Project progress report (5%): Around week 13, each project team will be asked to submit a very brief progress report with three short paragraphs to cover, respectively, 1) what has been done, 2) what remains to be done, and 3) any barrier(s) in achieving the proposed project goal. Progress report would be graded based on completion and should be submitted on Piazza.

**What Has Been Done:**

* Retrieved, cleaned and explored the Enron email corpus <https://www.kaggle.com/zichen/explore-enron/code>
* Created training and testing subsets of corpus. We chose to use a subset of 25,000 uniform randomly chosen e-mails instead of the entire corpus, since using the entire corpus has been impractical to us.
* Categorized 900 emails by hand to indicate their importance, to use it as a training set
* Created a prototype importance classifier using SpaCy NLP library. This classifier takes an e-mail as input and measures its importance. To determine importance, we use a weighted average of the e-mail’s similarity to the important e-mails and its dissimilarity with the unimportant e-mails

**What Remains to be Done:**

* Analyze classifier performance
* Consider improving the classifier
* Create a document summarizer (maybe leverage TF-IDF)
* Explore quality phrase mining for use in the document summarizer

**Any Barriers in Achieving the Proposed Project Goal:**

* The biggest problem is that currently graph theory is a bit hard to apply to the importance metric, considering that there are just so many potential senders and receiver