**WEIGHTING LEXICON FILES**

**(20.06.17)**

**2 sources of lemmas:**

**List A**:

frequency list retrieved from corpus 'rrc-x'.

For more information visit <http://corpus.leeds.ac.uk/list.html>

The frequency list is available at: <http://corpus.leeds.ac.uk/serge/frqlist/rnc-modern-lpos.num.html>

The frequency distribution for attribute 'lemma' in corpus 'rrc-x'

- corpus size: 109115810 tokens

- lexicon size: 576591 types

**List B**:

lemmas and their frequencies from Sharoff’s 2008 frequency list

corpus size - 16336972 words

lexicon size - 32617 types

**Assign weights to other files:**

abbreviations.lexc: List A, tag Ncmsvn

conjunctions.lexc: List A, tag C

interjections.lexc: List A, tag I

prepositions.lexc: List A, tag Sc

propernouns.lexc: List A, tag Np

punctuation.lexc: ??

subjunctions.lexc: List A, empty

adverb.lexc: List B

other.lexc: ?? pronouns (tag P), determiners, particles (tag Q)

numerals.lexc: List A, tag Mc, Mo

nouns.lexc: List B

adjectives.lexc: List B

verbs.lexc: List B

**Reasons**:

For adverbs, nouns, adjectives and verbs, the number of lemmas is higher in **List B** than in **List A**, cf.:

|  |  |  |
| --- | --- | --- |
| tag | sharoff 2008 freq list, Nb lemmas **list B** | 'rrc-x' freq list, Nb lemmas  **list A** |
| verb | 9421 | 6709 |
| noun | 15177 | 14842 |
| adv | 1510 | 1062 |
| pron | 67 | 508 |

Weights to parts of speech:

1. reorder the lemma list: lemma -> weight -> tag
2. extract only instances of C tag and save it to a separate file:
   1. conjunctions: save to a file: rnc-modern-lpos.num.ord.weight.conj.csv

none

* 1. abbreviations: rnc-modern-lpos.num.ord.weight.abb.csv

none

* 1. interjections: rnc-modern-lpos.num.ord.weight.int.csv

none

* 1. **prepositions**:

$ cat rnc-modern-lpos.num.ord.weight.prep.csv | grep -E "\b,Sc" | sort | cut -d',' -f1 | uniq -c | grep -v '^ \*1 '

rnc-modern-lpos.num.ord.weight.prep.csv, duplicate forms: в, на, с, во, из-за, надо

* 1. **propernouns**:

cat rnc-modern-lpos.num.ord.weight.csv | grep -E "\b,Np" | sort | cut -d',' -f1 | uniq -c | grep -v '^ \*1 '

**output**: ли, вид, борис, сергей, андрей, ельцин, пушкин, василий, александр, александр, мейерхольд, маяковский, черномырдин, достоевский, ходорковский, станиславский

* 1. **numerals**:

**$** cat rnc-modern-lpos.num.ord.weight.numo.csv | grep -E "\b,Mo" | cut -d',' -f1 | sort | uniq -d

**$**

**Mc**: (rnc-modern-lpos.num.ord.weight.numc.csv)

00, 01, 02, 03, 04, 07, 1, 10, 12, 2, 2004, 2005, два, оба, сто, три, двое, мало, один, пять, семь, трое, много, сорок, шесть, восемь, двести, девять, десять, немало, триста, тысяча, четыре, немного, полтора, пятьсот, сколько, столько, четверо, двадцать, тридцать, девяносто, несколько, пятьдесят, семьдесят, двенадцать, пятнадцать, семнадцать, тринадцать, шестьдесят, восемьдесят, одиннадцать, шестнадцать, восемнадцать, четырнадцать

**Mo**: (rnc-modern-lpos.num.ord.weight.numo.csv)

е 3083, го 6921, один 13112, пятый 3873, второй 46061, первый 113630, третий 21978, шестой 2370, восьмой 964, девятый 1288, седьмой 2488, четвертый 3031

* 1. **pronouns**: rnc-modern-lpos.num.ord.weight.pron.csv

cat rnc-modern-lpos.num.ord.weight.pron.csv |grep -E "\b,P" |  cut -d',' -f1 | sort | uniq -c | grep -v '^ \*1

