Luminoso Python Test

Marc Dupuis

1 Objective

The objective of this project was to determine how many times a word appears in its plural form in a given text. The text came from the Brown Corpus composed of just over $1,000,000$ words which were labelled by category.

2 Approach and Assumptions

The overall approach was the following:

1. Load complete text files by placing all individual text samples in one large text list variable.
2. Remove all punctuation and additional odd labels which can cause issues, such as *1-1/2/cd*
3. Categorize data according to plural and singular
4. Convert plural words to singular, and “clean up” the singular words (eg. translating the $ sign to the word *dollar* and removing the possessive “*’s*”)
5. Find all words that are present more frequently in their plural form and those that aren’t at all present in their singular form (discarding all other words)
6. Sorting the words according to how frequently they appear in their plural form and storing the data in CSV files.

The words of interest chosen for this study were sorted using a custom *singular()* function (extract of code shown below):

- nouns (NN, NNS)
- determiners and quantifiers (DT, DTS)
- reflexive and intensive personal prounouns (PPL, PPLS)
- nominal possessive pronouns (PP$,PP$$)

```
def singular(word):
    word = word.lower()
    if re.match("^[A-Za-z_-]*$", word):
        if word.endswith("s"):
            if word.endswith("ies"):
                word = (word[:-3] + "y")
                singularForm = word
            elif word.endswith("hes"):
                word = (word[:-3] + "h")
                singularForm = word

                ETC.

    return singularForm
```

The *singular()* function can easily be modified to add or change conditions to change a word from plural to singular form. However, a number of simplifying assumptions were made. First, the rule that affects the most items was chosen in situations such as the pluralization of words finishing in *o*. This for example, works for *shoes*, but fail for *toes* (the other chosen rules for this program can be easily inferred from the code and comments). Second, it was assumed that there were no typographical errors in the Brown Corpus, which does not seem to be the case. The word *employe* with only one *e* at the end for example, appears a number of times. Finally, it was assumed that all words were correctly labelled, and that the definition of the word was irrelevant (homonyms accepted: train station vs. radio station).

3 Results

The data output was saved to *output_only_plural.csv* and *output_plural_and_singular.csv*.

The top ten words that appear more frequently in their plural form than in their singular form are in Table 1 of the Appendix and those that are not present in their singular form in Table 2.

4 Conclusion

The code is designed to correctly identify as many plural words as possible and written in a way that allows the user to easily modify the classifying rules. However, looking at the list of words that only appear in their plural form, it is obvious that a number of ad hoc rules need to be setup for words such as: *people*, *children*, *teeth* etc.

Appendix

Word	Percent of plural form	Number of plural forms
stair	95.83	46
headquarter	95.65	22
relative	95.24	20
tear	94.44	34
employe	94.44	17
survivor	92.86	26
stockholder	92.86	13
investor	92.86	13
microorganism	92.31	12
rib	91.67	11

Table 1: Words according to frequency of appearance in plural form.

Word	Number of plural forms
people	864
children	344
other	320
feet	283
one	113
teeth	102
cattle	94
personnel	70
live	50
million	49

Table 2: Words ranked by number of occurrences.