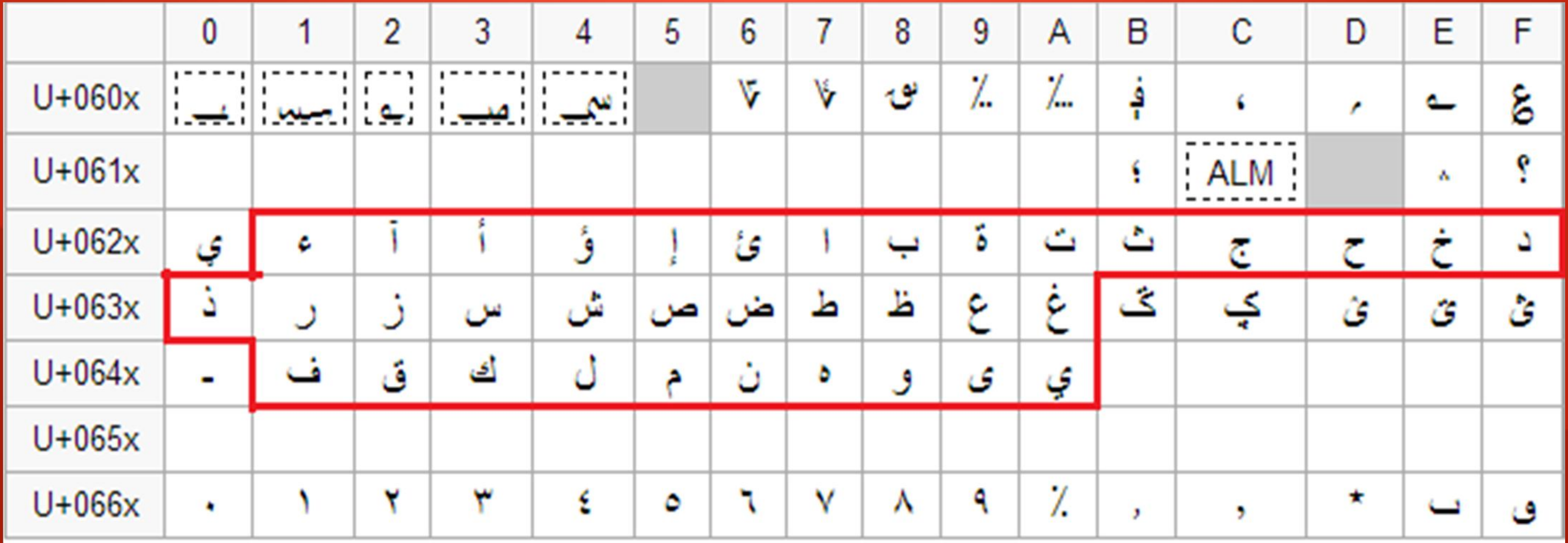
Arabic language characters and Internet domain names

Arabic language standard charset

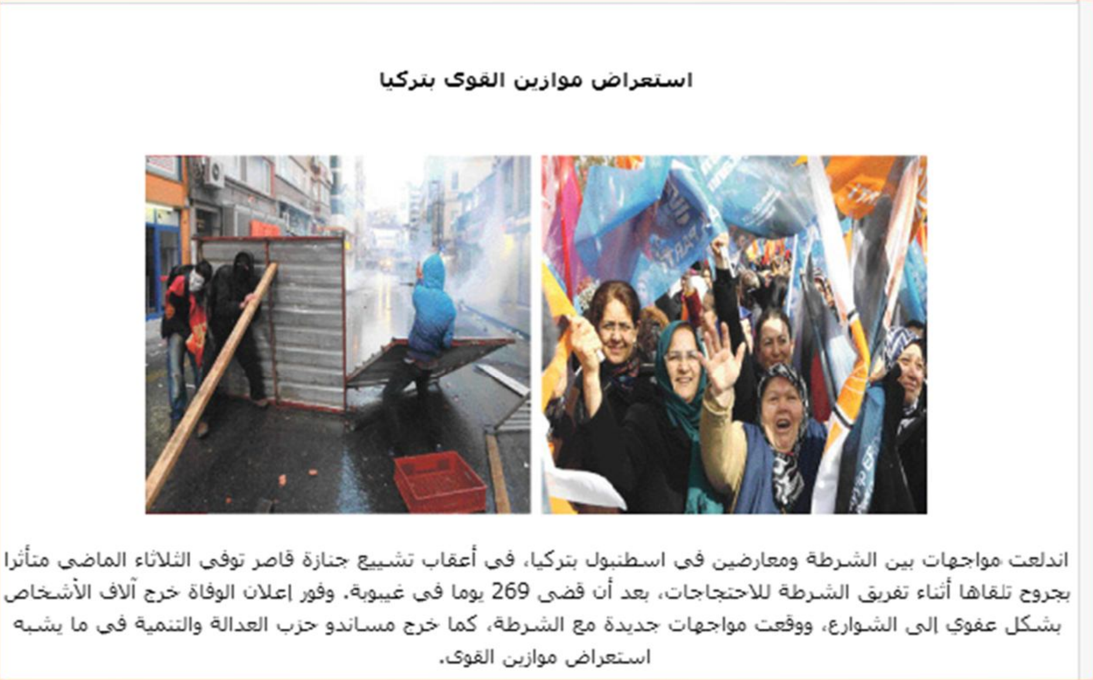
- Digits are not included since they can NOT be used in LGR for root zones

- U+002D: The Hyphen (-)



