A Natural Language URL-Shortener

David Branner
Hack and Tell
eBay, New York
20140610

URL shorteners and their benefits are well known –

URL shorteners and their benefits are well known –

create very short pointer to a long URL:

URL shorteners and their benefits are well known –

create very short pointer to a long URL:

http://bit.ly/TGlQtH

URL shorteners and their benefits are well known —

create very short pointer to a long URL:

http://bit.ly/TGlQtH

➔ *https://www.hackerschool.com/manual#sec-history*

URL shorteners and their benefits are well known —

create very short pointer to a long URL:

http://bit.ly/TGlQtH

➔ *https://www.hackerschool.com/manual#sec-history*

less susceptible to corruption in transmission

URL shorteners and their benefits are well known —

create very short pointer to a long URL:

http://bit.ly/TGlQtH

➔ *https://www.hackerschool.com/manual#sec-history*

less susceptible to corruption in transmission
concealment of final destination

URL shorteners and their benefits are well known —

create very short pointer to a long URL:

http://bit.ly/TGlQtH

➔ *https://www.hackerschool.com/manual#sec-history*

less susceptible to corruption in transmission
concealment of final destination
can be used to track visitors to the main site

key element: variable part of a shortened URL (“path”)

key element: variable part of a shortened URL (“path”)

- http://7.ly/jMau
- http://2.gp/zkSE
- http://qr.net/Bozx
- http://bit.ly/TGlQtH
- http://x.vu/KC4s4e
- http://ow.ly/xQNbx
- https://t.co/ZfUR4euiph

key element: variable part of a shortened URL ("path")

- http://7.ly/jMau
- http://2.gp/zkSE
- http://qr.net/Bozx
- http://bit.ly/TGlQtH
- http://x.vu/KC4s4e
- http://ow.ly/xQNbx
- https://t.co/ZfUR4euiph

ugglesome in the extreme

key element: variable part of a shortened URL ("path")

- http://7.ly/jMau
- http://2.gp/zkSE
- http://qr.net/Bozx
- http://bit.ly/TGlQtH
- http://x.vu/KC4s4e
- http://ow.ly/xQNbx
- https://t.co/ZfUR4euiph

ugglesome in the extreme
compressed, hence rarely readable

key element: variable part of a shortened URL ("path")

- http://7.ly/jMau
- http://2.gp/zkSE
- http://qr.net/Bozx
- http://bit.ly/TGlQtH
- http://x.vu/KC4s4e
- http://ow.ly/xQNbx
- https://t.co/ZfUR4euiph

ugglesome in the extreme
compressed, hence rarely readable
easy to remember?

"custom" shortened URLs can be easier to remember
because they rely on natural language ability

"custom" shortened URLs can be easier to remember
because they rely on natural language ability

http://bit.ly/HStory

"custom" shortened URLs can be easier to remember
because they rely on natural language ability

http://bit.ly/HStory

➔ *https://www.hackerschool.com/manual#sec-history*

"custom" shortened URLs can be easier to remember
because they rely on natural language ability

http://bit.ly/HStory

➔ *https://www.hackerschool.com/manual#sec-history*

easier to remember

"custom" shortened URLs can be easier to remember
because they rely on natural language ability

http://bit.ly/HStory

➔ *https://www.hackerschool.com/manual#sec-history*

easier to remember
short only if you get there first (tend to be long)

Why not have URL pointers that are always *both* readable *and* very short?

Why not have URL pointers that are always
both readable *and* very short?

(Talon-sharpening exercise at Hacker School recently.)

Why not have URL pointers that are always
both readable *and* very short?

(Talon-sharpening exercise at Hacker School recently.)

The trick to picking always-readable short URL-paths:

Why not have URL pointers that are always
both readable *and* very short?

(Talon-sharpening exercise at Hacker School recently.)

The trick to picking always-readable short URL-paths:

use the characters for the most common Chinese words.

