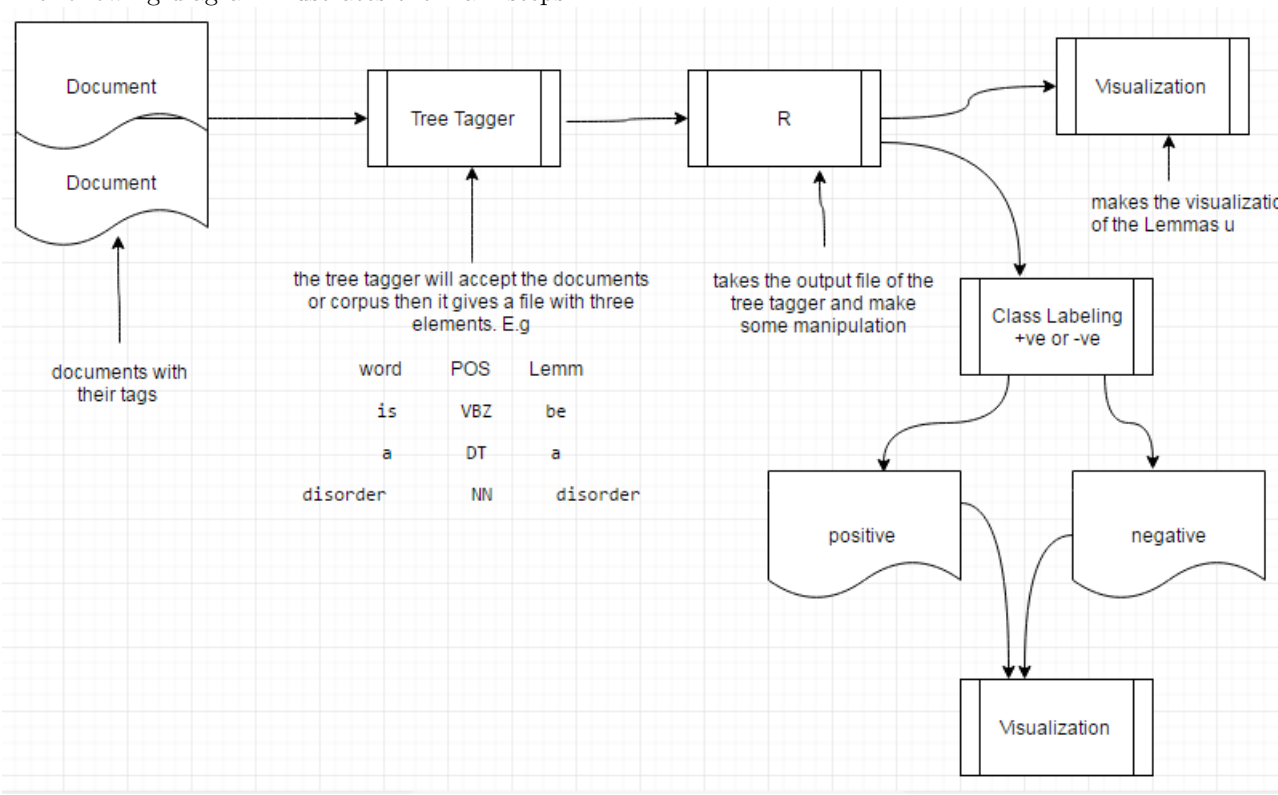
Text Classification

Aman Berhe

April 28, 2016

1 Data Preprocessing

The following diagram illustrates the main steps



As shown in the picture:

1. Documents

   Documents are collected from the website from the page source and then fed to the tree tagger.Currently, I have collected the data for pompe disease of 10 and 30 articles.

2. Tree Tagger

   The tree tagger takes the collected documents and gives an out put file which is three parts: which are word, POS, lemma and the command used was

   cmd/tree-tagger-english 'pompeDatawithtag30' tee /outputfile2.txt

E.g: part of output file

word	POS	lemma
is	VBZ	be
a	DT	a
lysosomal	JJ	lysosomal
storage	NN	storage
disorder	NN	disorder
caused	VBN	cause
by	IN	by
a	DT	a
deficiency	NN	deficiency
of	IN	of
the	DT	the
enzyme	NN	enzyme
acid	NN	acid
alpha-glucosidase	NN	<unknown>
.	SENT	.

3. Data Manipulation using R The output of tree tagger will be used, we will use only the lemmas of the file and visualize the item frequencies.
   It is done using the itemfrequencyplot(): which is a function inside the rpart library in R. as well using the item frequency: itemfrequency() and histogram: hist().

4. Labeling Words

   The labeling of the sentences into positive and negative is done using SENT POS which indicates the end of a sentence and "" which indicates a sentence has a word which is a symptom of a disease.