitude de uma parada ou estação. O valor do campo deve ser uma latitude WGS 84 válida.

Continua na próxima página

Tabela 10 – continuação da página anterior

Nome do campo	Condicional	Descrição
stop_lon	Obrigatório	Contém a longitude de uma parada ou estação. O valor do campo deve ser uma latitude WGS 84 válida entre -180 e 180.
zone_id	Opcional	Define a zona tarifária do ID de uma parada. Os IDs de zonas são obrigatórios para fornecer informações sobre tarifas usando *fare_rules.txt*. Se esse ID de parada representa uma estação, o ID de zona é ignorado.
stop_url	Opcional	Contém o URL de uma página da Web sobre uma parada específica. Ele deve ser diferente dos campos *agency_url* e *route_url*.
location_type	Opcional	Identifica se este ID de parada representa uma parada ou uma estação. Se nenhum tipo de local for especificado ou se o campo *location_type* estiver em branco, os IDs de parada serão tratados como paradas. As estações podem ter propriedades diferentes das paradas quando são representadas em um mapa ou usadas em planejamento de viagens. O campo de tipo de local pode ter os seguintes valores: 0 ou em branco (para parada) e 1 (estação).

Continua na próxima página

Tabela 10 – continuação da página anterior

Nome do campo	Condicional	Descrição
parent_station	Opcional	Para paradas que estejam fisicamente localizadas dentro de estações, o campo *parent_station* identifica a estação associada à parada. Para usar este campo, o arquivo *stops.txt* também deve conter uma linha em que esse ID de parada tenha o tipo de localização=1.

Continua na próxima página

Tabela 10 – continuação da página anterior

Nome do campo	Condicional	Descrição
stop_timezone	Opcional	Contém o fuso horário em que a parada ou estação está localizada. Se omitido, assume-se que a parada está localizada no fuso horário especificado por *agency_timezone* no arquivo *agency.txt*. Quando uma parada tem uma estação principal, considera-se que a parada esteja no fuso horário especificado pelo valor *stop_timezone* da estação principal. Se uma parada específica possui um valor *parent_station*, qualquer valor *stop_timezone* especificado para essa parada deve ser ignorado. Mesmo que os valores de *stop_timezone* sejam fornecidos no arquivo *stops.txt*, os horários em *stop_times.txt* devem continuar a ser especificados como horários desde a meia-noie no fuso horário especificado por *agency_timezone* em *agency.txt*. Isso garante que os valores de tempo em uma viagem sempre aumentam durante uma viagem, independentemente dos fusos horários pelos quais uma viagem passa.

Continua na próxima página

Tabela 10 – continuação da página anterior

Nome do campo	Condicional	Descrição
wheelchair_boarding	Opcional	Identifica se é possível o embarque de passageiros em cadeira de rodas na parada ou estação especificada. O campo pode ter os seguintes valores: 0 (ou vazio) - indica que não há informações sobre acessibilidade para a parada; 1 - indica que, pelo menos, alguns veículos nesta parada possibilitam o embarque de passageiros em cadeira de rodas; 2 - o embarque de pessoas em cadeiras de roda não é possível nesta parada. Quando uma parada faz parte de um complexo de estações maiores, como indicado por uma para com um valor *parent_station*, o campo *wheelchair_boarding* da parada possui a seguinte semântica adicional: 0 (ou vazio) - a parada herdará o valor para *wheelchair_boarding* da estação principal, se especificado; 1 - existem vias de acesso na parte externa da estação para a parada/plataforma específic; 2 - não há vias de acesso na parte externa da estação para a parada/plataforma específica

Fonte: Google Transit (adaptada)[1]

Tabela 11 – Detalhamento dos campos do arquivo *routes.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
route_id	Obrigatório	Contém um ID que identifica um trajeto.
agency_id	Opcional	Define uma agência para o trajeto especificado. Este valor é indicado no arquivo *agency.txt*. Campo destinado para quando for fornecido dados para trajetos de mais de uma agência.
route_short_name	Obrigatório	Contém o nome abreviado de um trajeto. Geralmente, será um identificador pequeno e abstrato, como, por exemplo "32", "100X"ou "Verde", que os passageiros usam para identificar um trajeto, mas que não fornece nenhuma identificação de quais lugares são atendidos pelo trajeto. Se o trajeto não tem um nome abreviado, especifique um *route_long_name* e use uma sequência vazia como o valor deste campo.
route_long_name	Obrigatório	Contém o nome completo de um trajeto. Em geral, esse nome é mais descritivo que *route_short_name* e incluirá o destino ou a parada do trajeto. Se o trajeto não tem um nome completo, especifique um *route_short_name* e use uma sequência vazia como o valor deste campo.
route_desc	Opcional	Contém uma descrição de um trajeto. Não basta repetir o nome do trajeto.