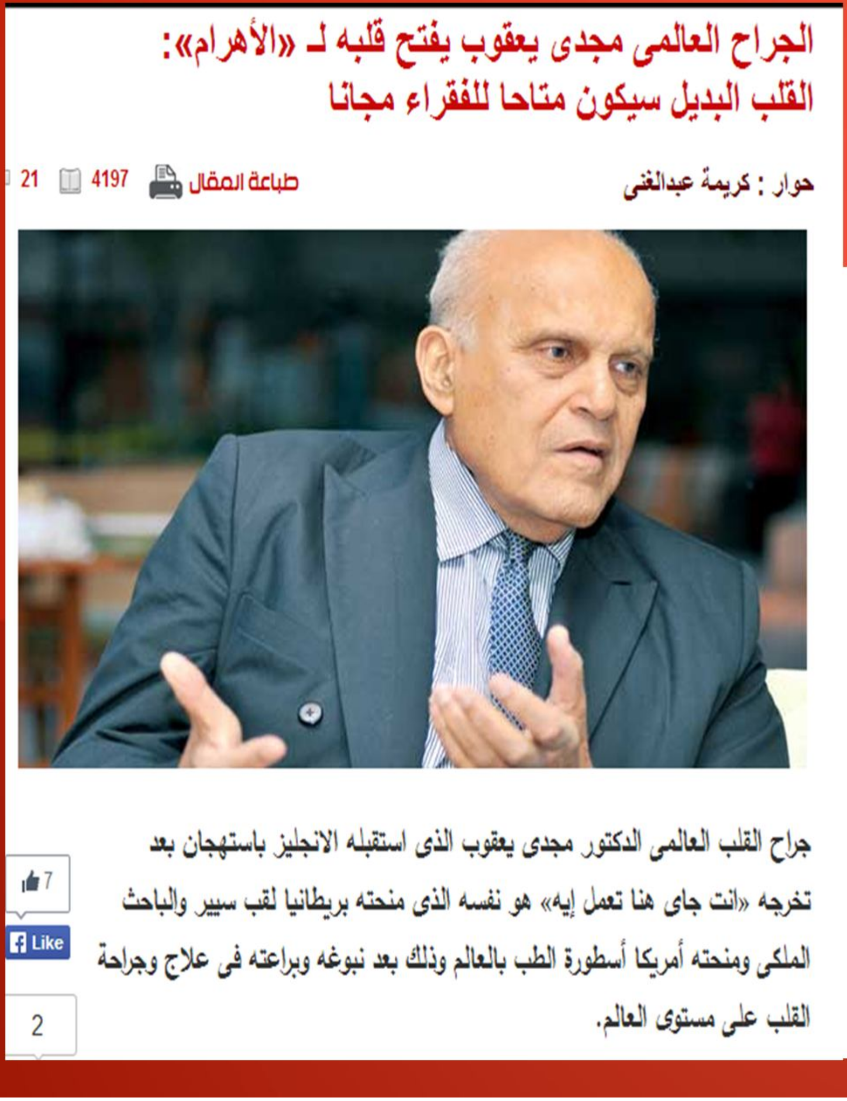
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+060x	؀	؁	؂	؃	؄		؆	؇	؈	؉	؊	؋	،	؍	؎	؏
U+061x												؛	ALM		؞	؟
U+062x	ؠ	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
U+063x	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ػ	ؼ	ؽ	ؾ	ؿ
U+064x	ـ	ف	ق	ك	ل	م	ن	ه	و	ى	ي					
U+065x																
U+066x	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	٪	٫	٬	٭	ٮ	ٯ

Arabic language standard charset



استعراض موازين القوى بتركيا

اندلعت مواجهات بين الشرطة ومعارضين فى اسطنبول بتركيا، فى أعقاب تشييع جنازة قاصر توفى الثلاثاء الماضى متأثرا بجروح تلقاها أثناء تفريق الشرطة للاحتجاجات، بعد أن قضى 269 يوما فى غيبوبة. وفور إعلان الوفاة خرج آلاف الأشخاص بشكل عفوي إلى الشوارع، ووقعت مواجهات جديدة مع الشرطة، كما خرج مساندو حزب العدالة والتنمية فى ما يشبه استعراض موازين القوى.

