Mestrado em Informática e Sistemas

Inferência das Atividades na Modelização de Escolhas de Destinos e seu impacto na mobilidade urbana

Relatório de projeto apresentado para a obtenção do grau de Mestre em Informática e Sistemas
Especialização em Desenvolvimento de Software

Autor
Rui Fernandes Ganhoto

Orientador
Doutora Ana Cristina da Costa Oliveira Alves
Professora do Departamento de Informática e Sistemas
Instituto Superior de Engenharia de Coimbra

Coimbra, Março de 2017

Abstract

The inference of human activities and the analysis of urban services distribution throughout a specific area are intrinsically related. This kind of analysis requires not only information from more than one source, as a single source usually is not able to provide all the required data, but also information that is constantly being updated.

This document proposes a method of integrating information from different Internet sources. This integration is performed by a learning process based on a classification algorithm that uses already interlinked data as the training dataset.

The proposal represented on this document used information from the Factual and the DBpedia Internet data sources in conjunction with the CityClusters platform to visualize the urban distribution of services created by the developed system. It also proposes a method to create new relations between the information provided by these sources by analyzing the preprocessing data and learning already existing relations.

Keywords: Ubiquitous computing, Activity Recognition, Trip Purpose, Destination Choice Modeling, semantic enrichment of a place

Resumo

Inferir as atividades realizadas pelos indivíduos está intimamente relacionado com a análise da distribuição de serviços urbanos numa determinada área. Esta análise necessita de dados constantemente atualizados e com mais contexto do que uma fonte única consegue fornecer. A integração de dados a partir de várias fontes disponíveis na Internet é proposta como um processo de aprendizagem de classificação para interligar a informação. A proposta efetuada neste documento utiliza dados do Factual e do DBpedia para efetuar análises de distribuição de serviços numa área urbana, com apoio da plataforma CityClusters. Assim como propõe um método para criar novas relações entre informação das duas fontes utilizando dados existentes já relacionados.

Palavras-chave: Computação Ubíqua, *Activity Recognition*, *Trip Purpose*, *Destination Choice Modeling*, Enriquecimento Semântico de Lugar

Índice

Lista de Figuras

Lista de Tabelas

Definições e Acrónimos

DUMP – Um *dump* é uma estrutura de dados referentes a uma base de dados. Estes são muito utilizados para cópias de segurança ou partilhas de dados de projetos com conteúdo livre.

GIS – *Geographic Information System* é um sistema informação geográfica que engloba o hardware, software, informação espacial, procedimentos computacionais e recursos humanos que permite e facilita a análise, gestão ou representação do espaço e dos fenómenos que nele ocorrem.

GPS – *Global Positioning System* é um sistema de posicionamento por satélite. Este sistema permite que qualquer pessoa na detenção de um dispositivo com esta capacidade consiga saber a sua localização em formato de coordenadas de latitude, longitude e altitude.

NAICS – *North American Industry Classification System* é um padrão utilizado pelos órgãos federais Norte-Americanos de estatística na classificação de estabelecimentos comerciais.

POI – *Point Of Interest* é um ponto numa localização que alguém considera útil ou interessante. Termo usado em cartografia com aplicações em sistemas de navegação, GIS e GPS.

SIC – *Standard Industrial Classification*, tal como o NAICS, é um sistema usado para classificar indústrias com códigos de quatro dígitos, usado para classificar áreas da indústria.

1 Introdução

Hoje em dia, são cada vez mais necessários sistemas de apoio à decisão relativamente a planeamento urbano, seja para melhorar os serviços de mobilidade disponíveis, detetar padrões de utilização e procura de serviços a fim de sugerir futuras localizações de novos recursos e instalações face às necessidades.

Poucos destes sistemas utilizam dados *online* para acompanhar a evolução e dinamismo da ocupação do espaço urbano. Por outro lado, existem imensos dados distribuídos pela Internet, em muitos casos, fontes de dados incompletas ou desatualizadas sem qualquer tipo de relação entre elas, o que dificulta uma correta avaliação da qualidade destes dados. Por exemplo, o Website das Páginas Amarelas[1] não está interligado muitas vezes com a página oficial das entidades anunciadas. Neste cenário em constante mudança, é imperativo trabalhar sobre dados corretamente atualizados, algo que muitas vezes não está disponível numa única fonte de dados.

Por forma a ultrapassar as dificuldades existentes quando se pretende efetuar uma análise a nível de planeamento urbano, seria necessário existir uma ferramenta capaz de interligar várias fontes de dados de forma automática, que possa ser aplicada em diferentes cidades com uma intervenção bastante reduzida por parte do investigador/urbanista facilitando assim o trabalho futuro com um conjunto de dados completo e atualizado.

A proposta a desenvolver envolve um sistema que consegue recolher de forma massiva dados sobre os serviços oferecidos numa dada cidade de forma a descobrir as prováveis atividades que os utilizadores procuram em certos destinos. Para isso é necessário visualizar os dados recolhidos, analisá-los e consequentemente interligar os dados obtidos de diferentes fontes. Neste trabalho além de apresentar o sistema que faz esta recolha e análise, realizaram-se casos de estudos, onde foi possível interligar duas fontes colaborativas de dados, efetuando a recolha de dados necessária e permitindo a visualização dos mesmos para identificar áreas com grandes concentrações de serviços.

1.1 Motivação

O conceito de um local tem evoluído ao longo da história, independentemente da cultura e desenvolvimento de um espaço urbano uma vez que este se encontra em constante mudança. Estas mudanças não são refletidas imediatamente num vasto conjunto de fontes, enquanto que outras conseguem ter informação quase em tempo real. Neste trabalho,

[1] http://www.pai.pt/ (acedido em 2016/12/08)

iremos denominar um local como Ponto de Interesse (POI de *Point of Interest*) que determina a localização de um serviço ou ponto turístico e que pode ter um determinado significado para um utilizador.

Já o conceito de mobilidade sustentável envolve a consideração de dimensões que não estão estritamente limitadas ao domínio do consumo de energia dos transportes e seus impactos sobre o meio ambiente: envolve perceber o modo e com que frequência os indíviduos se deslocam para satisfazer as suas diversas necessidades do imprescindível (escola/trabalho) ao ocasional/opcional (lazer, social, turismo, etc.).

Além dos dados disponíveis sobre serviços na cidade, existe atualmente um grande foco no desenvolvimento de métodos de recolha colaborativa para determinar os padrões de mobilidade urbana no uso de transportes públicos. Entender estes padrões permite aos operadores de transportes planear a oferta voltada para suprir as necessidades e eventuais lacunas que não conseguem dar respostas aos utentes destes transportes.

Dois estudos estão em estreita colaboração com o trabalho aqui apresentado. O primeiro, o projeto Future Mobility Survey (Cottrill, et al., 2013), é sistema de inquéritos de viagem disponível em dispositivos móveis que recolhe dados de forma voluntária e ubíqua sobre a utilização de transportes públicos. Este sistema recolhe dados em 4 fases: no registo do utilizador onde o seu perfil é indicado, pré-questionário onde são introduzidos dados sócio-económicos do agregado familiar, diário de atividades preenchido durante o período em estudo e modos de transporte detectados automaticamento pelo aplicativo, e finalmente um inquérito final onde o utilizador responde a questões sobre a utilização do sistema.

O segundo, o projeto URBY.SENSE (Referência P2020-PTDC/ECM-TRA/6803/2014[2]), que teve início em junho de 2016, pretende extrair padrões de mobilidade fora da rotina (de lazer, sociais, etc.) a partir de múltiplas fontes de dados através da recolha, fusão e análise destes dados. Os padrões que se pretendem estabelecer são diversos, tais como locais de interesse, modos de transporte, rotas comuns e atividades baseadas em localização.

Como estes estudos são feitos em uma amostra da população, importa perceber: quais as características dos destinos procurados, o que atrai as pessoas; e que atividades são aí desenvolvidas. Além das variáveis mais óbvias para inferir o padrão de mobilidade tais como período do dia, rotas já existentes, zona de residência de origem, outras variáveis mais semânticas e sociais tais como a popularidade de certos locais, as categorias de serviços oferecidos por determinados destinos poderão criar um nova dimensão na criação dos modelos de mobilidade.

De forma a estudar intensivamente um espaço urbano e acompanhar a evolução dos locais de interesse aí presentes, é importante existir uma ferramenta, com a capacidade de agrupar as várias fontes e fornecer um conjunto de dados associados aos POIs existentes numa

[2] https://www.cisuc.uc.pt/projects/show/217 (acedido em 2016/12/11)

região para que os dados sejam o mais atuais, corretos e completos no momento em que é necessária a sua utilização.

1.2 Objetivos

É proposto desenvolver uma ferramenta realiza um processamento não supervisionado capaz de analisar cidades, regiões ou países por forma a adicionar contexto e informação a pontos de interesse em áreas urbanas. Este contexto é inferido através da ligação entre duas fontes de dados que disponibilizam os seus dados de forma gratuita.

O principal objetivo do trabalho é identificar possíveis atividades realizadas pelo utilizador de transportes públicos nos destinos identificados previamente, nomeadamente através:

- Identificar a localização de serviços oferecidos pela cidade;
- Identificar áreas com predominância em determinados grupos de serviços;
- Reclassificar serviços através de uma categoria agregadora.

Dada uma cidade, região ou país a estudar o sistema a desenvolver deve ser capaz de identificar a localização de áreas onde predominem determinados serviços oferecidos na área em estudo, ao mesmo tempo que propõe um conjunto de interceções entre fontes de dados para enriquecer e aprender a interligar novos dados no futuro.

Todo este processo deve ser de fácil utilização e acesso, assim como deverá ser o mais completo possível em termos de cobertura geográfica. A extração de dados também deverá ser efetuada dentro de um tempo limite aceitável para futura utilização por investigadores, decisores e urbanistas que pretendam estudar a localização de novos serviços e oferta de transportes.

1.3 Estrutura do Documento

Este documento está organizado do seguinte modo: no capítulo 2 são apresentadas várias fontes de dados, assim como formas disponíveis para recolha dos mesmos; de seguida é feita uma seleção de fontes de dados para o estudo, onde é detalhado o desenvolvimento de uma ferramenta com a capacidade de obter os dados dessas mesmas fontes, são apresentadas e selecionadas áreas específicas para as quais são obtidos dados para efetuar um conjunto de análises.

No capítulo 3 é feito um estudo de análise espacial e organização urbana utilizando algoritmos de *clustering*, assim como é aplicado a um conjunto de casos de estudo para visualizar os dados gerados na plataforma CityClusters. Neste capítulo, também é proposta

a classificação de dados entre as várias fontes utilizando métodos de comparação de *strings* e classificadores binários do Weka.

No capítulo 4 são apresentadas as conclusões globais do projeto de investigação e do estudo efetuado nos vários pontos abordados.

2 Recolha de Dados

Este capítulo descreve a análise de fontes de dados e implementação da ferramenta desenvolvida para recolha dos dados estando organizado do seguinte modo: na secção 2.1 são descritos que dados podem ser considerados e como podem ser estruturados e classificados num estudo exploratório de extração de dados. Na secção 2.2 são descritas algumas fontes atualmente disponíveis e, comparativamente, quais as suas características principais. De forma a conceber uma visão geral do problema em si, alguns trabalhos de investigação são apresentados na secção 2.3. Um conjunto de requisitos e funcionalidades serão enumerados na secção seguinte, 2.4, assim como a arquitetura proposta como abordagem ao problema de recolha dos dados para o sistema proposto é realizada e descrita em detalhe. Na secção 2.5 são discutidas as principais questões quanto à aplicabilidade da solução implementada a diferentes áreas geográficas de estudo.

2.1 Introdução

A quantidade de informação georreferenciada cresce a um ritmo impressionante, no entanto essa informação está espalhada pela Internet em diferentes páginas web, sistemas e serviços, por este motivo, é difícil obter uma base de dados com a informação integrada, atual, completa e correta.

Tal como descrito por (Rodrigues F. , POI Mining and Generation, 2010), um ponto de interesse (POI), é uma localização no mapa que alguém pode considerar útil ou interessante. Estes podem ser usados para navegação, caracterização de um local, estudos sociológicos ou análise da dinâmica de uma cidade.

Existem centenas, senão milhares de fontes de dados com POIs na Internet, em que cada uma usa o seu formato para representar os POIs assim como uma taxonomia específica para os classificar. Ao invés de considerar todos os POIs presentes nestas fontes, mesmo os que façam sentido apenas para um utilizador (informação pessoal), neste trabalho o foco serão POIs que tenham uma utilidade pública e sejam reconhecidos como uma entidade que tenha uma atividade principal.

O trabalho nesta fase consiste na análise de várias fontes de POIs, para de seguida desenvolver uma metodologia para extração, normalização e tratamento dos dados, importando-os para um sistema de armazenamento local de forma a simplificar o acesso futuro aos mesmos.

2.1.1 Geração dos dados

Existem várias formas de gerar os dados para as fontes existentes, estas podem ser alimentadas pelos próprios proprietários de empresas e serviços que necessitem de publicitar o seu negócio (e.g. Páginas Amarelas[3], diretoria de empresas Manta[4]), autoridades de transportes entre outras (e.g. Boston[5], Singapura[6], Lisboa[7]) ou construídas de forma colaborativa com base em redes sociais (e.g. Facebook Places[8], Foursquare[9]), na secção 2.2 cada uma das fontes estudadas serão descritas de forma sistemática quanto a um conjunto de características tais como os apresentados na Tabela 1.

Tabela 1 – Exemplos de caraterísticas de fontes com diferentes formas de manutenção de dados

	Aberto/Colaborativo Factual[10]	***Privada/Mantido por uma entidade Manta[11]***
Número de POIs	Superior a 100.000.000	Superior a 35 000 000
Frequência de Atualização	Mais de 4 milhões de novos registos e registos removidos no 3º Trimestre de 2016	Não existe nenhum valor publicado, no entanto são publicitados, em média, 1000 novos negócios inseridos por dia
Cobertura Geográfica	50 Países	4 Países anglo-saxónicos
Precisão	Na sua maioria contém coordenadas de GPS devido ao crescimento de dipositivos com esta capacidade	Morada completa disponível, ficando com nome de rua e número de porta quando existe.
Tipos de fontes	Negócios, museus, parques.	Apenas relacionado com negócios

Verificou-se que a maior parte das fontes utilizam um conceito misto que retira um pouco o melhor dos dois mundos, que são fontes geradas de forma colaborativa com validação por um conjunto de responsáveis. Assim os dados são disponibilizados mais rapidamente que a recolha privada, com uma fiabilidade superior aos registos de fontes puramente sociais.

[3] http://www.pai.pt/ (acedido em 2016/12/08)
[4] http://www.manta.com/ (acedido em 2016/12/08)
[5] http://www.mbta.com/rider_tools/developers/ (acedido em 2016/12/08)
[6] https://data.gov.sg/ (acedido em 2016/12/08)
[7] http://dados.cm-lisboa.pt/dataset (acedido em 2016/12/08)
[8] https://www.facebook.com/places/ (acedido em 2016/12/08)
[9] https://foursquare.com/ (acedido em 2016/12/08)
[10] http://factual.com/ (acedido em 2016/12/08)
[11] http://www.manta.com/world (acedido em 2016/12/08)

2.1.2 Taxonomia das Fontes

Normalmente em cada fonte de dados, existem categorias para classificar os POIs nelas presentes, no anexos 6.1 e 6.2 estão dois exemplos de taxonomias de duas fontes estudadas neste trabalho. Estas categorias podem ser detalhadas e nomeadas de forma distinta para cada fonte. Com o intuito de uniformizar e criar um padrão internacional, a União Europeia e a América do Norte estabeleceram sistemas internacionais de categorização, respetivamente SIC (Standard Industrial Classification) e NAICS (North American Industry Classification System).

Tendencialmente a taxonomia de POIs está estruturada de forma hierárquica, tendo na sua base categorias mais genéricas, como por exemplo: 35 - Industrial And Commercial Machinery And Computer Equipment[12]; 54 - Professional, Scientific, and Technical Services[13]; e de onde descendem categorias mais específicas, tais como: 3571 - Electronic Computers; 541511 - Custom Computer Programming Services; o que facilita a tarefa de agrupar entidades com características distintas, mas pertencentes a um único grupo comum.

2.2 Estudo comparativo de fontes

Esta secção mostra os diferentes tipos de fontes de dados, assim como um conjunto de fontes de dados estudadas com várias características das mesmas.

2.2.1 Fontes de dados geográficos

Fontes ricas em georreferenciação, podem conter alguma informação contextual mas são tendencialmente fontes com pouca informação da atividade efetiva que se pratica na área.

Estas fontes são maioritariamente utilizadas em sistemas de navegação, em que o contexto deve ser simples e conciso.

2.2.2 Fontes de dados contextuais

Estas fontes adicionam contexto a um local, evento, pessoa ou objeto, estes também podem estar presentes em fontes de dados geográficos.

[12]https://www.osha.gov/pls/imis/sic_manual.html (acedido em 2016/12/08)
[13] http://www.census.gov/eos/www/naics/ (acedido em 2016/12/08)

Utilizadas maioritariamente por sistemas de turísticos e de lazer, como o Booking[14] e Traveleye, ou sistemas de *reviews* como Yelp e Tripadvisor e outros sistemas em que o contexto seja de grande interesse.

2.2.3 Fontes de dados de mobilidade

Fontes de dados obtidas através da monitorização voluntária de pessoas, seja por aplicações móveis, análise da navegação da web ou inquéritos, normalmente relacionados com um utilizador ou um grupo de utilizadores registados no decorrer de uma atividade, ação ou do dia.

Estas fontes de dados permitem registar atividades, rotinas, percursos entre outas informações. São dados muito explorados na área do marketing e publicidade mas podem ser usados noutros contextos, como detetar eventos ou interesses de uma cidade.

Estas fontes são mais restritas e mais dirigidas a um objetivo, por este motivo, existem vários projetos que criam os seus próprios dados, sem necessitar de fontes de dados exteriores, ou integrando a outras fontes de dados contextuais e geográficas, por forma a efetuar os seus estudos.

Embora existam vários tipos de dados, é muito difícil qualificar uma fonte como sendo apenas geográfica ou contextual, pois existe, na maioria das fontes, contexto e localização associados a um registo.

2.2.4 Fontes analisadas

Existem centenas de fontes de POIs distribuídas por todo o mundo, foram analisadas algumas destas fontes para verificar a sua facilidade de acesso e disponibilidade dos dados por forma a selecionar uma das fontes para o caso de estudo.

O Factual Places[15] é uma fonte de dados, embora também bastante rica contextualmente, contém uma grande quantidade de dados geográficos, com isto podemos concluir que se enquadra em duas categorias de fontes de dados (ver seções 2.2.1 e 2.2.2).

[14] http://www.booking.com (acedido em 2016/12/08)
[15] http://www.factual.com/data/t/places/schema (acedido em 2016/12/08)

Contento um número considerável de POIs em cada país do mundo, está bastante orientada para uso empresarial, através de uma API acessível por um vasto leque de dispositivos e contém um conjunto de funcionalidades bastante práticas para acesso direto de acordo com a necessidade. Esta API fornece os atributos apresentados na Tabela 2, embora a maior parte dos registos não esteja completa quanto a estes dados.

Para registar novas entidades é necessário efetuar um pedido, que será validado pela equipa ou por pessoas responsáveis distribuídas em cada ponto do planeta.

Esta é uma fonte de dados que não tem nenhuma aplicação integrada pela equipa que a mantém, servindo assim, apenas para uso pelo público em geral, ou por entidades que pretendam utilizar as APIs fornecidas.

Tabela 2 - Atributos disponíveis no *Factual Places*

Atributo	***Descrição***
factual_id	GUID (*globally unique identifier*) identificando cada POI do Factual
name	Nome do *place*
address	Morada simples
address_extended	Morada completa
locality	Localidade
region	Região
postcode	Código postal
country	País
tel	Número de telefone
fax	Número do fax
website	Página oficial
latitude	Latitude no formato WGS84
longitude	Longitude no formato WGS84
hours_display	Estrutura em JSON que indica as horas de abertura
hours	Estrutura em JSON que indica as horas de operação
post_town	Cidade ou Vila onde se encontra o POI
chain_name	Nome da cadeia do POI (caso tenha)
chain_id	ID da cadeia do POI
category_labels	Nomes das categorias às quais o POI pertence
category_ids	Ids das categorias às quais o POI pertence
email	Email principal da organização

Facebook Places[16] é uma fonte de dados muito simplista que oferece pouca informação sobre um local. Tal como a restante informação disponível no Facebook, os dados são criados pelos seus utilizadores, pelo que, podem existir locais sem interesse para o estudo, assim como locais fictícios. A partir do Facebook Places é possível interligar com muitas

[16] https://developers.facebook.com/docs/graph-api/reference/place/ (acedido em 2016/12/08)

outras fontes do Facebook, enriquecendo desta forma a fonte geográfica com contexto como fotos, descrição, horários de funcionamento entre outros.

Esta fonte disponibiliza uma API que contém apenas os campos apresentados pela Tabela 3, e a sua função principal é facilitar o processo de *check-in*[17] com uma API que fornece os POIs disponíveis nas redondezas de uma determinada localização.

Tabela 3 - Atributos presentes no Facebook Places

Atributo	***Descrição***
Id	Identificador único de um POI, este identificador é unívoco entre todas as funcionalidades do Facebook disponíveis para este POI
name	Nome do POI
locaion	Coordenadas GPS
overall_rating	Classificação de 1 a 5, dado pelos utilizadores, caso não tenha classificação, este valor será zero

O Foursquare[18] é uma rede social que tem como funcionalidade principal registar as visitas (*check-ins*) dos utilizadores a um POI, disponível através de uma aplicação móvel. Nesta aplicação, o utilizador poderá classificar o local, adicionar fotos ou comentários. Esta rede social disponibiliza muitos POIs no entanto também possui muitos registos mal classificados, inexistentes, incorretos ou repetidos. Estes erros devem-se à forma como os dados são gerados sem qualquer tipo de controlo ou validação. Associados a uma entidade do Foursquare, existem vários atributos disponíveis, representadas na Tabela 4.

Esta plataforma também permite o acesso aos dados disponibilizando uma API para o efeito, uma das principais funcionalidades é fornecer até 50 pontos nas redondezas de uma localização. Dificultando assim a extração massiva dos dados para futura análise.

Tabela 4 - Atributos presentes nas entidades do Foursquare

Atributo	***Descrição***
id	Identificador único
name	O nome pelo qual a entidade é conhecida
contact	Objeto que contém dados como telefone, conta do Twitter e outros contactos
location	Objeto que contém dados como rua, código postal, latitude e longitude
categories	Lista de categorias de uma entidade
verified	Identifica se o *owner* de um estabelecimento verificou os dados inseridos na plataforma
stats	Contem estatísticas como número de *check-ins*, número de utilizadores e número de dicas
url	Website da entidade

[17] Funcionalidade disponível em algumas fontes que permite o utilizador indicar onde está.
[18] https://developer.foursquare.com/docs/responses/venue (acedido em 2016/12/08)

hours	Horas de funcionamento
popular	Contém as horas mais populares quando as pessoas normalmente visitam a entidade
menu	Um endereço url com uma página a apresentar o menu da entidade
price	Objeto que contem a categoria de preços de 1 a 4, barato a caro respetivamente e uma mensagem a descrever esta classificação
rating	Rating numérico atribuído pelos utilizadores, de 0 a 10
specials	Um conjunto de ofertas e como as obter no estabelecimento
hereNow	Informação sobre os utilizadores que estão presentemente no local
description	Descrição fornecida pelo *owner* da entidade
createdAt	Data em que a entidade foi criada
mayor	Utilizador que é o frequentador mais assíduo (com mais *check-ins* no local nos últimos 60 dias) inclui também o número de *check-ins* no local nos últimos 60 dias
tips	Conjunto com dicas que os utilizadores da aplicação partilharam
listed	Um conjunto de listas de grupos onde esta entidade está presente
tags	Um conjunto de tags associadas a esta entidade
beenHere	Um contador com o número de vezes que o utilizador conectado visitou a entidade
shortUrl	Um url reduzido para aceder aos dados da entidade
canonicalUrl	O url completo para aceder aos dados da entidade
specialsNearby	Conjunto de ofertas disponíveis nas redondezas
photos	Lista de fotos associadas à entidade
likes	Número de utilizadores que fizeram *like* à entidade
like	Informação a indicar se o utilizador atual fez *like* à entidade
dislike	Informação a indicar se o utilizador atual fez *dislike* à entidade
phrases	Conjunto de frases vistas normalmente nas *tips* fornecidas pelos utilizadores
attributes	Informações adicionais associadas à entidade, tal como se aceita reservas, se tem estacionamento
roles	Informação associada ao utilizador, caso este seja funcionário na entidade
page	Página associada ao grupo da entidade, se for uma cadeia, apresenta uma página relativa à cadeia, senão apresenta a página da própria entidade

O Twitter Places foi lançado em 2010 em parceria com a empresa de sistemas de navegação TomTom[19] e Localeze[20] com o intuito de melhorar o sistema base do Twitter. Com a principal funcionalidade de inferir a localização de onde é emitido um *tweet*, também permite colocar uma *tag* com um local específico ou criar novos registos de localizações para serem utilizados da mesma forma. Outra funcionalidade disponível através do Twitter places é encontrar os *tweets* gerados e com *tags* registados num dado local.