Chinese characters

Chinese characters

In the current official “HSK” Chinese proficiency exam:

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

$|\{x : x \in \text{traditional}\}| = 2671$

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

$|\{x : x \in \text{traditional}\}| = 2671$

$|\{x : x \in \{\text{simp} \cap \text{trad}\}\}| = 1692$

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

$|\{x : x \in \text{traditional}\}| = 2671$

$|\{x : x \in \{\text{simp} \cap \text{trad}\}\}| = 1692$

compare to paired Roman letters (of either case):

$|\{x : x \in \{\text{upper} \cap \text{lower}\}\}| = 52 \quad (52^2 = 2704)$

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

$|\{x : x \in \text{traditional}\}| = 2671$

$|\{x : x \in \{\text{simp} \cap \text{trad}\}\}| = 1692$

compare to paired Roman letters (of either case):

$|\{x : x \in \{\text{upper} \cap \text{lower}\}\}| = 52 \quad (52^2 = 2704)$

kind	random 字	random Roman digraph
#	2635	2704

Chinese characters

In the current official "HSK" Chinese proficiency exam:

$|\{x : x \in \text{simplified}\}| = 2635$

$|\{x : x \in \text{traditional}\}| = 2671$

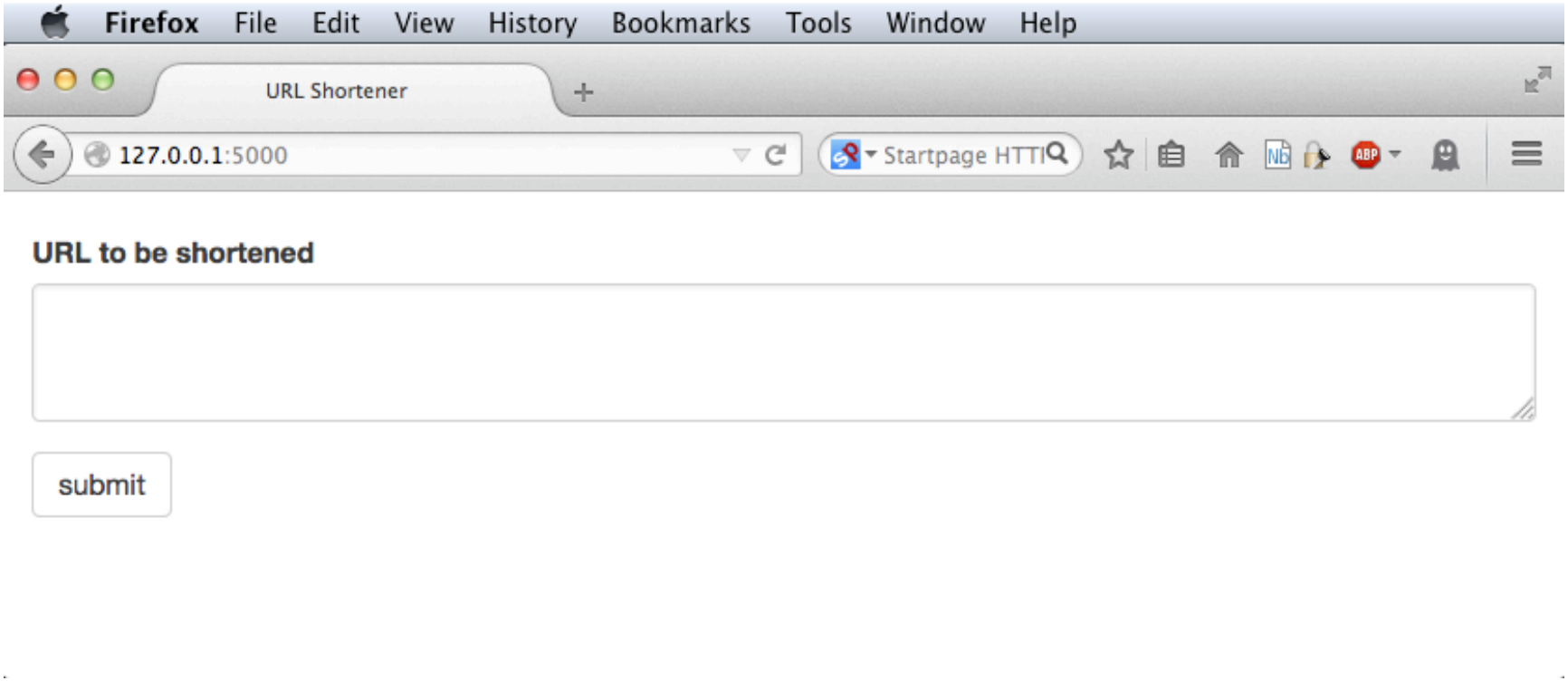
$|\{x : x \in \{\text{simp} \cap \text{trad}\}\}| = 1692$

