Text Classification

Aman Berhe

May 13, 2016

1 Visualizations

the following figures show the visualization of both Pompe diseases and Amytrophic Lateral Diseases for 60 articles each

1.1 Amytrophic Lateral Disease

The data collected has been preprocessed.the "¡unknown¿" lemma has been replaced by the word and also the different punctuation symbols has also been removed since they do not make any sense.

The Dimention of the Positive and negative words:

Table 1: Amytrophic Lateral Disease Dimension

negative	positive	negative
Words	2295	9774
vector	1:only lemmas	1:only lemmas

Table 2: Pompe Disease Dimension

	positive	negative
Words	3291	9247
vector	1:only lemmas	1:only lemmas

Here is the Item Frequency of the words(Items)

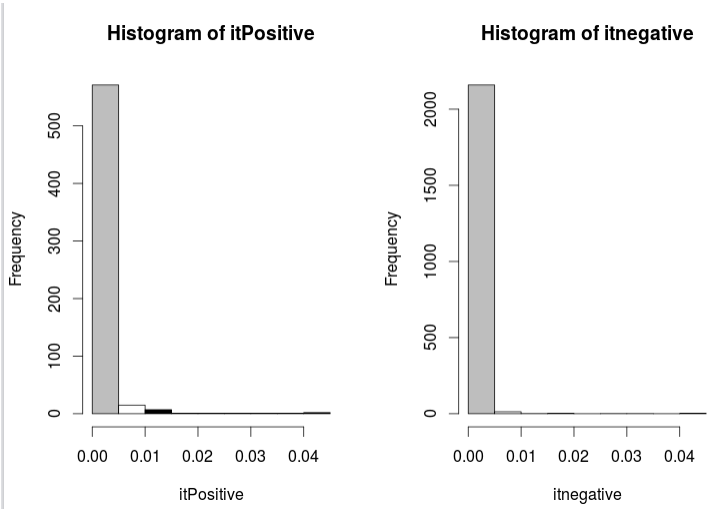
positive

0.021270131	0.000303859	0.003342449	0.000303859
V1=accompany	V1=accumulation	V1=acid	V1=action
0.002127013	0.001519295	0.001823154	0.000607718
V1=activation	V1=activity	V1=acute	V1=additional
0.000303859	0.001519295	0.001823154	0.000911577
V1=adductor	V1=adjunctive	V1=adolescent-onset	V1=adult
0.000303859	0.000303859	0.000303859	0.001215436
V1=adulthood	V1=affect	V1=age	V1=aged-matched
0.000303859	0.000911577	0.001519295	0.000911577
V1=agonist	V1=aim	V1=aldehyde	V1=all
0.000303859	0.000911577	0.000303859	0.000607718
V1=All	V1=allow	V1=alone.Propranolol-treated	V1=alpha-glucosidase
0.000303859	0.000607718	0.000607718	0.000607718
V1=ALS	V1=also	V1=alteration	V1=although
0.000911577	0.000911577	0.000911577	0.000607718
V1=.Although	V1=ambulation	V1=aminotransferase	V1=amyotrophic
0.000303859	0.000607718	0.000303859	0.000911577
V1=an	V1=analyses.Enzyme	V1=analysis	V1=analysis.We
0.003038590	0.000303859	0.000303859	0.000911577
V1=analyze	V1=and	V1=and/or	V1=aneurysm
0.000303859	0.043451838	0.000303859	0.001215436
V1=angiography	V1=antero-posterior	V1=appear	V1=arrhythmia