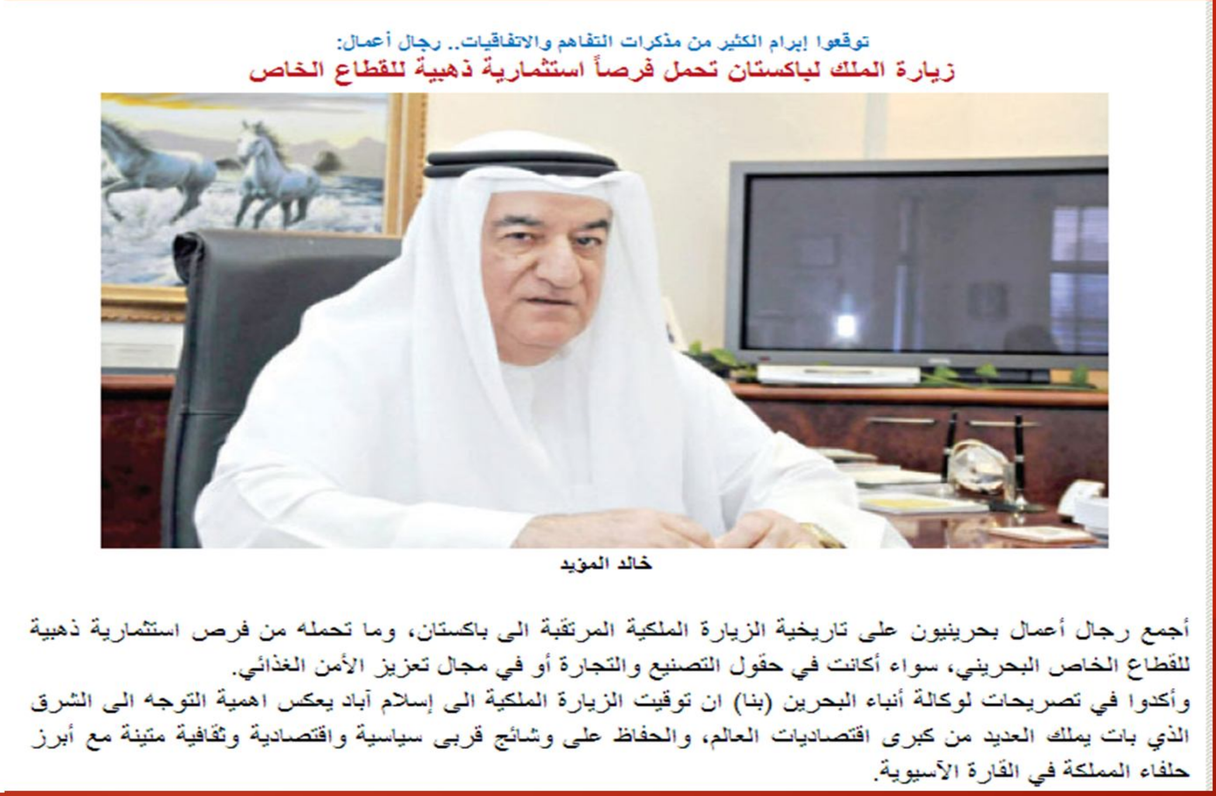


الجراح العالمى مجدى يعقوب يفتح قلبه لـ «الأهرام»:
القلب البديل سيكون متاحا للفقراء مجانا

حوار : كريمة عبدالغنى

طباعة المقال 4197 21

جراح القلب العالمى الدكتور مجدى يعقوب الذى استقبله الانجليز باستهجان بعد تخرجه «انت جاى هنا تعمل إيه» هو نفسه الذى منحته بريطانيا لقب سيير والباحث الملكى ومنحته أمريكا أسطورة الطب بالعالم وذلك بعد نبوغه وبراعته فى علاج وجراحة القلب على مستوى العالم.



توقعوا إبرام الكثير من مذكرات التفاهم والاتفاقيات.. رجال أعمال:

زيارة الملك لباكستان تحمل فرصاً استثمارية ذهبية للقطاع الخاص

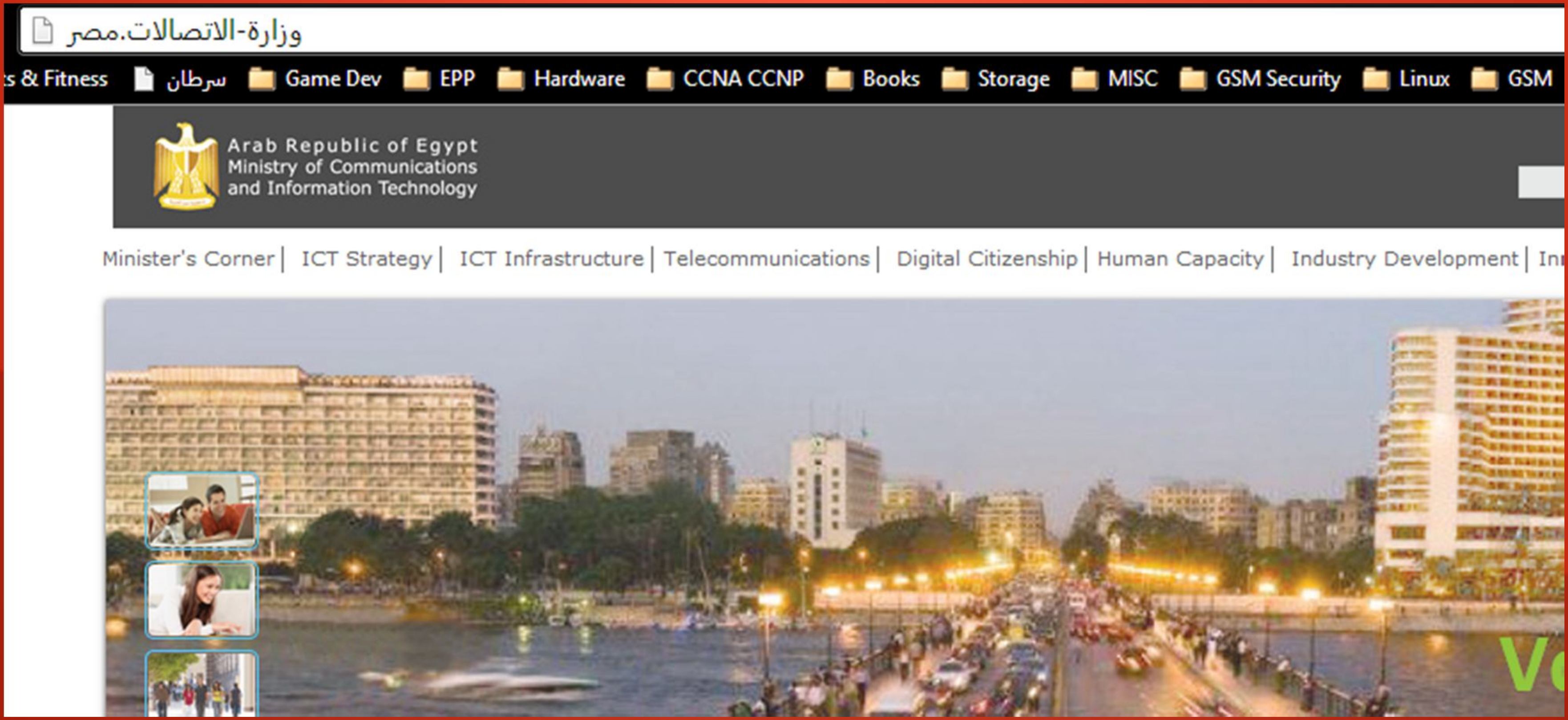
خالد المؤيد

أجمع رجال أعمال بحرينيون على تاريخية الزيارة الملكية المرتقبة الى باكستان، وما تحمله من فرص استثمارية ذهبية للقطاع الخاص البحريني، سواء أكانت في حقول التصنيع والتجارة أو في مجال تعزيز الأمن الغذائي.

