1º Trabalho – Mineração de Dados – Tópicos Avançados em BD
Campus UFC – Quixadá
Professora: Ticiana Linhares

Data de Entrega: 16 de Outubro (às 15:30)
O que deve ser entregue: Para cada questão, os arquivos de saída especificados no documento. Bem como, o código fonte. Deve ser feito o upload dos arquivos de saída no SIPPA.

Dados de mobilidade têm sido fomentados devido à ampla difusão de tecnologias sem fio, tais como os registros de detalhes de chamadas a partir de telefones móveis e GPS a partir de dispositivos de navegação. Estes dados abrem novas oportunidades para descobrir os padrões e modelos ocultos capazes de caracterizar as trajectórias humanas durante a sua atividade diária, analisar o fluxo de tráfego de rede e tempo de viagem, entre outras aplicações. Esses padrões não caracterizam apenas a mobilidade individual, mas também de grupos que compartilham trajetórias semelhantes por um determinado período de tempo.

Entre estes padrões de mobilidade, neste trabalho considere a clusterização de objetos móveis e como eles evoluem ao longo do tempo. Este tipo de padrão pode ser utilizado para capturar grupos que apresentam rotas semelhantes em um determinado intervalo de tempo, podendo auxiliar em aplicações como carona coletiva e reengenharia de tráfego.

Os dados que você utilizará neste trabalho foram coletados a partir de táxis na cidade de Beijing. Deseja-se descobrir se existe um padrão de cobertura das regiões que são atendidas pelos taxistas durante **a hora *H*, comparando esta mesma hora nos dias de semana**. Além disso, para que se possa afirmar que uma região é coberta, é necessário garantir a presença de pelo menos ***minPoints*** taxistas dentro de **um raio de distância de, no máximo, *eps***.

Em aplicações reais, muitos objetos se movem na rede de ruas. A conectividade na rede é geralmente modelada usando a representação em grafos, composta por um conjunto de vértices (nós) e um conjunto de arestas (conexões). Dependendo da aplicação, o grafo pode ter pesos (custo associado para cada aresta) e direcionado (cada aresta tem uma orientação). Neste trabalho, a rede de Beijing tem pesos (custo das arestas igual ao seu comprimento) e é um grafo direcionado. Existem algumas diferenças entre o movimento no Espaço Euclidiano e no Espaço da Rede:

1. A distância entre dois objetos móveis na rede deve ser o menor caminho que os conecta (na rede), ao invés de considerar uma linha reta.

2. Além disso, os dados de GPS devem ser mapeados na rede, portanto com informações sobre qual segmento da rede (ou aresta) cada posição reportada pertence. Este mapeamento é conhecido como *Map Matching*.

O algoritmo de *Map Matching* mais conhecido é o que associa um ponto P(longitude, latitude) a aresta com menor distancia perpendicular.

Colete o conjunto de dados de taxistas de determinada hora ***H em 3 dias diferentes (sendo tais dias consecutivos)***, aplique o algoritmo de *Map Matching*

nele sobre a rede de Beijing. Além disso, apresente e implemente um algoritmo que receba de entrada a hora ***H***, ***minPoints e eps*** e gere como saída as regiões no mapa que são cobertas pelos taxistas durante a hora **H** para cada dia da semana, de acordo com a densidade mínima a ser atendida (***minPoints***) dentro do raio (**eps**). Realize a clusterização desses dados na rede de ruas de Beijing, porém utilizando como distância entre os pontos a distância de rede, ou seja, o menor caminho entre quaisquer dois pontos dados de entrada. **Reporte todas as regiões cobertas por esses objetos móveis em formato de clusters, seguindo as instruções abaixo sobre os arquivos de saída.**

Formato dos arquivos de saída:

(student_id, id_taxista, weekday, hour, latitude, longitude, cluster, iscore)

1. **Student_id: escolha a matrícula de um dos membros da equipe.**
2. **Id_taxista: identificador do taxista presente no conjunto de dados de taxi**
3. **weekday: dia da semana (1-segunda, 2-terça, 3-quarta, 4-quinta, 5-sexta)**
4. **hour: inteiro entre 0 e 23 indicando a hora do dia escolhida para ser analisada**
5. **latitude: latitude da coordenada onde o taxista se encontra**
6. **longitude: longitude da coordenada onde o taxista se encontra**
7. **cluster: id do cluster (utilize -1 quando for outlier)**
8. **iscore: true-é um core point, false- caso contrário**

Observações:

1. **Ajuste os parâmetros para que você tenha a presença de pelo menos 5 clusters.**
2. **Topologia da Rede: rede de ruas de Beijing**
3. **Posições (long, lat) dos vertices: metadados vertices**
4. **Tdrive: dados dos taxistas**

Alternativa:

Se você quiser trabalhar com outros tipos de dados, por exemplo, com textos ou outro dataset — taxis em Fortaleza, de pessoas, por exemplo. É necessário conversar comigo sobre esse tipo de dado, principalmente se for texto. Pois é necessário avaliar o pré-processamento a ser realizado (como no caso do Map Matching que é um pré-processamento também). Ao utilizar texto, você pode clusterizá-los aplicando como medida de similaridade Jaccard.