Yijing Barry Zhang
IEMS 308

**Process:**

CEOs and Companies:

- First we identified all people and organizations using SpaCy's pretrained model. This gives us a nicely defined set of people and organizations that can be labeled and natively fed into a blank model
- Out of all people and organizations, we use the given positive labels to train a SpaCy NER model from scratch. The two tags that the new model is trained for are "CEO" and "Company". Training the NER model takes a fair amount of formatting and initialization but is a clean process.
- Due to time constraints and the large amount of data that we have, we are only training for 10 epochs. There is a noticeable plateauing in the decrease of the loss function, which also suggests that the marginal benefit of additional training is small

Since SpaCy is an end-to-end deep learning platform and the words are transformed into embeddings pretty early on in the process, the specific training process and internal features that are used are not all transparent. From the documentation, however, we can gather a couple of major types of features that go into the embedding:

- Normalized string
- Prefix of the string (length 3)
- Suffix (length 3)
- Shape of the word (what composes of a word: digit, uppercase letter, lowercase letter, etc.)

Percentage:

Percentages are a bit trickier because they can be numbers, texts, phrases, as well as a combination of those. Since all percentages have a strong identifier (% or 'percent), we decided to use regex as a clean way to approach this.

A list of regex used to detect the percentages in these articles are as follows:

- For all numeric representation: "(\+*\-*\d*\.*\,*\-*\d*\/*\\d*s*%)"
- For a mixture of textual and numeric representations: "(\-*\d*\.*\,*\-*\d*\/*\d*\s*percent)", (\-*\d*\.*\,*\-*\d*\s*percentage), (\-*\d*\.*\,*\-*\d*\s*percentage point)
- For any text-based representations: ((one|two|twe|thr|thi|for|fou|fif|fiv|six|sev|eig|nin|ten|zer|hun|hal)[a-z|-]*\spercent)
- Specifically for one hundred percent: 'one hundred percent'

**Model:**
We used regex to parse through the percentages so no model was involved in that process. The model that we used is a blank model from Spacy trained specifically for NER tasks. This model is a deep convolutional neural network that uses subword features and bloom embeddings. The optimizer is stochastic gradient descent and the loss function is multi-label log loss.

More details can be found here: **https://spacy.io/universe/project/video-spacys-ner-model**

**Performances:**
Since we do not have all positive labels and manually checking through the final output is simply too arduous, we used a training-testing split to do assessment on our model for CEO and Companies, and did a custom testing for percentages.

For CEO and Companies, we trained a blank model on 80% of the training data for 10 epochs. This model was used to find the named entities of the rest of the data. Since the rest of the data is already labelled from our previous iteration, we could use the total number of named entities and the number of correctly identified (both the entity and the exact text) to calculate a recall metric. We did not calculate the full confusion matrix because a quick check showed that there were very few false positives by our model at least based on the given labels. Given the much larger size of negative samples, this result makes sense and it suggests that the metric we would want to focus on is the recall rate.

For percentages, we combined all of the positive labels (the ones that include either a percent sign of "perc" somewhere in the text) and calculated the recall rate based on those samples. The reasoning behind calculating recall is similar. With regex, the false positives are very tightly controlled since I made sure that the regex would not capture anything without percent sign or a textual "percent".

CEO:

- Recall: 93.89%
- From spotchecking we see that there's a fair amount of false positives: names that are not necessarily CEOs but other executives (like CFOs, CIOs, Chairman) and even entirely different entities (like Arnold Schwarzenegger)

Company:

- Recall: 96.71%
- False positives are a lot less prevalent with companies and most of the entities extracted were indeed companies. There are still some mistakes (Oscar, USSR, etc.) but these were truthfully fairly hard to distinguish even for an uninformed human.

Percentages:

- Recall: 99.39%

- Looking at the set difference, there are certain edge cases that I missed but most of these cases were not actually present in the corpus (a combination of textual numbers followed by a %, for example) There are also some cases where the inclusion of a certain format might be too nitpicking (for example, “three-quarters of a percentage point” would have required us to allow either a really specific format or a wider format that included some unwanted results.)
- Overall, a recall of 99.39% should suffice for our task.

**Results:**

- The extracted percentages, CEO names, and companies are stored in csv files in the github. The CEOs and Companies do not include duplicates but the percentage file does.
- In total, we found:
    - 84822 percentages
    - 2902 CEO names
    - 3785 Companies