Continua na próxima página

Tabela 11 – continuação da página anterior

Nome do campo	Condicional	Descrição
route_type	Obrigatório	Descreve o tipo de transporte usado em um trajeto. Os valores válidos deste campo são: 0 - Bonde, ônibus elétrico, veículo leve sobre trilhos; 1 - Metrô, trem subterrâneo; 2 - Via férrea; 3 - Ônibus; 4 - Balsa; 5 - Teleférico; 6 - Gôndola, teleférico suspenso; 7 - Funicular.
route_url	Opcional	Contém o URL de uma página da Web sobre esse trajeto específico. Ele deve ser diferente de *agency_url*.
route_color	Opcional	Define uma cor que corresponda ao trajeto. A cor deve ser informada como um número hexadecimal de seis caracteres. Se nenhuma cor é especificada, a cor padrão de trajetos é branca (FFFFFF). A diferença de cores entre *route_color* e *route_text_color* deve fornecer contraste suficiente quando visualizado em uma tela em preto e branco.
route_text_color	Opcional	Usado para especificar uma cor legível para usar em desenho de texto contra um plano de fundo de *route_color*.

Fonte: Google Transit (adaptada)[1]

Tabela 12 – Detalhamento dos campos do arquivo *trips.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
route_id	Obrigatório	Contém um ID que identifica um trajeto. Este valor é indicado no arquivo *agency.txt*.
service_id	Obrigatório	Contém um ID que identifica um conjunto de datas em que o serviço está disponível para um ou mais trajetos. Este valor é indicado no arquivo *calendar.tx*t ou *calendar_dates.txt*.
trip_id	Obrigatório	Contém um ID que identifica uma viagem.
trip_headsign	Opcional	Contém o texto que aparece em uma sinalização que identifica o destino da viagem para os passageiros. Use este campo para distinguir diferentes padrões de serviço no mesmo trajeto. Se a placa muda durante uma viagem, você pode substituir o campo *trip_headsign*, especificando valores para o campo *stop_headsign* em *stop_times.txt*.
trip_short_name	Opcional	Contém o texto que aparace em programações e placas de sinalização para identificar a viagem para os passageiros, por exemplo, para identificar números de trens para viagens de trens suburbanos. Se os passageiros não recorrem normalmente aos nomes da viagem, deixe este campo em branco. Um valor de *trip_short_name*, se possível, deve identificar, com exclusividade, uma viagem em um dia de serviço; ele não deve ser usado para nomes de destino ou designações limitadas/expressas.

Continua na próxima página

Tabela 12 – continuação da página anterior

Nome do campo	Condicional	Descrição
direction_id	Opcional	Contém um valor binário que indica a direção de uma viagem. Use este campo para distinguir viagens bidirecionais com o mesmo *route_id*. Este campo não é usado na criação de trajetos; ele fornece uma maneira de separar viagens por direção durante a publicação de tabelas de horário. Você pode especificar nomes para cada direção com o campo *trip_headsign*. 0 - viagem em uma única direção (por exemplo, só ida); 1 - viagem na direção oposta (por exemplo, de volta),os campos *trip_headsign* e *direction_id* podem ser usados juntos para atribuir um nome a uma viagem em cada direção "1234".
block_id	Opcional	Identifica o quadro a que a viagem pertence. Um bloco consiste em duas ou mais viagens sequenciais feitas usando o mesmo veículo, em que um passageiro pode passar de uma viagem para a próxima permanecendo no veículo. O campo *block_id* deve ser indicado por duas ou mais viagens no *arquivo trips.txt*.
shape_id	Opcional	Contém um ID que define a forma da viagem. Este valor é indicado no arquivo *shapes.txt*. O arquivo *shapes.txt* permite definir como será traçada uma linha no mapa para representar uma viagem.

