IEMS 308 – HW 3

Alex Merryman

The model uses spaCy, a state-of-the-art Natural Language Processing package for Python that has been optimized for speed and accuracy (https://spacy.io/).

First, the text was imported as Unicode to remove non-Unicode characters (likely vestiges from CSS formatting). Then spaCy was used to tokenizes the corpus, identifying separate words and characters. Since the data is already plain-text and now in Unicode, there is minimal preprocessing to be done.

Percentage Identification

Since all percent-related numbers to be extracted are succeeded by "%," "percent," "percentage," "percentile," or some variation thereof, Regex can be used. Three classes of varying complexity of the desired percentage numbers were determined:

- those of the simplest *digit(s) + percent sign* format (for example, 20%, or -1.4%)
- those of a *digit(s) + "percent"/"percentage"/etc*. format (for example, 20 percent, or -1.4 percent)
- and those of the most complex *"number(s)" + "percent"/etc*. format (for example, twenty percent, or negative one percent).

Additionally, each of these classes can be used to specify a range of percentages, for example 20-35% / 20-35 percent / twenty to thirty-five percent.

Matching the first class simply required matching a simple Regex pattern: *"(\d*\-*\d*\.*\d*%)"*

The second class required a bit more complex Regex pattern:
"((negative)\s*\d*\s*(to)*\s*\d*\-*\d*\.*\d+\s*\-*(percent)+((age)|(ile))*\s*(point(s*))*)"*

Finally, the third class required an algorithm:

1. Iterating across all tokens' lemma, whenever a "stop pattern" ("percent", "percentage", etc.) was detected:
2. Iterate backwards across tokens and append them to a new list, until the next previous token POS is not NUM (number), HYPH (hyphen; –), or the token lemma is not "to".

This captures all desired percentages.

CEO Identification

To identify CEOs, the algorithm again iterates across all tokens. Whenever the token lemma is also one found in the CEO keywords list, or two neighboring tokens are proper nouns, the algorithm searches for and identifies proper nouns (based on POS tagging) within 5 tokens of the CEO keyword.

Company Identification

Similar to CEO identification, the algorithm searches for and identifies proper nouns within 7 tokens on either side of a token that is also found within the company keywords list.

My model identified:

- 77523 percentages
- 467255 CEOs
- 471223 companies

And it processed all 730 text files in less than 30 minutes.