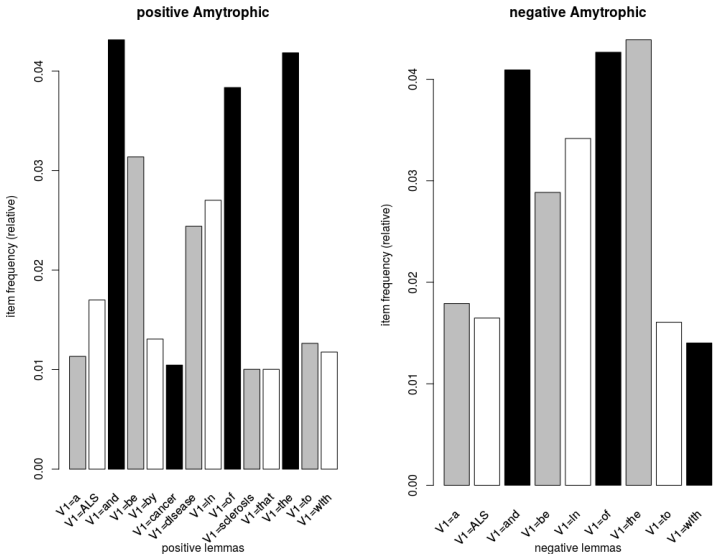
negative

V1=differentiation	V1=diminish	V1=dioxide
0.0002162864	0.0001081432	0.0001081432
V1=diphosphate	V1=direct	V1=disappearance
0.0001081432	0.0002162864	0.0001081432
V1=disaster	V1=discharge	V1=discordant
0.0001081432	0.0001081432	0.0001081432
V1=discrete	V1=discriminate	V1=discuss
0.0001081432	0.0001081432	0.0003244295
V1=disease	V1=disease)	V1=Disease
0.0187087704	0.0001081432	0.0023791500
V1=DISEASE)	V1=disease.In	V1=Diseases
0.0001081432	0.0001081432	0.0002162864
V1=disease-specific	V1=dish	V1=disorder
0.0001081432	0.0001081432	0.0037850114
V1=distance	V1=distinct	V1=distribution
0.0002162864	0.0001081432	0.0002162864
V1=diurnal	V1=divide	V1=Division
0.0001081432	0.0002162864	0.0001081432
V1=DNA	V1=do	V1=document
0.0002162864	0.00064488591	0.0002162864
V1=Dongjin	V1=dorsal	V1=dose
0.0001081432	0.0001081432	0.0015140045

The Histogram of the words and their frequencies

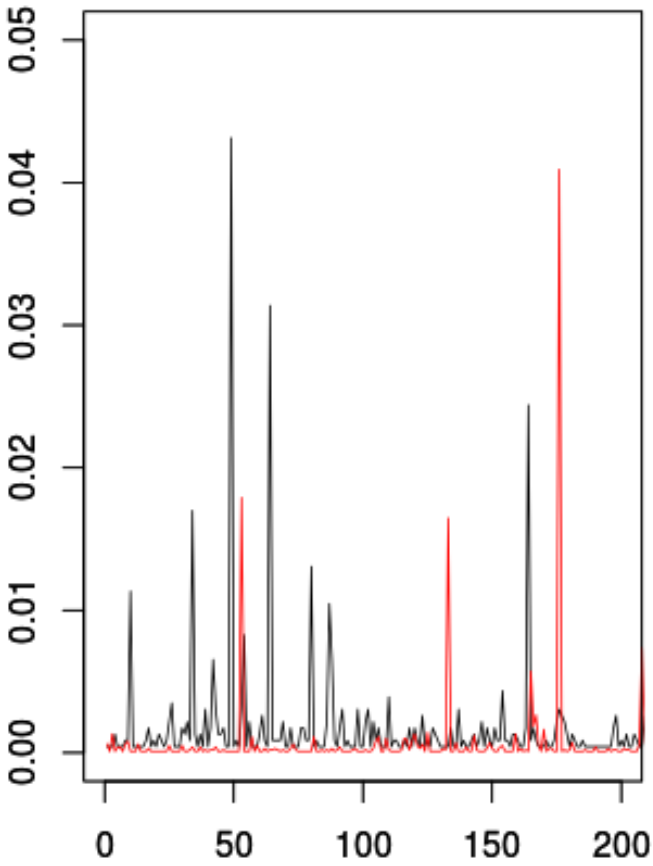


ItemFrequency Plot Showing the words along the histogram



on the positive words the 'and' word is most frequent and in the negative the 'the' is more frequent

