Universidade do Estado do Amazonas

Escola Superior de Tecnologia

Algoritmos e Estrutura de Dados II (Code: ESTCMP012)

17 de outubro de 2019

Tempo: 14 dias

Trabalho Prático II

Especificação do Trabalho

1. Os usuários da Web postam com frequência vários comentários sobre produtos em sites de comércio eletrônico. Os comentários geralmente são relativos a diversos aspectos dos produtos. Por exemplo, na sentença *great battery life*, o usuário está informando que gostou (*great*) da duração da bateria (*battery life*) do aparelho. Neste caso, o aspecto que ele gostou foi a duração da bateria (*battery life*).

   Existem diversos métodos na literatura que fazem extração de aspectos [1]. Porém, os aspectos podem ter como alvo partes específicas do produto. No exemplo anterior a opinião era sobre a bateria. Uma maneira de organizar os aspectos é usar técnicas de agrupamento (*clustering*). Por exemplo, os usuários podem comentar sobre diversos aspectos da bateria, tais como a duração, a velocidade de carregamento, a substituição da bateria, entre outros. Assim, como podem fazer comentários sobre diversos aspectos da dimensão do aparelho, tais como o peso, o tamanho, a largura, entre outros. A hipótese que surge é que aspectos pertencentes a um mesmo grupo têm maior grau de similaridade quando comparados com pares de aspectos de grupos distintos.

   Diante desse cenário, o trabalho consiste em implementar um método de agrupamento de grafos explicado a seguir. Seja o grafo `G = (V, A)`, onde `V` é o conjunto de vértices e `A` é o conjunto de arestas. Os vértices serão os aspectos e as arestas serão a ligação entre cada um dos vértices. As arestas serão ponderadas (pesos) pela similaridade entre os aspectos. A Figura 1 mostra um exemplo de grafo `G`.

   A geração dos grupos (*clusters*) deverá ser realizada em quatro passos sequenciais:

   1. Gerar o grafo `G` conforme a apresentação inicial. Vide Figura 1.
   2. Obter a árvore de extensão mínima (*Minimum Spanning Tree* - MST) a partir do grafo `G`.
   3. Remover as k-1 arestas com menor peso.
   4. Gerar os k grupos (*clusters*) de aspectos.

   Para a geração do grafo `G` será necessário calcular os pesos entre os vértices. Esse peso pode ser obtido através do cálculo de similaridade entre os pares de aspectos. Para o cálculo da similaridade deve-se utilizar a ferramenta Spacy [2]. Quando não for possível calcular a similaridade entre dois termos, deve-se assumir que o valor é zero.

   Existem diversos algoritmos para a geração da árvore de extensão mínima (AEM). Pesquise os métodos existentes e escolha o método que seja apropriado (adequado) para o problema.

   O valor de k será igual 10 e corresponde aos 10 alvos possivelmente encontrados entre os aspectos (`Processador, Bateria, Preço, Memória, Geral, Câmera, Outros, Dimensão, Tela e Software`).

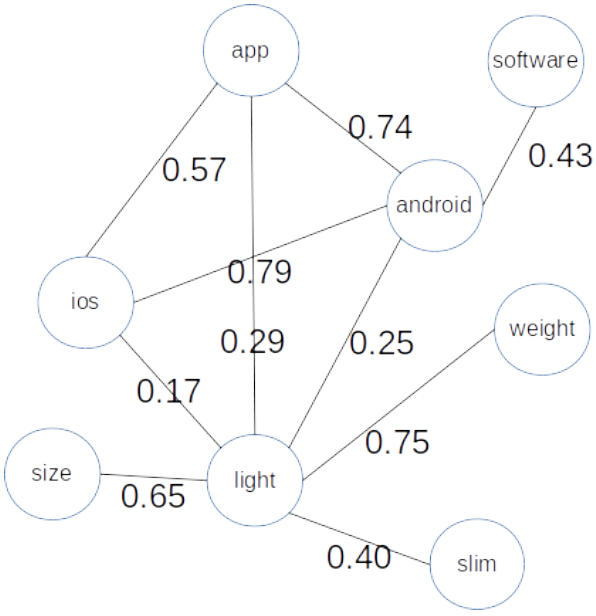

Figura 1: Exemplo de grafo resultante.

De posse do resultado dos grupos (*clusters*), avalie se os aspectos foram corretamente agrupados. Faça uma análise descritiva dos resultados obtidos comparando com o gabarito. Se for o caso, tente identificar/explicar os possíveis erros de agrupamento e o que poderia ser feito para gerar melhores resultados.

[1] Schouten, Kim, and Flavius Frasincar. "Survey on aspect-level sentiment analysis."IEEE Transactions on Knowledge and Data Engineering 28.3 (2015): 813-830.

[2] https://spacy.io/usage/vectors-similarity

Regras:

- Os trabalhos deverão ser realizados em dupla.
- O início do trabalho será no dia **17 de outubro de 2019** (12:00 horas) e a entrega será no dia **31 de outubro de 2019** (até as 10 horas).
- Cada dupla terá **até 5 minutos** para apresentar o trabalho. A defesa deve ser preparada para esse tempo. Nesta defesa, a dupla deverá apresentar e justificar os algoritmos escolhidos para geração do resultado. Além disso, deverá apresentar o resultado e explicar o quanto ele ficou próximo do gabarito.
- Deverá ser encaminhado para o e-mail tmelo@uea.edu.br do professor o relatório, os slides e o código-fonte. Os três arquivos deverão estar compactados (.zip) e o título do e-mail deverá ser *"Trabalho Pratico II - AED 2"*.
- A defesa do trabalho deverá acontecer no dia **05 de novembro de 2019** no horário da aula. A ordem das apresentações será por sorteio. Portanto, todas as duplas deverão estar presentes no início das apresentações.
- Esse trabalho será composto por duas notas: a implementação e a defesa. Estilo de programação e técnicas adotadas serão consideradas na avaliação da implementação.

Segurança e conhecimento demonstrado na defesa serão considerados como critérios de avaliação.

- A discussão sobre o trabalho deverá acontecer **somente** entre a dupla.
- A versão padrão do Python será a 2.7. Caso a implementação tenha sido em outra versão, a dupla deverá informar no relatório.