* + 1. **duplicates**: этот, тот, весь, такой, один, который, другой, мой, наш, любой, ваш, некоторый, твой, некий
    2. **duplicates**: т, я, Вы, вы, ее, их, мы, он, то, ты, че, ваш, все, его, кто, мой, наш, она, они, оно, сам, сей, тот, чей, что, это, кто-то, тот-то, чей-то, что-то, кое-что, кто-либо, кое-какой, кто-нибудь, что-нибудь, весь, иной, один, свое, свой, себя, твой, этот, каков, какой, любой, некий, нечто, никто, ничто, самый, таков, такой, какой-то, какой-либо, какой-нибудь, всякий, данный, другой, каждый, многие, многий, многое, нечего, прочий, который, никакой, таковой, некоторые, некоторый, остальной
  1. particles: rnc-modern-lpos.num.ord.weight.part.csv

none

* 1. adverbs

**Duplicate forms**

1. Check if there are duplicate forms, e.g.:

cat rnc-modern-lpos.num.ord.weight.csv | grep -E "\b,Q" | cut -d',' -f1 | uniq -d

this set of commands checks if there are duplicates, if yes, it displays them.

1. If there are duplicate forms, convert them into one form and sum their weights, e.g.:

rnc-modern-lpos.num.ord.weight.propn.csv

пушкин, 2570,Npmsny

пушкин, 2552,Npmsgy

Пушкин, 545,Npm--y

пушкин, 430,Npmsay

пушкин, 386,Npmsdy

Пушкин, 345,Npmsny

пушкин, 278,Npmsiy

go to the source file

lst=[2570, 2552, 545, 430, 386, 345, 278]

corpusSizePm=109115810.0/1000000.0 #corpus size per million

typesNbPm=576591.0/1000000.0 #number of types per million

s = sum(lst)

perMillionFreqSmooth=(float(s)/1000000.0)+1.0

weight=math.log((perMillionFreqSmooth/(corpusSizePm + typesNbPm)), 10) + 3.0

print(weight)

0.9628986407230835

==> final weight for пушкин is 0.9628986407230835

1. replace all the instances of пушкин with one instance with the weight of 0.9628986407230835, e.g.:

пушкин, 0.9628986407230835, Npm

борис:

8742 борис Npmsny

2255 борис Npmsgy

765 борис Npfsny

760 борис Npmsdy

715 борис Npmsiy

601 борис Npmsay

=> weight is 0.9657920224954193

борис, 0.9657920224954193, Np

1. rnc-modern-lpos.num.ord.weight.prep.csv

в

1674758 в Sc-l

787198 в Sc-a

187563 в Sc

10658 в Sc-n

4854 В Sc

335 в Sc-g

>>> lst=[1674758, 787198, 187563, 10658, 4854, 335]

>>> corpusSizePm=109115810.0/1000000.0 #corpus size per million

>>> typesNbPm=576591.0/1000000.0 #number of types per million

>>> s = sum(lst)

>>> perMillionFreqSmooth=(float(s)/1000000.0)+1.0

>>> weight=math.log((perMillionFreqSmooth/(corpusSizePm + typesNbPm)), 10) + 3.0

>>> print(weight)

1.5239408043777625

=> в, 1.5239408043777625, Sc

на

686372 на Sc-a

557107 на Sc-l

16430 на Sc

2078 на Sc-n

lst=[686372, 557107, 16430, 2078]

>>> s = sum(lst)

>>> perMillionFreqSmooth=(float(s)/1000000.0)+1.0

>>> weight=math.log((perMillionFreqSmooth/(corpusSizePm + typesNbPm)), 10) + 3.0

>>> print(weight)

1.3143135620037047

=>

на, 1.3143135620037047, Sc

**Numerals**:

8110 первый Mofsn

6977 первый Momsi

6180 первый Mofsg

5160 первый Mo-pg

4970 первый Momsl

4762 первый Mo-pa

4365 первый Monsn

3995 первый Mo-pn

3131 первый Monsa

3080 первый Mofsl

2288 первый Mo-s

2242 первый Mofsi

2024 первый Moms

1845 первый Monsl

1742 первый Mo-pi

1666 первый Mo-pl

1479 первый Monsi

1321 первый Monsg

1234 первый Momsd

1215 первый Mo-p

513 первый Mofsd

311 первый Mo-pd

=> 68610 первый Mo