Processamento de Linguagem Natural

Trabalho Prático 1

Objetivo: Exercitar e materializar o conceito de “representações distribuídas, densas, altamente informativas, através de fatores latentes”, ou simplesmente, “word embeddings.”

Recursos: Você pode escolher a implementação/pacote, dados de avaliação e o corpus para treinamento a serem utilizados neste trabalho prático. Apenas como sugestão de implementação aponto o repositório: https://github.com/nicholas-leonard/word2vec

No mesmo repositório você também encontrará os dados de analogia necessários para avaliação dos seus modelos de linguagem:
https://github.com/nicholas-leonard/word2vec/blob/master/questions-words.txt

Por fim, sugiro o corpus para treinamento disponível em: http://mattmahoney.net/dc/text8.zip

Atividades: Sucintamente, você deverá produzir e avaliar modelos de linguagem neural. A qualidade desses modelos depende dos hiperparâmetros utilizados. A habilidade a ser exercitada neste trabalho prático é a escolha adequada dos hiperparâmetros de um modelo de linguagem. Portanto, mais especificamente, você deverá produzir diversos modelos de linguagem, cada um deles obtidos com diferentes hiperparâmetros, e em seguida você deverá avaliar cada modelo através de uma aplicação de analogias (descrita a seguir). Por fim, você deverá concluir a respeito dos melhores hiperparâmetros para o modelo de linguagem. Os hiperparâmetros a serem considerados são:

- CBOW ou Skip-gram
- Tamanho da janela de contexto
- Tamanho do embedding
- Quantidade de iterações de treinamento

Avaliação do modelo: Cada modelo produzido (ou seja, uma combinação específica de hiperparâmetros) deverá ser avaliado através de operações algébricas com vetores, e para isso você utilizará os dados de analogia.

Você perceberá que um modelo de linguagem será um arquivo contendo um vetor representando cada palavra presente no corpus de treinamento. Ou seja, cada palavra será tratada através de seu respectivo vetor.

No arquivo de analogias você encontrará linhas com quatro palavras. Suponha que sejam as seguintes: Paris France Berlin Germany. Nesse caso, a operação a ser feita é:

vector(‘France’) + vector(‘Paris’) - vector(‘Berlin’)

Repare que essa operação resultará em um vetor resultante R. Portanto, calcule a distância (distância do cosseno) entre R e o vetor da palavra esperada, que nesse exemplo é Germany. Repita esse mesmo procedimento para todas as linhas (ou algumas) no arquivo de analogias e calcule a distância média. Escolha o modelo de linguagem que resulte na menor distância média.

Entregável: Você deverá entregar o notebook com o seu código e resultados. Não é necessário uma documentação, toda a explicação e o conjunto de gráficos poderá estar no próprio notebook.