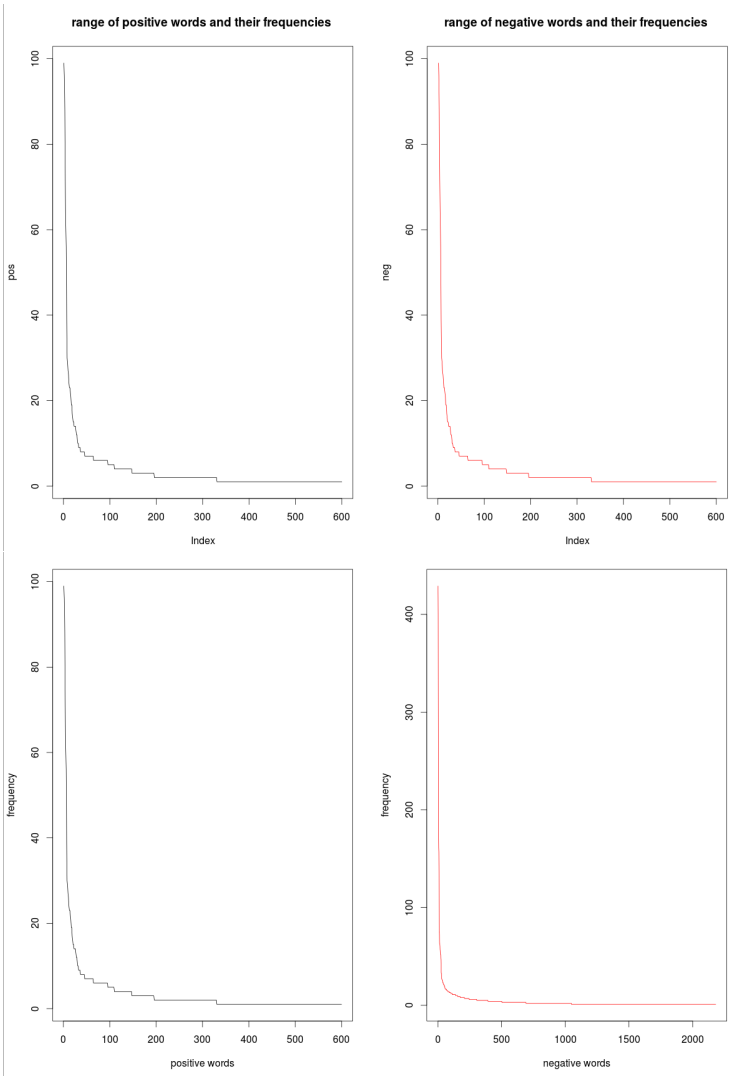
this figure shows the combination of the two, as where they occur



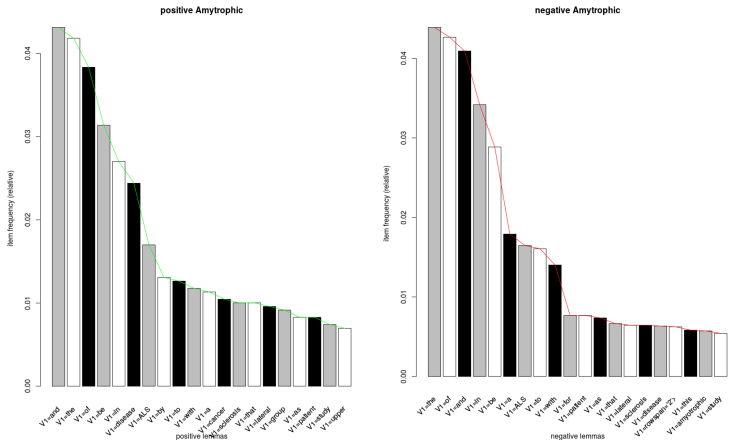
the following figure shows some example of the words and their number of occurrence for the positive words

and	the	of
99	96	88
be	in	disease
72	62	56
ALS	by	to
39	30	29
with	a	cancer
27	26	24
sclerosis	that	lateral
23	23	22
group	as	patient
21	19	19
study	upper	amyotrophic
17	16	15
@card@	motor	neurodegenerative
15	14	14
neuron	these	or
14	14	13
reactive	specie	include
12	12	11
diabetes	from	common
10	10	9

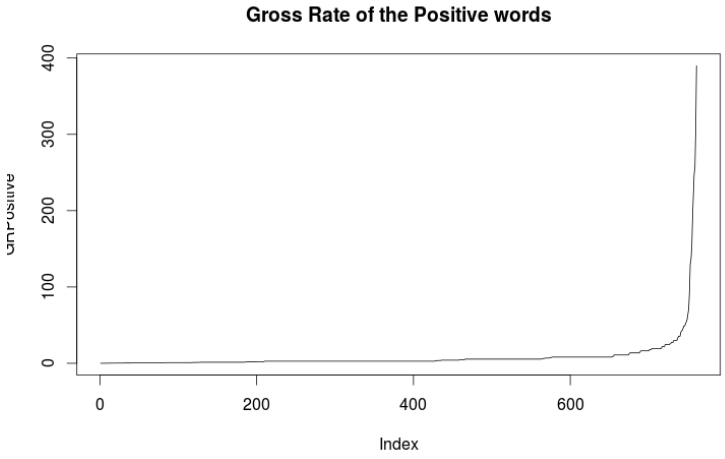
The following two images shows the number of occurrence and number of words that occupy that amount of frequency which is the same as the zipf's law