Continua na próxima página

Tabela 12 – continuação da página anterior

Nome do campo	Condicional	Descrição
wheelchair_accessible	Opcional	0 (ou vazio) - indica que não há informações sobre acessibilidade para a viagem; 1 - indica que o veículo que está sendo usado nesta viagem específica pode acomodar, pelo menos, um passageiro em cadeira de rodas; 2 - indica que não é possível acomodar passageiros em cadeiras de rodas nesta viagem

Fonte: Google Transit (adaptada)[1]

Tabela 13 – Detalhamento dos campos do arquivo *stop_times.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
trip_id	Obrigatório	Contém um ID que identifica uma viagem. Este valor é indicado no arquivo *trips.txt*.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
arrival_time	Obrigatório	Especifica o horário de chegada em uma parada específica de uma viagem específica de um trajeto. No caso de horários que ocorram após a meia-noite na data do serviço, digite o horário como um valor maior que 24:00:00 em horário local HH:MM:SS para o dia em que começa a programação da viagem. Se não há horários separados para chegada e partida em uma parada, insira o mesmo valor para *arrival_time* e *departure_time*. É necessário especificar os horários de chegada para a primeira e a última paradas de uma viagem. Se essa parada não for programada, use uma sequência vazia para os campos *arrival_time* e *departure_time*. As paradas sem horário de chegada são programadas conforme a parada programada anterior mais próxima. Para garantir trajetos precisos, forneça horários de chegada e de partida para todas as paradas programadas. Não intercale as parada, ou, preencha os horários com espaços. Observação: as viagens que abrangem várias datas terão horários de parada maiores que 24:00:00. Por exemplo, se uma viagem começa às 10:30:00 p.m e termina às 2:15:00 a.m. do dia seguinte, os horários de parada seriam 22:30:00 e 26:15:00. A inclusão desses horários de parada como 22:30:00 e 02:15:00 não produzem os resultados desejados.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
departure_time	Obrigatório	Especifica o horário de partida de uma parada específica para uma viagem específica de um trajeto. O horário é medido de "meio-dia menos 12h"(efetivamente meia-noite, exceto para dias do horário de verão), no início da data do serviço. No caso de horários que ocorram após a meia-noite na data do serviço, digite o horário como um valor maior que 24:00:00 em horário local HH:MM:SS para o dia em que começa a programação da viagem. Se não há horários diferentes para a chegada e a saída em uma parada, insira o mesmo valor para *arrival_time* e *departure_time*. É necessário especificar os horários de partida da primeira e da última paradas em uma viagem. Se essa parada não for programada, use uma sequência vazia para os campos *arrival_time* e *departure_time*. As paradas sem horário de chegada são programadas conforme a parada programada anterior mais próxima. Para garantir trajetos precisos, forneça horários de chegada e de partida para todas as paradas programadas. Não intercale as paradas. Os horários devem ter oito dígitos no formato HH:MM:SS (o formato H:MM:SS também é aceito, se a hora iniciar com 0). Não preencha os horários com espaços.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
stop_id	Obrigatório	Contém um ID que identifica uma parada. Diversos trajetos podem usar a mesma parada. O campo *stop_id* é indicado no arquivo *stops.txt*. Se *location_type* é usado no arquivo *stops.txt*, todas as paradas indicadas em *stop_times.txt* deverão ter *location_type* igual a 0. Onde possível, os valores de *stop_id* devem permanecer consistentes entre as atualizações de feed. Se uma parada não está programada, digite valores em branco para *arrival_time* e *departure_time*.
stop_sequence	Obrigatório	Identifica a ordem das paradas de uma viagem específica. Os valores de *stop_sequence* devem ser números inteiros positivos e devem aumentar ao longo da viagem.
stop_headsign	Opcional	Contém o texto que aparece em uma sinalização que identifica o destino da viagem para os passageiros. Use este campo para substituir o *trip_headsign* padrão quando as placas mudarem durante as viagens. Se esta placa está associada a uma viagem inteira, use *trip_headsign* no lugar.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
pickup_type	Opcional	Indica se os passageiros são embarcados em uma parada como parte da programação normal ou se não há embarque disponível na parada. Este campo também permite que a agência de transporte público indique se os passageiros devem ligar para a agência ou notificar o motorista para agendar um embarque em uma parada específica. Os valores válidos deste campo são: 0 - Embarque no horário normal; 1 - Sem embarque disponível; 2 - Deve ligar para a agência a fim de agendar o embarque; 3- Deve combinar com o motorista para agendar o embarque. O valor padrão deste campo é 0.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
drop_off_type	Opcional	Indica se há desembarque de passageiros em uma parada, como parte da programação normal ou se não há desembarques na parada. Este campo também permite que a agência de transporte público indique se os passageiros devem ligar para a agência ou notificar o motorista para agendar um desembarque em uma determinada parada. Os valores válidos deste campo são: 0 - Desembarque no horário normal; 1 - Desembarque não disponível; 2 - Deve telefonar para agendar o desembarque; 3 - Deve combinar com o motorista para agendar o desembarque. O valor padrão deste campo é 0.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condicional	Descrição
shape_dist_traveled	Opcional	Quando usado no arquivo *stop_times.txt*, o campo *shape_dist_traveled* posiciona uma parada como uma distância a partir do primeiro ponto de forma. O campo *shape_dist_traveled* representa uma distância real percorrida ao longo do trajeto em unidades como, por exemplo, pés ou quilômetros. Essas informações permitem que o planejador da viagem determine o quanto da forma deve ser desenhado ao exibir parte de uma viagem no mapa. Os valores usados para *shape_dist_traveled* devem aumentar juntamente com *stop_sequence*. As unidades usadas para *shape_dist_traveled* no arquivo *stop_times.txt* devem corresponder às unidades usadas para este campo no arquivo *shapes.txt*.