Places são localizações com nome e as respetivas coordenadas geográficas. Conseguem estar associadas aos *tweets* a partir de um *place_id.* Este é associado quando se está a criar um novo *tweet* embora não seja obrigatoriamente emitido dessa localização. Para além dos atributos básicos, disponíveis na Tabela 5, os *twitter places* podem possuir atributos mais detalhados (Tabela 6), para melhor descrever um POI, embora não seja obrigatório o seu

[19] http://www.tomtom.com/pt_pt/ (acedido em 2016/12/08)
[20] https://www.neustarlocaleze.biz/ (acedido em 2016/12/08)

preenchimento dando origem a lacunas a nível de dados contextuais associados a cada *place*.

Tabela 5 - Atributos de um Twitter place

Propriedade	***Descrição***
country	Nome do país
country_code	Código do país
full_name	Nome completo do *Place*
id	ID único do *Place*
name	*Short-name* do *Place*
place_type	Tipo de *Place*
url	Este endereço web dá acesso direto aos atributos adicionais de um *Place*

Tabela 6 - Atributos detalhados disponíveis para associar a um Twitter place

Propriedade	***Descrição***
street_address	Nome da rua do POI
locality	A cidade onde está o POI
region	A região do país
iso3	O código Alpha-3[21] do país de acordo com a norma ISO 3166
postal_code	O código postal formatado de acordo com o país
phone	O número de telefone formatado de acordo com o país incluindo o indicativo do país
twitter	ID de uma conta Twitter associada ao local
url	URL oficial do POI
app:id	ID ou conjunto de Ids de aplicações registadas no Twitter associadas ao Local
bounding_box	Conjunto de coordenadas que geram um polígono para definir a área de uma entidade

A Wikipedia[22] é a enciclopédia online mais usada a nível mundial, um exemplo perfeito de uma fonte de dados contextual, que associa informação histórica e de senso comum a um determinado local, evento, personagem, música entre outros. Como esta fonte é mantida pela comunidade de seus utilizadores, por vezes os dados carecem de confirmação, podendo por vezes conter erros ou falsos argumentos. Como a Wikipedia disponibiliza informação em linguagem natural de forma pouco estruturada, torna-se muito difícil extrair dados assim como efetuar pesquisas sobre os mesmos. Ao contrário das enciclopédias comuns, a Wikipedia é constantemente atualizada, podendo levar apenas minutos para adicionar uma entrada ao invés de meses ou anos.

[21] http://www.iso.org/iso/home/standards/country_codes.htm (acedido em 2016/12/08)
[22] https://en.wikipedia.org/wiki/Wikipedia:About (acedido em 2016/12/08)

Dada a limitação estrutural existente na Wikipedia, foi criada a DBpedia (Lehmann, et al., 2012) em que o objetivo é manter um conjunto de dados mais estruturados para facilitar pesquisas específicas utilizando os dados da Wikipedia tendo como base especialmente os dados presentes na *infobox*.

Uma das razões pela qual a qualidade dos dados da DBpedia tem vindo a crescer, deve-se ao sistema ser mantido pela comunidade, que cria mapeamentos das páginas do Wikipedia para serem importados pela DBpedia, conseguindo cada vez mais extrair novos dados e com maior relevância.

O Flickr[23] é uma plataforma de partilha de fotografias, com a missão de disponibilizá-las aos utilizadores assim como fornecer novas formas de as organizar. As fotos enviadas pelos utilizadores podem ter meta-dados introduzidos pelos dispositivos que capturaram a imagem, ou mesmo pelo utilizador. Esses meta-dados normalmente contém o instante em que a fotografia foi tirada e, por vezes, as coordenadas de GPS. Normalmente, o local não está incluído nos meta-dados, podendo ser associado a outras fontes pela localização ou pela descrição dada pelo utilizador à fotografia em questão.

O Yelp[24] foi desenvolvido com o intuito de efetuar *reviews* sobre estabelecimentos. Os utilizadores registados na plataforma têm a possibilidade de criar *reviews* dando uma classificação a um negócio, os outros utilizadores poderão ver essa *review* e classificá-la quanto à sua utilidade. Com estes dados esta plataforma consegue partilhar eventos e por em contato os seus vários utilizadores. Existe também uma funcionalidade para indicar novos estabelecimentos e eventos perto da localização do utilizador. É disponibilizada uma API para efetuar pesquisas de empresas, através de uma palavra-chave, localização ou número de telefone. A resposta a pedidos à API é dada com os dados presentes na Tabela 7.

Tabela 7 – Atributos de um negócio da API do Yelp

Propriedade	***Descrição***
rating	Classificação do negócio, valor decimal de 1 a 5
price	Nível de preços do negócio, de $ a $$$$ conforme o preçário.
hours	Horas de funcionamento
photos	Endereços de até 3 fotografias do estabelecimento
image_url	Endereço da foto de miniatura do negócio
name	Nome
url	Endereço para a página do negócio no Yelp
review_count	Número de *reviews* registadas
coordinates	Localização GPS

[23] https://www.flickr.com/about (acedido em 2016/09/05)
[24] https://www.yelp.com/about (acedido em 2016/12/11)

id	Id único do negócio
categories	Lista de categorias
location	Morada completa do negócio

No âmbito do projeto Future Cities (Rodrigues, Aguiar, & Barros, 2014), foi desenvolvida uma aplicação que recolhe vários dados de um dispositivo móvel Android, tais como: movimento (acelerómetro e giroscópio), posição (proximidade, magnético), dados ambientais (luminosidade, pressão, temperatura e humidade), telefonia, localização (GPS e redes WIFI), comunicação e outros sensores externos.

Este projeto foi criado com o intuito de recolher informação para estudos em curso e futuros, também permite ao utilizador ver os seus dados de uma forma amigável através de uma página web, constituindo assim uma fonte de dados geográficos e de mobilidade.

Já o projeto One Stop Transport[25] (Braga, Santos, & Moreira, 2014) agrega dados fornecidos pelos provedores de transportes para disponibilizá-los de forma aberta e estruturada às várias equipas que se podem apoiar sobre este conceito para desenvolverem as suas aplicações com dados de transportes urbanos. Estes dados são fornecidos diretamente pelos prestadores dos serviços, o que garante que estão corretos e deverão estar devidamente atualizados, utilizando estes dados é possível fazer algumas análises relacionadas com redes de transportes em diferentes cidades do território nacional.

O projeto estabelece várias parcerias com empresas de transportes (Serviços Municipalizados de Transportes Urbanos de Coimbra, MoveAveiro - Empresa Municipal de Mobilidade, E.M., CP - Comboios de Portugal, Transdev), eventos (Município do Porto) e pontos de interesse (Município do Porto, Distritos, Municípios e Freguesias de Portugal) permitindo a quem acede os dados reunir informações complementares para uma dada cidade.

[25] www.ost.pt (acedido em 2016/09/05)

Tabela 8 - Quadro comparativo das fontes de dados analisadas

	Contribuição	***Disponibilidade***	***Cobertura Geográfica***	***Atualização***	***Dados Disponíveis***	***Facilidade de Utilização***
Factual	Colaborativa	NR POIS	50 Países	4 Milhões de registos no 3º trimestre de 2016	Coordenadas GPS, horário, nome e website de um POI. POIs de diversas categorias	API disponível gratuitamente*
Wikipedia	Colaborativa		Global	10 Milhões de edições de artigos a cada 50 dias	Dados em linguagem natural sobre os mais diversos temas	API e *Database Dump*
Facebook Places	Colaborativa		Global			API
Foursquare	Colaborativa		Global	9 Milhões de checkins por dia	Coordenadas, UserID, dados gerais do POI	API *
DBpedia	Colaborativa		Global		Coordenadas GPS, Categorias, Morada, Dados Gerais	*Database Dump*
Twitter Places	Colaborativa		Global		Coordenadas GPS, Nome, Website	API
Flickr	Colaborativa		Utilizadores em 63 países **	1 Milhão de fotos por dia	Fotografias com coordenadas GPS e informações adicionadas manualmente	API
Yelp	Colaborativa		Global	115 mihões de reviews no 3º trimestre de 2016	Coordenadas GPS, Nome e Categorias de negócios	API
Sense My City	Colaborativa	N/D	Porto, Portugal	N/D	N/D	N/D
One Stop Transport	Colaborativa		Portugal	N/D	Dados de Transportes urbanos, POI's	API
Manta	Privada		4 Países Anglo Saxónicos	N/D	Dados relativos a negócios locais	API paga disponível
Google Maps	Colaborativa		Global		Coordenadas GPS, Nome, Contactos, WebSite	API

* Com limitações de acesso
** Dados gerados em todo o mundo

2.3 Trabalhos relacionados

Existem vários estudos que efetuam recolha de dados de outras fontes, criando assim o seu conjunto de dados (*dataset)* para análise e testes necessários a fim de verificar e alcançar o objetivo de cada estudo. Para todos os trabalhos aqui apresentados, são indicadas as fontes utilizadas, e, em casos específicos, a forma como foi efetuada a recolha de dados.

Em (Canneyt, Schockaert, Laere, & Dhoedt, 2011) é desenvolvido um sistema de recomendações turísticas baseadas no momento em que um utilizador visita uma cidade com o principal intuito de melhorar o sistema atual que se baseia apenas na popularidade de um local. Para tal, foram descarregados meta-dados (id de utilizador, localização, data e hora em que foi tirada) de fotografias do Flickr[26], de Janeiro 2000 a Setembro 2010. Destes dados foram selecionados apenas os que tinham coordenadas que se encontrassem na mesma rua de um dos *Points-of-Interest* (POIs). Estes POIs foram selecionados a partir do Traveleye[27] como *Top 100 Best Travel Destinations* (em Setembro 2010). São removidas as fotos que um mesmo utilizador tenha no mesmo local, pois apenas uma delas é considerada, com o objetivo de contar o número de turistas a visitar um local e não quantas fotos foram tiradas. Também foram removidas todas as fotos de utilizadores que não sejam consideradas turísticas, ou seja, as fotos na cidade só podem ser tiradas em dois períodos de 14 dias, com o mínimo de um mês de intervalo entre esses períodos. Restaram 664,330 meta-dados de fotografias para o estudo final onde se avalia a popularidade de um local, tendo em conta a hora, dia da semana e mês.

Em (Sakai, Tamura, & Kitakami, 2014) é feita uma análise espacial e semântica dos *posts* numa região. A fonte utilizada foi o Twitter[28] de onde foi recolhido um *dataset* com 480.000 *tweets*, de Novembro de 2011 a Fevereiro de 2012, todos estes com coordenadas geográficas, indicando a localização onde o *tweet* foi escrito.

Em (Feldman, Sugaya, Sung, & Rus, 2013), é apresentada uma forma de converter os registos de coordenadas GPS num diário, assim como criar um sistema de pesquisa textual sobre esses dados. Com os dados de coordenadas GPS recolhidos através de um *smartphone*, é possível converter num log em formato de texto que indica as ações do utilizador no seu dia-a-dia, permitindo mais tarde ao utilizador efetuar pesquisas acerca da

[26] www.flickr.com (acedido em 2016/12/08)
[27] www.traveleye.com (acedido em 2016/12/08)
[28] www.twitter.com (acedido em 2016/12/08)

ação efetuada (e.g.: "onde fui comprar livros?"). O utilizador transporta um *smartphone* que regista dados de GPS para efetuar um log inicial. Posteriormente, este log é analisado e reduzido a entradas e saídas de ruas (através do *timestamp* e coordenada GPS), assim como entradas e saídas de edifícios. Com estes dados é possível interligar com as *reviews* do *dataset* académico do Yelp[29], efetuando *reverse geocoding*[30], fornecendo assim a atividade associada a um negócio local, ou ao Google Maps para inferir o nome das ruas percorridas e os restantes pontos de paragem.

No trabalho realizado em (Rodrigues F. , POI Mining and Generation, 2010) é desenvolvida uma ferramenta para extrair informação de fontes como Yahoo, Manta e Páginas Amarelas através de *web scrapping* agregando-os numa grande base de dados. Assim podem ser utilizadas para caracterizar um local, navegação, georreferenciação de textos e análise de utilização do espaço.

O sistema também tenta classificar os POIs recolhidos segundo a classificação NAICS. Como as várias fontes de dados utilizadas não têm uma taxonomia *standard*, é importante uniformizar os dados, para conseguir responder a questões como o número de POIs do tipo "Restaurante" existentes em determinada área.

Por fim também foram obtidos dados do Flickr, incidindo maioritariamente na área de Boston, para identificar locais de interesse através das fotografias partilhadas com coordenadas GPS e tags associadas a um POI.

2.4 Arquitetura proposta

Após a análise das fontes foram selecionadas o DBpedia e o Factual para fontes em estudo, o motivo é que o DBpedia interliga de forma direta com o Wikipedia, mantendo os dados de forma estruturada, acessível e legível por sistemas que não interpretam linguagem natural.

Por outro lado, o Factual oferece o serviço Crosswalk que disponibiliza a interligação dos dados da sua fonte com outras, nomeadamente o Wikipedia, fornecendo-nos, de forma simples, alguns exemplos de ligação. Por sua vez o Wikipedia e a DBPedia estão diretamente relacionados já que esta apresenta os artigos da enciclopédia de forma estruturada. Todos estes dados interligados podem ser utilizados como exemplos de treino para um processo de aprendizagem para interligar dados que ainda não estejam presentes no serviço, tal como será descrito no próximo capítulo.

Especificamente na recolha de dados, foi desenvolvida uma ferramenta em Java para importar estes dados, cada fonte é acedida ou importada de acordo a sua metodologia de

[29] https://www.yelp.com/dataset_challenge (acedido em 2016/10/6)
[30] Funcionalidade para converter uma localização no mapa numa morada textual.

disponibilização dos dados. Depois de importados, os dados são registados numa única base de dados MySQL para facilitar o posterior acesso, filtragem e processamento.

Nesta fase apresentam-se as APIs do Factual e do DBpedia e o método utilizado para efetuar a recolha dos dados.

2.4.1 Fontes

Para efetuar a recolha de dados foi necessário verificar a forma como estes são disponibilizados.

A API do Factual[31] disponibiliza três serviços úteis para descarregar as categorias, os *Places* e o *Crosswalk*.

Para usar esta API é necessário gerar um *token*, obtido a partir da página do Factual, mediante um *login*, esse *token* será utilizado para inferir qual o utilizador e respetivos limites estão associados à conta que está a efetuar as operações.

Estes serviços assentam na tecnologia HTTP REST que funcionam nos seguintes *endpoints*:

- ***Endpoint: places***

Este *endpoint* disponibiliza os POIs tendo como parâmetros de entrada: país, região (opcional), cidade (opcional), iniciais do nome, offset e número de linhas de resultado. Apresenta como resultado um conjunto de POIs, e o número de linhas efetivas de POIs retornados. Cada POI pode ser acedido individualmente e retornar um grande conjunto de campos, entre estes encontram-se: o identificador do POI - *factual_id*, nome, endereço, categorias, website e coordenadas gps.

- ***Endpoint: categories***

Este *endpoint* disponibiliza as categorias pelas quais estão organizados os POIs não tendo nenhum parâmetro de entrada. Apresenta como resultado um conjunto de categorias, em que cada categoria consiste: num id, id da categoria pai – *parent*, e nome da categoria em várias línguas.

- ***Endpoint: crosswalk***

O Crosswalk é uma funcionalidade diferenciadora do Factual que, por si, contém algumas ligações a fontes externas, como o Yelp, e o Wikipedia, embora esteja muito incompleto é

[31] http://developer.factual.com/api-docs/ (acedido em 2016/10/24)

um ponto de partida para a recolha de dados. Como parâmetro de entrada é necessário indicar: o país, o *factual_id* e a fonte de dados – *namespace*. Caso os filtros enviados resultem numa pesquisa com sucesso, o retorno é precisamente o *factual_id*, o *namespace* e o url relativo à fonte em questão.

A DBpedia oferece as *dumps* da sua base de dados em vários formatos e várias línguas. O método apresentado suporta esta característica de abrangência desta fonte. A língua deverá ser selecionada de acordo com a cidade em estudo.

Estas *dumps* são disponibilizadas em dois formatos: *turtle-triplet* em que cada linha identifica uma propriedade com o seu valor e os dados com o formato "<sujeito><predicado><objeto>"; ou *quad-turtle*, equivalente ao anterior e adiciona ao primeiro informação de onde foi obtido aquele valor a partir do site da Wikipedia ficando com um registo com o formato "<sujeito><predicado><objeto><graph/contexto >" tal como é possível ver nos exemplos apresentados na Tabela 9. Neste formato, o sujeito é o *resource* do DBpedia, ou seja, o artigo em questão, o predicado é a propriedade e o objeto é o valor da propriedade. A *dump* está organizada alfabeticamente por sujeito e é comum uma propriedade repetir-se para o mesmo sujeito com diferentes valores.

Tabela 9 - Exemplos dos mesmos registos no formato *turtle-triplet* e *quad-turtle*

Turtle-Triplet	<http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/author> "Sicile-Kira, C"@en . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/title> "Autism spectrum disorder : the complete guide to understanding autism"@en . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/date> "2014"^^<http://www.w3.org/2001/XMLSchema#integer> . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/publisher> "Perigee"@en . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/location> "New York, New York"@en .
Quad-Turtle	<http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/author> "Sicile-Kira, C"@en <http://en.wikipedia.org/wiki/Autism?oldid=682116152#section=Further_readin&relative-line=2&absolute-line=242> . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/title> "Autism spectrum disorder : the complete guide to understanding autism"@en <http://en.wikipedia.org/wiki/Autism?oldid=682116152#section=Further_readin&relative-line=2&absolute-line=242> . <http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/date> "2014"^^<http://www.w3.org/2001/XMLSchema#integer>

<http://en.wikipedia.org/wiki/Autism?oldid=682116152#section=Further_readin&relative-line=2&absolute-line=242> .

<http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/publisher> "Perigee"@en <http://en.wikipedia.org/wiki/Autism?oldid=682116152#section=Further_readin&relative-line=2&absolute-line=242> .

<http://dbpedia.org/resource/Autism> <http://dbpedia.org/property/location> "New York, New York"@en <http://en.wikipedia.org/wiki/Autism?oldid=682116152#section=Further_readin&relative-line=2&absolute-line=242> .

2.4.2 Pesquisa

Utilizando a API disponibilizada pelo Factual e após uma análise preliminar, os filtros básicos necessários para obter todos os registos de um local seriam *country*, *region* (opcional) e *locality* (opcional).

Para recolher dados do *Factual Places* foi desenvolvido um método em Java no sistema de recolha de dados chamado ***getPlaces*** com a funcionalidade de efetuar uma chamada à API e tratar a resposta para registar os dados na base de dados MySQL. Durante o processo de desenvolvimento e testes este método foi evoluindo de acordo com as necessidades encontradas.

A partir daí as limitações inerentes da própria API, apresentadas na Tabela 10, com a conta gratuita disponível fizeram-se notar. As contas *premium* têm limites personalizados de acordo com a necessidade do cliente, não havendo uma especificação geral.

Tabela 10 - Limites da conta gratuita do Factual no *endpoint* ***places***

Limitação	***Limite***	
Limite de Linhas por Query	500	Dado um conjunto de filtros, no máximo poderão ser lidas 500 linhas utilizando a API
Limite de Linhas por Pedido	50	Um pedido com uma determinada query, só pode retornar 50 linhas de cada vez, fornecendo um offset para conseguir, a cada 50, obter as 500 linhas
Limite de pedidos por minuto	500	Só podem ser efetuados 500 pedidos por minuto, dificilmente esta operação é excedida porque não foram executados processos em paralelo
Limite de pedidos por dia	10 000	A partir dos 10 000 pedidos por dia o sistema deixa de responder, independentemente do número de registos retornados.

Tendo em conta as limitações, a primeira a ultrapassar foi o limite de 500 linhas por *query*, para tal foi necessário adicionar um novo filtro que permitisse uma filtragem parcial e sequencial, o único parâmetro de pesquisa que ofereceu estas condições foi o nome.

Foi criado um dicionário com números, letras e caracteres comuns existentes em nomes dos pontos de interesse a pesquisar.

```
static final String Dicionario = "0123456789abcdefghijklmnopqrstuvwxyz
_-.,'&$#()=/";
```

Com este dicionário é iniciada uma pesquisa, usando combinações de duas letras de forma sequencial para descarregar todos os dados de pontos de interesse. Com isto registou-se um aumento de pedidos ao *endpoint* com as várias combinações, mas com a vantagem de reduzir o número de registos retornados em cada pedido. Nesta fase, o método ***getPlaces*** foi melhorado para receber como input, o filtro de nome construído através do dicionário.

```
for (char i : Dicionario.toCharArray()) {
	for (char j : Dicionario.toCharArray()) {
		getPlaces(String.valueOf(i) + String.valueOf(j), 0);
	}
}
```

Mesmo com a condição desenvolvida pelo ponto anterior, o limite não foi devidamente ultrapassado, porque em inúmeras situações, existem mais de 500 registos para uma combinação de letras. Em cada situação que o limite de 500 registos for ultrapassado, uma nova letra do dicionário é adicionada à pesquisa anterior, criando uma pesquisa mais específica, podendo existir um número ilimitado de letras para definir a pesquisa.

```
if (Response.getTotalRowCount() >= MAX_ROW_COUNT_OFFSET) {
	for (char c : Dicionario.toCharArray()) {
		getPlaces(SearchString + String.valueOf(c), 0);
	}
}
```

Ultrapassada a primeira limitação da API encontra-se a próxima, de 50 linhas por pedido que foi facilmente ultrapassada, tendo em conta a funcionalidade de *offset*, que permite, a partir de uma pesquisa com N registos, descarregar gradualmente em pequenos grupos de registos. No nosso caso, foram utilizados os 50 registos, para evitar ao máximo efetuar pedidos desnecessários.

```
if (Response.getTotalRowCount() > Offset + MAX_ROW_COUNT) {
      Offset += MAX_ROW_COUNT;
      getPlaces(SearchString, Offset);
}
```

Por fim, existem os limites impostos num determinado período de tempo. Como estes limites lançam exceções no código, a abordagem adotada foi: o tratamento dessas exceções, estabelecer pausas e, após este tempo em espera, repetir de seguida o pedido para obter os dados em falta.

Na primeira exceção, normalmente relativa a demasiados pedidos por minuto, é feita uma espera de 15 minutos e o pedido é feito de novo. Caso este novo pedido também resulte numa exceção, é feita uma espera de 24 horas e 5 minutos.

Este tratamento de exceções para além de suprimir estas limitações da API pode resolver falhas temporárias de comunicação ou outras falhas inerentes ao próprio sistema uma vez que estas executam o mesmo pedido.

No caso da DBpedia, tendo em conta que cada linha da *dump* representa uma propriedade e o seu valor, mas com um grande volume de registos, foi desenvolvido um algoritmo de procura exaustiva que percorre todas as linhas em busca de um termo identificado pelo investigador. Cada vez que esse termo é encontrado o nome do artigo é adicionado a uma lista de combinações. Por fim a *dump* é percorrida de novo e cada vez que um dos artigos existe na lista, a propriedade é registada na base de dados.

Durante a implementação, verificou-se a performance da migração da *dump* para a base de dados era baixa e que o algoritmo poderia ser alterado para reduzir o tempo e processamento necessário para efetuar esta pesquisa. Assim o algoritmo foi alterado para funcionar da seguinte forma:

Foram adicionadas duas novas variáveis.

- Lista de *turtle-triplets* (Artigo, Propriedade e Valor)
- Variável booleana que indica se o artigo contém o termo de pesquisa

O *dump* é percorrido linha a linha, em cada linha existem duas verificações:

- Se o artigo é o mesmo que o anterior
    - Adiciona o registo à lista de *turtle-triplets*

- Se o artigo for diferente do anterior,
    - Se a variável de estado for verdadeira, regista a lista completa na base de dados
    - Coloca a variável de estado a falso
    - Limpa a lista e adiciona este novo registo.
- Se a linha contém o termo de pesquisa (coloca a variável de estado a verdadeiro)

Esta forma de trabalhar reduziu a espera para cerca de metade devido ao facto de não percorrer a *dump* duas vezes e por cada propriedade não ter que percorrer uma lista com milhares de artigos para verificar se este se encontra na lista ficando o código de leitura da *dump* como apresentado abaixo.

```
FileReader fr = new FileReader(fileToRead);
BufferedReader br = new BufferedReader(fr);
String line;
while ((line = br.readLine()) != null) {
    //convert line in turtle-triplet var
    TurtleTripplet ttl = ParseTripplet(line);
    if (!CurrentArticleName.equals(ttl.Name)) {
         if (AddArticle) {
            _ArticleNames.add(CurrentArticleName);
            //save in database
            SaveArticleData(CurrentArticleData, mysql);
        }
        CurrentArticleName = ttl.Name;
        AddArticle = false;
        CurrentArticleData.clear();
    }

    CurrentArticleData.add(ttl);

    // process the line.
    if (line.toUpperCase().contains(SearchTerm)) {
        AddArticle = true;
    }
}
```

2.4.3 Armazenamento

Para armazenar os dados recolhidos foi utilizada uma base de dados relacional em MySQL. Este é constituída por um conjunto de tabelas com as relações apresentadas pelo modelo disponível na Figura 1.

