Text Encoding (utf-8)

Table of contents

Character encoding: What is it?

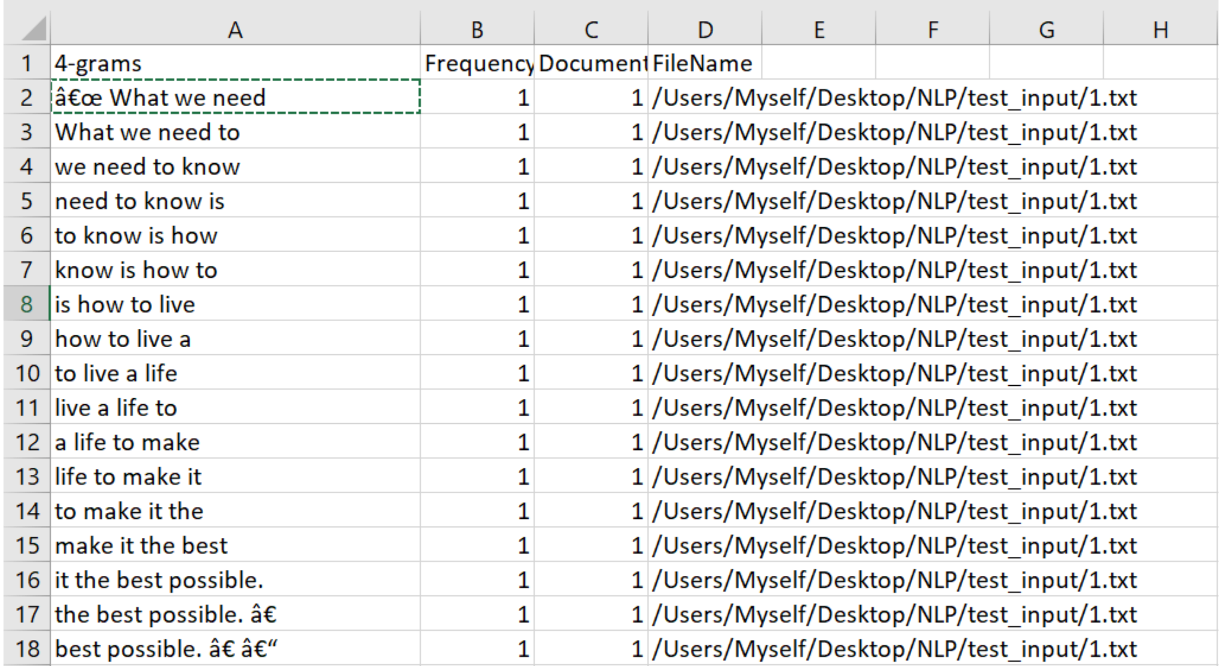
A character encoding is used in computation, data storage, and transmission of **textual data**.

From Wikipedia (https://en.wikipedia.org/wiki/Character_encoding) we read:

> Early character codes associated with the optical or electrical telegraph could only represent a subset of the characters used in written languages, sometimes restricted to upper case letters, numerals and some punctuation only. The low cost of digital representation of data in modern computer systems allows more elaborate character codes (such as Unicode) which represent most of the characters used in many written languages. Character encoding using internationally accepted standards permits worldwide interchange of text in electronic form.

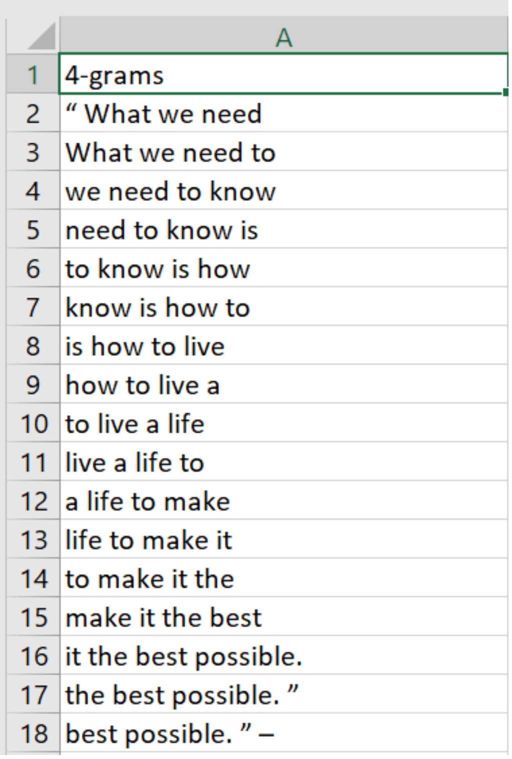
Weird characters in my csv file opened in Windows: Why do I get this?

