Homework

Reading

- word2vec paper: https://arxiv.org/abs/1301.3781
- Optional: start reading Hands on ML: Chapter 15 (Processing Sequences with RNN and CNN) – this will be next week's reading, but you could get a head start if you want.

Coding

- Build an embedding model for the Sentiment Analysis dataset.
  - see this colab for tips on how to access; replace dataset/username/key as needed
- Use a similarity metric of choice (e.g., L2, Cosine) or other tooling (e.g., in gensim) to evaluate the 10-20 most similar words in your embedding model to some movie-related terms:
  - popcorn, terminator, movie, great, terrible
  - [at least 5 of your choice]
  - Do word subtraction/addition element-wise on some word embedding vectors you generated. E.g.,
    - king-female=[some_vector]. Find most similar words to [some_vector] - hopefully 'queen' is near the top!
    - Think of a few other words, just as an experiment.
- Discuss your approach. How did you tokenize your text? What are your observations of the word similarity? Are obvious terms missing? Do some make no sense?
- Don't forget to make your code readable and annotate it with text blocks describing why you're doing what you're doing. Without it, I don't know if you actually understand your code.