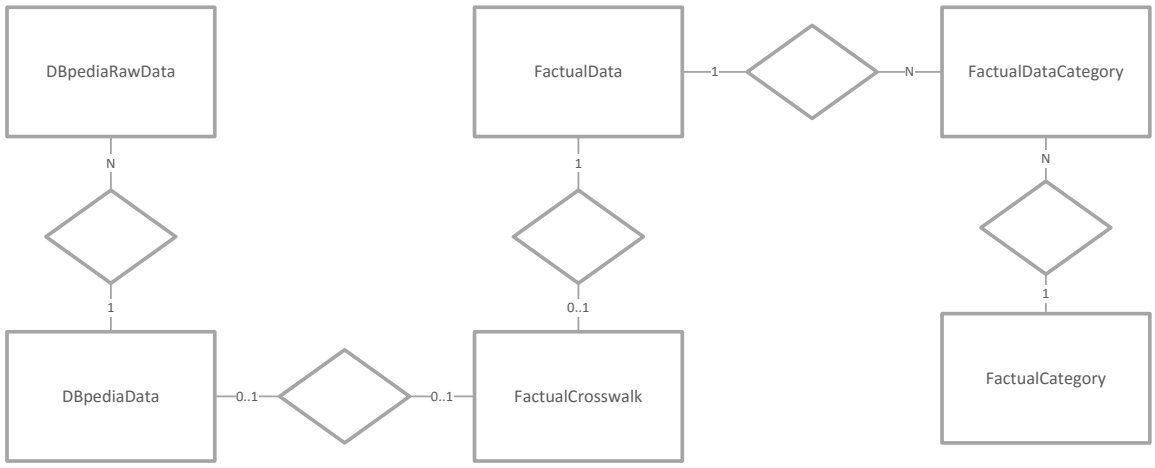


Figura 1 - Modelo conceptual de dados

Para utilizar esta estrutura, em alguns casos foi necessário efetuar algum tipo de normalização e processamento dos dados. No caso do Factual, como um *place* pode estar classificado em mais do que uma categoria, foi necessário distribuir os dados por duas tabelas a FactualData, estruturada de acordo com a Tabela 11, que inclui todos os dados disponibilizados pela API do Factual, e a FactualDataCategory, com a estrutura apresentada na Tabela 12, que regista as várias categorias para cada Factual *place*.

Tabela 11 - Estrutura de dados da tabela FactualData

Campo	
factual_id	Identificador único do POI
name	Nome do POI
tel	Número de telefone do POI
locality	Localidade
region	Regiao
latitude	Latitude
longitude	Longitude
Category_ids	Ids das categorias às quais o POI pertence
Category_labels	Descrições das categorias às quais o POI pertence
Postcode	Código postal
Address	Morada simples
Address_extendeded	Morada completa
Website	Página web do POI
Hours	Horas de funcionamento do Estabelecimento
Email	Email principal da organização
Chain_id	ID de uma cadeia
Chain_name	Nome da cadeia

Tabela 12 - Estrutura de dados da tabela FactualDataCategory

Campo	
factual_id	Id do POI
Category_id	Id da Categoria

Por sua vez a taxonomia do factual (ver anexo 6.1) fica registada na tabela Factual Category com a estrutura disponível na Tabela 13.

Tabela 13 - Estrutura de dados da tabela FactualCategory

Campo	
category_id	Id da categoria
Parents	Categoria imediatamente acima na hierarquia
En	Nome da categoria em Inglês
Pt	Nome da categoria em Português
Base_category_id	Código da categoria inicial na árvore da hierarquia

Através de uma pesquisa das categorias hierarquicamente superiores é possível chegar recursivamente à raiz e indicar a categoria base de cada subcategoria existente, presentes na Tabela 14, desta forma é possível criar grandes grupos de POIs com a mesma categoria base, com o intuito de desenvolver análises futuras sobre esses grandes grupos.

Tabela 14 - Categorias base do factual para futura análise espacial por categoria

ID	***EN***	***PT***
1	Factual Places	Lugares Factual
2	Automotive	Automóvel
20	Community and Government	Comunidade e Governo
62	Healthcare	Assistência Médica e Sanitária
107	Landmarks	Marcos
123	Retail	Venda a Retalho
177	Businesses and Services	Empresas e Serviços
308	Social	Social
372	Sports and Recreation	Desportos e Lazer
415	Transportation	Transporte
430	Travel	Viagem
467	NoExport	NoExport

Numa primeira fase, o armazenamento dos dados da DBpedia é feito na tabela DBpediaRawData que é apresentada na Tabela 15. Nesta tabela, cada registo equivale a um valor de uma propriedade, mapeando diretamente os dados do *dump* (*turtle-triplet*)

com a tabela. Neste caso existe um pós-processamento para identificar o nome do artigo, o valor e o tipo de dados da propriedade.

Tabela 15 - Estrutura da tabela DBpediaRawData

Campo	***Mapeamento***	
ID		Contador com identificação única para garantir que não existem registos sobrepostos
Name	***Artigo***	Endereço para a *resource* do artigo
Property	***Propriedade***	Identificação do endereço da propriedade
Value	***Valor***	Valor do campo com tags que identificam o tipo de dados
Clean_name		Nome da *resource* sem endereço
Clean_value		Valor sem tags
Clean_datatype		A partir das tags presentes em *Valor*, é detetado o tipo de dados

No momento de conversão dos registos para a tabela final do DBpedia deverão ser eliminados todos os registos que não são POIs, assim, foi feita uma análise manual de registos classificados como não sendo POI e propriedades únicas desses registos, apresentadas na Tabela 16 (neste caso só foi efetuado este trabalho para a língua inglesa, para outras línguas terá de ser feito o mesmo trabalho de forma manual).

Tabela 16 - Indicação de propriedades que excluem registos do DBpedia

Propriedade	***Relacionado com***
Artist	Quadros famosos ou álbuns musicais
BirthDate	Pessoas
DateOfBirth	Pessoas
Score	Jogos de competição como futebol e basquetebol.
Referee	Jogos de competição como futebol e basquetebol
Coach	Equipas e jogos de competição como futebol e basquetebol
League	Equipas e temporadas de futebol
Season	Jogos de competição como futebol e basquetebol
Author	Livros
Edition	Jogos e Livros
Collegeteam	Atletas universitários
Article	Artigos de jornais, revistas e generalistas
Champion	Campeonatos de jogos de competição
Rd1Team	Campeonatos de jogos de competição
Team	Campeonatos de jogos de competição

Cada vez que um registo contém uma destas propriedades este é ignorado no passo de conversão, o que reduz substancialmente o número de registos na tabela final mas também aumenta a qualidade dos restantes.

Na fase seguinte, os dados são analisados e distribuídos numa tabela com uma estrutura mais simplificada, a tabela DBpediaData, em que cada registo equivale a um artigo do DBpedia, as propriedades presentes na Tabela 17 foram utilizadas para mapear cada campo.

Tabela 17 - Estrutura de dados da tabela DBpediaData

Campo	***Descrição***	***Propriedade***
article	Código do artigo (faz parte do endereço)	Baseado no endereço da tabela
name	Nome do artigo	<http://dbpedia.org/property/name>
website	Website oficial	<http://dbpedia.org/property/website> <http://xmlns.com/foaf/0.1/homepage>
location	Rua ou Cidade	<http://dbpedia.org/property/location>
address	Morada	<http://dbpedia.org/property/address>
latitude	Latitude	<http://dbpedia.org/property/latd> <http://dbpedia.org/property/latm> <http://dbpedia.org/property/lats> <http://dbpedia.org/property/latns>
longitude	Longitude	<http://dbpedia.org/property/longd> <http://dbpedia.org/property/longm> <http://dbpedia.org/property/longs> <http://dbpedia.org/property/longew>
category	Categoria do artigo	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

A partir deste momento, a tabela DBpediaRawData não é mais utilizada pelo sistema, podendo ser acedida esporadicamente para pesquisas de novas propriedades

Por fim os dados do *Crosswalk* que interligam o Factual com o DBpedia são registados na tabela FactualCrosswalk, apresentada pela Tabela 18. Estes dados também passam por um processo de normalização porque, como a API retorna um endereço do Wikipedia, é necessário extrair o título do artigo desse endereço, que retirando a parte final do mesmo, equivalente ao artigo do DBpedia.

Tabela 18 - Estrutura de dados da tabela FactualCrosswalk

Campo	
factual_id	Id do POI do factual
Url	Endereço do artigo da Wikipedia
Namespace	Fonte de dados onde liga o URL (apenas é usada a *namespace* Wikipedia)
Article	Código do artigo da Wikipedia, equivale ao final do endereço

2.4.4 Evolução

De acordo com o histórico da DBpedia, as *dumps* no formato escolhido já são um standard deste a versão 3.0 de 2008[32], tendo-se mantido estável ao longo do tempo. Em paralelo surgiram e foram extintos outros formatos. Tendo em conta estes fatores, pode-se assumir que o formato se manterá no futuro sendo também do interesse da equipa manter a estabilidade para os seus utilizadores. Para utilizar novas versões dos dados, simplesmente será necessário descarregar do site e utilizar estas versões na configuração da aplicação desenvolvida.

Em paralelo com o desenvolvimento deste trabalho o Factual iniciou o desenvolvimento e publicação de uma nova versão da API descontinuando uma funcionalidade importante na recolha de dados deste projeto, o *endpoint* de obtenção categorias, estes dados estão agora disponíveis no GitHub[33] da equipa do factual em formato JSON. Para já os restantes *endpoints* não sofreram alterações, mas prevê-se a necessidade de atualização da biblioteca oficial utilizada na recolha, assim como possivelmente alguma alteração do código para suportar esta nova realidade.

2.4.5 Caso de estudo

Foram selecionadas algumas cidades de estudo de acordo com as necessidades dos projetos que estão relacionados com este.

Inicialmente selecionou-se o país de Singapura por ser o projeto onde está focado o projeto FMS e onde existem registos de paragens não rotineiras recolhidas através de inquéritos aos utilizadores de transportes públicos desse país. Sendo uma cidade-estado onde a língua inglesa é usada de forma exclusiva como língua de trabalho e na conversação em geral, existe uma grande utilização por parte da população dos transportes públicos.

De forma a testar a aplicabilidade do sistema desenvolvido em diferentes regiões, outra cidade onde a língua inglesa é falada foi escolhida para caso de estudo. Entre as fontes utilizadas, nomeadamente o Factual, a cidade norte americana de Nova Iorque é a que possui mais registos.

Apesar do projeto URBY.SENSE ainda não ter tido o seu início durante o desenvolvimento deste trabalho, selecionou-se a maior cidade em território nacional, Lisboa, de forma a testar o sistema em uma outra região onde a língua nativa não é a inglesa, mas que permite avaliar em larga medida qual será a utilidade deste trabalho para os objetivos do projeto.

[32] http://wiki.dbpedia.org/services-resources/datasets/previous-releases/data-set-30 (acedito em 2016/12/11)
[33] https://github.com/Factual/places/tree/master/categories (acedito em 2016/12/11)

Para efetuar cada caso de estudo foi alterado o ficheiro de configurações da ferramenta para acomodar cada um dos casos, na Figura 2 é possível ver o exemplo para recolher os dados da região de Lisboa.- Exemplo de ficheiro de configurações para efetuar recolha da região de Lisboa

```
#Propriedades do projeto Destination Choice Model
mysql_address=127.0.0.1
mysql_port=3306
mysql_user=root
mysql_password=
mysql_database=rui_mis
mysql_table_prefix=PT

#[factual_read]
factual_key=cuedatKTlasErXDXmDMwWa3TOWe7BEzkNJ*****
factual_secret=QQPsjfy3SA6BL5rp8tJnnVOuwMPX0Qx4lXR*****
factual_country=PT
factual_region=Lisboa
factual_locality=

#[dbpedia]
dbpedia_word=Lisboa
dbpedia_path=D:/
dbpedia_file1=infobox_properties_pt.ttl
dbpedia_file2=instance_types_pt.ttl
dbpedia_file3=homepages_pt.ttl
dbpedia_file4=article_categories_pt.ttl
dbpedia_file5=
dbpedia_file6=
dbpedia_file7=
dbpedia_file8=
dbpedia_file9=
dbpedia_file10=
```

Figura 2 - Exemplo de ficheiro de configurações para efetuar recolha da região de Lisboa

Para maior controlo cada processo foi executado de forma individual:

- Recolha do Factual Places
- Recolha do Factual Crosswalk
- Recolha da DBpedia

Inicialmente efetuou-se a recolha dos dados do Factual Places, devido às limitações previamente mencionadas, os tempos de recolha incluem tempos de espera inerentes à própria limitação da API e foram recolhidos em média 300 mil registos por área de estudo, tal como indica a Tabela 19.

Tabela 19 - Resultados da recolha de dados do *endpoint* Factual Places

Local	***Tempo de recolha***	***Número de Registos***
Singapura	~53 horas	349,749
Região de Lisboa	~94 horas	281,037
Cidade de Nova Iorque	~84 horas	334,300

De seguida iniciou-se a recolha do Factual Crosswalk, como descrito na secção 2.4.2 esta recolha utiliza como filtros os Ids gerados pela operação anterior e dadas as limitações da API, apenas é possível analisar cerca de 100,000 registos por dia. Verifica-se que apesar de ser a primeira cidade em estudo, Singapura não contém registos que permitem, só por si, utilizar esta fonte. Desta forma, a cidade de Nova Iorque poderá contribuir para uma solução para o enriquecimento dos dados necessários para que a análise espacial seja possível em Singapura, tal csomo será discutido na secção 3.3.5.

Tabela 20 – Estatísticas da recolha de dados do Factual Places

Local	*Tempo de recolha*	*Número de Registos*
Singapura	~79 horas	0
Região de Lisboa	~54 horas	32
Cidade de Nova Iorque	~79 horas	2,073

Uma vez que a recolha para Singapura não apresentou resultados no Crosswalk houve a necessidade de utilizar outra cidade de língua inglesa para mais tarde conseguir criar a relação entre os pontos. Foi assim selecionada a cidade de Nova Iorque por ser a cidade dos Estados Unidos com mais POIs registados nesta plataforma.

Analisando os dados do Factual Crosswalk verificou-se que a relação existente entre o Factual e a Wikipedia para Lisboa indicava os artigos em português, assim sendo, foi necessário descarregar, para além dos *dump* em inglês, também os portugueses.

A recolha final do DBpedia gerou um total de 3,206,797 propriedades, após limpeza dos dados e retirados registos associados a entidades que não eram POIs válidos, referentes a pessoas, eventos, livros e artigos, foram removidas 2,170,511 propriedades. A Tabela 21 apresenta esses dados distribuídos por cada cidade em estudo, com o tempo de extração das propriedades, o número de propriedades antes e após remoção de registos extra e o número de artigos final.

Tabela 21 - Métricas do processo de recolha de dados do DBpedia

Local	*Tempo de recolha*	*Número de Propriedades*	*Número de Artigos*
Singapura	~3 horas	338,222/93,752	2,709
Região de Lisboa	~1 hora	113,055/110,019	4,205
Cidade de Nova Iorque	~14 horas	2,755,520/832,515	15,536

2.5 Discussão

A recolha foi feita com sucesso, conseguindo obter grande parte dos registos do Factual para cada área de estudo. No entanto, devido às limitações, a utilização da API demorou demasiado tempo, pois não podem ser executadas várias recolhas em paralelo, uma vez que as limitações são impostas à conta com que se acede.

Entretanto também foi possível verificar que os métodos de recolha são válidos para fontes tanto de língua inglesa assim como fontes de outras línguas, o que é uma grande vantagem. No entanto será necessário efetuar algum trabalho para melhorar os resultados obtidos nas fontes de diferentes línguas, especialmente no momento de excluir registos que não sejam identificados como POIs do DBpedia, uma vez que a as *dumps* em diferentes línguas, contém os nomes das propriedades traduzidas e os filtros já criados foram definidos para a língua Inglesa.

Uma vez que esta recolha não necessita de muita interação por parte do utilizador para além da configuração inicial é bastante simples extrapolar para várias cidades e recolher dados de forma exaustiva para diferentes estudos.

3 Análise espacial

Este capítulo descreve a análise espacial e o relacionamento de dados assim como a análise da ferramenta desenvolvida com esta finalidade estando organizado do seguinte modo: na secção 3.1 são descritos os problemas existentes devido à grande dispersão dos dados. Na secção 3.2 é feita uma análise a trabalhos já existentes que abordam o tema de planeamento urbano e *clustering* de dados assim como a correspondência entre fontes de dados. A secção 3.3 apresenta a análise da arquitetura assim como são feitos testes com vários casos de estudo e analisados os seus valores. Na secção 3.4 são discutidas as principais questões quanto à aplicabilidade da solução implementada a diferentes áreas geográficas de estudo.

3.1 Introdução

Na última década, devido ao crescimento das aplicações com tecnologia GPS, a partilha de dados espaciais, mapeamento, investigação e serviços tem crescido. No entanto, a aplicação destes dados à análise social e de classificação do uso do solo (Mennis & Guo, 2010), assim como a existência de sistemas disponíveis para apresentação da distribuição de serviços em espaços urbanos são bastante limitados, não havendo muitas ferramentas a agrupar os serviços por categorias específicas. Por vezes para identificar a falta ou concentração de um determinado tipo de serviço é necessário avaliar um conjunto de pontos um a um tornando este trabalho muito complexo e propenso a erros.

Existem também centenas de fontes de dados, com dados que se complementam e bastante úteis quando se pretende analisar informação sobre o espaço urbano, no entanto não existe nenhuma relação entre os dados de cada fonte, o que incapacita a análise de determinadas informações devido à dispersão dos dados.

Aqui propõe-se desenvolver um método capaz de criar *clusters* por categorias específicas por forma a facilitar a síntese da distribuição de serviços de forma espacial sobre uma área em estudo. Assim como é proposta a utilização de algoritmos de aprendizagem, nomeadamente algoritmos de classificação existentes na literatura para aprender relacionar informações de duas fontes de dados de forma a complementar a informação existente em cada uma delas.

3.2 Trabalhos relacionados

Em (Canneyt, Schockaert, Laere, & Dhoedt, 2011) é possível verificar que feita a análise espacial (e neste caso, temporal) é possível descobrir pontos de interesse de cidades

utilizando as fotografias partilhadas numa rede social, uma vez que cada vez mais equipamentos possuem a capacidade de adicionar uma tag georreferenciada às fotografias.

No trabalho realizado por (Alves, Rodrigues, & Pereira, 2011) que agrupa semanticamente os *clusters* de Pontos de Interesse (POIs), utiliza técnicas de processamento de linguagem natural e a taxonomia dos mesmos para integrar fontes de dados (oriundos da plataforma Yahoo! Local e Gowalla onde era inferida a popularidade dos POIs).

No artigo (Shi, Mamoulis, Wu, & Cheung, 2014) é proposta uma análise de *clusters* de forma social, usando os *checkin's* dos utilizadores de redes sociais como o Facebook[34], o Foursquare[35] e o Gowalla. Neste trabalho, os autores efetuaram uma extensão ao algoritmo DBSCAN de nome DCPGS (Density-based Clustering Places in Geo-Social Networks) que para além de considerar a distância entre dois pontos, também considera a distância social. A distância social é calculada com base em utilizadores que tenham feito *checkin* nos vários POIs, quando um conjunto de POIs contém *checkin* de um número de utilizadores comuns estes POIs são considerados socialmente similares.

Já em (Sakai, Tamura, & Kitakami, 2014) é justificada a necessidade de construir um novo algoritmo de *clustering* de densidade, com o objetivo de descobrir áreas relacionadas com um determinado tópico ou evento a partir de "documentos" com coordenadas geográficas. Esse algoritmo deve ser capaz de criar *clusters* através de uma melhoria ao DBSCAN, que dado um conjunto de documentos georreferenciados, agrupa semanticamente os POI's de acordo com o texto presente nos documentos, utilizando as palavras mais frequentes para definir essa semântica, para gerar *clusters* com uma maior especificidade e excluindo do cluster pontos com um contexto diferente. Na analise efetuada utilizados registos do Twitter como documentos, pois devido à massificação de dispositivos com GPS existem muitos registos com esta informação.

Como os dados relativos a Pontos de interesse em áreas urbanas são difíceis de visualizar, em (Polisciuc, Alves, & Machado, 2015) é proposto o desenvolvimento de uma ferramenta capaz de apresentar os *clusters* em formato de polígonos em vez de mostrar os pontos sobrepostos. Estes polígonos contém dados textuais como os POIs neles contidos e as categorias predominantes. Uma vez que os mapas tipográficos contém informações importantes (tal como nomes de ruas, rios e locais), os polígonos são apresentados de forma semitransparente para não ocultar esta informação ao contrário dos métodos de visualização atual que, em muitas situações a escondem. No contexto deste projeto foi desenvolvida uma plataforma web, CityClusters[36] para visualização dos mapas da forma proposta. Esta plataforma fornece um mapa interativo onde se pode ver a informação de cada *cluster* e visualizar os clusters que partilhem POIs de uma determinada categoria. Também inclui até três níveis de zoom que podem apresentar diferentes níveis de detalhe

[34] https://www.facebook.com/about/location (acedido em 2016/08/02)
[35] https://foursquare.com/ (acedido em 2016/12/08)
[36] http://ubiquo.dei.uc.pt/cclusters/ (acedido em 2016/12/11)

de acordo com a aproximação do mapa, ao mesmo tempo que é possível aplicar filtros para reduzir o número de clusters visíveis e encontrar uma informação específica de forma mais rápida.

Em (Cohen, Ravikumar, & Fienberg, 2003) foi utilizada uma ferramenta *open source* desenvolvida em java com métodos de *name-matching* para verificar com os vários métodos disponíveis, qual a melhor forma de classificar pares de entidades como sendo *matching* ou *non-matching*.

Foram estudados algoritmos de distância de *strings* em que objetivo é obter o valor máximo durante a comparação das mesmas, estes algoritmos, Jacard, Jaro, Levenstein e Monge and Elkan, classificam um par de *strings* pela sua semelhança, quanto mais semelhantes forem os nomes maior será o valor retornado pelo algoritmo.

Estes algoritmos foram utilizados para avaliar qual deles teria uma melhor performance e menos erros na avaliação de pares de entidades.

Já na análise desenvolvida por (Milne & Witten, 2008) são exploradas técnicas para associar endereços da Wikipedia a textos de linguagem natural, são utilizados métodos de *machine learning* para enriquecer textos não estruturados com ligações para a Wikipedia.

Este trabalho é dividido em duas fases distintas, inicialmente é proposto um método para retirar ambiguidade a termos encontrados no texto, utilizando *machine learning* e casos de exemplo onde existem endereços da Wikipedia ambíguos e endereços que não o são, aqui são utilizados os segundos para procurar referências de endereços ambíguos e garantir o relacionamento entre todos os possíveis endereços do documento.

A segunda fase, consiste na deteção de endereços, que se divide em quatro avaliações, também desenvolvidas utilizando *machine learning*:

- *Link Probability* – As instâncias são analisadas para encontrar referências. Neste caso é analisada a probabilidade entre as várias ocorrências de palavras ou conjuntos de palavras para formar um link. Tendo vários pares relacionados, qual deles o mais provável de ser aquele que se identifica no contexto do documento.
- *Relatedness* – Já houve um processo para encontrar o relacionamento entre vários tópicos e a desambiguação praticada na primeira fase, no entanto aqui efetua-se uma avaliação entre todos os endereços associados ao documento para efetuar exclusões daqueles que não contenham relações.
- *Disambiguation Confidence* – O classificador de desambiguação verifica se um determinado termo faz sentido naquele tópico, assim como cria uma probabilidade de confiança em relação ao do documento. Neste caso os tópicos são todos avaliados de novo de acordo com o método anterior, para verificar se um dos tópicos ambíguos não foi mal selecionado.

- *Generality* – Criada uma forma de ignorar tópicos demasiado comuns, que o utilizador já conheça à partida, o ideal será fornecer links que o utilizador necessite de ter mais informação.
- *Location and Spread* – Aqui considera-se mais importante um link sobre um tópico que apareça várias vezes no documento, no inicio ou final do mesmo, pois é uma área de destaque, como a introdução e a conclusão.

Utilizando o método foi conseguida uma precisão aproximadamente 75% a encontrar e desambiguar links do Wikipedia em textos de linguagem natural. Outros links encontrados correspondiam a links que apontam para outros dados, falha no algoritmo de desambiguação *ou* endereços não relevantes.