Fonte: Google Transit (adaptada)[1]

Tabela 14 – Detalhamento dos campos do arquivo *calendar.txt* da GTFS

Nome do campo	Condicional	Descrição
service_id	Obrigatório	Contém um ID que identifica um conjunto de datas em que o serviço está disponível para um ou mais trajetos. Cada valor de *service_id* pode aparecer, no máximo, uma vez em um arquivo *calendar.txt*. Este valor é um conjunto de dados exclusivo. Ele é indicado pelo arquivo *trips.txt*.
monday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as segundas-feiras. O valor 1 indica que o serviço está disponível todas as segundas-feiras durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível às segundas-feiras no período. Observação: você pode listar exceções para datas específicas, como, por exemplo, feriados, no arquivo *calendar_dates.txt*.
tuesday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as terças-feiras. O valor 1 indica que o serviço está disponível todas as terças-feiras durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível às terças-feiras no período.

Continua na próxima página

Tabela 14 – continuação da página anterior

Nome do campo	Condicional	Descrição
wednesday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as quartas-feiras. O valor 1 indica que o serviço está disponível todas as quartas-feiras durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível às quartas-feiras no período.
thursday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as quintas-feiras. O valor 1 indica que o serviço está disponível todas as quintas-feiras durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível às quintas-feiras no período.
friday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as sextas-feiras. O valor 1 indica que o serviço está disponível todas as sextas-feiras durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível às sextas-feiras no período.

Continua na próxima página

Tabela 14 – continuação da página anterior

Nome do campo	Condicional	Descrição
saturday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas os sábados. O valor 1 indica que o serviço está disponível todos os sábados durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível aos sábados no período.
sunday	Obrigatório	Contém um valor binário que indica se o serviço é válido para todos os domingos. O valor 1 indica que o serviço está disponível todos os domingos durante o período. O período é especificado utilizando-se os campos *start_date* e *end_date*. O valor 0 indica que o serviço não está disponível aos sábados no período.
start_date	Obrigatório	O *campo start_date* contém a data de início do serviço. O valor do campo *start_date* deve estar no formato YYYYMMDD.
end_date	Obrigatório	O campo *end_date* contém a data final do serviço. Essa data está incluída no intervalo do serviço. O valor do campo end_date deve estar no formato AAAAMMDD.