When I open a csv file created by an NLP Suite Python script (e.g., n-grams) I get some weird characters, such as â€œ on my Windows laptop (not on a Mac). Such characters are not in the text file processed by the script in input.



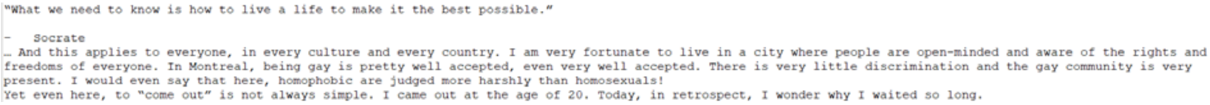
	A	B	C	D	E	F	G	H
1	4-grams	Frequency	Document	FileName				
2	â€œ What we need	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
3	What we need to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
4	we need to know	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
5	need to know is	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
6	to know is how	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
7	know is how to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
8	is how to live	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
9	how to live a	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
10	to live a life	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
11	live a life to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
12	a life to make	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
13	life to make it	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
14	to make it the	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
15	make it the best	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
16	it the best possible.	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
17	the best possible. â€	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
18	best possible. â€ â€“	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				

Oddly, when the same csv file is used to create an Excel chart, the weird character disappears



	A
1	4-grams
2	“ What we need
3	What we need to
4	we need to know
5	need to know is
6	to know is how
7	know is how to
8	is how to live
9	how to live a
10	to live a life
11	live a life to
12	a life to make
13	life to make it
14	to make it the
15	make it the best
16	it the best possible.
17	the best possible. ”
18	best possible. ” –

And the weird character is not in the input text file either.



```
“What we need to know is how to live a life to make it the best possible.”

-   Socrate
… And this applies to everyone, in every culture and every country. I am very fortunate to live in a city where people are open-minded and aware of the rights and
freedoms of everyone. In Montreal, being gay is pretty well accepted, even very well accepted. There is very little discrimination and the gay community is very
present. I would even say that here, homophobic are judged more harshly than homosexuals!
Yet even here, to “come out” is not always simple. I came out at the age of 20. Today, in retrospect, I wonder why I waited so long.
```

The weird character â€œ seems to be a csv display of the character “.

while quotation marks and apostrophes seem innocuous, they can cause problems when the CSV is opened, as opposed to imported into Excel, partly because quotation marks or other symbols such as apostrophes come in different forms. Some are curly “, some are straight ". It can depend on what application generated the characters to begin with. It can be hard to explain exactly why Excel has an issue with some characters but read on for some solutions.
A **solution** is to replace all curly quote marks to straight ones before saving a csv file. (https://help.shotfarm.com/hc/en-us/articles/115004652968-Why-are-there-odd-characters-when-I-open-a-CSV-in-Excel-)

ISO & IEC: What are they?

Most character encodings bear acronyms such as ISO/IEC 8859-1:1998.

ISO, for International Organization for Standardization, “is an independent, non-governmental international organization with a membership of 164 national standards bodies” located in Geneva, Switzerland (https://www.iso.org/about-us.html). It was founded in 1947 and promotes worldwide industrial and commercial standards.

IEC, the International Electrotechnical Commission, is a nonprofit organization that develops and publishes standards concerning electrical technologies, of which a truly wide variety exists in today's modern world. Headquartered in Geneva, Switzerland, IEC standards reach over 150 countries. https://webstore.ansi.org/sdo/iec?gclid=EAIaIQobChMIy-CGjPvB6QIVBL3ICh0BDA80EAAYASAAEgIKQ_D_BwE

ASCII codes

From the ASCII website (https://www.ascii-code.com/) we read:

> ASCII (American Standard Code for Information Interchange is a 7-bit character code where every single bit represents a unique character. On this webpage you will find 8 bits, 256 characters, ASCII table according to Windows-1252 (code page 1252) which is a superset of ISO 8859-1 in terms of printable characters. In the range 128 to 159 (hex 80 to 9F), ISO/IEC 8859-1 has invisible control characters, while Windows-1252 has writable characters. Windows-1252 is probably the most-used 8-bit character encoding in the world.
>
> The first 32 characters in the ASCII-table (0-31) are unprintable control codes and are used to control peripherals such as printers.
>
> Codes 32-127 are common for all the different variations of the ASCII table, they are called printable characters, represent letters, digits, punctuation marks, and a few miscellaneous symbols. You will find almost every character on your keyboard. Character 127 represents the command DEL.

Extended ASCII or Latin-1

The extended ASCII code (128-255) contains many special symbols (e.g., £, ©, ®) and characters used in different languages, both uppercase and lowercase (e.g., Ã, É, à, æ, ó, ü).

There are several different variations of the 8-bit ASCII table. The extended Windows ASCII set is according to Windows-1252 (CP-1252) which is a superset of ISO 8859-1, also called **ISO Latin-1**, in terms of printable characters. It is the basis for most popular 8-bit character sets and the first block of characters in Unicode.

utf-8 code (Unicode Transformation Format)

From Wikipedia (https://en.wikipedia.org/wiki/UTF-8) we read:

> UTF-8 (8-bit Unicode Transformation Format) is a variable width character encoding capable of encoding all 1,112,064 valid character code points in Unicode using one to four one-byte (8-bit) code units. The encoding is defined by the Unicode Standard, and was originally designed by Ken Thompson and Rob Pike [in 1992/3]. The name is derived from Unicode (or Universal Coded Character Set) Transformation Format – 8-bit.
>
> It was designed for backward compatibility with ASCII. Code points with lower numerical values, which tend to occur more frequently, are encoded using fewer bytes. The first 128 characters of Unicode, which correspond one-to-one with ASCII, are encoded using a single byte with the same binary value as ASCII, so that **valid ASCII text is valid UTF-8**-encoded Unicode as well.

utf-8 encoding encompasses the characters of most known languages. Thus, Chinese

characters are utf-8 compliant.

utf-8 compliance: What does it mean?

A text is utf-8 compliant when all of the characters in the text are valid utf-8 codes.

utf-8 compliance: Why bother?

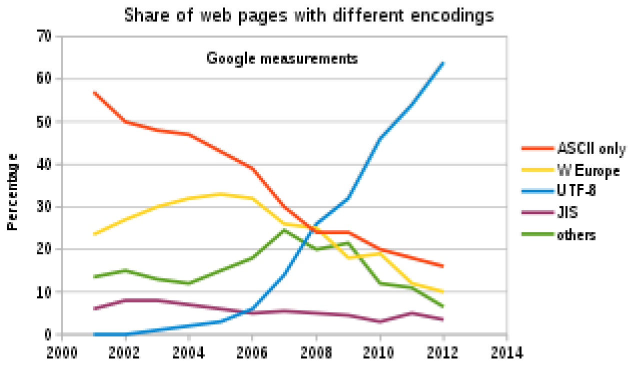
There are at least two good reasons for making sure that a text is utf-8 compliant.

First good reason: The NLP Suite and Stanford CoreNLP

All NLP Suite scripts read, process, and save data in utf-8 encoding standard. In particular, the Stanford CoreNLP algorithms which form the basis of many of the NLP Suite algorithms will break with non utf-8 text data.