3.3 Arquitetura proposta

Nesta secção é apresentada a arquitetura proposta para a análise de utilização dos espaços urbanos apresentando os dados de entrada, efetuando vários tipos de *clusters* espaciais. Da mesma forma é apresentado o método utilizado para classificar a relação entre registos das duas fontes de dados em estudo.

3.3.1 Dados de Entrada

Para efetuar análises espaciais serão utilizados os dados recolhidos na fase anterior provenientes do *Factual places* e do *Factual categories*.

Como o *factual places* contém na sua grande parte POIs com coordenadas GPS, pretende-se criar e analisar um conjunto de *clusters* espaciais.

3.3.2 Pré processamento

Para criar estes *clusters* será necessário determinar a categoria base de cada POI recolhido para criar vários grandes grupos, pois existe todo um interesse distribuir os dados de uma forma semelhante a (Sakai, Tamura, & Kitakami, 2014). Devido à distribuição hierárquica da taxonomia do Factual, foi inferida a categoria principal de cada subcategoria, com essa categoria principal é possível criar *clusters vetoriais* distribuídos por áreas generalistas como serviços de saúde, marcos, etc.

3.3.3 *Clustering* Espacial

Inicialmente, para perceber como estão distribuídos os serviços numa determinada área, foi usado o Quantum GIS (QGIS) para criar *heatmaps* (representação bidimensional de dados utilizando cores para representar os valores tal como representado na Figura 3) com os POIs descarregados, assim permitiu uma primeira análise espacial através da perceção da densidade de POIs existente em cada zona de acordo com a área em estudo.

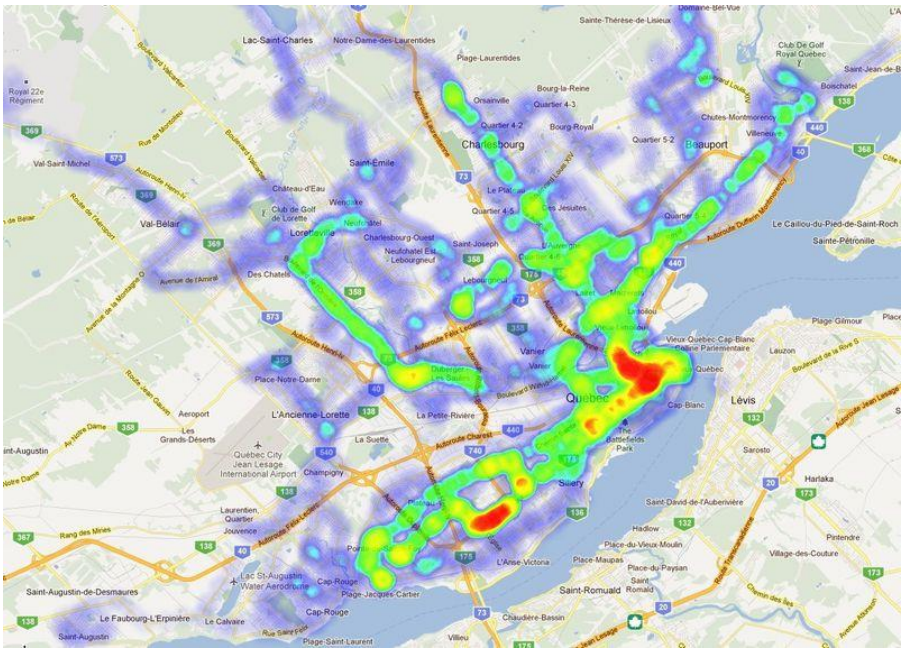


Figura 3 - Exemplo de um *heatmap*[37]

Vários estudos com os parâmetros de *heatmaps*, presentes na Figura 4, permitem uma melhor perceção da verdadeira densidade dos pontos e de valores que podem ser utilizados mais tarde em algoritmos de *clustering* vetorial. Os parâmetros utilizados consistem na dimensão da matriz e o raio de cada pequeno *cluster* gerado nesta ferramenta.

[37] http://humantransit.org/2012/04/heatmaps-of-service-intensity.html (acedido em 2016/12/08)

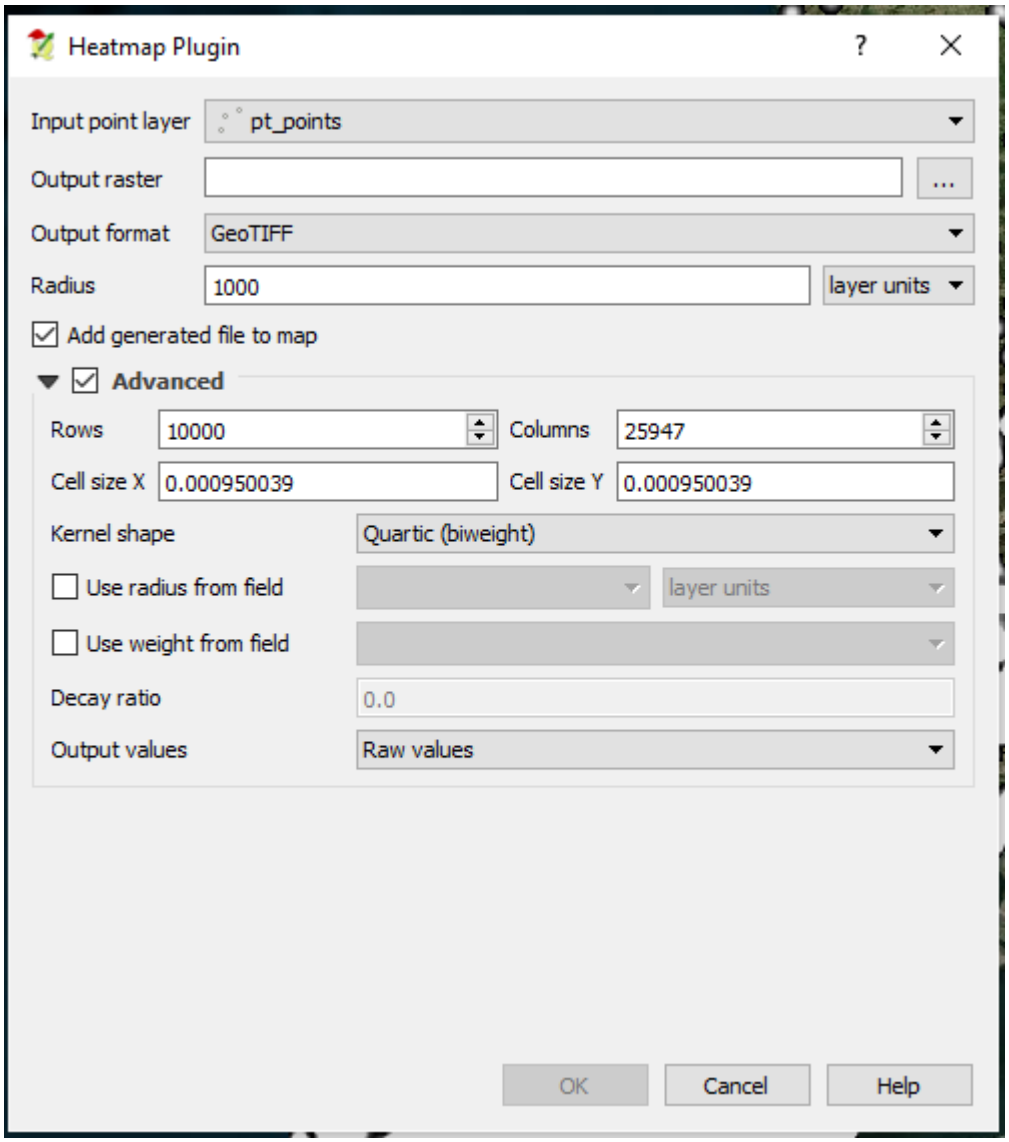


Figura 4 - Parâmetros existentes no *plugin* de *heatmaps* do Quantum GIS (QGIS)

Para o sistema funcionar corretamente, a dimensão da matriz tem de ser superior ao somatório de *clusters* possíveis no eixo. Por exemplo, se a área de estudo tiver 50,000 metros no eixo do X e o raio de um *cluster* for de 1,000 metros a matriz deverá ter pelo menos 50 colunas. Esta definição matriz corresponde à resolução do *heatmap*, quanto maior a matriz, maior a resolução e melhores se tornam os resultados finais, no entanto também incrementa o tempo para a ferramenta analisar todos os valores a colocar na matriz. Na Figura 5, é possível ver a mesma análise de *clusters* com o raio de 1,000 metros em que num caso temos uma matriz de 60 colunas por 37 linhas (em cima) e no segundo caso uma resolução de 600 colunas por 369 linhas (em baixo).

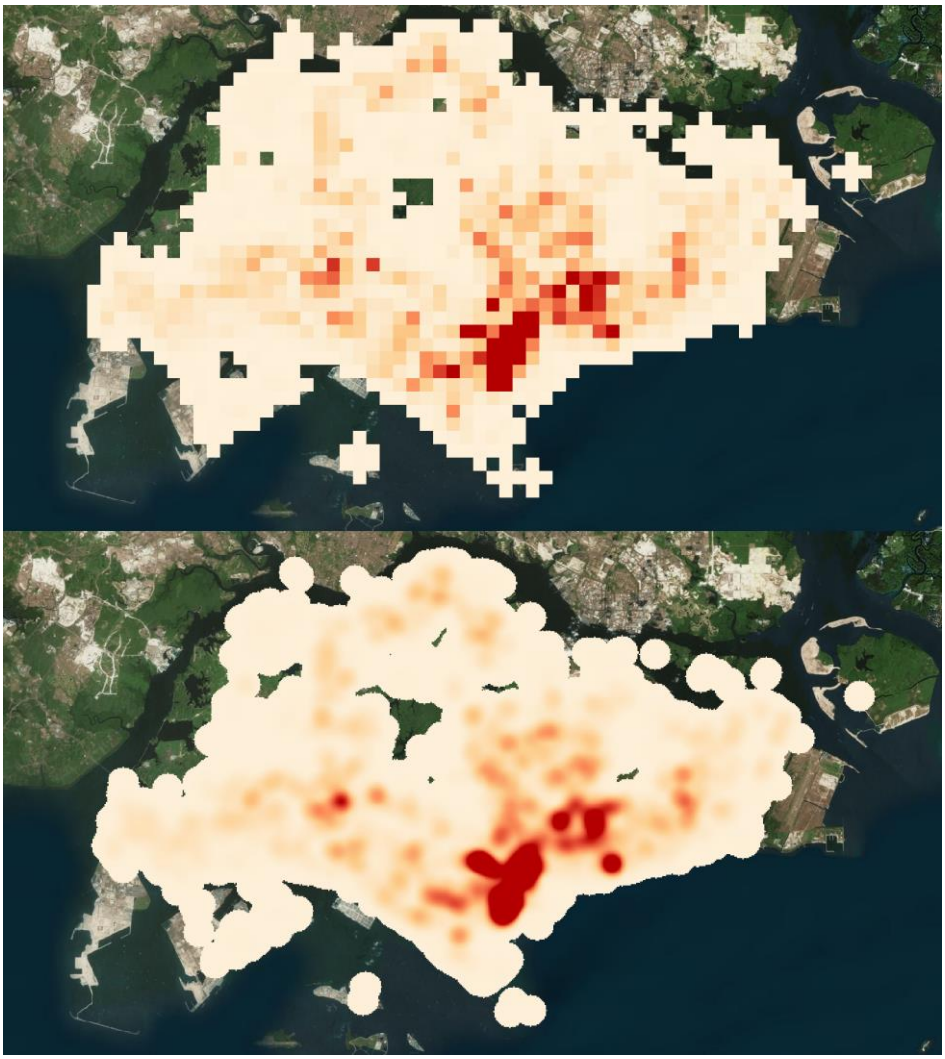


Figura 5 - Diferentes resoluções na geração de *heatmaps*

De seguida partiu-se para a criação de clusters vetoriais utilizando o algoritmo DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), este utiliza duas variáveis. O número mínimo de pontos, e a distância máxima (épsilon) entre este número de pontos para formar um *cluster*. Em zonas com uma densidade de pontos elevada, é criado um *cluster*, porque existe um mínimo de N pontos numa vizinhança dentro da distância máxima definida.

O algoritmo assim define três regras:

- Um ponto é definido como ponto central se tem, dentro dos limites de distância máxima, o número mínimo de pontos definido.
- Um ponto é alcançável se estiver à distância máxima de um ponto central.
- Todos os pontos não alcançáveis são *outliers*.

Um cluster gerado com DBSCAN satisfaz duas propriedades:

- Todos os pontos de um *cluster* estão conectados entre si.
- Se um ponto é alcançável por um ponto central, este também pertence ao *cluster*.

Ao agrupar os pontos desta forma, é possível identificar zonas com grande densidade de pontos e zonas de baixa densidade, tal como acontecia com os *heatmaps*, só que agora em vez de apenas termos imagens (formato matricial – *raster*), pretende-se criar uma camada vetorial com os polígonos que formam os *clusters* identificados pelo algoritmo. Tendo em conta que, à semelhança do (Sakai, Tamura, & Kitakami, 2014) é efetuada a distribuição por categoria, adiciona-se a capacidade de identificar com mais facilidade, tendo em conta a zona e a categoria de uma paragem.

É necessário definir as variáveis épsilon e o número mínimo de pontos do DBSCAN, nesta fase existem várias soluções propostas.

No caso da ferramenta Spark DBSCAN[38] disponibilizada no GitHub é apresentada uma solução em que é calculado o vizinho mais próximo de cada *cluster* para determinar o valor de épsilon e de seguida, tendo em conta esse valor, encontrar o número de vizinhos de cada ponto que se encontrem dentro da distância previamente determinada. Tendo em conta que o nosso *cluster* tem uma terceira variável, as categorias, este algoritmo não se mostrou capaz de resolver o problema apresentado, embora pudesse ser refeito com esta nova variável.

No livro (Tan, Steinbach, & Kumar, 2005) foi apresentada outra solução semelhante, em que inicialmente se define o número mínimo N de vizinhos pretendido para gerar um cluster. Tendo em conta esse valor, é calculada distância do N vizinho mais próximo. Por fim, com estas distâncias é criado um gráfico que apresenta a distância do N vizinho mais próximo com o número de pontos a essa distância para selecionar um épsilon. Após selecionado o épsilon é necessário ter em conta que todos os registos que tenham a distância menor ou igual ao pretendido, vão ser pontos centrais, todos os que não conseguirem estar dentro dos limites serão pontos apenas pontos alcançáveis ou *outliers*.

Este algoritmo foi melhorado para ter em conta as categorias dos POIs quando efetua os cálculos das distâncias, assim sendo, se dois pontos não forem da mesma categoria, não será calculada qualquer distância entre eles.

Na tentativa de obter os melhores resultados possíveis houve uma pesquisa sobre algoritmos de avaliação dos clusters, por forma a conseguir garantir uma boa qualidade

[38] https://github.com/alitouka/spark_dbscan/wiki/Choosing-parameters-of-DBSCAN-algorithm

dos resultados, no entanto não foi possível encontrar informação sobre métodos não supervisionados de avaliação de clusters nos trabalhos encontrados.

Deverão ser feitos alguns testes utilizando o Weka e o algoritmo de *clustering* K-Means (Hartigan & Wong, 1979) onde será introduzido o número de clusters gerados pelo DBSCAN e se analisam os dados oferecidos pelo mesmo para, de acordo com estes, verificar a qualidade do *clustering*.

O K-Means, é um método utilizado para criação de *clusters* e tem como objetivo a distribuição de n pontos em k clusters. Este é um algoritmo heurístico, não garantindo o melhor resultado. Num primeiro passo, k pontos são selecionados de forma aleatória, a partir daí, todos os pontos são avaliados para se associarem ao *cluster* com o ponto médio mais próximo. De seguida os centróides são calculados, descobrindo os novos pontos médios para efetuar avaliação anterior, este ciclo ocorre até ocorrer um máximo local (os pontos não são alterados de uma iteração para a seguinte)

Após criação dos *clusters* é possível efetuar uma importação para a plataforma *CityClusters*, utilizando uma ferramenta desenvolvida pelo criador da mesma, que recebe um objeto com os clusters em JSON e cria um ficheiro com o *convex hull* de cada cluster (polígono que contém todos os pontos de um *cluster* para o representar) .

Com estes dois ficheiros (o *input* e *output* da ferramenta) a plataforma consegue criar um mapa interativo onde são apresentados os vários clusters gerados agrupados pelo tipo de cluster, tal como apresentado na Figura 6, também com a capacidade de incluir vários níveis de zoom de acordo com a aproximação mostrar dados mais específicos.

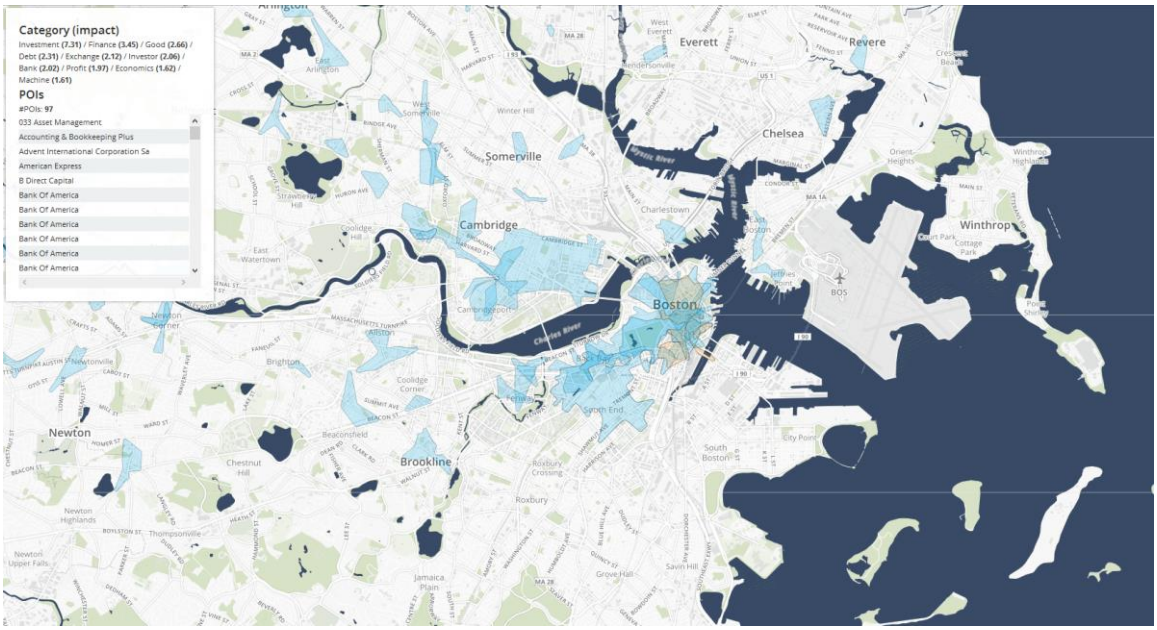


Figura 6 - Exemplo de *clusters* apresentados pela plataforma CityClusters[39]

3.3.4 Classificação

O passo seguinte será relacionar os POIs do factual com os artigos do Wikipedia, para tal foram utilizados os dados do *crosswalk*. No entanto verificou-se que apenas uma pequena parte dos POIs está presente no *crosswalk*. Assim pretende-se criar um classificador para avaliar uma ligação entre um POI e um artigo do DBpedia através da utilização da informação estruturada do DBpedia (que contém um conjunto de atributos da entidades mencionadas no Wikipedia) e o POIs do factual.

Para criar esse classificador serão necessários implementar métodos de comparação entre as duas fontes utilizando os dados disponíveis para tal.

Foram utilizados métodos de comparação de ***strings*** que vão além da simples verificação se uma é igual a outra, o primeiro fator de exclusão é o *Soundex* (MySQL SOUNDEX() function, 2016), fornecido pelo MySQL, método que converte um texto num valor fonético e indexa-o de acordo com a pronuncia e sonoridade em Inglês, assim como vários métodos de comparação (Jaccard, Jaro-Winkler, Levenshtein, Monge-Elkan, Smith-Waterman e UnsmoothedJS) utilizados por (Cohen, Ravikumar, & Fienberg, 2003). Estes métodos de comparação foram utilizados sobre o nome e sobre o *website* de cada caso de teste entre o Factual e o DBpedia.

[39] http://ubiquo.dei.uc.pt/cclusters/ (acedido em 2016/12/08)

Em conjunto com a comparação de nomes, será calculada a distância entre o POI do Factual e o artigo do DBpedia, no caso de ambos incluírem coordenadas GPS. Para cirar o conjunto de treinos para o classificador fazer a sua análise, foi adicionada uma propriedade IsMatch que indica se o par de registos é ou não coincidente.

A ferramenta Weka contém um conjunto de classificadores standard, que analisam um conjunto de dados (*training set*) para criar um conjunto de regras, regras essas que serão mais tarde aplicadas a outros conjuntos de dados para efetuar o *match* ente as várias fontes de dados.

Para fornecer o *training set* serão utilizados os dados do Factual Crosswalk, que possui a ligação entre um artigo do DBpedia e um POI do Factual, nesta fase é necessário criar um conjunto de registos que se relacionem e que não se relacionem, para criar estes registos utiliza-se a ligação do Crosswalk para determinar os dados relacionados e utilizam-se os mesmos registos, utilizando artigos aleatórios, com os mesmos POIs, que não correspondam à relação anterior, inferindo assim os casos que não são relacionados.

. Um primeiro passo foi recolher um conjunto de exemplos positivos através dos dados presentes no *crosswalk*, de seguida foram utilizados algoritmos de classificação binários disponíveis no Weka para identificar a relação entre um artigo e um POI como *match* ou não com o objetivo de avaliar qual o classificador que oferece melhores resultados.

Por fim, com os vários modelos gerados por esses classificadores é possível classificar novos datasets e descobrir desta forma um conjunto de *matches* útil para criar novas ligações entre as duas fontes.

3.3.5 Caso de Estudo

Após a recolha de dados foram criados os *heatmaps* para Singapura – Figura 7, Lisboa Figura 8, e Nova Iorque, Figura 9, estes permitiram-nos avaliar a densidade de serviços uma vez que é apresentada uma escala de cores no mapa de acordo com a concentração de POIs do Factual. Nesta fase verificou-se que embora no Factual os registos estejam classificados como pertencendo à região de Lisboa, as coordenadas GPS estão espalhadas por todo o país, o que indica que existe um erro nos dados existentes na plataforma, criando alguns *clusters* fora da região de Lisboa.