وأكدوا في تصريحات لوكالة أنباء البحرين (بنا) ان توقيت الزيارة الملكية الى إسلام آباد يعكس اهمية التوجه الى الشرق الذي بات يملك العديد من كبرى اقتصاديات العالم، والحفاظ على وشائج قربى سياسية واقتصادية وثقافية متينة مع أبرز حلفاء المملكة في القارة الآسيوية.

Arabic language standard charset

- Used in Arabic IDN Domain names
  - وزارة-الاتصالات.مصر
  - xn----ymcbaaajlc6dj7bxne2c.xxn--wgbh1c



وزارة-الاتصالات.مصر

Arab Republic of Egypt
Ministry of Communications
and Information Technology

s & Fitness
سرطان
Game Dev
EPP
Hardware
CCNA CCNP
Books
Storage
MISC
GSM Security
Linux
GSM
Minister's Corner | ICT Strategy | ICT Infrastructure | Telecommunications | Digital Citizenship | Human Capacity | Industry Development | In
V

Diacritics and Shadda

- ـًـU+064B ARABIC FATHATAN
- ـٌـU+064C ARABIC DAMMATAN
- ـٍـU+064D ARABIC KASRATAN
- ـَـU+064E ARABIC FATHA
- ـُـU+064F ARABIC DAMMA
- ـِـU+0650 ARABIC KASRA
- ـْـU+0652 ARABIC SUKUN
- ـّـU+0651 ARABIC SHADDA

Diacritics and Shadda (Cont'd)

- Diacritics:
  - مَلِك: King
  - مُلك: Property
  - مَلَك: Angel
- Shadda:
  - صُوَر: Photos
  - صُوِّر: To be photographed
  - صَوَّر: To take a photo

Diacritics and Shadda (Cont'd)

- It is recommended that they should NOT be permitted in zone files
- They can be supported, not supported, or ignored on the client side
- They should be removed before processing (Querying a DNS)

Kasheeda

- Example:
  - ايكـــانAnd ايكان
  - العربيـــةAnd العربية

Kasheeda (Cont'd)

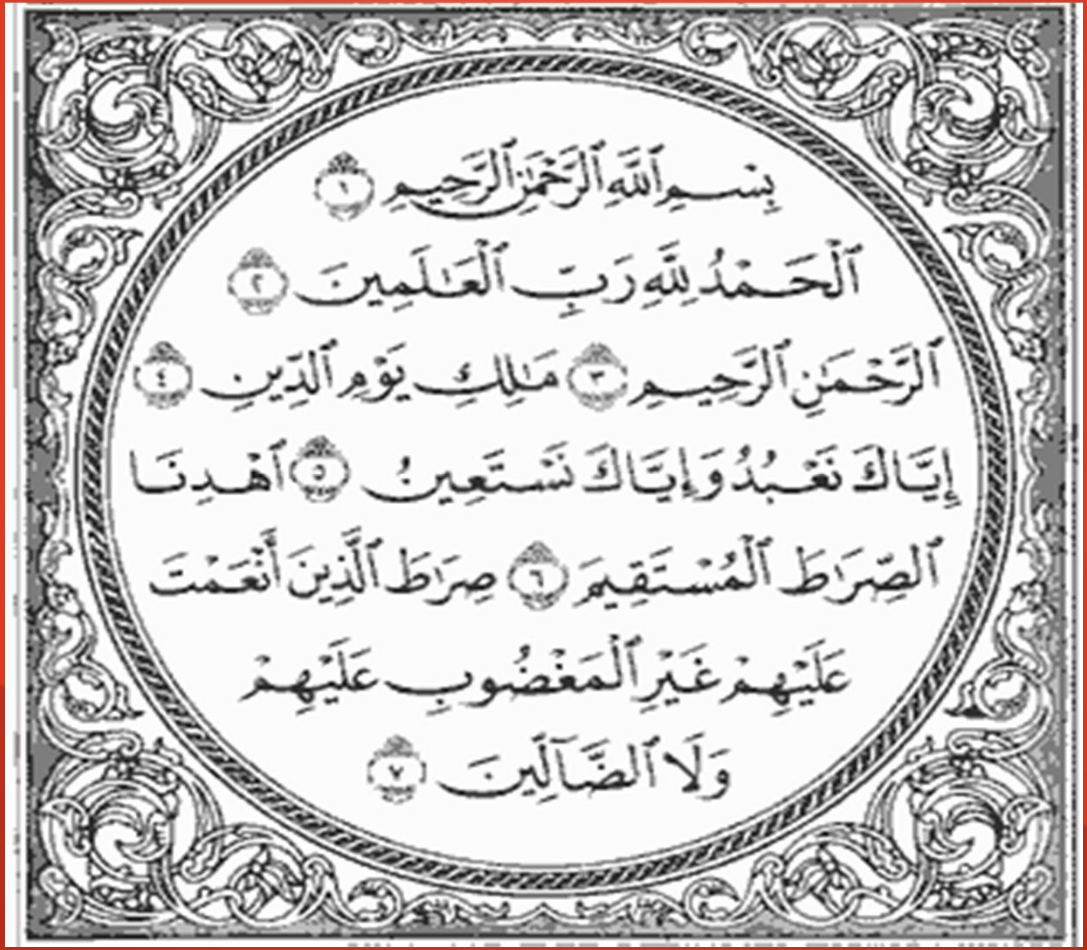
- U+0640 ARABIC TATWEEL
- Not a letter
- Does not change how a letter/word is pronounced
- Only for graphical presentation

