Primeiro Trabalho Prático

Modelos de Linguagem

O objetivo deste trabalho prático é a avaliação intrínseca de modelos de linguagem neuronais. Para tanto você deverá

1. Obter e preparar o corpus para a modelagem da linguagem.
2. Produzir um modelo word2vec.
3. Avaliar o modelo.

Todo o código a necessário para produzir o modelo de linguagem está disponível em https://code.google.com/archive/p/word2vec/source/default/source

No link acima há disponível também os dados de validação no arquivo questions-words.txt.

Existem diversas outras implementações livremente disponíveis, e sinta-se livre para escolher a implementação mais conveniente para você.

O corpus está em Inglês e disponível para download em: http://mattmahoney.net/dc/text8.zip. Você deve descompactar o arquivo, de forma que você possa trabalhar com um arquivo textual.

Você deverá produzir e avaliar diferentes modelos de linguagens, obtidos a partir da variação dos seguintes parâmetros:

1. Diferentes tamanhos de corpus.
2. Diferentes tamanhos de contexto.
3. CBOW e Skip-gram.

A avaliação de cada modelo de linguagem produzido será da seguinte forma:

- Forneça as três primeiras palavras de cada linha do arquivo question-words.txt ao programa word-analogy.exe.
- O resultado correto é a quarta palavra.
- Calcule o erro com base na diferença entre as distâncias da palavra retornada no topo do ranking e da palavra correta. Apresente os resultados com gráficos ou tabelas.

Você pode fazer uma documentação em pdf, ou usando um jupyter notebook.