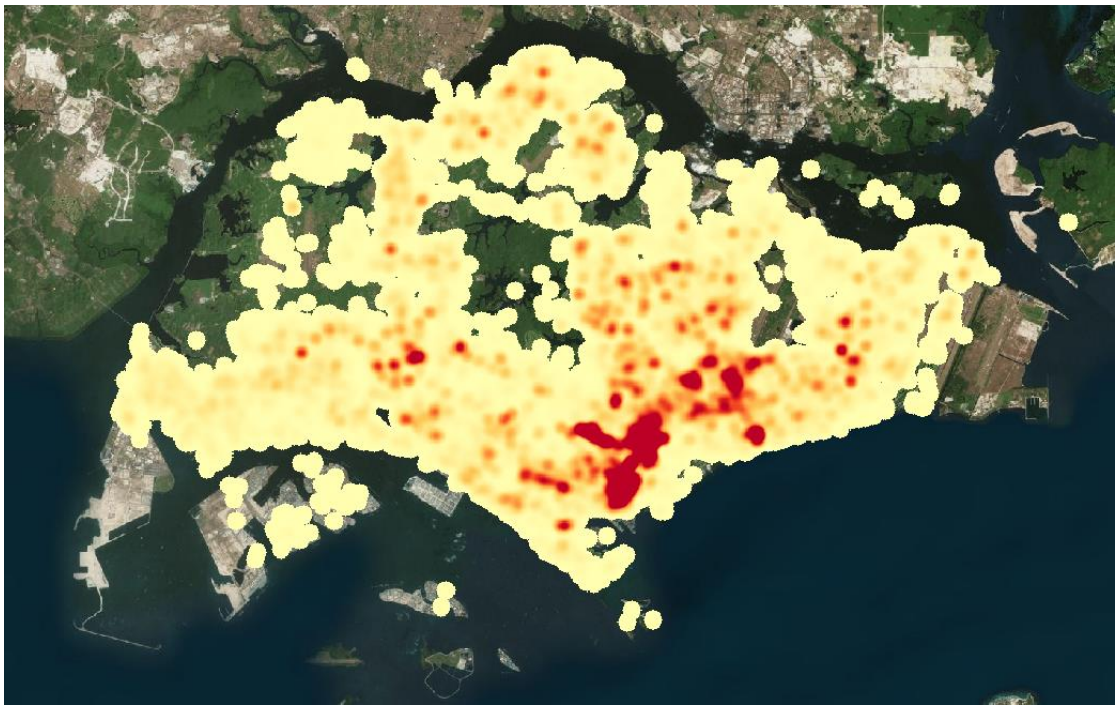


Figura 7 - *Heatmap* de densidade de POIs em Singapura

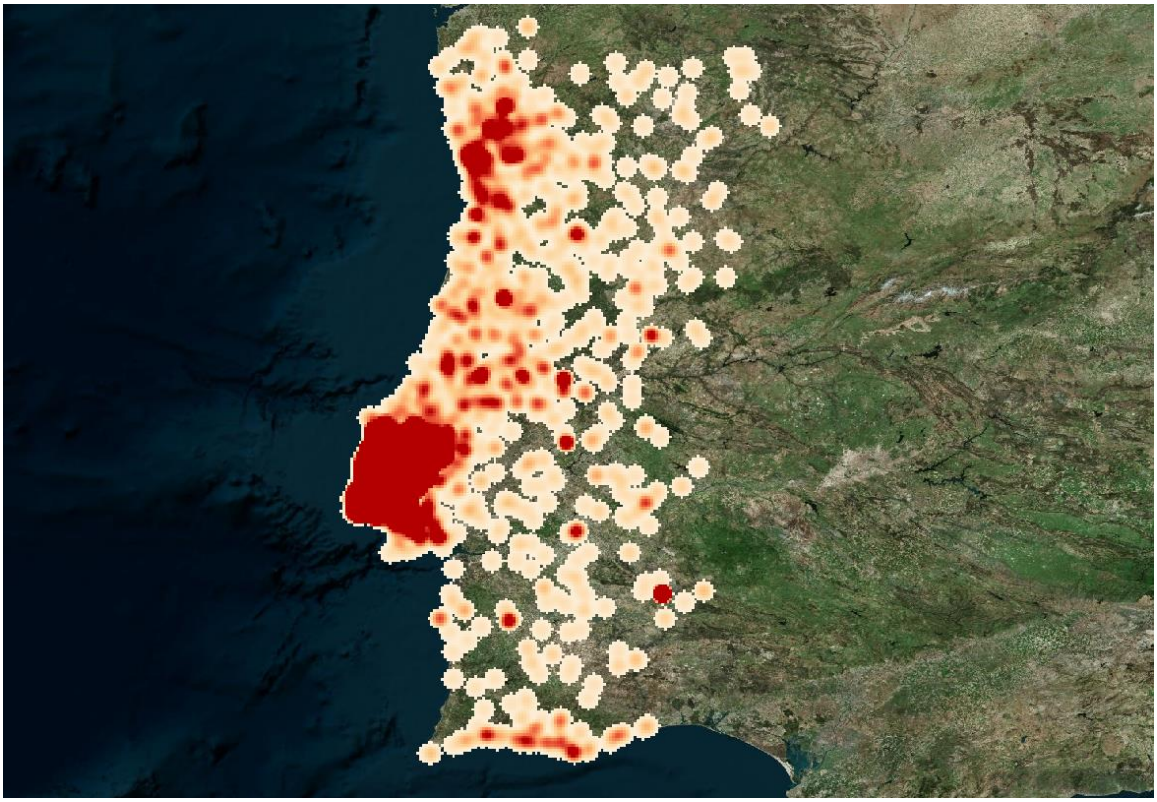


Figura 8 - *Heatmap* de densidade de POIs classificados na região de Lisboa

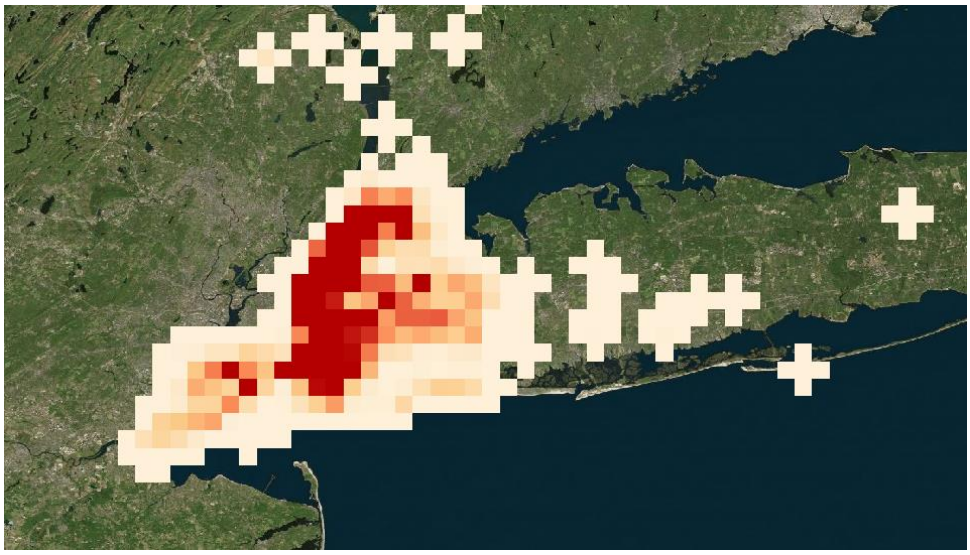


Figura 9 - *Heatmap* de densidade de POIs cidade de Nova Iorque

Com esta análise espacial baseada em *heatmaps* foi possível verificar que existem vários casos de classificação incorreta no Factual, seja erro na região ou erro nas coordenadas encontradas no POI. Por exemplo, uma empresa localizada na Lousã, distrito de Coimbra, está classificada no Factual como pertencendo à região de Lisboa[40], assim como um POI com morada em Oeiras, contém a coordenada de GPS nos arredores da Figueira da Foz[41].

Antes da criação dos *clusters* utilizando o método do DBSCAN foram efetuadas várias análises de *heatmaps* para cada área em estudo para verificar a distribuição de serviços presentes em cada local.

Na Figura 10, Figura 11 e Figura 12 é possível verificar os vários estudos efetuados com raios distintos para criação dos *clusters*, esta validação apoia a compreensão a distribuição dos dados.

[40] http://www.factual.com/518280f0-e2e1-4d25-be8f-9b2812979a39 (acedido em 2016/12/09)
[41] http://www.factual.com/c6d60e76-8b7e-463f-aa61-c0bc4a7f767c (acedido em 2016/12/09)

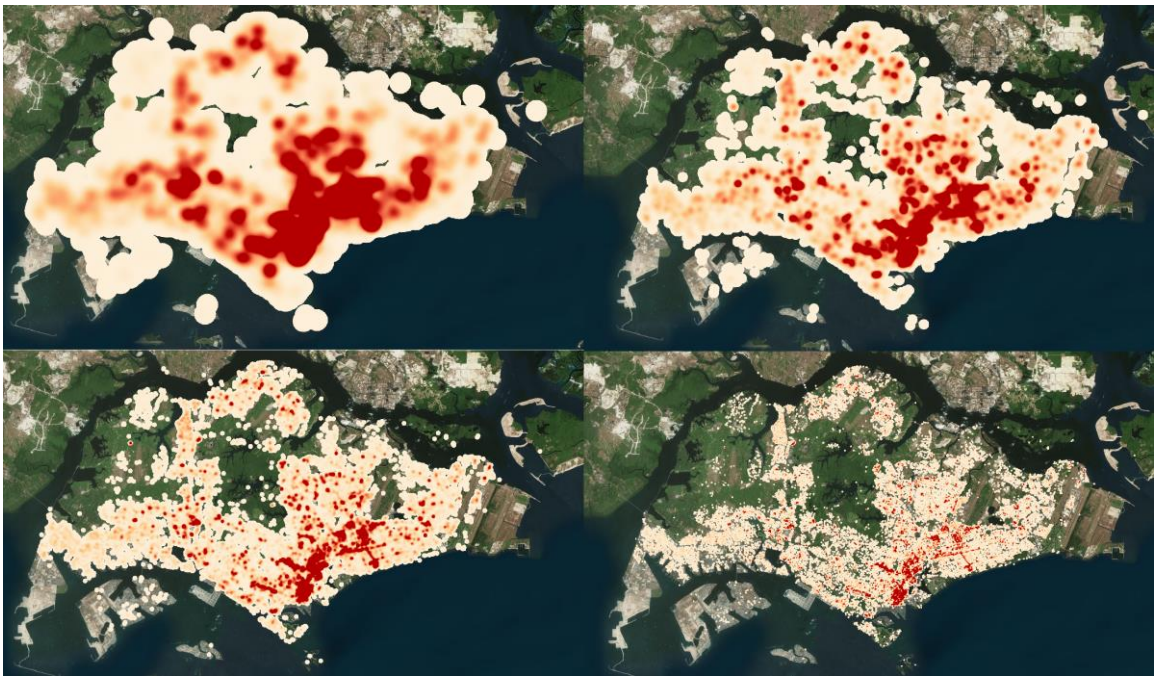


Figura 10 - Estudo de vários raios de *heatmaps* para Singapura

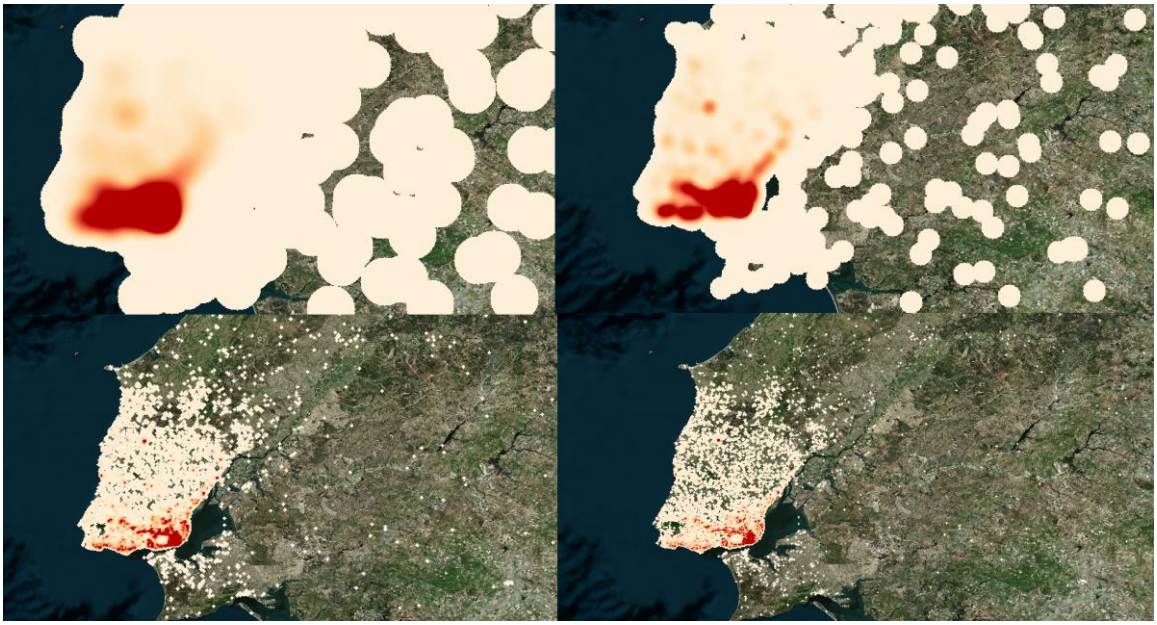


Figura 11 - Estudo de vários raios de *heatmaps* para a região de Lisboa

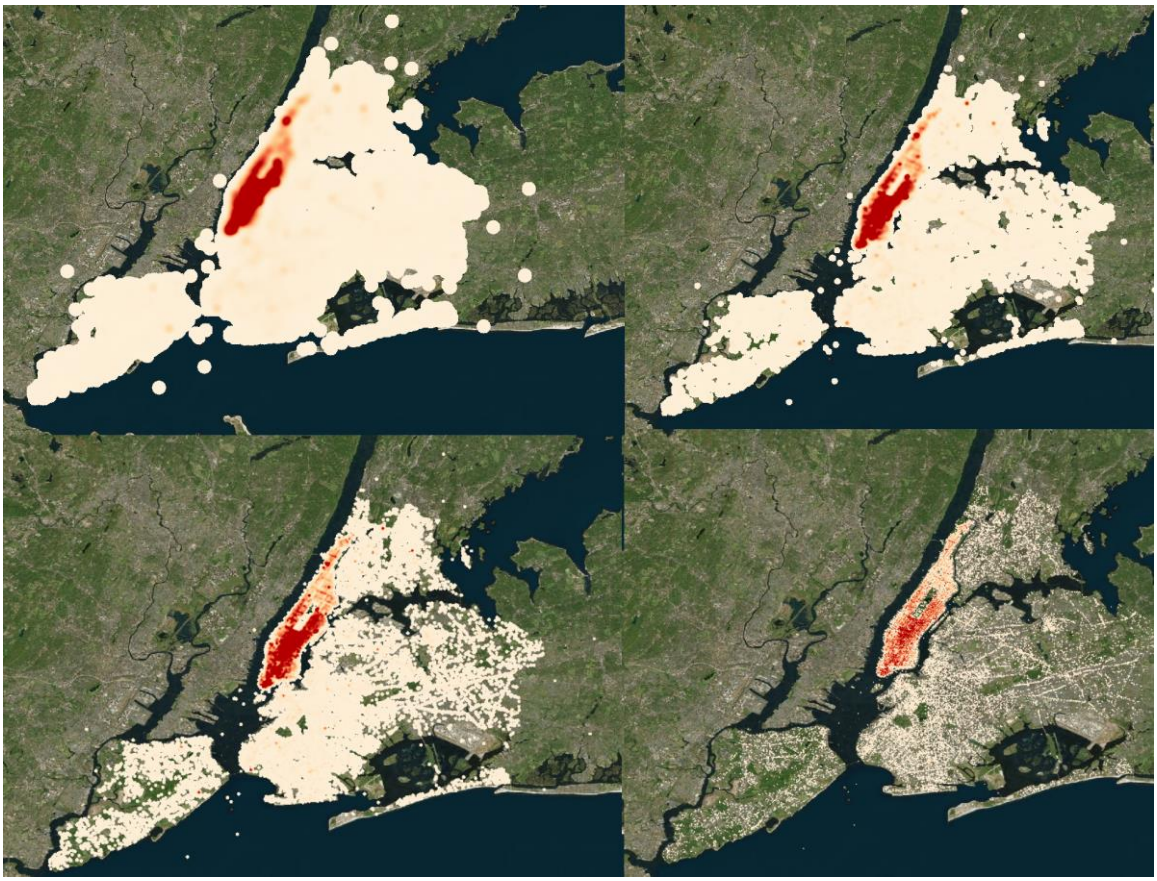


Figura 12 - Estudo de vários raios de *heatmaps* para a cidade de Nova Iorque

Na figura Figura 10 é possível ver a análise efetuada a Singapura onde foram criados *heatmaps* com *clusters* com variações do raio mínimo do cluster de 1,000 metros (canto superior esquerdo), 500 metros (canto superior direito), 250 metros (canto inferior esquerdo) e 100 metros de raio (canto inferior direito).

No caso de Lisboa, na Figura 11, a avaliação com um raio de 10,000 metros (no canto superior esquerdo) cria um único grande *cluster*, não havendo muitas separações, pelo que foi necessário reduzir essa área para clusters mais pequenos, assim, foram também efetuadas experiências com clusters com raio de 5,000 metros (no canto superior direito), 1,000 metros (canto inferior esquerdo) e 500 metros (canto inferior direito).

Em Nova Iorque é possível verificar que os POIs estão muito concentrados, especialmente em Manhattan à volta do Central Park é possível ver a área vermelha, esta análise está representada pela Figura 12 que, no canto superior esquerdo, apresenta o *heatmap* de *clusters* com um raio de 1000 metros, imediatamente à direita, com um raio 500 metros, no canto inferior direito, 250 metros e no canto inferior esquerdo 100 metros.

O DBSCAN utiliza duas variáveis para além dos POIs para gerar os clusters, tendo em conta o algoritmo apresentado por (Tan, Steinbach, & Kumar, 2005) o primeiro passo é decidir o número mínimo de pontos que devem ser utilizados para criar um cluster. Uma

vez que a plataforma CityClusters permite a introdução de 3 níveis de *zoom*, cada nível irá ter um número de pontos diferentes.

Para o nível de *zoom* mais próximo, foi considerado que o número mínimo de POIs para gerar um cluster deveria ser 10, para o nível de *zoom* intermédio consideram-se 20 pontos, para um zoom mais geral, serão considerados 40 pontos. A ferramenta desenvolvida permite calcular a distância o N Vizinho mais próximo de cada ponto, após efetuar este cálculo, analisam-se os dados, com apoio de um gráfico para inferir um valor de épsilon.

No caso de Singapura, apresentado pela Figura 13, Figura 14 e Figura 15, ao verificar a curvatura de distâncias de vizinhos 10, 20 e 40 foram inferidos os valores de épsilon 0.00138, 0.00232 e 0.00333 respetivamente, onde todos os pontos à esquerda deste valor, serão pontos centrais de *clusters*, todos os valores à direita serão pontos alcançáveis ou *outliers*.

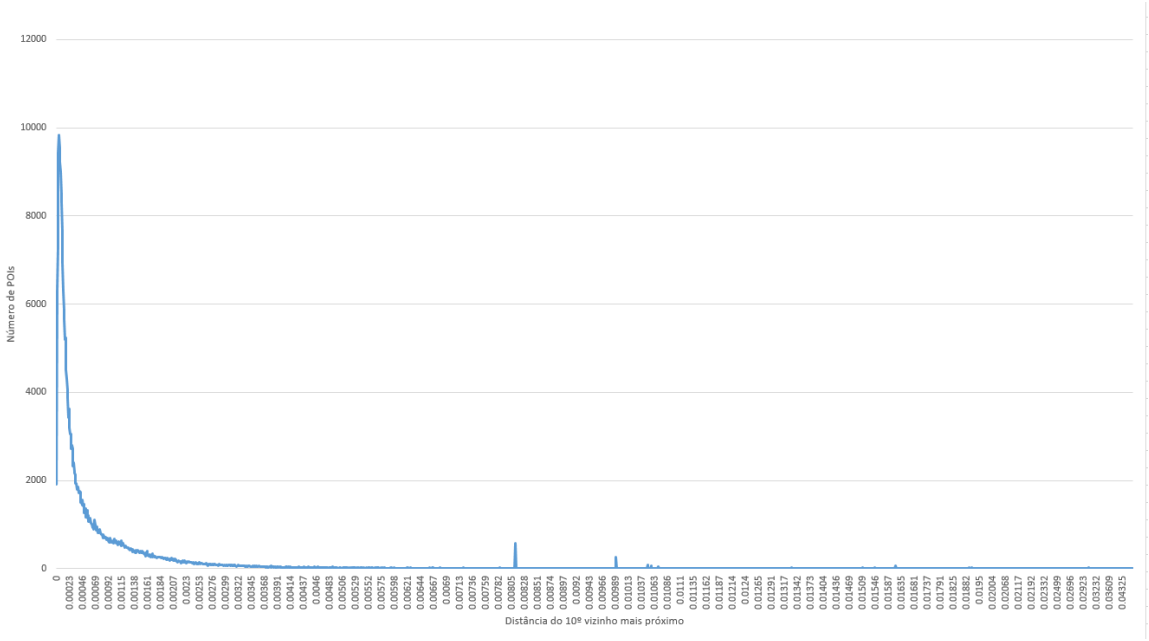


Figura 13 - Gráfico do número de POIs à distância do 10º vizinho mais próximo em Singapura

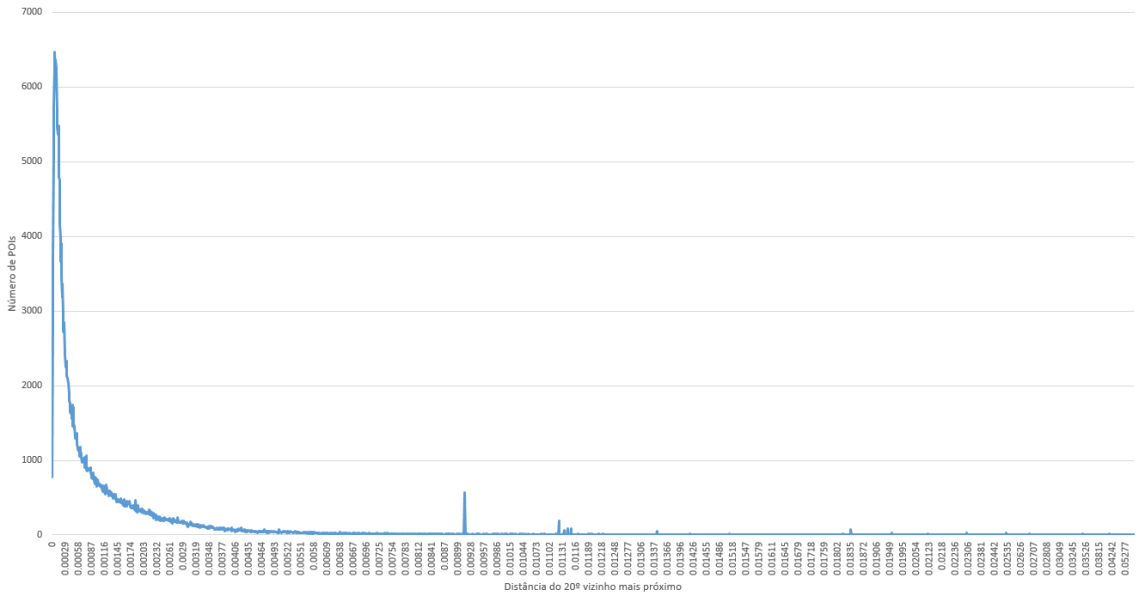


Figura 14 - Gráfico do número de POIs à distância do 20º vizinho mais próximo em Singapura

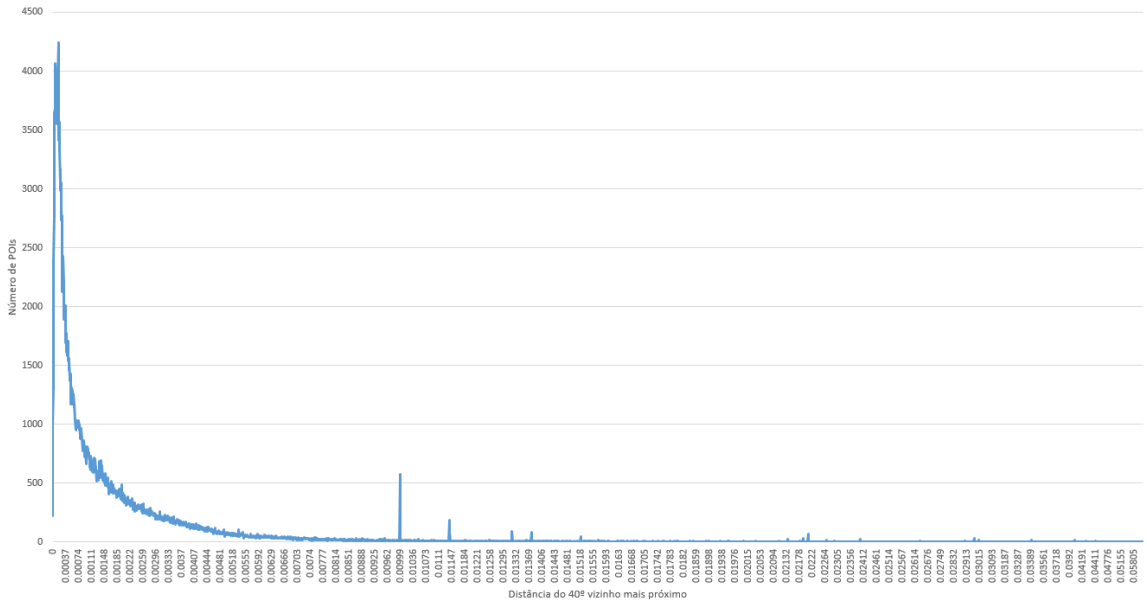


Figura 15 - Gráfico do número de POIs à distância do 40º vizinho mais próximo em Singapura

Para a região de Lisboa foi efetuado o mesmo teste e assim foram gerados os gráficos presentes nas figuras Figura 16, Figura 17 e Figura 18. Após a análise dos dados definiram-se os valores de épsilon 0.0036, 0.00474 e 0.00776 para 10, 20 e 40 vizinhos respetivamente.