Fonte: Google Transit (adaptada)[1]

Tabela 15 – Detalhamento dos campos do arquivo *calendar_dates.txt* da GTFS

service_id	Obrigatório	Contém um ID que identifica identifica um conjunto de datas em que uma exceção ao serviço está disponível para um ou mais trajetos. Cada par (*service_id, date*) pode aparecer somente uma vez em *calendar_dates.txt*. Se um valor de *service_id* aparace nos arquivos *calendar.txt* e *calendar_dates.txt*, as informações contidas em *calendar_dates.txt* modifica as informações de serviço especificadas em *calendar.txt*. Este campo é indicado pelo arquivo *trips.txt*.
date	Obrigatório	Especifica uma data específica em que a disponibilidade do serviço é diferente do normal. Você pode usar o campo *exception_type* para indicar se o serviço está disponível na data especificada. O valor do campo date deve estar no formato AAAAMMDD.
exception_type	Obrigatório	Indica se o serviço está disponível na data especificada no arquivo date. O valor 1 indica que o serviço foi adicionado para a data especificada. O valor 2 indica que o serviço foi removido para a data especificada.

Fonte: Google Transit (adaptada)[1]

Tabela 16 – Detalhamento dos campos do arquivo *fare_attributes.txt* da GTFS

fare_id	Obrigatório	Contém um ID que identifica uma classe de tarifas.
price	Obrigatório	Contém o preço da tarifa, na unidade especificada por *currency_type*.
currency_type	Obrigatório	Define a moeda usada para pagar a tarifa. Use os códigos de moeda em ordem alfabética ISO 4217.
payment_method	Obrigatório	Indica quando a tarifa deve ser paga. Os valores válidos deste campo são: 0 - A tarifa é paga a bordo; 1 - A tarifa deve ser paga antes do embarque.
transfers	Obrigatório	O campo *transfers* especifica o número de baldeações permitidas nesta tarifa. Os valores válidos deste campo são: 0 - Não são permitidas baldeações nesta tarifa; 1 - Os passageiros só podem fazer uma baldeação; 2 - Os passageiros podem fazer duas baldeações; (*empty*) - Se o campo estiver vazio, não há limites para o número de baldeações.
transfer_duration	Opcional	Especifica a duração, em segundos, antes da expiração da baldeação. Quando usado com um valor 0 para *transfers*, o campo *transfer_duration* indica por quanto tempo uma passagem é válida para uma tarifa quando as baldeações não são permitidas. A menos que você pretenda usar este campo para indicar a validade da passagem, *transfer_duration* deve ser omitido ou deve ficar em branco, quando *transfers* é definido como 0.

Fonte: Google Transit (adaptada)[1]

Tabela 17 – Detalhamento dos campos do arquivo *fare_rules.txt* da GTFS

fare_id	Obrigatório	Contém um ID que identifica uma classe de tarifas. Este valor é indicado no arquivo *fare_attributes.txt*.
route_id	Opcional	Associa o ID da tarifa a um trajeto. Os IDs de trajetos são indicados no arquivo *routes.txt*. Se você tem diversos trajetos com os mesmos atributos de tarifa, crie uma linha no arquivo *fare_rules.txt* para cada trajeto.
origin_id	Opcional	Associa o ID da tarifa a um ID de zona de origens. Os IDs de zona são indicados no arquivo *stops.txt*. Se há vários IDs de origem com os mesmos atributos, crie uma linha no arquivo *fare_rules.txt* para cada ID de origem.
destination_id	Opcional	Associa o ID da tarifa a um ID de zona de destino. IDs de zona são indicados no arquivo *stops.txt*. Se há vários IDs de destino com os mesmos atributos de tarifa, cria-se uma linha no arquivo *fare_rules.txt* para cada ID de destino.
contains_id	Opcional	Associa o ID da tarifa a um ID de zona ID, indicado no arquivo *stops.txt*. O ID da tarifa é, então, associado a itinerários que transmitem cada zona de *contains_id*.

