UFRPE - Pró-Reitoria de Pesquisa e Pós-Graduação
Disciplina: PLN/MT – Prof. Rinaldo Lima

Data de Entrega: 03/12/2018 - Enviar ao email: rjlima01@gmail.com

UFRPE Universidade Federal Rural de Pernambuco

7ª. LISTA DE EXERCÍCIOS

1. Word Embedding (WE) usando W2Vec

Usando um Word Embedding de sua escolha (W2Vec, Glove ou NumberBatch) implemente um programa que forneça as seguintes funcionalidades:

1. Retornar a similaridade entre 2 palavras quaisquer presentes no vocabulário abaixo.

   {banana, potato, pear, pineapple, apple, turtle, peacock, dog, cat, duck, swan, elephant, pig, lion, penguin, cup, bowl, kettle, spoon, car, truck, ship, helicopter, boat, pen, pencil, knife, scissors, screwdriver}

2. Projetar, em um gráfico 2D, os pares de palavras como pontos usando apenas as 2 primeiras dimensões do vetor.
3. Usar este word embedding como entrada para o algoritmo *k-means* com o objetivo de gerar os clusters do vocabulário do item (1). Tente os seguintes valores de k={2,3,4,5}.
4. Quais deles fazem mais sentido? Analise os clusters encontrados quanto a similaridade entre as palavras pertencentes a um mesmo grupo, e a dissimilaridade inter-grupos.
5. Use a ferramenta **t-SNE** para projetar os word embeddings das palavras em (3), diferenciando cada cluster encontrado usando cores diferentes.

2. Classificação de sentenças de reviews de filmes usando um WE de sua escolha.

1. Faça o seguinte pré-processamento do dataset **rt-polaritydata.tar** usando apenas 1.000 sentenças positivas e 1.000 sentenças negativas.
   - Remoção de stopwords
   - Conversão para minúscula
   - Remoção de símbolos especiais.
   - Tokenização

   Em vez de construir os vetores *esparsos* de todas as sentenças do item (1) acima, gere apenas 1 vetor *w2vec* para cada sentença adotando os seguintes procedimentos para "combinar" todos os tokens de uma sentença, em apenas 1 vetor w2vector final:

   a. **Somar** todos os vetores correspondente às palavras de uma sentença, onde cada vetor assim gerado é acrescido de uma última coluna com seu label (pos, neg)
   b. Calcular a **média** de todos os vetores correspondente às palavras de uma sentença.
   c. Ponderar cada palavra por seu TF-IDF correspondente, antes de somar todos os vetores como em (a).

d. Ponderar cada palavra seu TF-IDF correspondente, antes de calcular a média de todos os vetores como em (b)

2. Use o Weka ou Python para construir os seguintes classificadores (NB, SVM, J48, PART, Simple_Logistic) e avaliando-os usando cross validation com 10 folds usando as versões dos datasets em (a) (b) (c) e (d) do item anterior.
    a. Qual método de combinação de representação apresentou melhor desempenho na classificação, independentemente do algoritmo usado?
    b. Qual classificador se beneficiou mais da representação w2vector?

3. Use as ferramentas GenSim e t-SNE para:

4.1. Calcular as expressões listadas a seguir. Quais palavras são mais próximas do vetor resultante?

a) vetor("king")-vetor("man")+vetor("woman") --> ?
b) vetor("paris")-vetor("france")+ vetor("italy") --> ?
c) vetor("recife")-vetor("brazil")+vetor("usa") --> ?
d) vetor("apple")-vetor("apple-tree")+vetor("strawberry") --> ?
e) vetor("lion")-vetor("africa")+vetor("kangaroo") --> ?

4.2 Para encontrar a palavra "intrusa" presentes nos conjuntos abaixo (Teste TOEFL):

a) evaluation, assessment, examination, supervision, verification
b) context, meaning, significance, perspective, emphasis
c) method, procedure, techinique, approach, model
d) result, outcome, effect,evidence, reason
e) conclusion, outcome, finding, assertion, explanation

Material extra de apoio para os exercícios

- Ver os links no arquivo **Links.txt**.