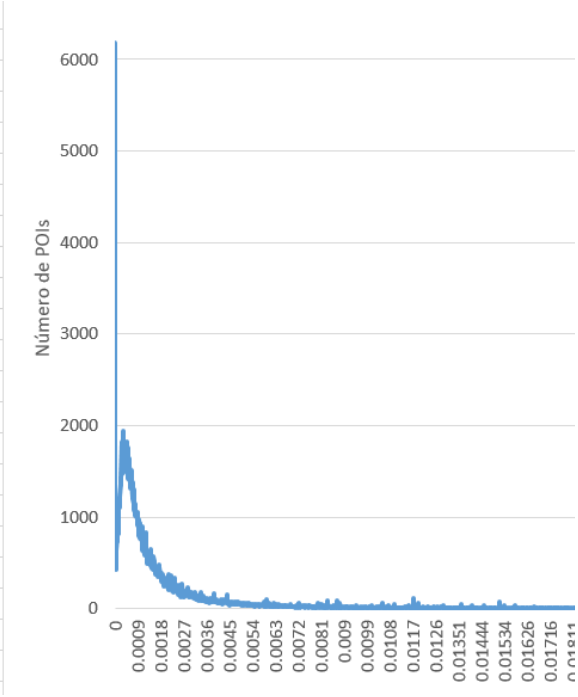


Figura 16 - Gráfico do número de POIs à distância do 10º vizinho mais próximo

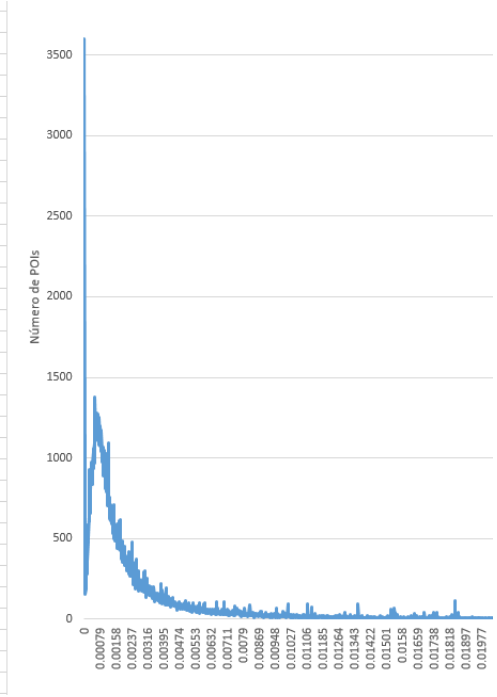


Figura 17 - Gráfico do número de POIs à distância do 20º vizinho mais próximo

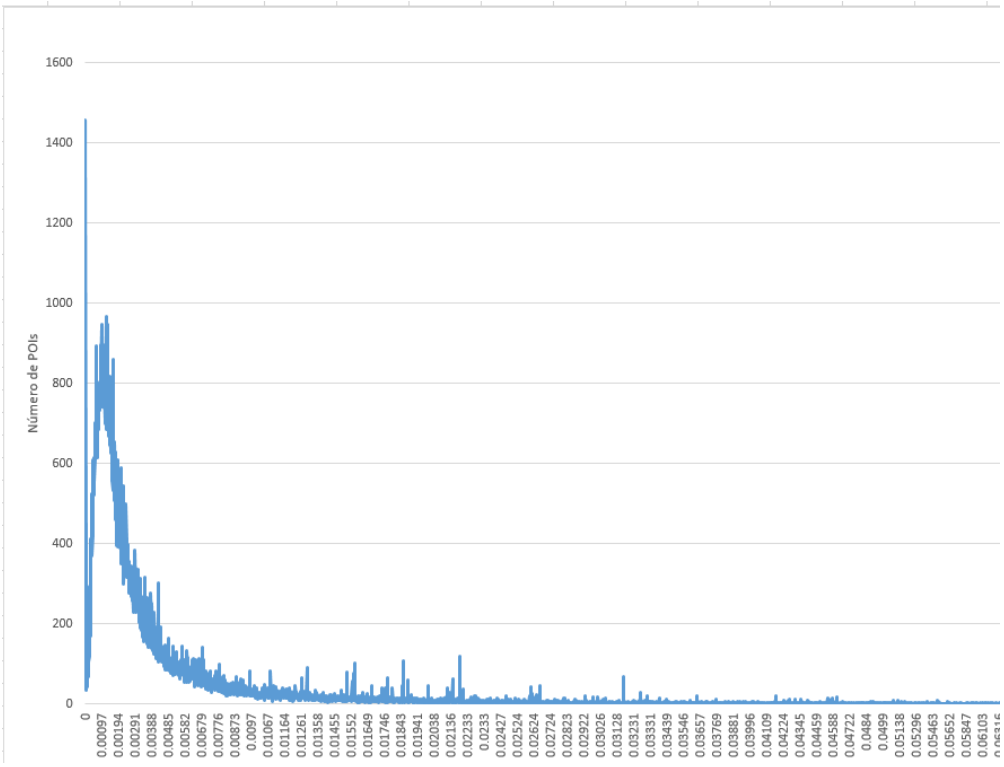


Figura 18 - Gráfico do número de POIs à distância do 40º vizinho mais próximo

Para a cidade de Nova Iorque, através da Figura 19, Figura 20, Figura 21, consegue se visualizar uma maior concentração dos registos gerando valores de épsilon 0.00108, 0.00141 e 0.00228 para a vizinhança correspondente.

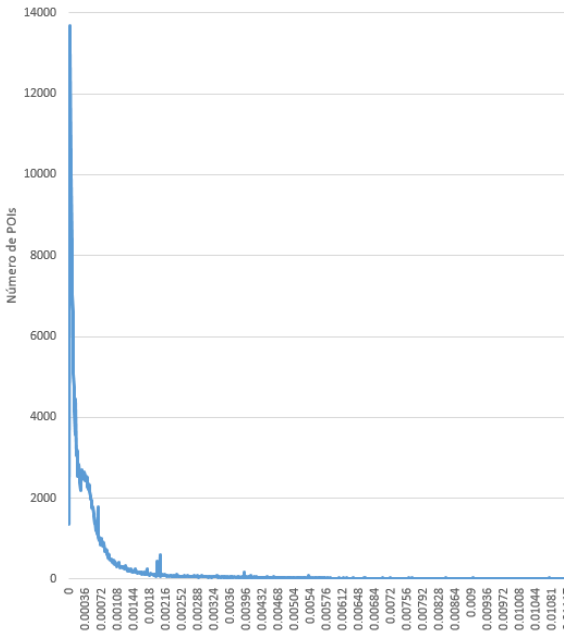


Figura 19 - Gráfico do número de POIs à distância do 10º vizinho mais próximo

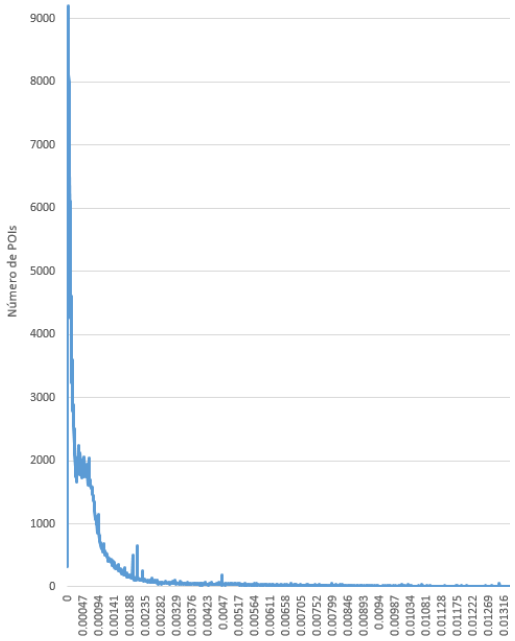


Figura 20 - Gráfico do número de POIs à distância do 20º vizinho mais próximo

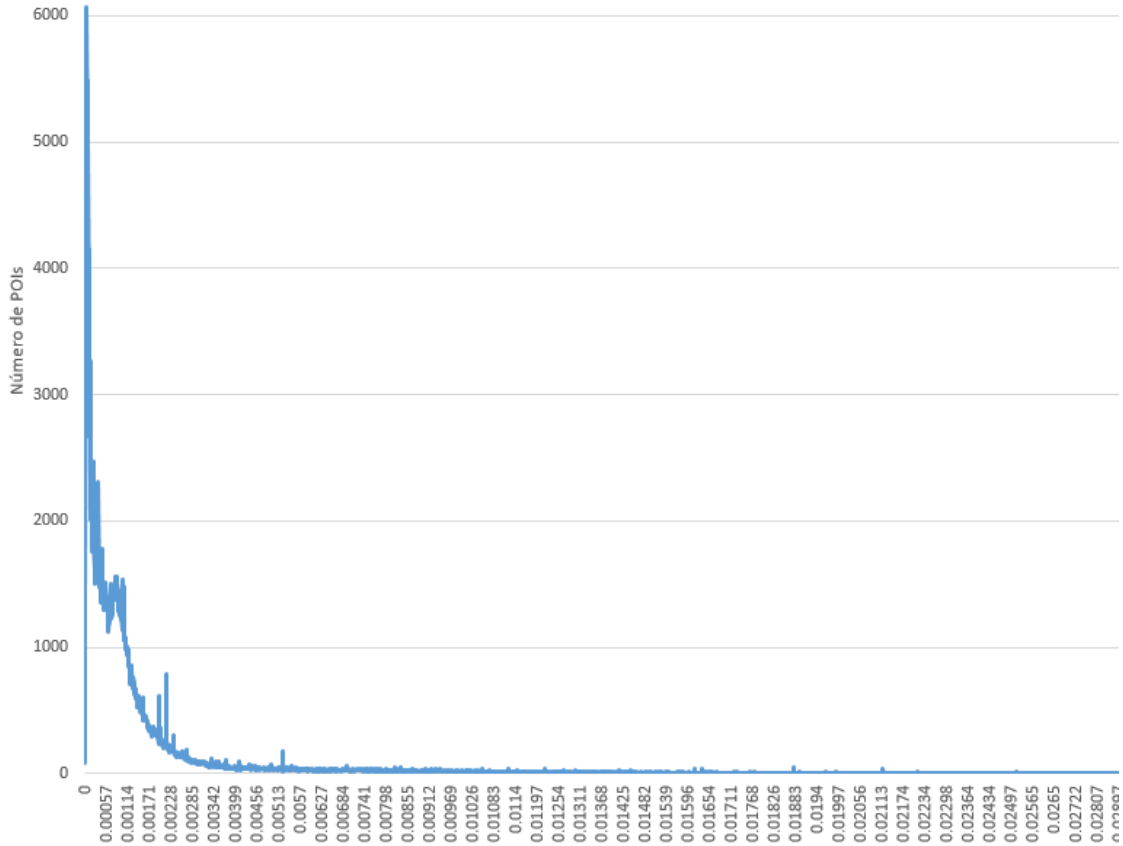


Figura 21 - Gráfico do número de POIs à distância do 40º vizinho mais próximo

Como o épsilon é um valor em ângulo, para melhor compreender os dados, efetuou-se uma conversão dos graus decimais para metros, tendo em conta que, em diferentes latitudes pode originar diferentes distâncias entre dois pontos. Esta diferença depende da curvatura da Terra, apresentada na Figura 22, por exemplo, na zona da linha do equador, dois pontos com a mesma diferença em graus estarão mais distantes entre si, pois trata-se de uma zona mais plana, do que nos trópicos, com uma curvatura mais acentuada.

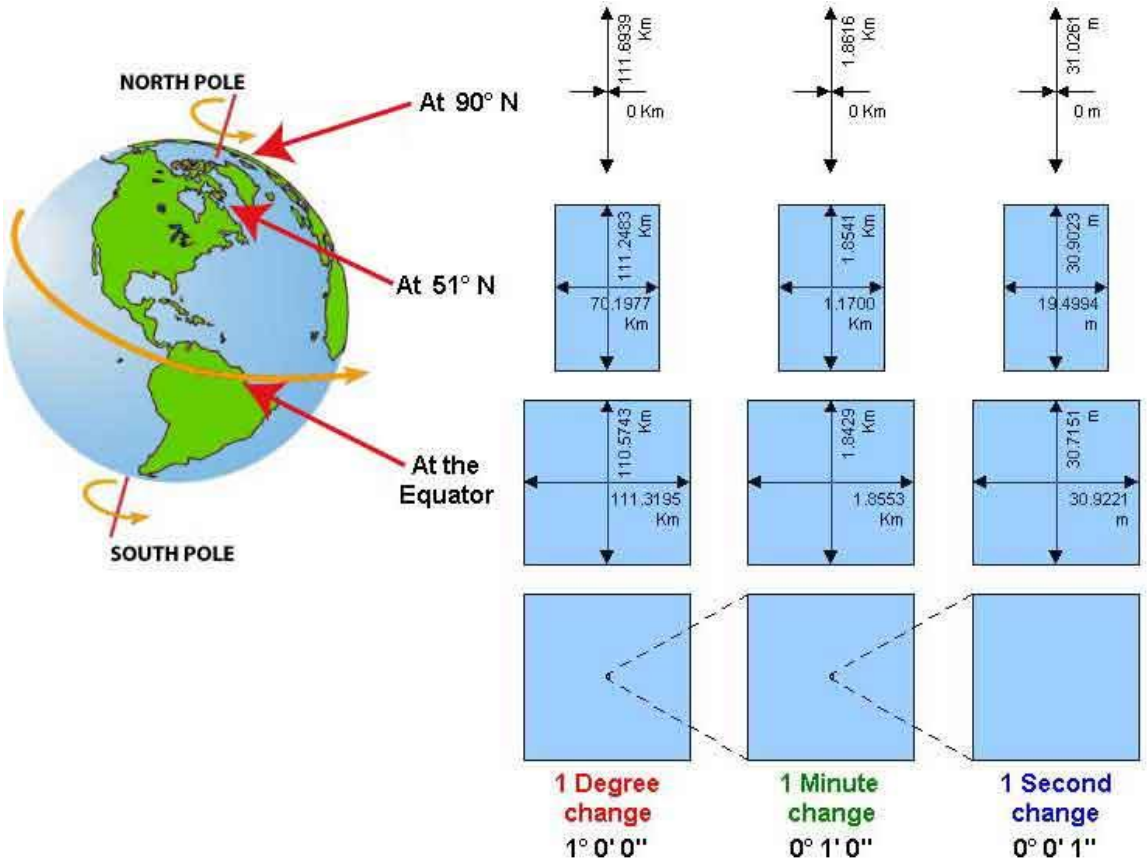


Figura 22 - Diferenças de distância de acordo com a curvatura da Terra[42]

A Tabela 22 apresenta os vários exemplos para criação dos POIs utilizando os dados inferidos através dos gráficos gerados pelo enésimo vizinho mais próximo.

Tabela 22 - Valores sugeridos criação de *clusters*

	Latitude	***Nível de Zoom***	***Raio (metros)***	***Épsilon***	***Número mínimo de POIs***
Singapura	1.3º	1	~788	0.00333	40
		2	~549	0.00232	20
		3	~326	0.00138	10
Lisboa	38.7º	1	~768	0.00776	40
		2	~469	0.00474	20
		3	~356	0.0036	10
Nova Iorque	40.7º	1	~222	0.00228	40
		2	~138	0.00141	20
		3	~106	0.00108	10

[42] http://www.longitudestore.com/how-big-is-one-gps-degree.html (acedido em 2016/12/12)

Utilizando os valores referidos na Tabela 22, iniciou-se a criação dos clusters vetoriais com o DBSCAN, a ferramenta desenvolvida recebe as variáveis através do ficheiro de configuração e inicia o processo de criação de clusters com base nos dados recolhidos previamente do Factual.

```
#[city_clusters]
#epsylon type can be [angle|kilometer|mile] (kilometer and mile are not precise values)
cc_epsilon_type=angle

cc_zoom1_minpoints=15
cc_zoom1_epsilon=0.051259099

cc_zoom2_minpoints=6
cc_zoom2_epsilon=0.01025182

cc_zoom3_minpoints=4
cc_zoom3_epsilon=0.00512591
```

No final foram gerados 5,599 *clusters* distribuídos da forma indicada na Tabela 23 com os vários níveis de zoom.

Tabela 23 - Resultado da criação de Clusters

	Nível de Zoom	***Tempo de processamento***	***Número de Clusters***
Singapura	1	52 minutos	662
	2	47 minutos	1344
	3	54 minutos	3028
Lisboa	1	38 minutos	11
	2	37 minutos	24
	3	32 minutos	35
Nova Iorque	1	44 minutos	49
	2	48 minutos	112
	3	41 minutos	334

De seguida deverá ser executada a ferramenta de geração dos polígonos para conseguir visualizar na plataforma CityClusters[43], esta criou os ficheiros necessários para visualizar os *clusters* através da página web.

Na Figura 23 é possível verificar os clusters do primeiro nível de zoom gerados para Singapura através da plataforma CityClusters, aqui só são visíveis os clusters com mais de 40 POIs, já na Figura 24 foi selecionado um cluster e é apresentada a informação sobre esse cluster, assim como todos os clusters que não sejam da mesma categoria são ocultados, facilitando a análise de planeamento urbano quando se trata de analisar sobre uma determinada categoria de dados.

[43] http://ubiquo.dei.uc.pt/cclusters/ (acedido em 2016/12/11)

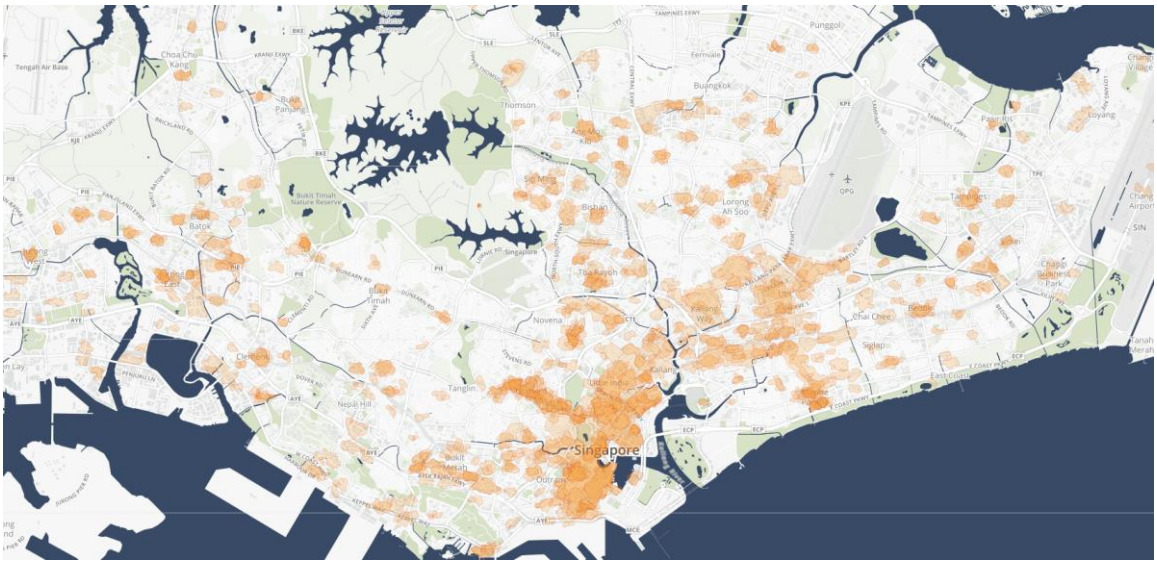


Figura 23 - *Clusters* de Singapura representados na plataforma CityClusters

Figura 24 - *Clusters* do tipo "Serviços de Medicina" na plataforma CityClusters

Uma vez que nova Iorque é o caso de estudo com mais registos no Crosswalk, que interligam as duas fontes, esta foi a opção utilizada para gerar os modelos no Weka com os classificadores binários disponíveis.

Utilizando os vários algoritmos de comparação de textos, foi criado um *dataset* para utilizar no Weka, para tal criou-se um caso de teste com 1713 registos, pois existem vários registos do Crosswalk que não contém relação com os registos existentes no DBpedia,

Motivos pelos quais o factual não ligou com DBpedia:

- Endereço incorreto
- Artigo do DBpedia não foi classificado como pertencendo a Nova Iorque
- Artigo não existe no DBpedia

Apos executar um conjunto de classificadores disponíveis, foram obtidos os resultados presentes na Tabela 24

Tabela 24 - Testes de vários classificadores no Weka

Modelo	***Número de Registos classificados corretamente***	***% de Sucesso***
NaiveBayesSimple	1451	84.7052
NaiveBayes	1452	84.7636
BayesNet	1465	85.5225
KStar	1663	97.0811
Bagging	1524	88.9667
Dagging	1468	85.6976
AdaBoostM1	1466	85.5809
ADTree	1481	86.4565
BFTree	1469	85.756
DecisionStump	1455	84.9387
FT	1492	87.0987
J48	1475	86.1062
J48graft	1475	86.1062
LADTree	1483	86.5733
LMT	1518	88.6165
NBTree	1468	85.6976
RandomForest	1660	96.906
RandomTree	1649	96.2639
REPTree	1490	86.9819
SimpleCart	1529	89.2586
ZeroR	1035	60.4203

Após verificar os dados, podem-se utilizar os modelos criados para avaliar novos registos tendo em conta os classificadores com melhor percentagem de sucesso. O classificador que apresentou melhor resultado foi o KStar com uma taxa de sucesso de 97%

3.4 Discussão

Na fase de criação de clusters, utilizando os dados recolhidos, é necessário definir previamente um conjunto de variáveis que serão usadas, o número mínimo de pontos do cluster e o épsilon, estas obrigam a um grande trabalho manual e requerem bastante processamento antes de iniciar a criação de clusters, tornando-se um processo algo moroso.

Também a criação de clusters e exportação para a plataforma CityClusters requer imenso processamento de dados.

Em áreas com uma densidade de POIs bastante elevada seria interessante distribuir os clusters por outros grupos de categorias em vez de ser apenas a categoria base.

Na fase de relação entre fontes de dados, como existem imensos registos em cada fonte torna-se complicado indicar dois registos que coincidam antes de passar pelo classificador, podendo ser criados novos casos de teste onde nenhum dos registos seja relacionado.

Devido à falta de classificação da Wikipedia com a taxonomia numa hierarquia conhecida seria interessante conseguir utilizar a informação gerada para associar categorias do Factual ou a um sistema internacional, como NAICS ou SIC, às categorias do Wikipedia, reclassificando todo este sistema.

Com esta associação feita de forma não supervisionada seria um avanço importante para futuros estudos e para classificar as várias categorias do Wikipedia num sistema devidamente estruturado.

4 Conclusões

No decorrer deste projeto foram analisadas várias fontes de dados disponíveis online, das fontes estudadas foram definidas para este estudo o Factual e DBpedia. Foi desenvolvida uma ferramenta capaz de efetuar a recolha e análise de dados de acordo com um país, região ou cidade. Esta recolha revelou-se bem-sucedida pois foi possível obter grande parte dos registos de forma não supervisionada e automática.

Quando se trata de planeamento urbano, as tarefas necessárias para tratar a informação são extremamente complicadas de fazer quando não existe conhecimento aprofundado do meio envolvente, sendo importante criar ferramentas que apoiem este estudo.

Desta forma o método abordado para criação de *clusters*, que melhora o DBSCAN com a capacidade de os distribuir por categorias, é bastante útil quando se necessita de efetuar trabalho na área de planeamento urbano, pois com essas categorias é possível inferir se um determinado tipo de serviço existe em demasia ou em falta numa determinada área de forma simplificada ao contrário dos algoritmos comuns que apenas indicam se existe ou não uma quantidade considerável de serviços naquela área.

Para melhorar este algoritmo, seria interessante efetuar testes com vários níveis de categorias base, para limitar mais os clusters gerados.

Com os dados recolhidos de ambas as fontes e utilizando um conjunto de métodos de comparação de *strings* recorreu-se ao Weka para gerar um modelo de caracterização binária que conseguiu efetuar o match entre as duas fontes com uma precisão de aproximadamente 97%.

5 Referências

Alves, A. O., Rodrigues, F., & Pereira, F. C. (2011). Tagging Space from Information Extraction and Popularity of Points of Interest. *Ambient Intelligence,* 115-125.

Braga, M., Santos, M. Y., & Moreira, A. (2014). Integrating Public Transportation Data: Creation and Editing of GTFS Data. *Advances in Intelligent Systems and Computing*, 53-62.

Canneyt, S. V., Schockaert, S., Laere, O. V., & Dhoedt, B. (2011). Time-dependent Recommendation of Tourist Attractions Using Flickr. *BNAIC : Belgian/Netherlands Artificial Intelligence Conference.*

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A Comparsion of String Distance Metrics for Name-Matching Tasks. 73-78.

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, 59–67.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 226-231.

Feldman, D., Sugaya, A., Sung, C. R., & Rus, D. (2013). iDiary: From GPS Signals to a Text-Searchable Diary. *ACM Transactions on Sensor Networks*, 11(4).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 100-108.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., . . . Bizer, C. (2012). DBpedia – A Large-scale, Multilingual. *IOS Press and the author*.

Mennis, J., & Guo, D. (2010). Spatial data mining and geographic knowledge discovery - An introduction. *Computers, Environment and Urban Systems*, 175.

Milne, D. N., & Witten, I. (2008). Learning to Link with Wikipedia.

MySQL SOUNDEX() function. (09 de 11 de 2016). Obtido de w3resource: http://www.w3resource.com/mysql/string-functions/mysql-soundex-function.php

Patankar, B., & Chavda, D. V. (2015). An Experiment to Improve Classification Accuracy Using Ensemble Methods. *International Journal of Scientific Research in Science and Technology*, 94-97.

Polisciuc, E., Alves, A. O., & Machado, P. (2015). Understanding Urban Land Use through the Visualization of Points of Interest. *Proceedings of the Fourth Workshop on Vision and Language.*

Rodrigues, F. (2010). *POI Mining and Generation.*

Rodrigues, F. (2010). *POI Mining and Generation.* Coimbra.

Rodrigues, J. G., Aguiar, A., & Barros, J. (2014). SenseMyCity: Crowdsourcing an Urban Sensor.