Kasheeda (Cont'd)

- Since it has no linguistic value, it should not be allowed in zone files.
- They can be supported, not supported, or ignored on the client side
- They should be removed before processing (Querying a DNS)

Diacritics and Shadda (Cont'd)

- Understanding how Diacritics and Shadda are pronounced is crucial in Arabic language



بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ١
ٱلۡحَمۡدُ لِلَّهِ رَبِّ ٱلۡعَٰلَمِينَ ٢
ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ٣ مَٰلِكِ يَوۡمِ ٱلدِّينِ ٤
إِيَّاكَ نَعۡبُدُ وَإِيَّاكَ نَسۡتَعِينُ ٥ ٱهۡدِنَا
ٱلصِّرَٰطَ ٱلۡمُسۡتَقِيمَ ٦ صِرَٰطَ ٱلَّذِينَ أَنۡعَمۡتَ
عَلَيۡهِمۡ غَيۡرِ ٱلۡمَغۡضُوبِ عَلَيۡهِمۡ
وَلَا ٱلضَّآلِّينَ ٧

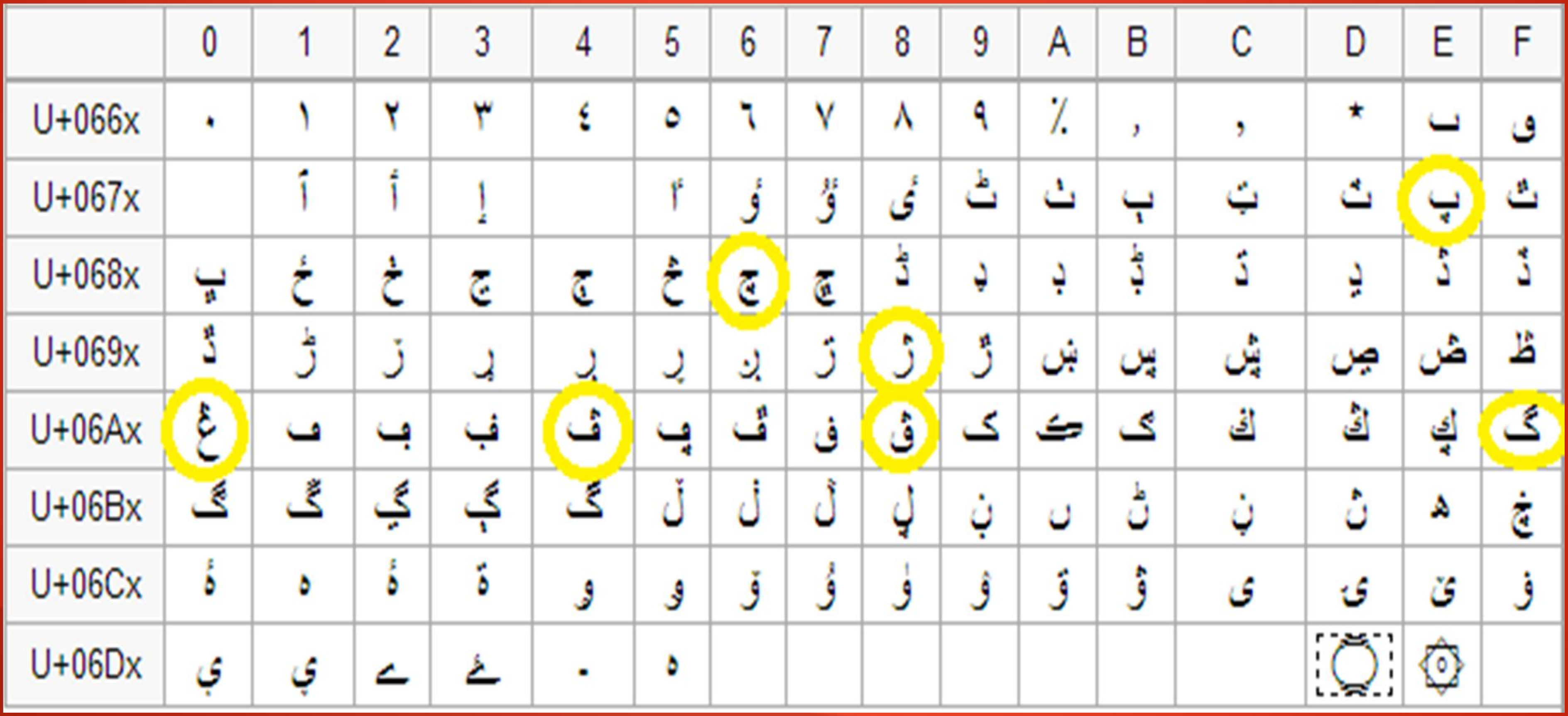
Space Character

- The breaking space U+0020 (ASCII 32)
- The non-breaking space U+00A0 (ASCII 160)
- Not allowed to be used in domain names, even in non-IDN domains.
- We can use the hyphen instead

Non-standard letters in Arabic language

- Using the letter چinstead of ج, or pronounced as CHE
- Using the letter ڨinstead of ج
- Using the letter ڤinstead of ف
- Using the letter گinstead of ج
- Using پ, pronounced as P
- Using ژ
- Using ڠ

