Document vectors

Explore how dense word/document embeddings can be used for document classification. You will be attempting to distinguish between documents from two different authors.

Use the provided script as a starting point. Before beginning, read and understand what it's doing. Then **implement two types of dense document vectors**:

1. using LSA on raw token counts
2. summing pretrained word2vec embeddings

Both should produce document vectors of length 300.

Show and discuss the results. The results/discussion should include

1. the percent correct for each method, and
2. a brief explanation of the relative performance, i.e. why might method A lead to better classification performance than method B, etc.

You may work in a group of 1 or 2. Submissions will be graded without regard for the group size. You should turn in a document (`.txt`, `.md`, or `.pdf`) answering all of the **red** items above. You should also turn in a Python scripts (`.py`) for the **blue** items . You may use only the standard library, `numpy`, `sklearn`, and `gensim`.

Resources

- https://radimrehurek.com/gensim/models/word2vec.html
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html