SPELL CHECKER IN INFORMATION RETRIEVAL

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,SRICITY

BY
E.P.S.Santhosh
UG-3
S20180010053

14-10-2020

- The big.txt is used as a vocabulary .
- This spell checker have a accuracy of above 80%
- In the given test cases we got all the correct words
- The peternorvigs is used as reference for writing code

Exercise: Understand Peter Norvig's spelling corrector

```
import re, collections
def words(text): return re.findall('[a-z]+', text.lower())
def train(features):
model = collections.defaultdict(lambda: 1)
for f in features:
model[f] += 1
return model
NWORDS = train(words(file('big.txt').read()))
alphabet = 'abcdefghijklmnopqrstuvwxyz'
def edits1(word):
splits     = [(word[:i], word[i:]) for i in range(len(word) + 1)]
deletes    = [a + b[1:] for a, b in splits if b]
transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b) gt 1]
replaces   = [a + c + b[1:] for a, b in splits for c in alphabet if b]
inserts    = [a + c + b     for a, b in splits for c in alphabet]
return set(deletes + transposes + replaces + inserts)
def known_edits2(word):
return set(e2 for e1 in edits1(word) for e2 in
edits1(e1) if e2 in NWORDS)
def known(words): return set(w for w in words if w in NWORDS)
def correct(word):
candidates = known([word]) or known(edits1(word)) or
known_edits2(word) or [word]
return max(candidates, key=NWORDS.get)
```

EXTENSIONS

1) Reductions
   if we have the word “Jjoobbbb” then it will convert it to job(which is the correct word)

```
def reductions(self,word):
    word = list(word)
    for index,i in enumerate(word):
        n = self.numberofduplicates(word, index)
        if n>1:
            flat_list = [i*(r+1) for r in range(n+1)][:3]
            for j in range(n):
                word.pop(index+1)
            word[index] = flat_list
    for p in product(*word):
        yield ''.join(p)

def list_reductions(self,word):
    x=[]
    for j in self.reductions(word):
        x.append(j)
    return x
```

EXTENSIONS

1) Vowel insertions

if we have the word "weke" then it will convert it to wake(which is the correct word)

```
def vowelinsertion(self,word):
    vowels = ["a","e","i","o","u"]
    word = list(word)
    for idx, l in enumerate(word):
        if type(l) == list:
            pass
        elif l in vowels:
            word[idx] = list(vowels)
    for p in product(*word):
            yield ''.join(p)

def list_voweladded(self,word):
    x=[]
    for j in self.vowelinsertion(word):
        x.append(j)
    return x
```

EXTENSIONS

1)Both reduction and vowel

if we have the word "cunspirrancy" then it will convert it to conspirancy(which is the correct word)

```
def both_reduction_vowel(self,word):
    x=[]
    for j in self.both(word):
        x.append(j)
    return x

def both(self,word):
    for reduction in self.reductions(word):
        for variant in self.vowelinsertion(reduction):
            yield variant
```

REFERENCES

1) Download the data set from here
https://norvig.com/big.txt

2) https://norvig.com/spell-correct.html

3)
https://amunategui.github.io/peter_norvig_magic_spell_checker/index.html