Atividade - Processamento de Linguagem Natural

Os questionamentos propostos aqui são baseados na implementação disponível em https://github.com/dimmykarson/aulanlp. Para resolver o exercício, o aluno deve baixar o código e realizar as modificações necessárias.

Obs. 1: Os experimentos podem ser realizados com uma porção do dataset (quando for inviável utilizá-lo por completo)
Obs. 2: Nos questionamentos, solicita-se que se reportem os resultados. Entende-se que o aluno deve descrever os valores de acurácia do modelo de classificação (no caso MLP) para cada um dos experimentos. Os valores de perda e F1 score médio também devem ser reportados. O valor de perda (loss) pode ser obtido no próprio modelo do classificador (no caso MLPClassifier). O valor de F1 score pode ser obtido usando a função f1_score do scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html).

1. Modifique o código do arquivo 'bow_tfidf.ipynb'. Implemente uma funcionalidade que remova as palavras **mais** utilizadas, isto é, remova do vocabulário, palavras que apareçam em mais de 80%, 70% e 60% dos documentos. Reporte os resultados.
2. Modifique o código do arquivo 'bow_tfidf.ipynb'. Implemente uma funcionalidade que remova as palavras **menos** utilizadas, isto é, remova do vocabulário, palavras que apareçam em menos de 30%, 20% e 10% dos documentos. Reporte os resultados.
3. Modifique o código do arquivo 'w2v.ipynb'. Implemente uma funcionalidade que remova as palavras **mais** utilizadas, isto é, remova do vocabulário, palavras que apareçam em mais de 80%, 70% e 60% dos documentos. Reporte os resultados.
4. Modifique o código do arquivo 'w2v.ipynb'. Implemente uma funcionalidade que remova as palavras **menos** utilizadas, isto é, remova do vocabulário, palavras que apareçam em menos de 30%, 20% e 10% dos documentos. Reporte os resultados.
5. Avalie e descreva se e porque a mudança na frequência de palavras melhoram o modelo baseado em TFIDF e Word2Vec.
6. Modifique o código do arquivo 'w2v.ipynb', alterando os valores de min_count e window do modelo Word2Vec. Varie os valores de min_count de 1 a 5, e window de 5 a 10. Reporte os resultados. Avalie e descreva se e porque a mudança nestes parâmetros melhoram o modelo. Os valores mudariam se os textos de cada revisão de filme fossem maiores? E se fossem menores?