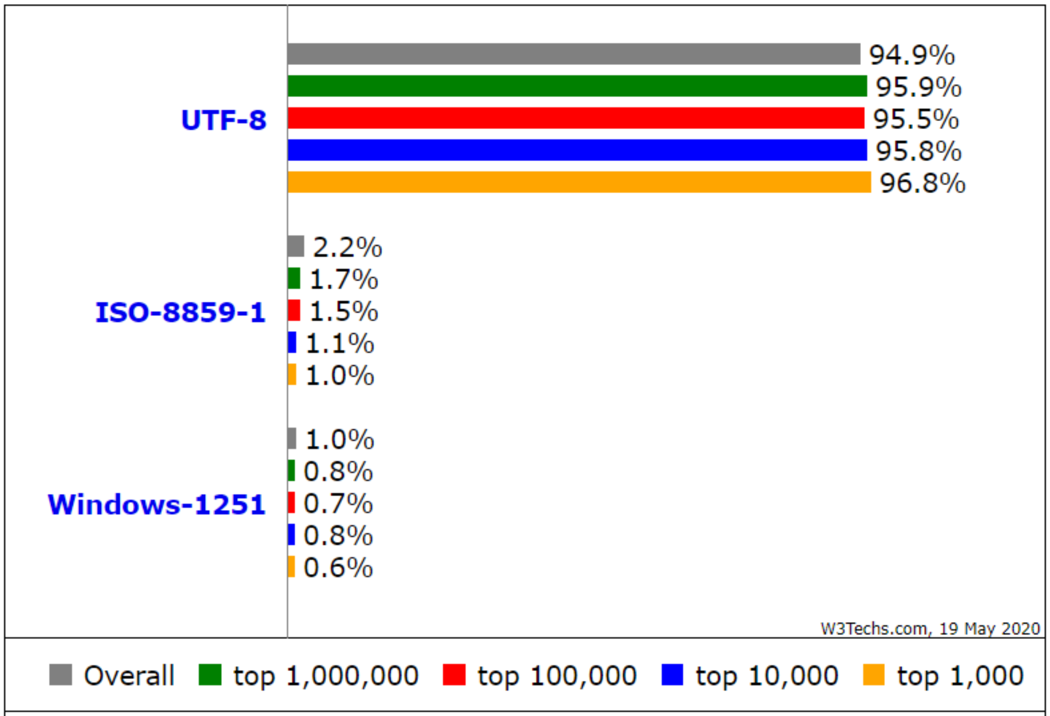
Second good reason: The encoding standard of the world's websites

UTF-8 is by far the most common encoding for the World Wide Web, accounting for 95% (global average), up to 100% for some languages, of all web pages as of 2020. **No encoding other than UTF-8 is very popular on the web for any country.**



Use of the main encodings on the web from 2001 to 2012 as recorded by Google, with UTF-8 overtaking all others in 2008 and over 60% of the web in 2012.

2020 usage of utf-8 encoding on websites



Percentages of websites using various character encodings broken down by ranking

Unicode (or Universal Coded Character Set)

Unicode is a 16-bit character encoding established by the *Unicode Consortium*, which describes the standard as follows (see http://unicode.org):

> The Unicode Standard encodes a single, very large set of characters, encompassing all the characters needed for worldwide use. This single repertoire is intended to be universal in coverage, containing all the characters for **textual representation in all modern writing systems, in most historic writing systems**, and for symbols used in plain text.

The current (May 2020) version of Unicode has 143,859 characters that cover the principal written languages of the world (current and old), plus a variety of special characters like emoji.

All character sets map onto the Unicode set. Thus, the ASCII set maps on Unicode.

> Unicode characters are represented in one of three encoding forms: a 32-bit form (UTF32), a 16-bit form (UTF-16), and an 8-bit form (UTF-8). The 8-bit, byte-oriented form, UTF-8, has been designed for ease of use with existing ASCII-based systems.
>
> Actual implementations in computer systems represent integers in specific code units of particular size—usually 8-bit (= byte), 16-bit, or 32-bit. In the Unicode character encoding model, precisely defined encoding forms specify how each integer (code point) for a Unicode character is to be expressed as a sequence of one or more code units. The Unicode Standard provides three distinct encoding forms for Unicode characters, using 8-bit, 16- bit, and 32-bit units. These are named UTF-8, UTF-16, and

UTF-32, respectively. The "UTF" is a carryover from earlier terminology meaning **Unicode (or UCS) Transformation Format**. Each of these three encoding forms is an equally legitimate mechanism for representing Unicode characters; each has advantages in different environments. All three encoding forms can be used to represent the full range of encoded characters in the Unicode Standard; they are thus fully interoperable for implementations that may choose different encoding forms for various reasons. Each of the three Unicode encoding forms can be efficiently transformed into either of the other two without any loss of data.