Fonte: Google Transit (adaptada)[1]

Tabela 18 – Detalhamento dos campos do arquivo *shapes.txt* da GTFS

shape_id	Obrigatório	Contém um ID que identifica uma forma.
shape_pt_lat	Obrigatório	Associa a latitude de um ponto de forma ao ID de uma forma. O valor do campo deve ser uma latitude WGS 84 válida. Cada linha do arquivo *shapes.txt* representa um ponto de forma em sua definição de formas.
shape_pt_lon	Obrigatório	Associa a longitude de um ponto de forma ao ID de uma forma. O valor do campo deve ser uma longitude WGS 84 de valor de -180 a 180. Cada linha do arquivo *shapes.txt* representa um ponto de forma em sua definição de formas.
shape_pt_sequence	Obrigatório	Associa a latitude e a longitude de uma forma de um ponto de formas com sua ordem sequencial juntamente com a forma. Os valores de *shape_pt_sequence* devem ser números inteiros positivos e devem aumentar com a viagem.
shape_dist_traveled	Opcional	Quando usado no arquivo *shapes.txt,* o campo *shape_dist_traveled* posiciona um ponto de forma como uma distância percorrida juntamente com uma forma a partir do primeiro ponto de forma. O campo *shape_dist_traveled* representa uma distância real percorrida ao longo do trajeto em unidades como, por exemplo, pés ou quilômetros. Esta informação permite que o planejador de viagens determine o quanto da forma deve ser desenhado ao mostrar parte de uma viagem no mapa. Os valores usados para *shape_dist_traveled* devem aumentar juntamente com *shape_pt_sequence*. As unidades usadas para *shape_dist_traveled* no arquivo *shapes.txt* devem corresponder às unidades usadas para este campo no arquivo *stop_times.txt*.

Fonte: Google Transit (adaptada)[1]

Tabela 19 – Detalhamento dos campos do arquivo *frequencies.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
trip_id	Obrigatório	Contém um ID que identifica uma viagem à qual a frequência especificada de serviço se aplica. Os IDs de viagem são indicados no arquivo *trips.txt*.
start_time	Obrigatório	Especifica o horário em que o serviço começa com a freqüência especificada. Para horários após a meia-noite, insira-os como um valor maior que 24:00:00 no horário local HH:MM:SS para o dia em que a programação das viagens começa.
end_time	Obrigatório	Especifica o horário em que o serviço muda para uma frequência diferente (ou é interrompido), na primeira parada da viagem. Para horários após a meia-noite, insira-os como um valor maior que 24:00:00 no horário local HH:MM:SS para o dia em que a programação das viagens começa.

Tabela 19 – continuação da página anterior

Nome do campo	Condicional	Descrição
headway_secs	Obrigatório	Indica o horário entre as saídas da mesma parada (intervalo entre as viagens) deste tipo de viagem, durante o intervalo de tempo especificado por *start_time* e *end_time*. O valor do intervalo de tempo entre duas viagens deve ser inserido em segundos. Períodos em que intervalos entre as viagens são definidos (as linhas no arquivo *frequencies.txt*) não devem ser sobrepostos para a mesma viagem, uma vez que é difícil determinar o que deve ser inferido de dois intervalos de viagem sobrepostos. No entanto, um período de intervalo entre viagens pode começar exatamente no mesmo horário em que outro termina.

Tabela 19 – continuação da página anterior

Nome do campo	Condicional	Descrição
exact_times	Opcional	Determina se viagens baseadas em frequência devem ser programadas com exatidão com base nas informações especificadas dos intervalos entre as viagens. Os valores válidos deste campo são: 0 ou (vazio) - Viagens baseadas em frequência não são programadas com exatidão. Este é o comportamento padrão; 1 - Viagens baseadas em frequência são programadas com exatidão. Para uma linha no *frequencies.txt*, as viagens são programadas com início com *trip_start_time* =*start_time* + *x* * *headway_secs* para todos x em (0, 1, 2, ...), em que *trip_start_time* < *end_time*. O valor de *exact_times* deve ser o mesmo para todas as linhas de *frequencies.txt* com o mesmo *trip_id*. Se *exact_times* for igual a 1, e uma linha de *frequencies.txt* tiver um *start_time* igual a *end_time*, nenhuma viagem deverá ser programada. Quando *exact_times* é 1, deve-se escolher um valor *end_time* que seja maior que o último horário de início da viagem programada, mas menor que o último horário de início da viagem desejada + *headway_secs*.

