Reducing Vocabulary Size

School of Information Studies
Syracuse University

Approaches for Reducing Vocabulary Size

Stemming

Case merging

Removing stop words

Word clustering

Stemming

Can be used in inflected languages like English

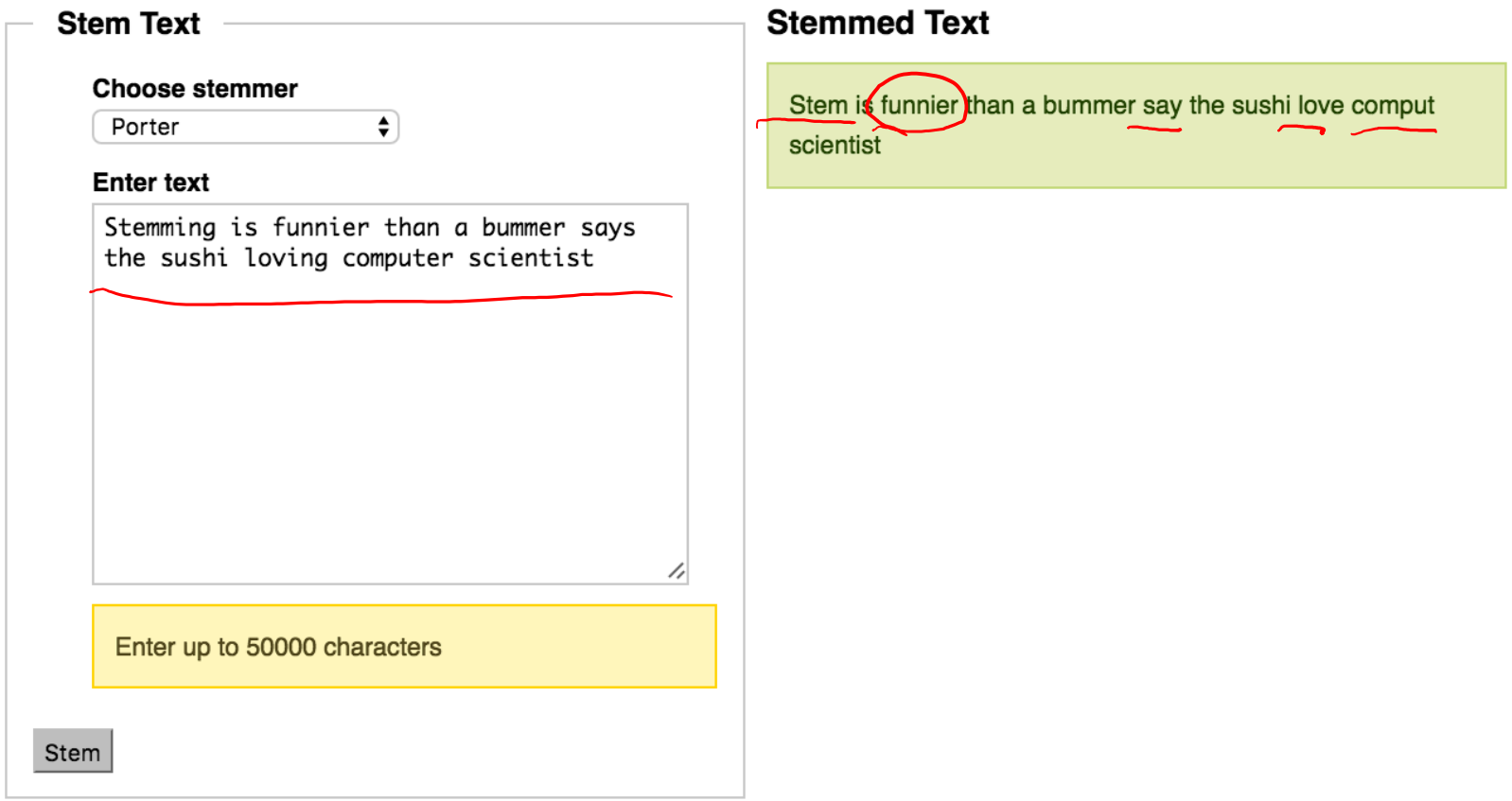
Stemmer: remove postfixes to find the root form

- "Applied" and "application" -> "appli"

Lemmatizer: transform the root to a real word

- "Applied" and "application" -> apply

NLTK Stemming Demo



Stemming Issues

How far should it go?

- “denormalization” -> denormalize -> denormal -> normal -> norm?

How accurate can it be?

- “bore”/ “He wanted to bore a hole” / “He bore the students on his heart”

How Useful Is Stemming?

No consistent conclusion

Information retrieval

- Search “dishwasher” to know how it works
- Search “dishwashers” to shop around

Text categorization

- Future tense vs. past tense in company performance report
  - “Will do” vs. “have done”

Convert Uppercase to Lowercase?

Emily Dickinson's poem

- "Joy" vs. "joy"
- "Love" vs. "love"

Uppercase

But pompous
Joy
Betrays us, as
his first
Betrothal
Betrays a
Boy.

The Treason
of an Accent
Might vilify
the Joy -
To breathe -
corrode the
rapture
Of Sanctity
to be

Boundlessness -
Expanse cannot
be lost -
Not Joy, but
a Decree
Is Deity -
His Scene,
Infinity -

Lowercase

Could she have guessed that it would be - Could but a Crier of the joy Have climbed the distant hill! -	I want to send you joy, I have half a mind to put up one of these dear little Robin's, and . . .	I cant believe you are coming - but when I think of it, and tell myself it's so, a wondrous joy comes over me, and my old fashioned life . . .

Remove Stop Words

Observation: words that occur in most documents are not useful of distinguishing documents

Stop words are usually function words that bear no specific meaning, compared to content words

Example of the Start of a Stop Word List

a
about
across
after
afterwards
again
against
all
alone
along
already
also
always
amongst
an
and
another
any
anyhow
anyone
anything
are
around
as
be
because
becomes
becoming
been
before
beforehand
behind
being
below
besides
between
beyond
but
can