Sakai, T., Tamura, K., & Kitakami, H. (2014). Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial Clustering Algorithm. *IAENG International Journal of Computer Science*, 41(3):185-192.

Shi, J., Mamoulis, N., Wu, D., & Cheung, D. W. (2014). Density-based place clustering in geo-social networks. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (pp. 99-110). Snowbird, Utah, USA: ACM New York, NY, USA.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining . Addison-Wesley Longman Publishing Co., Inc.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage., (pp. 354-359).

6 Anexos

6.1 Anexo 1 – Taxonomia do Factual

De seguida apresenta-se a taxonomia do Factual pela qual os POIs são categorizados quanto ao seu tipo principal de serviço,

Foi descarregada utilizando a API do Factual pelo *endpoint* Categories em Fevereiro de 2016.

ID	*Parent*	*Descrição EN*	*Descrição PT*	*Base*
1	0	Factual Places	Lugares Factual	1
2	1	Automotive	Automóvel	2
3	2	Car Appraisers	Peritos de Automóveis	2
4	2	Car Dealers and Leasing	Aluguer e Leasing de Automóveis	2
5	4	Used Cars	Carros Usados	2
6	2	Car Parts and Accessories	Partes de Automóveis e Acessórios	2
7	2	Car Wash and Detail	Lavagem e Limpeza Automóvel	2
8	2	Classic and Antique Car	Carros Clássicos e Antigos	2
9	2	Maintenance and Repair	Manutenção e Reparação	2
10	9	Oil and Lube	Óleo e Lubrificante	2
11	9	Smog Check	Controlo de Poluição	2
12	9	Tires	Pneus	2
13	9	Transmissions	Transmissões	2
14	2	Motorcycles, Mopeds and Scooters	Motociclos, Ciclomotores e Velocípedes	2
15	14	Repair	Reparação	2
16	14	Sales	Vendas	2
17	2	RVs and Motor Homes	Roulottes e Autocaravanas	2
18	2	Salvage Yards	Ferro-velho	2
19	2	Towing	Reboque	2
20	1	Community and Government	Comunidade e Governo	20
21	20	Animal Shelters and Humane Societies	Abrigos de Animais e Associações Humanitárias	20
22	20	Cemeteries	Cemitérios	20
23	20	Day Care and Preschools	Creches e Jardins de Infância	20
24	20	Disabled Persons Services	Serviços de Apoio a Pessoas com Necessidades Especiais	20
25	20	Drug and Alcohol Services	Serviços de Tratamento de Alcoolismo e Toxicodependência	20
26	20	Education	Ensino	20
27	26	Adult Education	Ensino para Adultos	20

28	26	Art Lessons and Schools	Escolas de Arte	20
29	26	Colleges and Universities	Faculdades, Universidades e Institutos Superiores	20
30	26	Computer Training	Formação em Informática	20
31	26	Culinary Lessons and Schools	Escolas de Culinária	20
32	26	Driving Schools	Escolas de Condução	20
33	26	Fraternities and Sororities	Repúblicas	20
34	26	Primary and Secondary Schools	Escolas Primárias e Secundárias	20
35	26	Tutoring and Educational Services	Cursos Particulares e Serviços Educativos	20
36	26	Vocational Schools	Escolas Profissionais	20
37	20	Government Departments and Agencies	Ministérios e Organismos Governamentais	20
38	20	Government Lobbyists	Lobbyistas Governamentais	20
39	20	Housing Assistance and Shelters	Assistência Habitacional e Abrigos	20
40	20	Law Enforcement and Public Safety	Forças de Segurança e Ordem Pública	20
41	40	Rescue Services	Serviços de Socorro	20
42	40	Fire Stations	Quarteis de Bombeiros	20
43	40	Police Stations	Esquadras de Polícia	20
44	20	Libraries	Bibliotecas	20
45	20	Military	Forças Armadas	20
46	45	Bases	Bases Militares	20
47	20	Organizations and Associations	Organizações e Associações	20
48	47	Charities and Non-Profits	Organizações de Beneficiência e sem Fim Lucrativo	20
49	47	Environmental	Ambiente	20
50	47	Youth Organizations	Organizações Juvenis	20
51	20	Post Offices	Postos de Correio	20
52	20	Public and Social Services	Serviços Públicos e Sociais	20
53	20	Religious	Religião	20
54	53	Buddhist Temples	Templos Budistas	20
55	53	Churches	Igrejas	20
56	53	Hindu Temples	Templos Hindus	20
57	53	Mosques	Mesquitas	20
58	53	Synagogues	Sinagogas	20
59	20	Senior Citizen Services	Serviços de Apoio a Pessoas da Terceira Idade	20
60	59	Retirement	Reforma	20
61	20	Utility Companies	Empresas de Serviços Públicos	20
62	1	Healthcare	Assistência Médica e Sanitária	62
63	62	AIDS Resources	Recursos relacionados com SIDA	62
64	62	Assisted Living Services	Serviços de Assistência Pessoal	62
65	62	Home Health Care Services	Início Serviços de Saúde	62
66	64	Facilities and Nursing Homes	Lares e Casas de Repouso	62
67	62	Blood Banks and Centers	Centros de Recolha e Bancos de Sangue	62
68	62	Chiropractors	Quiropráticos	62
69	62	Dentists	Dentistas	62
70	62	Emergency Services	Serviços de Emergência	62
71	70	Ambulance	Ambulância	62

72	62	Holistic, Alternative and Naturopathic Medicine	Medicinas Alternativas e Naturais	62
73	72	Acupuncture	Acupunctura	62
74	62	Hospitals, Clinics and Medical Centers	Hospitais, Clínicas e Centros Médicos	62
75	62	Medical Supplies and Labs	Material Médico e Laboratórios	62
76	62	Mental Health	Saúde Mental	62
77	76	Counseling and Therapy	Aconselhamento e Terapia	62
78	76	Psychologists	Psicólogos	62
79	62	Nurses	Enfermeiros	62
80	62	Pharmacies	Farmácias	62
81	62	Physical Therapy and Rehabilitation	Terapia Física e Reabilitação	62
82	81	Sports Medicine	Medicina Desportiva	62
83	62	Physicians	Médicos	62
84	83	Anesthesiologists	Anestesistas	62
85	83	Cardiologists	Cardiologistas	62
86	83	Dermatologists	Dermatologistas	62
87	83	Ear, Nose and Throat	Otorrinolaringologistas	62
88	83	Family Medicine	Médicos de Família	62
89	83	Gastroenterologists	Gastrenterologistas	62
90	83	General Surgery	Cirurgia Geral	62
91	83	Internal Medicine	Medicina Interna	62
92	83	Neurologists	Neurologistas	62
93	83	Obstetricians and Gynecologists	Obstetras e Ginecologistas	62
94	83	Oncologists	Oncologistas	62
95	83	Ophthalmologists	Oftalmologistas	62
96	83	Orthopedic Surgeons	Cirurgiões Ortopédicos	62
97	83	Pathologists	Patologistas	62
98	83	Pediatricians	Pediatras	62
99	83	Plastic Surgeons	Cirurgiões Plásticos	62
100	83	Psychiatrists	Psiquiatras	62
101	83	Radiologists	Radiologistas	62
102	83	Respiratory	Pneumologia	62
103	83	Urologists	Urologistas	62
104	62	Podiatrists	Podólogos	62
105	62	Pregnancy and Sexual Health	Gravidez e Saúde Sexual	62
106	62	Weight Loss and Nutritionists	Perda de Peso e Nutricionistas	62
107	1	Landmarks	Marcos	107
108	107	Buildings and Structures	Edifícios e Estruturas	107
109	107	Gardens	Jardins	107
110	107	Historic and Protected Sites	Lugares Históricos e Protegidos	107
111	107	Monuments and Memorials	Monumentos Históricos e Comemorativos	107
112	107	Natural	Natureza	107
113	112	Beaches	Praias	107

114	112	Mountains	Montanhas	107
115	112	Forests	Florestas	107
116	112	Lakes	Lagos	107
117	112	Rivers	Rios	107
118	107	Parks	Parques	107
119	118	Natural Parks	Parques Naturais	107
120	118	Picnic Areas	Parques de Merendas	107
121	118	Playgrounds	Parques Infantis	107
122	118	Urban Parks	Parques Urbanos	107
123	1	Retail	Venda a Retalho	123
124	123	Adult	Adulto	123
125	123	Antiques	Antiguidades	123
126	123	Arts and Crafts	Artesanato	123
127	123	Auctions	Leilões	123
128	123	Beauty Products	Produtos de Beleza	123
129	123	Bicycles	Bicicletas	123
130	123	Bookstores	Livrarias	123
131	123	Cards and Stationery	Artigos de Papelaria	123
132	123	Children	Crianças	123
133	123	Computers and Electronics	Informática e Electrónica	123
134	133	Cameras	Câmeras	123
135	133	Mobile Phones	Telemóveis	123
136	133	Video Games	Jogos de Vídeo	123
137	123	Construction Supplies	Material de Construção	123
138	123	Convenience Stores	Drogarias	123
139	123	Costumes	Máscaras e Disfarces	123
140	123	Dance and Music	Música e Dança	123
141	123	Department Stores	Lojas de Departamento	123
142	123	Fashion	Moda	123
143	142	Clothing and Accessories	Roupa e Acessórios	123
144	142	Jewelry and Watches	Jóias e Relógios	123
145	142	Shoes	Sapatos	123
146	142	Swimwear	Fatos de Banho	123
147	123	Flea Markets	Mercados	123
148	123	Florists	Floristas	123
149	123	Food and Beverage	Comida e Bebida	123
150	149	Beer, Wine and Spirits	Cerveja, Vinho e Licores	123
151	149	Candy Stores	Confeitarias	123
152	149	Cheese	Queijo	123
153	149	Chocolate	Chocolate	123
154	149	Farmers' Markets	Mercados ao Ar Livre	123
155	149	Health and Diet Food	Comida Saudável e Dietética	123
156	149	Kosher	Kosher	123
157	123	Furniture and Decor	Mobiliário e Decoração	123
158	123	Gift and Novelty	Presentes e Lembranças	123
159	123	Glasses	Óptica	123
160	123	Hobby and Collectibles	Coleccionismo e Passatempos	123
161	123	Luggage	Bagagem e Malas	123
162	123	Music, Video and DVD	Música, Vídeo e DVD	123

163	123	Newsstands	Quiosques	123
164	123	Nurseries and Garden Centers	Viveiros e Centros de Jardinagem	123
165	123	Outlet	Outlet	123
166	123	Pawn Shops	Lojas de Penhores	123
167	123	Pets	Animais Domésticos	123
168	123	Photos and Frames	Fotografias e Molduras	123
169	123	Shopping Centers and Malls	Centros Comerciais	123
170	123	Sporting Goods	Artigos Desportivos	123
171	123	Supermarkets and Groceries	Supermercados e Mercearias	123
172	123	Tobacco	Tabaco	123
173	123	Toys	Brinquedos	123
174	123	Vintage and Thrift	Clássicos e em Segunda Mão	123
175	123	Warehouses and Wholesale Stores	Armazéns e Feiras Grossistas	123
176	123	Wedding and Bridal	Casamentos e Acessórios para a Noiva	123
177	1	Businesses and Services	Empresas e Serviços	177
178	177	Business and Strategy Consulting	Consultoria em Negócios e Estratégia	177
179	177	Industrial Machinery and Vehicles	Máquinas e Veículos Industriais	177
180	177	Logging and Sawmills	Madeireiros e Serrações	177
181	177	Metals	Metalurgia	177
182	177	Packaging	Embalamento e Empacotamento	177
183	177	Petroleum	Petróleo	177
184	177	Plastics	Plásticos	177
185	177	Refrigeration and Ice	Gelo e Refrigeração	177
186	177	Rubber	Borracha	177
187	177	Scientific	Científico	177
188	177	Security and Safety	Segurança	177
189	177	Telecommunication Services	Serviços de Telecomunicações	177
190	177	Textiles	Têxteis	177
191	177	Water and Waste Management	Gestão de Água e de Resíduos	177
192	177	Welding	Soldagem	177
193	177	Advertising and Marketing	Publicidade e Marketing	177
194	193	Advertising Agencies and Media Buyers	Agências de Publicidade e Compradores de Meios de Comunicação	177
195	193	Creative Services	Serviços Criativos	177
196	193	Direct Mail and Email Marketing Services	Correio e Serviços de Marketing Postal	177
197	193	Market Research and Consulting	Estudo e Consultoria de Mercado	177
198	193	Online Advertising	Publicidade Online	177
199	193	Print, TV, Radio and Outdoor Advertising	Publicidade Impressa, para Televisão, Rádio e Outdoors	177
200	193	Promotional Items	Artigos Promocionais	177
201	193	Public Relations	Relações Públicas	177

202	193	Search Engine Marketing and Optimization	Optimização e Marketing para Motores de Pesquisa	177
203	193	Writing, Copywriting and Technical Writing	Escrita, Redacção e Redacção Técnica	177
204	177	Agriculture and Forestry	Agricultura e Silvicultura	177
205	177	Art Restoration	Restauro de Arte	177
206	177	Audiovisual	Audiovisual	177
207	177	Automation and Control Systems	Automação e Controlo de Sistemas	177
208	177	Chemicals and Gasses	Químicos e Gases	177
209	177	Computers	Informática	177
210	177	Corporate HQ	Sede de Empresa	177
211	177	Electrical Equipment	Equipamento Eléctrico	177
212	177	Employment Agencies	Centros de Emprego	177
213	177	Engineering	Engenharia	177
214	177	Entertainment	Entretenimento	177
215	214	Media	Meios de Comunicação	177
216	177	Equipment Rental	Aluguer de Equipamento	177
217	177	Events and Event Planning	Eventos e Planeamento de Eventos	177
218	221	ATMs	Caixas Automáticas Multibanco	177
219	177	Financial	Finanças	177
220	219	Accounting and Bookkeeping	Contabilidade	177
221	219	Banking and Finance	Bancos e Finanças	177
222	219	Business Brokers and Franchises	Corretores de Negócios e Franquia	177
223	219	Check Cashing	Depósito de Cheques	177
224	219	Collections	Colecções	177
225	219	Financial Planning and Investments	Planeamento Financeiro e Investimentos	177
226	219	Fund Raising	Angariação de Fundos	177
227	219	Loans and Mortgages	Empréstimos e Hipotecas	177
228	219	Stock Brokers	Corretores de Bolsa	177
229	219	Student Aid and Grants	Auxílio e Bolsas para Estudantes	177
230	177	Food and Beverage	Comida e Bebida	177
231	230	Catering	Catering	177
232	230	Distribution	Distribuição	177
233	177	Funeral Services	Agências Funerárias	177
234	177	Geological	Geologia	177
235	177	Home Improvement	Renovação Habitacional	177
236	235	Architects	Arquitectos	177
237	235	Carpenters	Carpinteiros	177
238	235	Carpet and Flooring	Tapetes e Pavimentos	177
239	235	Contractors	Empreiteiros	177
240	239	Bathrooms	Quartos de Banho	177
241	239	Deck and Patio	Deck e Pátio	177
242	239	Sewer	Esgoto	177
243	235	Doors and Windows	Portas e Janelas	177
244	235	Electricians	Electricistas	177
245	235	Fences, Fireplaces and Garage Doors	Vedações, Lareiras e Portões de Garagem	177
246	235	Hardware and Services	Ferramentas e Serviços	177

247	235	Heating, Ventilating and Air Conditioning	Aquecimento, Ventilação e Ar Condicionado	177
248	123	Home Appliances	Electrodomésticos	123
249	235	Home Inspection Services	Serviços de Inspecção de Edifícios	177
250	123	Housewares	Equipamento para Cozinha	123
251	235	Interior Design	Design de Interiores	177
252	235	Kitchens	Cozinhas	177
253	235	Landscaping and Gardeners	Jardinagem	177
254	235	Lighting Fixtures	Iluminação	177
255	291	Mobile Homes	Casas Móveis	177
256	235	Movers	Mudanças	177
257	235	Painting	Pintura	177
258	235	Pest Control	Controlo de Pragas	177
259	235	Plumbing	Canalização	177
260	235	Pools and Spas	Piscinas e Spas	177
261	235	Roofers	Conserto de Telhados	177
262	177	Storage	Armazenamento	177
263	235	Swimming Pool Maintenance and Services	Serviços e Manutenção de Piscinas	177
264	235	Tree Service	Serviço de Arborização	177
265	235	Upholstery	Estofamento	177
266	177	Human Resources	Recursos Humanos	177
267	177	Import and Export	Importação e Exportação	177
268	177	Leather	Couro	177
269	177	Legal	Jurídico	177
270	269	Credit Counseling and Bankruptcy Services	Aconselhamento de Crédito e Serviços de Falência	177
271	269	Immigration	Imigração	177
272	177	Insurance	Seguros	177
273	177	Machine Shops	Oficinas Mecânicas	177
274	177	Management	Administração	177
275	177	Manufacturing	Manufactura	177
276	177	Paper	Papel	177
277	177	Personal Care	Cuidados Pessoais	177
278	277	Dry Cleaning, Ironing and Laundry	Limpeza a seco, Engomadoria e Lavandaria	177
279	277	Hair Removal	Depilação	177
280	277	Beauty Salons and Barbers	Salões de Beleza e Barbeiros	177
281	277	Manicures and Pedicures	Manicure e Pedicure	177
282	277	Massage Clinics and Therapists	Clínicas de Massagem e Terapeutas	177
283	277	Piercing	Piercing	177
284	277	Skin Care	Cuidados com a Pele	177
285	277	Spas	Spas e Termas	177
286	277	Tanning Salons	Solários	177
287	277	Tattooing	Tatuagens	177

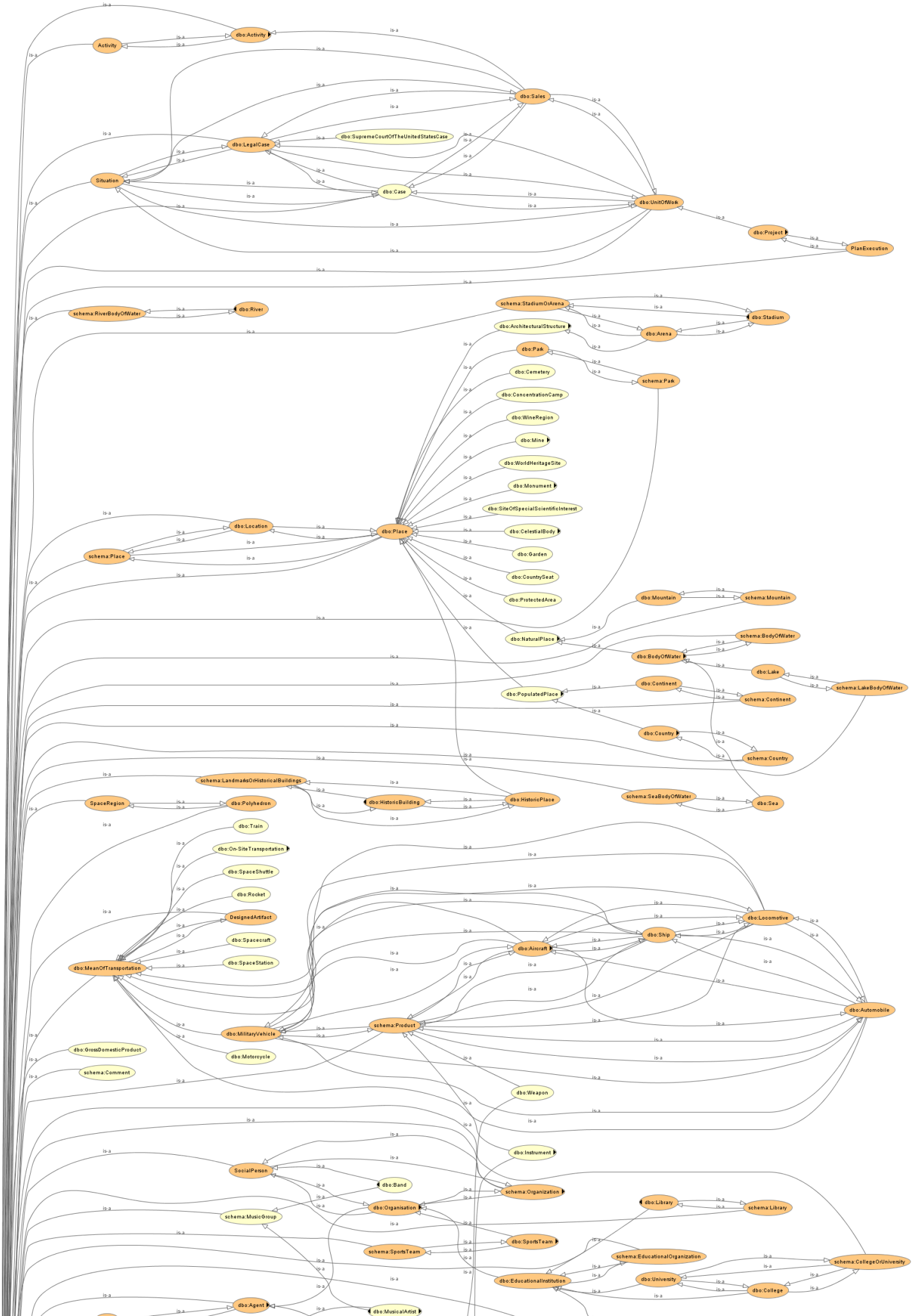
288	177	Printing, Copying and Signage	Impressão, Cópia e Sinalização	177
289	177	Professional Cleaning	Limpeza e Higiene Profissional	177
290	177	Publishing	Editoras	177
291	177	Real Estate	Ramo Imobiliário	177
292	291	Property Management	Gestão de Propriedades	177
293	291	Real Estate Agents	Agentes Imobiliários	177
294	291	Real Estate Appraiser	Avaliador Imobiliário	177
295	291	Real Estate Development and Title Companies	Empresas Imobiliárias e de Títulos	177
296	291	Apartments, Condos, and Houses	Apartamentos, Condomínios e Casas	177
297	291	Boarding Houses	Pensões	177
298	291	Building and Land Surveyors	Construção e Topógrafos	177
299	291	Commercial Real Estate	Comércio Imobiliário	177
300	291	Corporate Housing	Condomínio Empresarial	177
301	177	Renewable Energy	Energia Renovável	177
302	177	Repair Services	Serviços de Reparação	177
303	177	Shipping, Freight, and Material Transportation	Expedição e Transporte de Materiais	177
304	177	Tailors	Alfaiates	177
305	177	Veterinarians	Veterinários	177
306	460	Web Design and Development	Design e Desenvolvimento para a Web	177
307	177	Wholesale	Comércio por Grosso	177
308	1	Social	Social	308
309	308	Arts	Artes	308
310	309	Art Dealers and Galleries	Negociantes e Galerias de Arte	308
311	309	Museums	Museus	308
312	308	Bars	Bares	308
313	312	Hotel Lounges	Salões de Hotel	308
314	312	Jazz and Blues Cafes	Cafés de Jazz e Blues	308
315	312	Sports Bars	Bares Desportivos	308
316	312	Wine Bars	Bares de Vinho	308
317	308	Entertainment	Entretenimento	308
318	317	Adult Entertainment	Entretenimento para Adultos	308
319	317	Amusement Parks	Parques de Diversões	308
320	317	Billiard and Pool	Salas de Bilhar	308
321	317	Bingo	Bingo	308
322	317	Bowling	Bowling	308
323	317	Carnivals	Carnavais	308
324	317	Casinos and Gaming	Casinos e Jogo	308
325	317	Circuses	Circos	308
326	317	Dance Halls and Saloons	Salões de Dança	308
327	317	Fairgrounds and Rodeos	Feiras e Rodeios	308
328	317	Go Carts	Karting	308
329	317	Hookah Lounges	Salões de Narguilé	308
330	317	Karaoke	Karaoke	308
331	317	Miniature Golf	Mini-golfe	308
332	317	Movie Theatres	Cinemas	308
333	317	Music and Show Venues	Salas de Música e Espectáculos	308
334	317	Night Clubs	Clubes Nocturnos e Discotecas	308