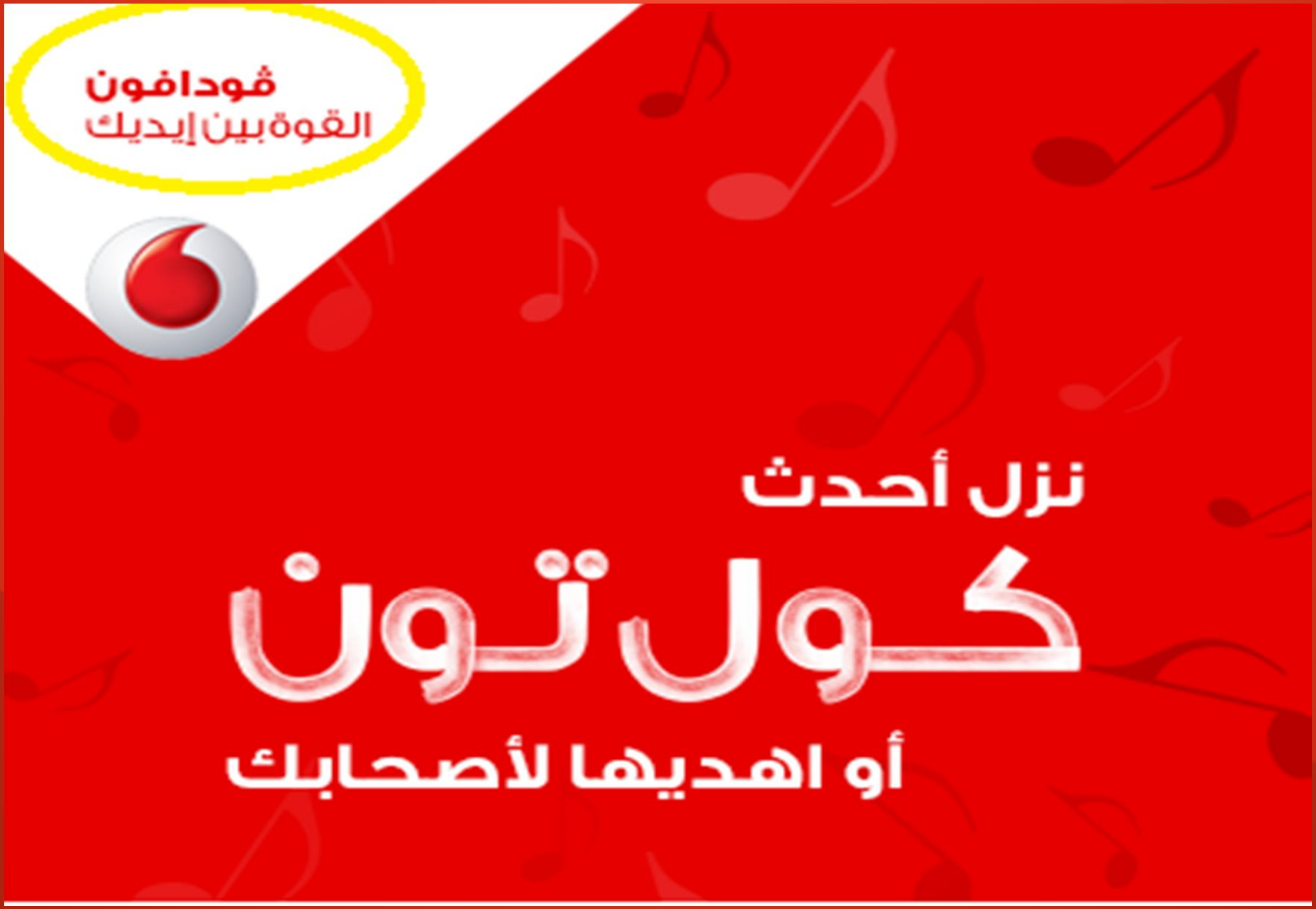
Non-standard letters in Arabic language



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+066x	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	٪	٫	٬	٭	ٮ	ٯ
U+067x		ٱ	ٲ	ٳ		ٵ	ٶ	ٷ	ٸ	ٹ	ٺ	ٻ	ټ	ٽ	پ	ٿ
U+068x	ڀ	ځ	ڂ	ڃ	ڄ	څ	چ	ڇ	ڈ	ډ	ڊ	ڋ	ڌ	ڍ	ڎ	ڏ
U+069x	ڐ	ڑ	ڒ	ړ	ڔ	ڕ	ږ	ڗ	ژ	ڙ	ښ	ڛ	ڜ	ڝ	ڞ	ڟ
U+06Ax	ڠ	ڡ	ڢ	ڣ	ڤ	ڥ	ڦ	ڧ	ڨ	ک	ڪ	ګ	ڬ	ڭ	ڮ	گ
U+06Bx	ڰ	ڱ	ڲ	ڳ	ڴ	ڵ	ڶ	ڷ	ڸ	ڹ	ں	ڻ	ڼ	ڽ	ھ	ڿ
U+06Cx	ۀ	ہ	ۂ	ۃ	ۄ	ۅ	ۆ	ۇ	ۈ	ۉ	ۊ	ۋ	ی	ۍ	ێ	ۏ
U+06Dx	ې	ۑ	ے	ۓ	۔	ە								۝	۞	

