Running with Data

FUN WITH DATA, MUSIC, AND VISUALIZATION

‘W’ Considered Harmful

Not the magazine and not even the former president. But the letter ‘W’ itself. The letter 'W’, 23rd in the English alphabet, is unique in two ways: it is the only letter whose name is more than one syllable, and also the only letter whose name doesn’t include the sound it makes.

The fact that 'W’ takes 3 syllables to say bothers me. Even Wikipedia’s entry on 'W’ points out, twice, that the abbreviation *www* requires *nine* syllables to say. Crazy. So I wondered, how often is it the case that words that start with W (hereafter W-words) have fewer syllables than the letter W (double-yew)?

Syllabification in general is a hard problem in English, but fortunately I don’t have to solve it. The Carnegie Mellon University (CMU) Pronouncing Dictionary provides the pronunciations for over 125,000 words. I say pronunciations, plural, because words can be pronounced in a variety of different ways (e.g. *fire* can be pronounced to rhyme with *higher*, or in a single syllable. Only 41 W-words in the CMU dict have pronunciations with different numbers of syllables). Using the CMU Pronouncing Dictionary, it’s possible to count syllables in a word in a short (if cryptic) Python function, courtesy of Jordan Boyd-Graber – I found it on the nltk-users google group:

```
from curses.ascii import isdigit
from nltk.corpus import cmudict

d = cmudict.dict() # get the CMU Pronouncing Dict

def nsyl(word):
    """return the max syllable count in the case of multiple pronunciations"""
    return max([len([y for y in x if isdigit(y[-1])]) for x in d[word.lower()]])
```

So, now that we’ve got a syllable counter, let’s get all the W-words in the CMU dictionary, and see what the syllable distribution looks like.

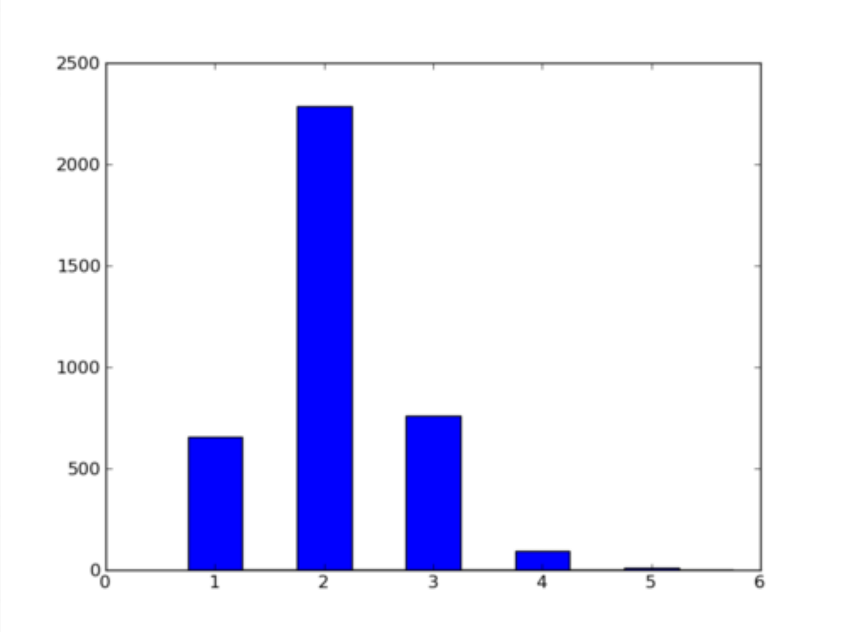
```
import pylab

w_words = dict([(w, nsyl(w)) for w in d.keys() if w[0] == 'w'])
worth_abbreviating = [(k,v) for (k,v) in w_words.iteritems() if v > 3]
pylab.hist(w_words.values())
```

About Jason Sundram

Elsewhere

twitter: @jsundram
About me



Only 101 W-words in the CMU dictionary (of 3805 total W-words) have more than 3 syllables. That's 2.6%. Here's a sampling of the words where using W to abbreviate them actually *saves* syllables: wagnerian, wallpapering, washingtonians, weatherperson, workaholic. So, by all means, call a meeting of the Wagnerian Wallpapering Workaholic Weatherperson Washingtonians the WWWWW. It will save time. Otherwise, consider not using an abbreviation. Or looking for synonyms.

Suggestions for further work:

- Take the data from Google n-grams viewer and count syllables for W-words using nltk's ~90%-accurate syllabification code.
- Get a list of acronyms (maybe netlingo) and see how many of them require more syllables to say than the phrase they stand for.

Monday, February 28, 2011 — 26 notes ()

Privacy Badger has replaced this Disqus widget

Allow once

Always allow on this site

edmure liked this

coltuldescris-blog reblogged this from runningwithdata

coltuldescris-blog liked this

vikdutt liked this

tophtucker liked this

czumikakooo-blog liked this

wakamiii-blog liked this

amigossso-blog liked this

gaussianmixture liked this

ashalynd reblogged this from runningwithdata

semicolons liked this

ryanberdeen liked this

yiah liked this

runningwithdata posted this

← Previous post Next post →

Search