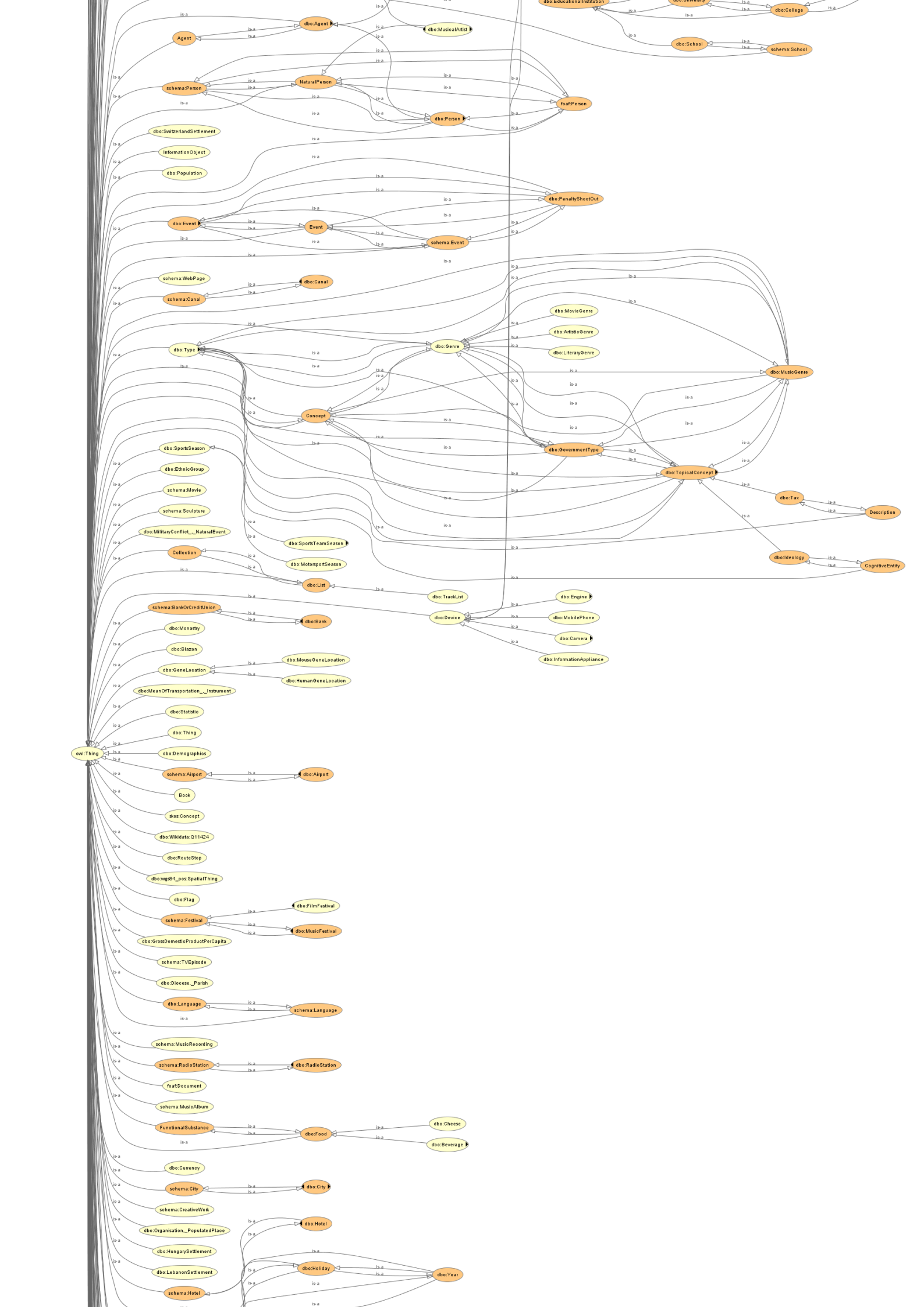
335	317	Party Centers	Centros de Festa	308
336	317	Psychics and Astrologers	Videntes e Astrólogos	308
337	317	Ticket Sales	Bilheteiras	308
338	308	Food and Dining	Restauração	308
339	338	Bagels and Donuts	Bagels e Donuts	308
340	338	Bakeries	Padarias e Pastelarias	308
341	338	Breweries	Cervejarias	308
342	338	Cafes, Coffee and Tea Houses	Cafés e Salas de Chá	308
343	338	Dessert	Sobremesas	308
344	338	Ice Cream Parlors	Gelatarias	308
345	338	Internet Cafes	Cibercafés	308
346	338	Juice Bars and Smoothies	Bares de Sumos e Batidos	308
347	338	Restaurants	Restaurantes	308
348	347	American	Americano	308
349	347	Barbecue	Barbecue	308
350	347	Buffets	Buffet	308
351	347	Burgers	Hambúrgueres	308
352	347	Chinese	Chinês	308
353	347	Delis	Casas de Produtos Gourmet	308
354	347	Diners	Cantinas	308
355	347	Fast Food	Fast Food	308
356	347	French	Francês	308
357	347	Indian	Indiano	308
358	347	Italian	Italiano	308
359	347	Japanese	Japonês	308
360	347	Korean	Coreano	308
361	347	Mexican	Mexicano	308
362	347	Middle Eastern	Do Médio Oriente	308
363	347	Pizza	Pizzarias	308
364	347	Seafood	Marisqueiras	308
365	347	Steakhouses	Churrascarias	308
366	347	Sushi	De Sushi	308
367	347	Thai	Tailandês	308
368	347	Vegan and Vegetarian	Vegan e Vegetarianos	308
369	308	Country Clubs	Clube Privado	308
370	308	Wineries and Vineyards	Adegas e Vinhas	308
371	308	Zoos, Aquariums and Wildlife Sanctuaries	Jardins Zoológicos, Aquários e Parques Biológicos	308
372	1	Sports and Recreation	Desportos e Lazer	372
373	372	Athletic Fields	Campos de Atletismo	372
374	372	Baseball	Basebol	372
375	374	Batting Ranges	Centros de Treino de Basebol	372
376	372	Basketball	Basquetebol	372
377	372	Combat Sports	Desportos de Combate	372

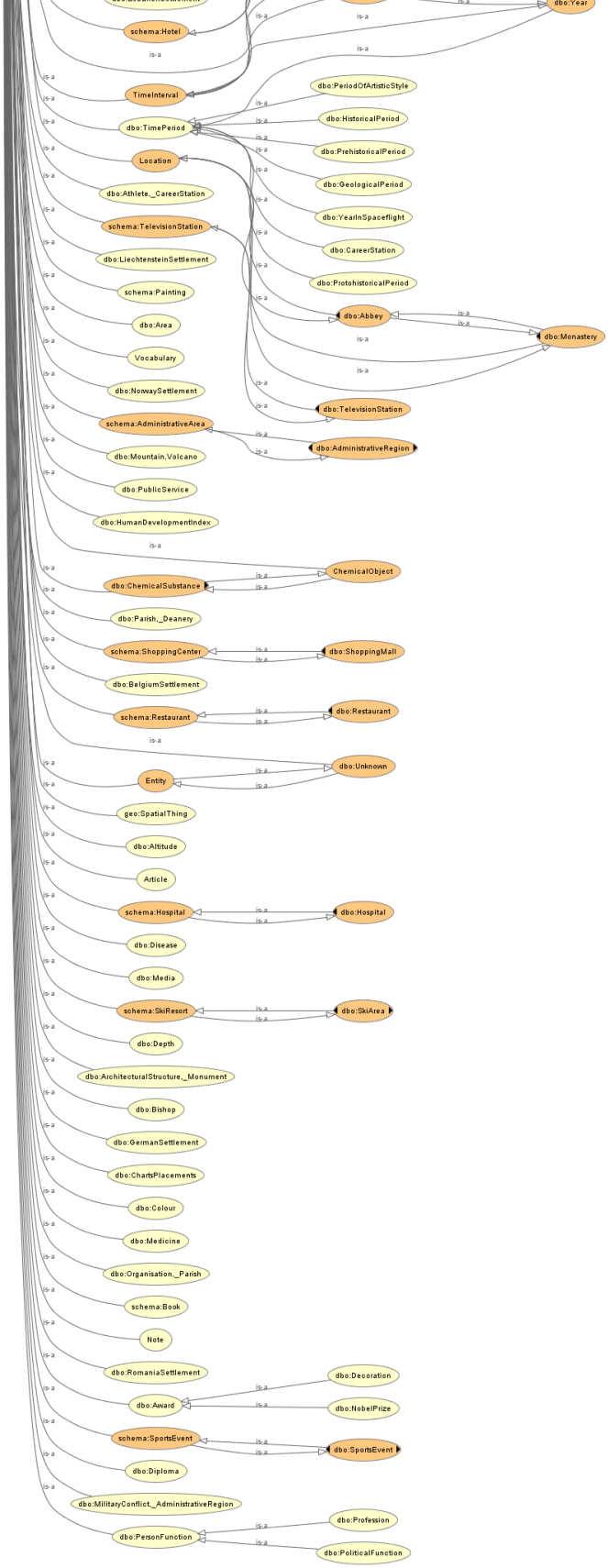
378	372	Cycling	Ciclismo	372
379	372	Dance	Dança	372
380	372	Equestrian	Equitação	372
381	372	Football	Futebol Americano	372
382	372	Golf	Golfe	372
383	372	Gun Ranges	Campos de Tiro	372
384	372	Gymnastics	Ginástica	372
385	372	Gyms and Fitness Centers	Ginásios e Centros de Fitness	372
386	372	Hockey	Hóquei	372
387	372	Outdoors	Ar Livre	372
388	387	Campgrounds and RV Parks	Parques de Campismo e de Caravanas	372
389	387	Hiking	Caminhadas	372
390	387	Hot Air Balloons	Balões de Ar Quente	372
391	387	Hunting and Fishing	Caça e Pesca	372
392	387	Rock Climbing	Escalada	372
393	387	Skydiving	Pára-quedismo	372
394	372	Paintball	Paintball	372
395	372	Personal Trainers	Treinadores Pessoais	372
396	372	Race Tracks	Pistas de Corrida	372
397	372	Racquet Sports	Desportos de Raquete	372
398	397	Racquetball	Raquetebol	372
399	397	Tennis	Ténis	372
400	372	Recreation Centers	Centros Recreativos	372
401	372	Running	Corrida	372
402	372	Skating	Patinagem	372
403	372	Snow Sports	Desportos de Inverno	372
404	372	Soccer	Futebol	372
405	372	Sports Clubs	Clubes Desportivos	372
406	372	Stadiums and Arenas	Estádios e Recintos Desportivos	372
407	372	Swimming Pools	Piscinas	372
408	372	Water Sports	Desportos Aquáticos	372
409	408	Boating	Passeios de Barco	372
410	408	Canoes and Kayaks	Canoas e Kayaks	372
411	408	Rafting	Rafting	372
412	408	Scuba Diving	Mergulho	372
413	408	Swimming	Natação	372
414	372	Yoga and Pilates	Yoga e Pilates	372
415	1	Transportation	Transporte	415
416	415	Airlines and Aviation Services	Companhias Aéreas e Serviços de Aviação	415
417	415	Gas Stations	Postos de Combustível	415
418	415	Parking	Estacionamento	415
419	415	Public Transportation Services	Serviços de Transporte Público	415
420	415	Taxi and Car Services	Serviços de Automóveis e Táxi	415
421	420	Car and Truck Rentals	Aluguer de Carros e Camiões	415
422	420	Charter Buses	Autocarros de Serviço Ocasional	415
423	420	Limos and Chauffeurs	Limusinas e Motoristas	415
424	415	Transport Hubs	Estações de Transporte	415
425	424	Airports	Aeroportos	415
426	424	Bus Stations	Paragens de Autocarro	415

427	424	Heliports	Heliportos	415
428	424	Ports	Portos	415
429	424	Rail Stations	Estações Ferroviárias	415
430	1	Travel	Viagem	430
431	430	Cruises	Cruzeiros	430
432	430	Lodging	Alojamento	430
433	432	Bed and Breakfasts	Bed and Breakfast	430
434	432	Cottages and Cabins	Chalés e Cabanas	430
435	432	Hostels	Albergues	430
436	432	Hotels and Motels	Hotéis e Motéis	430
437	432	Lodges and Vacation Rentals	Albergues e Alugueres para Férias	430
438	432	Resorts	Resorts	430
439	430	Tourist Information and Services	Serviços de Informação Turística	430
440	430	Travel Agents and Tour Operators	Agências de Viagens e Guias Turísticos	430
441	83	Geriatrics	Geriatria	62
442	123	Discount Stores	Lojas de Desconto	123
443	149	Meat and Seafood	Carnes e Frutos do Mar	123
444	123	Office Supplies	Materiais de Escritório	123
445	123	Party Supplies	Fontes do Partido	123
446	177	Career Counseling	Aconselhamento de Carreira	177
447	177	Construction	Construção	177
448	269	Notary	Notário	177
449	177	Photography	Fotografia	177
450	177	Translation Services	Serviços de Tradução	177
451	382	Golf Courses	Campos de Golfe	372
452	408	Surfing	Surfe	372
453	37	Embassies	Embaixadas	20
454	460	Infrastructure	Infra-estrutura	177
455	460	Mobile	Móvel	177
456	460	Advertising	Publicidade	177
457	347	Asian	Asiático	308
458	347	Food Trucks	Food Trucks	308
459	415	Rest Areas	Áreas de Descanso	415
460	177	Technology	Tecnologia	177
461	118	Dog Parks	Parques Cão	107
462	425	International Airports	Aeroporto Internacional	415
463	317	Arcades	Vídeojogos	308
464	347	International	Internacional	308
465	217	Convention Centers	Centros de Convenções	177
466	62	Optometrist	Optometria	62
467	1	NoExport	NoExport	467

6.2 Anexo 2 – Taxonomia das categorias do DBPedia




dbo:Agent
Agent
dbo:MusicalArtist
dbo:EducationalInstitution
dbo:College
dbo:School
schema:School
schema:Person
NaturalPerson
foaf:Person
dbo:Person
dbo:SwitzerlandSettlement
InformationObject
dbo:Population
dbo:PenaltyShootOut
dbo:Event
Event
schema:Event
schema:WebPage
schema:Canal
dbo:Canal
dbo:Type
dbo:Genre
dbo:MovieGenre
dbo:ArtisticGenre
dbo:LiteraryGenre
dbo:MusicGenre
Concept
dbo:GovernmentType
dbo:TopicalConcept
dbo:Tax
Description
dbo:Ideology
CognitiveEntity
dbo:SportsSeason
dbo:EthnicGroup
schema:Movie
schema:Sculpture
dbo:MilitaryConflict_-_NaturalEvent
Collection
dbo:SportsTeamSeason
dbo:MotorsportSeason
dbo:List
dbo:TrackList
schema:BankOrCreditUnion
dbo:Bank
dbo:Device
dbo:Engine
dbo:MobilePhone
dbo:Camera
dbo:InformationAppliance
dbo:Monastry
dbo:Blazon
dbo:GeneLocation
dbo:MouseGeneLocation
dbo:HumanGeneLocation
dbo:MeanOfTransportation_-_Instrument
dbo:Statistic
dbo:Thing
owl:Thing
dbo:Demographics
schema:Airport
dbo:Airport
Book
skos:Concept
dbo:Wikidata:Q11424
dbo:RouteStop
dbo:wgs84_pos:SpatialThing
dbo:Flag
schema:Festival
dbo:FilmFestival
dbo:MusicFestival
dbo:GrossDomesticProductPerCapita
schema:TVEpisode
dbo:Diocese_-_Parish
dbo:Language
schema:Language
schema:MusicRecording
schema:RadioStation
dbo:RadioStation
foaf:Document
schema:MusicAlbum
FunctionalSubstance
dbo:Food
dbo:Cheese
dbo:Beverage
dbo:Currency
schema:City
dbo:City
schema:CreativeWork
dbo:Organisation_-_PopulatedPlace
dbo:HungarySettlement
dbo:LebanonSettlement
schema:Hotel
dbo:Hotel
dbo:Holiday
dbo:Year
is-a


schema:Hotel
dbo:Year
TimeInterval
dbo:PeriodOfArtisticStyle
dbo:TimePeriod
dbo:HistoricalPeriod
dbo:PrehistoricalPeriod
Location
dbo:GeologicalPeriod
dbo:Athlete_CareerStation
dbo:YearInSpaceflight
schema:TelevisionStation
dbo:CareerStation
dbo:LiechtensteinSettlement
dbo:ProtohistoricalPeriod
schema:Painting
dbo:Abbey
dbo:Area
dbo:Monastery
Vocabulary
dbo:NorwaySettlement
dbo:TelevisionStation
schema:AdministrativeArea
dbo:AdministrativeRegion
dbo:Mountain,Volcano
dbo:PublicService
dbo:HumanDevelopmentIndex
ChemicalObject
dbo:ChemicalSubstance
dbo:Parish,_Deanery
schema:ShoppingCenter
dbo:ShoppingMall
dbo:BelgiumSettlement
schema:Restaurant
dbo:Restaurant
dbo:Unknown
Entity
geo:SpatialThing
dbo:Altitude
Article
schema:Hospital
dbo:Hospital
dbo:Disease
dbo:Media
schema:SkiResort
dbo:SkiArea
dbo:Depth
dbo:ArchitecturalStructure,_Monument
dbo:Bishop
dbo:GermanSettlement
dbo:ChartsPlacements
dbo:Colour
dbo:Medicine
dbo:Organisation,_Parish
schema:Book
Note
dbo:RomaniaSettlement
dbo:Decoration
dbo:Award
dbo:NobelPrize
schema:SportsEvent
dbo:SportsEvent
dbo:Diploma
dbo:MilitaryConflict,_AdministrativeRegion
dbo:Profession
dbo:PersonFunction
dbo:PoliticalFunction
is-a

6.3 Anexo 3 – Proposta de Projeto

De seguida apresenta-se a proposta que deu origem a este projeto. Desenvolvida pela Doutora Ana Cristina Oliveira Alves em Setembro de 2014.

PROPOSTA DE PROJETO

MESTRADO em INFORMÁTICA E SISTEMAS
especialização em Desenvolvimento de Software

Ano Lectivo de 2014/2015

TEMA

Inferência das Atividades na Modelização de Escolhas de Destinos e seu impacto na mobilidade urbana

SUMÁRIO

Palavras-chave: Computação Ubíqua, *Activity Recognition, Trip Purpose*, *Destination Choice Modeling*, Enriquecimento Semântico de Lugar

1 ÂMBITO

Um problema ainda em aberto que surge da investigação da área da mobilidade urbana, é definir com uma dada probabilidade um conjunto de alternativas de destino para diferentes indivíduos ou grupos de pessoas. A este problema é dado na literatura o nome de *Destination Choice Modeling* e uma das abordagens é inferir as atividades realizadas (*Activity recognition*) por este(s) indivíduos em dados destinos de forma a se perceber qual o objectivo de uma dada viagem (*trip purpose*). Esta inferência é feita atualmente de uma forma algo direta de acordo com a predominância de serviços de uma dada categoria (ex. Lazer, desporto, compras) e não permite diferenciar a nível mais detalhado zonas urbanas que ofereçam estes serviços e consequentemente ter uma maior precisão na possível escolha do utilizador para a realização de uma dada atividade.

Atualmente o factor de atratividade de um destino é calculado baseado na sua dimensão, categoria, e dados recolhidos de forma voluntária (*crowdsourced*). Este trabalho pretende detalhar a representação dos destinos utilizando dados oportunísticos associados a um lugar (recolhido das Web, Wikipedia, Twitter) e extair os tópicos mais relevantes a si associados. A procura de informação e extração destes tópicos constitui o processo de enriquecimento semântico[1].

A melhor representação de destinos permitiria compreender o porquê da escolha de determinado lugares de uma mesma categoria na cidade. Como exemplo, para dois lugares da categoria Desporto/Exercício tais como "Mata Nacional do Choupal[44]" e "Estádio Cidade de Coimbra[45]" de áreas superiores a 20 hectares, somente humanos percebem a diferença entre estes lugares que não está na categoria nem nas dimensões.

Apesar de parecerem exemplos e associações óbvias, ainda não existem um método automático para extração destes tópicos associados a lugares a sua utilização na criação do modelo de escolha de destinos. Para o exemplo apresentado acima, a distribuição dos tópicos varia consoante o lugar (e.g. o estádio teria maior peso dos termos (jogo, Liga, concerto,...) enquanto o Choupal teria (churrasqueira, lazer, verde, ...).

2 OBJECTIVOS

O presente projeto pretende atingir os seguintes objectivos genéricos:

- Estudar métodos de representação de lugares em modelos de escolhas de destinos (*Destination Choice Modeling).*
- Avaliar a disponibilidade de dados a recolher de diversas fontes online para 2 vertentes fundamentais do projeto: enriquecimento da representação de lugares, recolha de viagens de voluntários (*crowdsourced*). Verificar ainda em que cidades tais dados estão disponíveis para definição da área urbana de estudo.
- Recolha efetiva dos dados e disponibilização através de API.
- Visualizar e agregar espacialmente os dados recolhidos de forma a identificar zonas de grande concentração de serviços.
- Analisar os dados recolhidos de forma a extrair tópicos que melhor descrevam os destinos possíveis dos dados de mobilidade recolhidos.

Estes objectivos intermédios vão permitir atingir o objectivo principal deste projeto, o estudo da contribuição de representação enriquecida de lugares na melhoria dos modelos de escolha de destinos.

3 PROGRAMA DE TRABALHOS

O Projeto consistirá nas seguintes atividades e respectivas tarefas:

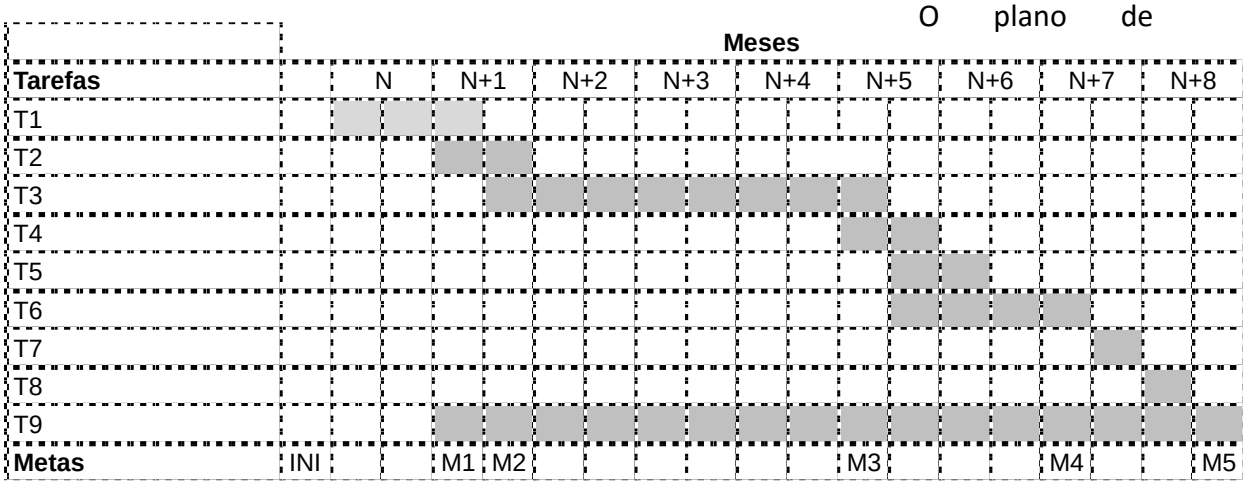
- T1 – Estudo e caracterização dos métodos representação de destinos
- T2 – Catalogação das fontes disponíveis para recolha de dados
- T3 – Recolha efetiva e integração de dados
- T4 – Desenvolvimento da API de disponibilização de dados

[44] http://pt.wikipedia.org/wiki/Mata_Nacional_do_Choupal
[45] http://pt.wikipedia.org/wiki/Estádio_Cidade_de_Coimbra

- T5 – Análise espacial e temporal dos dados para definição da representação de destinos
- T6 – Utilização da representação proposta num modelo de escolha de destinos
- T7 – Documentação (manuais, artigo e dissertação)

4 CALENDARIZAÇÃO DAS TAREFAS

As Tarefas acima descritas, incluindo os testes de validação de cada módulo, serão executadas de acordo com a seguinte calendarização:



O plano de escalonamento dos trabalhos é apresentado em seguida:

INI		Início dos trabalhos
M1	(INI + 6 Semanas)	Tarefa T1 terminada
M2	(INI + 8 Semanas)	Tarefa T2 terminada
M3	(INI + 22 Semanas)	Tarefa T3 terminada
M4	(INI + 28 Semanas)	Tarefa T6 terminada
M5	(INI + 36 Semanas)	Tarefa T9 terminada

5 RESULTADOS

Os resultados do estágio serão consubstanciados num conjunto de documentos a elaborar pelo estagiário de acordo com o seguinte plano:

M1:

Relatório técnico com a descrição da tecnologias estudadas

M2:

Relatório técnico com a análise de requisitos e especificação da aplicação e serviço a desenvolver.

M3:

Relatório técnico com as etapas de desenvolvimento e metodologia seguida.

M4:

Documento a descrever os testes efectuados

M5:

Relatório final de estágio incluindo o manual de instalação e utilização da aplicação, assim como a documentação do funcionamento do serviço desenvolvido.

6 LOCAL DE TRABALHO

DEIS

7 METODOLOGIA

Organização de um Dossier de Projeto e reuniões semanais.

8 ORIENTAÇÃO

Ana Cristina Oliveira Alves (aalves@isec.pt)
Professora Adjunta

Orientando: Rui Fernandes Ganhoto
Aluno do Mestrado de Informática e de Sistemas

9 CARACTERIZAÇÃO

- Data de início: Outubro de 2014
- Data de fim: Julho de 2014

10 REFERÊNCIAS

(1) Alves, A. Semantic Enrichment of Places - Understanding the Meaning of Public Places from Natural Language Texts. PhD Thesis. University of Coimbra, 2012 https://www.cisuc.uc.pt/publication/show/2992