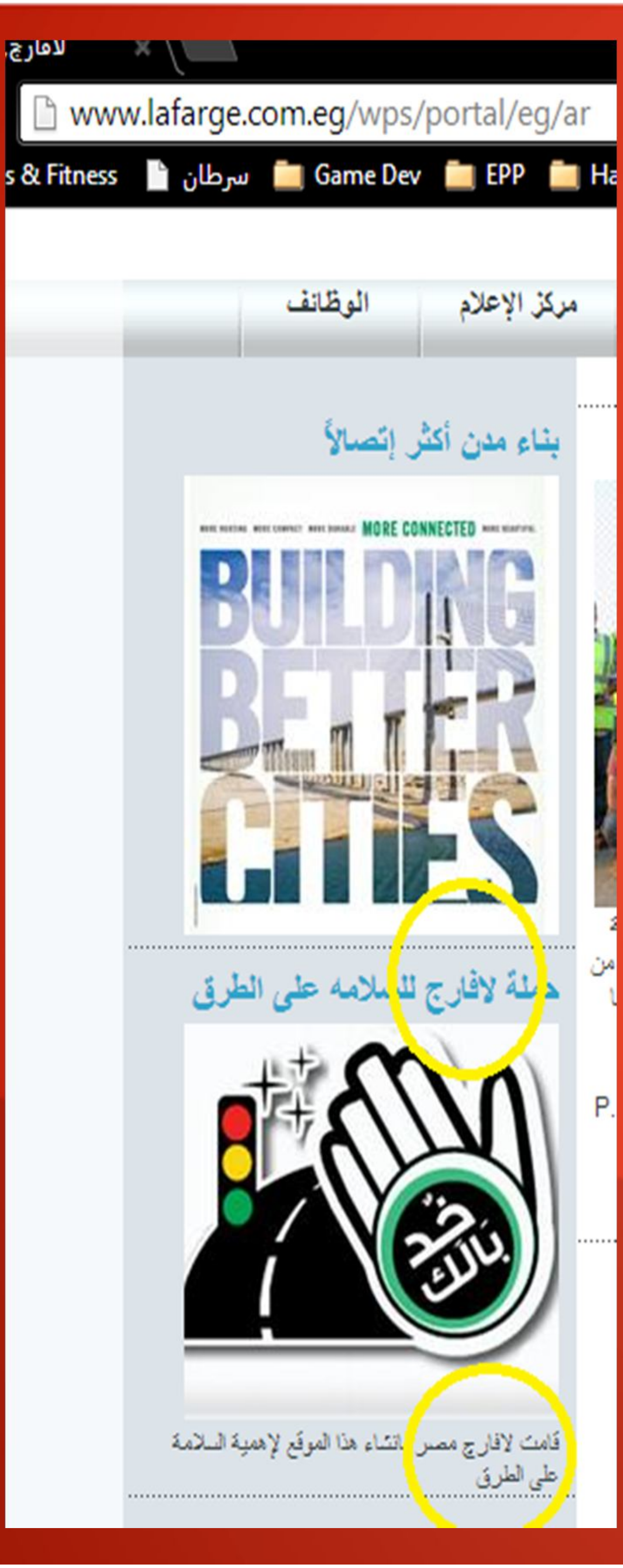
Non-standard letters in Arabic language

- Vodafone ad (Egypt)



Non-standard letters in Arabic language

- From a construction company official web page (Egypt)