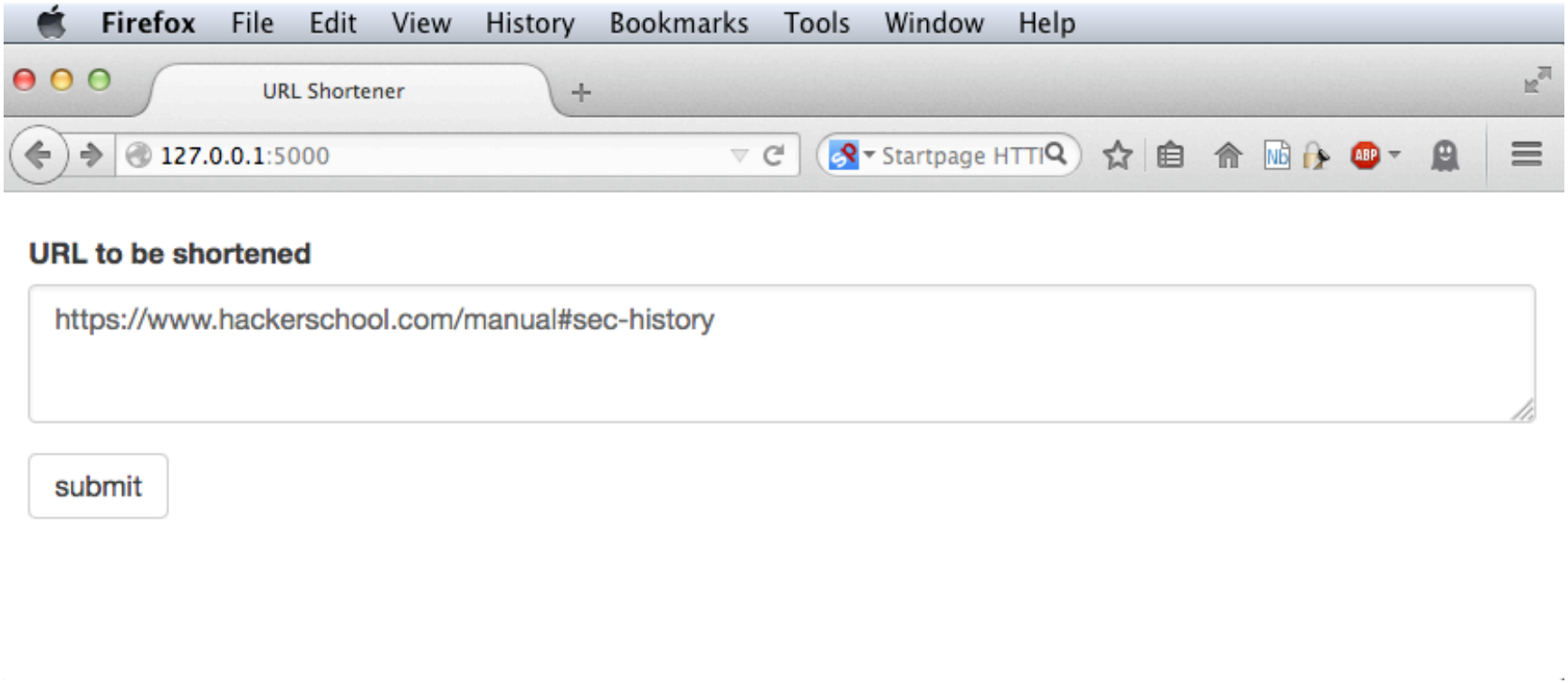
compare to paired Roman letters (of either case):