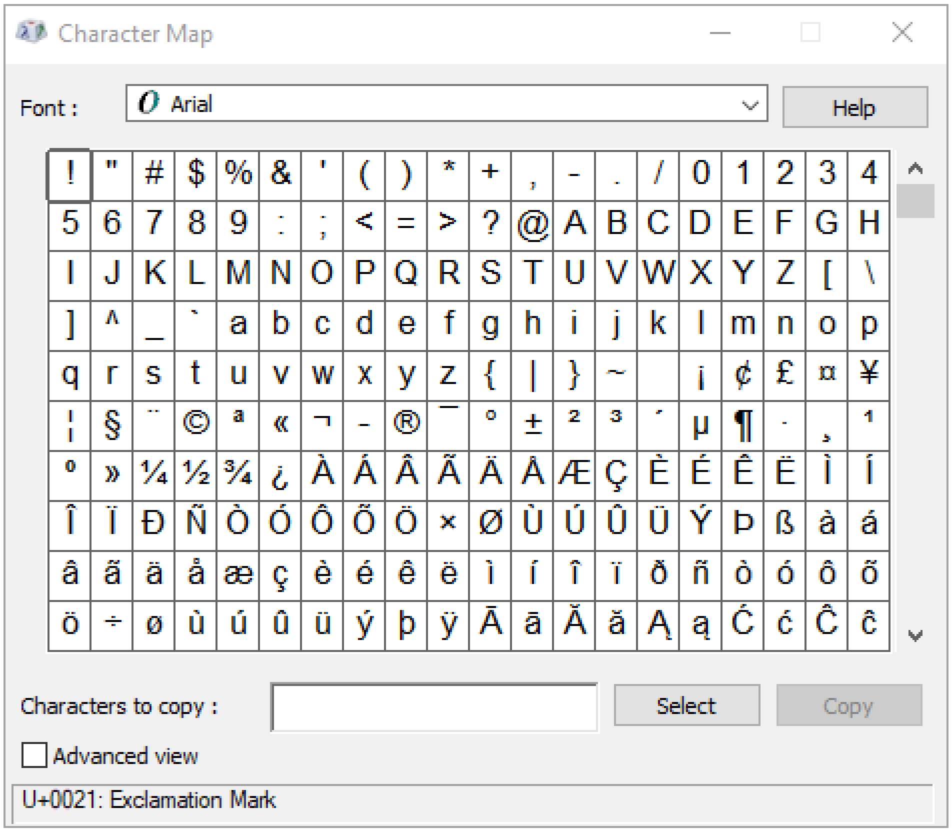
Unicode characters are represented in one of three encoding forms: a 32-bit form (UTF32), a 16-bit form (UTF-16), and an 8-bit 1-byte form (UTF-8). The 8-bit, byte-oriented form, UTF-8, has been designed for ease of use with existing ASCII-based systems.

Unicode is not an encoding. Unicode is a standard for representing text, and encoding (actually, several encodings) is part of the standard.

Windows Character Map

https://support.office.com/en-us/article/insert-ascii-or-unicode-latin-based-symbols-and-characters-d13f58d3-7bcb-44a7-a4d5-972ee12e50e0#bmcharactermap

Character Map is a program built into Microsoft Windows that enables you to view the characters that are available in a selected font.



Font : Arial

!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	0	1	2	3	4
5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H
I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[	\
]	^	_	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
q	r	s	t	u	v	w	x	y	z	{	|	}	~		¡	¢	£	¤	¥
¦	§	¨	©	ª	«	¬	-	®	¯	°	±	²	³	´	µ	¶	·	¸	¹
º	»	¼	½	¾	¿	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í
Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	à	á
â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ
ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	Ā	ā	Ă	ă	Ą	ą	Ć	ć	Ĉ	ĉ

Characters to copy :

□ Advanced view

U+0021: Exclamation Mark

Using Character Map, you can copy individual characters or a group of characters to the

Clipboard and paste them into any program that can display them. To open Character Map:

In Windows 10: Type "character" in the search box on the task bar, and choose Character Map from the results.

In Windows 8: Search for the word "character" on the Start screen and choose Character Map from the results.

In Windows 7: Click Start, point to All Programs, point to Accessories, point to System Tools, and then click Character Map.

Characters are grouped by font. Click the fonts list to choose a set of characters. To select a character, click the character, click Select, click the right mouse button in your document where you want the character, and then click Paste.