REDUCING VOCABULARY SIZE

SYRACUSE UNIVERSITY
School of Information Studies

APPROACHES TO REDUCE THE VOCABULARY SIZE

Stemming

Case merging

Removing stop words

Word clustering

STEMMING

Characteristic of inflected language like English

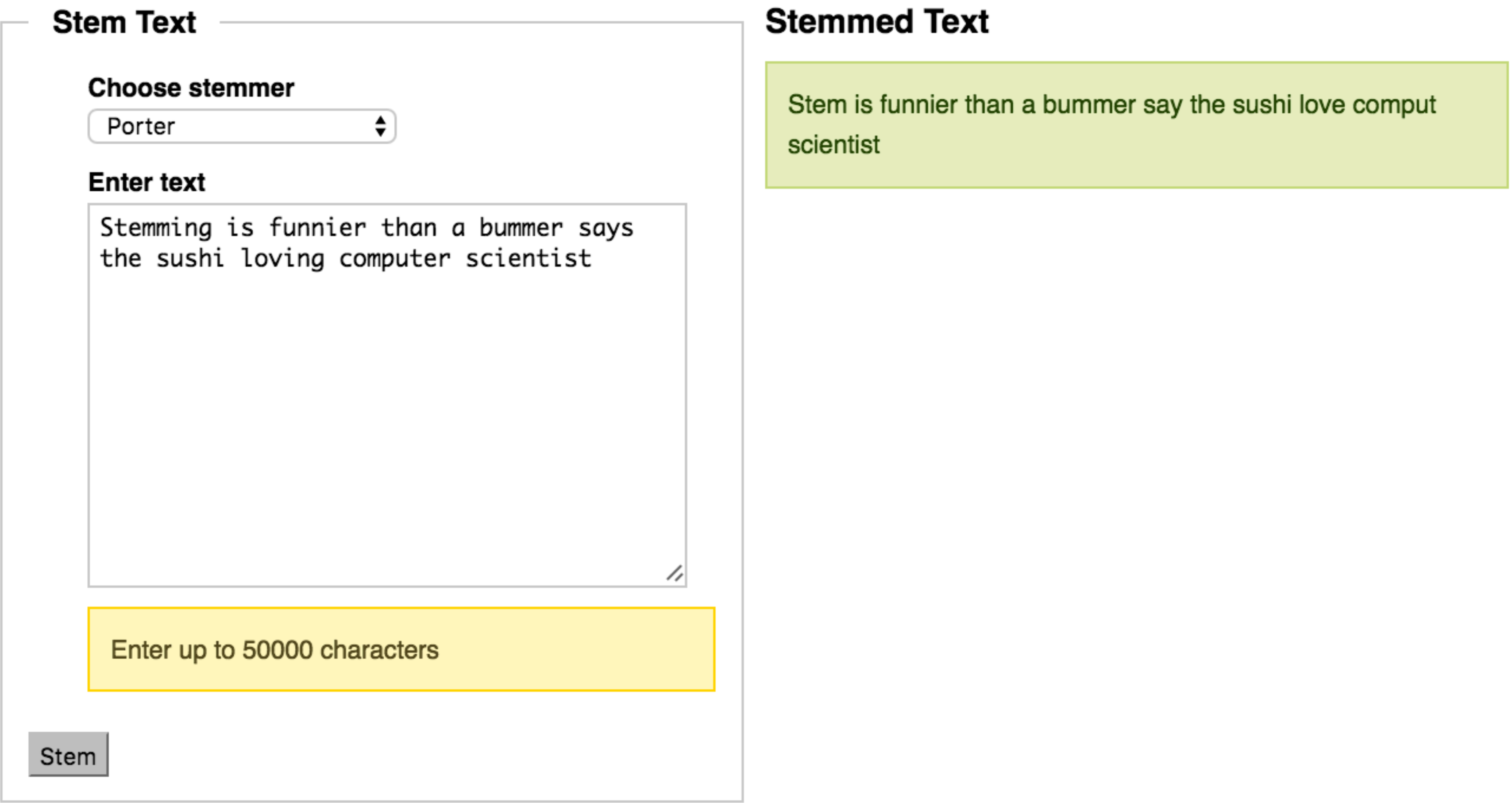
Stemmer: Remove postfixes to find the root form

“applied” and “application” -> “appli”

Lemmatizer: Transform the root to a real word

“applied” and “application” -> apply

NLTK STEMMING DEMO



Stem Text

Choose stemmer: Porter

Enter text

```
Stemming is funnier than a bummer says
the sushi loving computer scientist
```

Enter up to 50000 characters

Stemmed Text

Stem is funnier than a bummer say the sushi love comput scientist

http://text-processing.com/demo/stem/

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 can't 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 occur in most documents that are not useful for 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

among

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