Fonte: Google Transit (adaptada)[1]

Tabela 20 – Detalhamento dos campos do arquivo *transfer.txt* da GTFS

Nome do campo	Condicional	Descrição
from_stop_id	Obrigatório	Contém um ID que identifica uma parada ou uma estação onde começa uma conexão entre trajetos. Os IDs de paradas são indicados no arquivo *stops.txt*. Se a ID de parada se refere a uma estação que contém várias paradas, essa regra de baldeação se aplica a todas as paradas nesta estação.
to_stop_id	Obrigatório	Contém um ID que identifica uma parada ou uma estação onde termina uma conexão entre trajetos. Os IDs de paradas são indicados no arquivo *stops.txt*. Se a ID de parada se refere a uma estação que contém várias paradas, essa regra de baldeação se aplica a todas as paradas nesta estação.
transfer_type	Obrigatório	Especifica o tipo de conexão para o par (*from_stop_id*, *to_stop_id*) especificado. Os valores válidos deste campo são: 0 ou (vazio) - Este é um ponto de baldeação recomendado entre dois trajetos; 1 - Este é um ponto de baldeação programado entre dois trajetos; 2 - Essa baldeação exige um tempo mínimo entre a chegada e a partida para garantir uma conexão. O tempo necessário para a baldeação é especificado por *min_transfer_time*; 3 - Não é possível fazer baldeações entre trajetos neste local.

Continua na próxima página

Tabela 20 – continuação da página anterior

Nome do campo	Condicional	Descrição
min_transfer_time	Opcional	Quando uma conexão entre trajetos exige um tempo entre a chegada e a partida (*transfer_type=2*), o campo *min_transfer_time* define o período de tempo que deve estar disponível em um itinerário para permitir uma baldeação entre trajetos nestas paradas. O *min_transfer_time* deve ser suficiente para que um passageiro típico se desloque entre as duas paradas, incluindo um tempo extra para variação na programação em cada trajeto. O valor de *min_transfer_time* deve ser inserido em segundos e deve ser um número inteiro positivo.

Fonte: Google Transit (adaptada)[1]

Tabela 21 – Detalhamento dos campos do arquivo *feed_info.txt* da GTFS

Nome do campo	**Condicional**	**Descrição**
feed_publisher_name	Obrigatório	Contém o nome completo da organização que publica o *feed*. Pode ser o mesmo que aquele definido pelos valores de *agency_name* no arquivo *agency.txt*. Aplicativos que utilizam GTFS podem exibir este nome ao concederem atribuições relacionadas aos dados de um *feed* específico.
feed_publisher_url	Obrigatório	Contém o URL do *website* da organização que está publicando o *feed*. Pode ser o mesmo que um dos valores de *agency_url* no arquivo *agency.txt*.
feed_lang	Obrigatório	Contém um código de idiomas IETF BCP 47 que especifica o idioma padrão usado para o texto neste *feed*. Esta configuração ajuda os consumidores de GTFS a escolherem regras para o uso de letras maiúsculas e minúsculas e outras configurações específicas do idioma para o *feed*.

Continua na próxima página

Tabela 21 – continuação da página anterior

Nome do campo	Condicional	Descrição
feed_start_date / *feed_end_date*	Opcional	O *feed* fornece informações completas e confiáveis sobre a programação de um serviço, no período entre o início do dia *feed_start_date* e o final do dia *feed_end_date*. As datas nos dois dias estão no formato AAAAMMDD, assim como no arquivo *calendar.txt*, ou são deixadas em branco se não estiverem disponíveis. A data *feed_end_date* não deve preceder a data *feed_start_date*, se ambas forem fornecidas. Os provedores de feeds são encorajados a oferecerem dados de programação fora desse período a fim de informarem sobre possíveis serviços no futuro, mas os consumidores de *feed* devem estar conscientes de seu status não autorizado. Se *feed_start_date* ou *feed_end_date* se estendem além das datas do calendário ativo definidas nos arquivos *calendar.txt* e *calendar_dates.txt*, o *feed* se torna uma afirmação explícita de que não há serviços para as datas entre *feed_start_date* ou *feed_end_date* que não estão incluídas nas datas do calendário ativo.

Continua na próxima página

Tabela 21 – continuação da página anterior

Nome do campo	Condicional	Descrição
feed_version	Opcional	O editor de *feeds* pode especificar uma sequência que indique a versão atual do *feed* GTFS. Os aplicativos que utilizam GTFS podem exibir este valor para ajudar os editores de *feed* a determinar se foi incorporada a versão mais recente do *feed*.

Fonte: Google Transit (adaptada)[1]