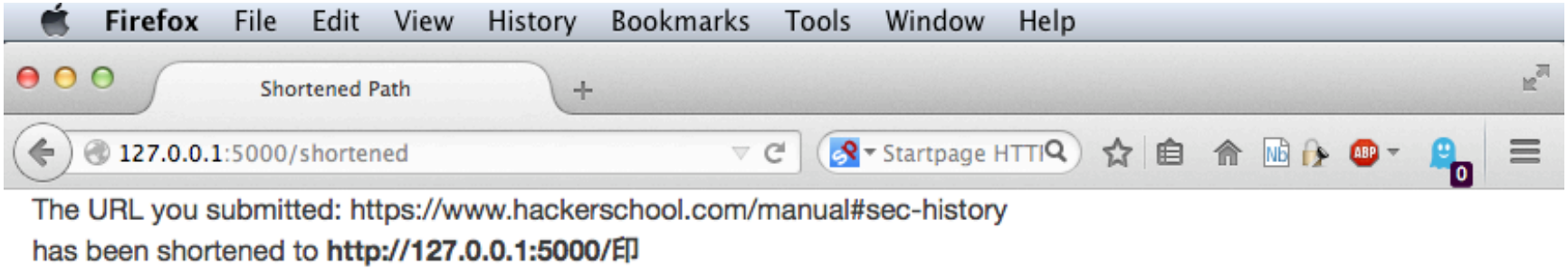
$|\{x : x \in \{\text{upper} \cap \text{lower}\}\}| = 52 \quad (52^2 = 2704)$

kind	random 字	random Roman digraph
#	2635	2704
readable?	guaranteed	probably not

Code for proof-of-concept on my public Git repository.




Firefox File Edit View History Bookmarks Tools Window Help
URL Shortener
127.0.0.1:5000
Startpage HTTI
URL to be shortened
https://www.hackerschool.com/manual#sec-history
submit


Firefox File Edit View History Bookmarks Tools Window Help
Shortened Path
127.0.0.1:5000/shortened
Startpage HTTI
The URL you submitted: https://www.hackerschool.com/manual#sec-history
has been shortened to http://127.0.0.1:5000/印

First c. 2650 URLs ➔ one character

http://127.0.0.1:5000/印

First c. 2650 URLs ➔ one character

http://127.0.0.1:5000/印

Next c. 2650 x 2650 = 7022500 URLs ➔ two characters

http://127.0.0.1:5000/厉吉

First c. 2650 URLs ➔ one character

http://127.0.0.1:5000/印

Next c. 2650 x 2650 = 7022500 URLs ➔ two characters

http://127.0.0.1:5000/厉吉

Next 18,609,625,000 URLs ➔ three characters

http://127.0.0.1:5000/天鼻歪

Always readable

Always readable (may not make sense, since random…)

印: ‘to print’; 厉吉: ‘to pass through auspiciousness’;

天鼻歪: ‘Heaven’s nose is crooked‘

Note:

Always readable (may not make sense, since random…)

印: ‘to print’; 厉吉: ‘to pass through auspiciousness’;

天鼻歪: ‘Heaven’s nose is crooked‘

Note: many custom shorteners allow Chinese characters:

http://bit.ly/史 (史: ‘history’)

Always readable (may not make sense, since random…)

印: ‘to print’; 厉吉: ‘to pass through auspiciousness’;

天鼻歪: ‘Heaven’s nose is off-center’

Note: many custom shorteners allow Chinese characters:

http://bit.ly/史 (史: ‘history’)

Also note:

Always readable (may not make sense, since random…)

印: ‘to print’; 厉吉: ‘to pass through auspiciousness’;

天鼻歪: ‘Heaven’s nose is crooked‘

Note: many custom shorteners allow Chinese characters:

http://bit.ly/史 (史: ‘history’)

Also note: the advantage of shortening to Chinese does not mean bandwidth savings:

Always readable (may not make sense, since random…)

印: ‘to print’; 厉吉: ‘to pass through auspiciousness’;

天鼻歪: ‘Heaven’s nose is crooked‘

Note: many custom shorteners allow Chinese characters:

http://bit.ly/史 (史: ‘history’)

Also note: the advantage of shortening to Chinese does not mean bandwidth savings:

http://bit.ly/史 may be sent from your browser as

http://bit.ly/%E5%8F%B2

One more thing to think about:

Someone commented that without knowing Chinese, she would find this system perhaps of limited use. Imagine!

An intermediate solution would be Korean Han'gŭl 한글, where fully pronounceable syllables are written in one-character-width glyphs composed of alphabetic subunits. There are 2100 recognized syllable-glyphs that can be built up of these sub-units by simple principles, although another 9000 are possible graphically. By no means all possible glyphs are meaningful, though. But learning the 40 subunits is a more manageable task than learning the two-plus thousand basic Chinese characters:

ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ

ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ .

劇
終