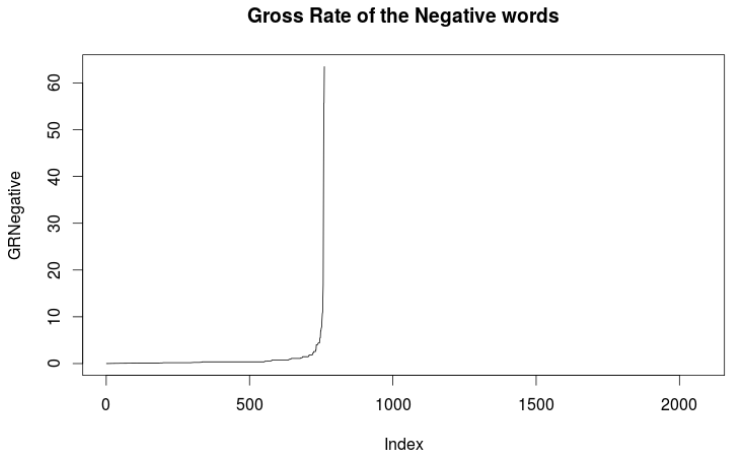


the following figure shows the Zipf's law on the histogram of top 20 words according to their occurrence



Gross Rate has been computed using the number of occurrence and total number of words. Here are the Graphs





2 Accepting Sentence

using the prompt I accept a sentence to be classified then I lemmatize it. E.eg

classify<-readSentence()

Enter your sentence:

Enzyme replacement therapy (ERT) is important for the treatment of lysosomal storage disorders. Hypersensitivity reactions with ERT have been reported, and in these cases, desensitisation with the enzyme is necessary. Here we report the cases of 3 patients with lysosomal storage disorders, including Pompe disease and mucopolysaccharidosis type I and VI, who had IgE-mediated hypersensitivity reactions and positive skin tests. Successful desensitisation protocols with the culprit enzyme solution were used for these patients. All 3 patients were able to safely receive ERT with the desensitisation protocol.

>cassifyTagged<-taggedText(classify)

The out put looks like this

token	tag	lemma	lttr	wclass	desc	stop	stem
Enzyme	NN	enzyme	6	noun	Noun, singular or mass	*NA*	*NA*
replacement	NN	replacement	11	noun	Noun, singular or mass	*NA*	*NA*
therapy	NN	therapy	7	noun	Noun, singular or mass	*NA*	*NA*
(	(	(	1	punctuation	Opening bracket	*NA*	*NA*
ERT	NN	ERT	3	noun	Noun, singular or mass	*NA*	*NA*
)	)	)	1	punctuation	Closing bracket	*NA*	*NA*
is	VBZ	be	2	verb	Verb, 3rd person singular present of "to be"	*NA*	*NA*
important	JJ	important	9	adjective	Adjective	*NA*	*NA*
for	IN	for	3	preposition	Preposition or subordinating conjunction	*NA*	*NA*
the	DT	the	3	determiner	Determiner	*NA*	*NA*
treatment	NN	treatment	9	noun	Noun, singular or mass	*NA*	*NA*
of	IN	of	2	preposition	Preposition or subordinating conjunction	*NA*	*NA*
lysosomal	JJ	lysosomal	9	adjective	Adjective	*NA*	*NA*
storage	NN	storage	7	noun	Noun, singular or mass	*NA*	*NA*
disorders	NNS	disorder	9	noun	Noun, plural	*NA*	*NA*
.	SENT	.	1	fullstop	Sentence ending punctuation	*NA*	*NA*
Hypersensitivity	NN	hypersensitivity	16	noun	Noun, singular or mass	*NA*	*NA*
reactions	NNS	reaction	9	noun	Noun, plural	*NA*	*NA*
with	IN	with	4	preposition	Preposition or subordinating conjunction	*NA*	*NA*

here is the Graph