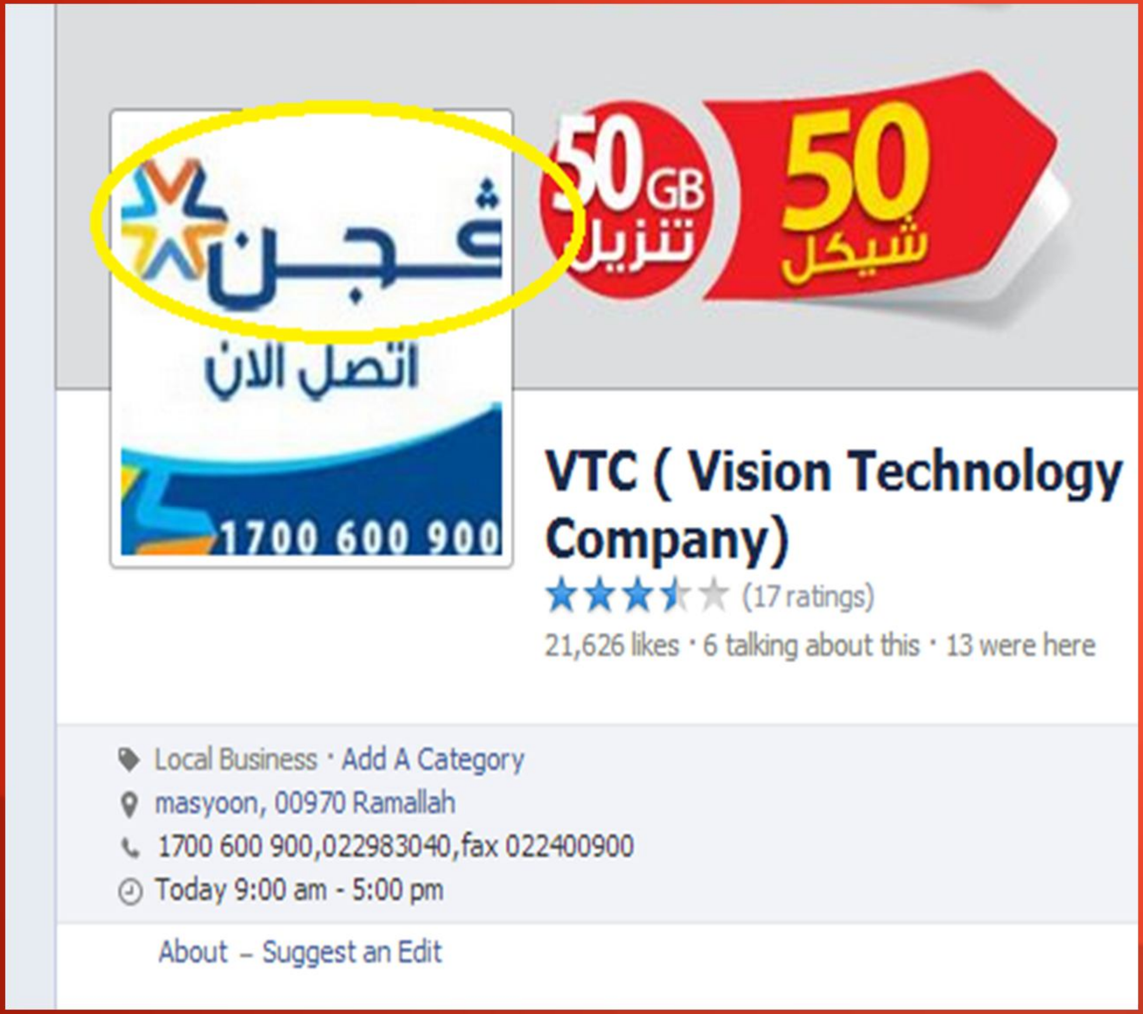


بناء مدن أكثر إتصالاً

حملة لافارج للسلامه على الطرق

قامت لافارج مصر بانشاء هذا الموقع لإهمية السلامة على الطرق

Non-standard letters in Arabic language



VTC (Vision Technology Company)

★★★★☆ (17 ratings)

21,626 likes · 6 talking about this · 13 were here

Local Business · Add A Category

masyoon, 00970 Ramallah

1700 600 900,022983040,fax 022400900

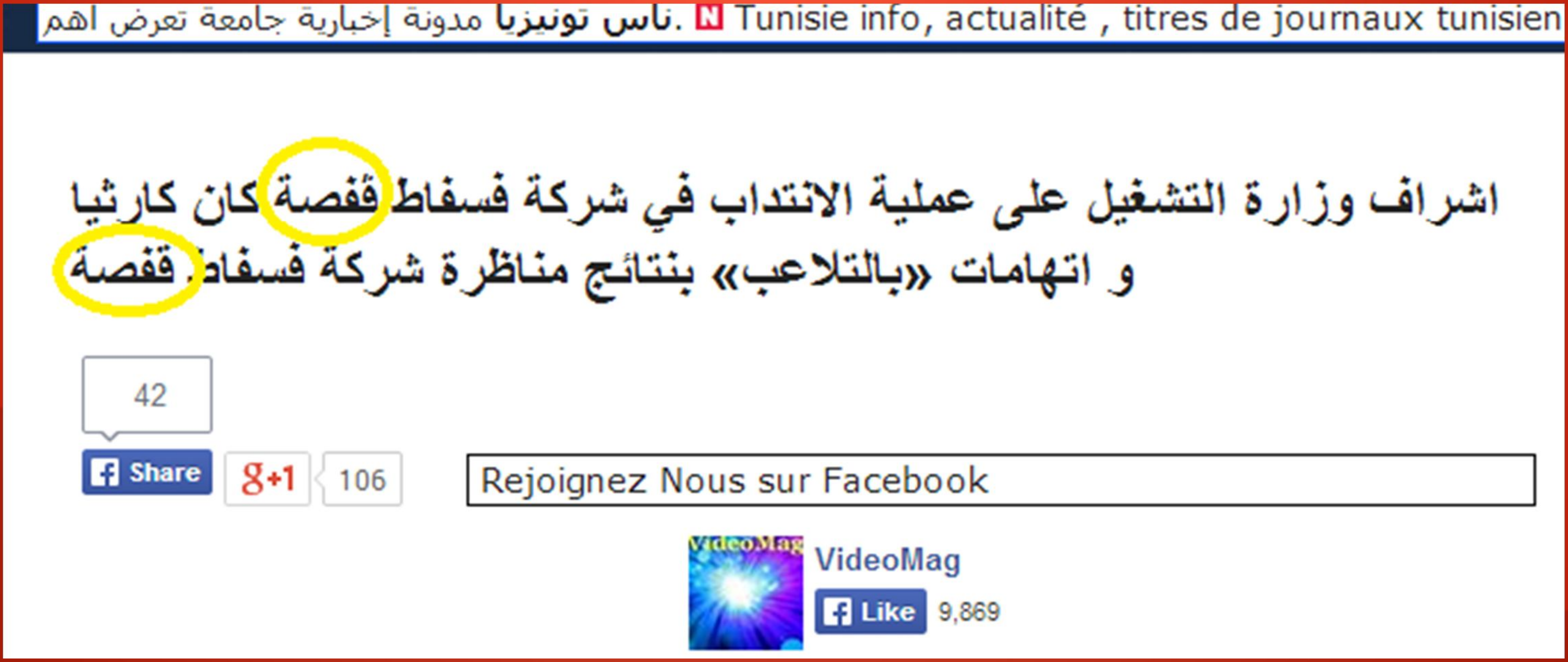
Today 9:00 am - 5:00 pm

About – Suggest an Edit

- A company in Ramallah facebook page

Non-standard letters in Arabic language

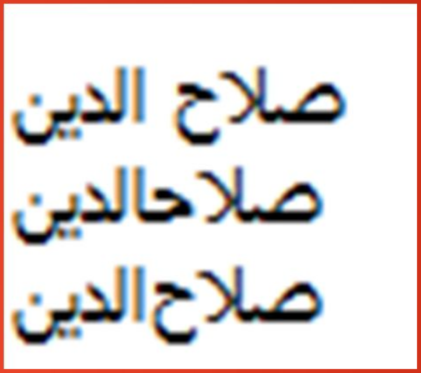
- From a Tunisian news blog



ناس تونيزيا. مدونة إخبارية جامعة تعرض اهم N Tunisie info, actualité , titres de journaux tunisien

اشراف وزارة التشغيل على عملية الانتداب في شركة فسفاط قفصة كان كارثيا
و اتهامات «بالتلاعب» بنتائج مناظرة شركة فسفاط قفصة

Zero-width non-joiner (ZWNJ)



صلاح الدين

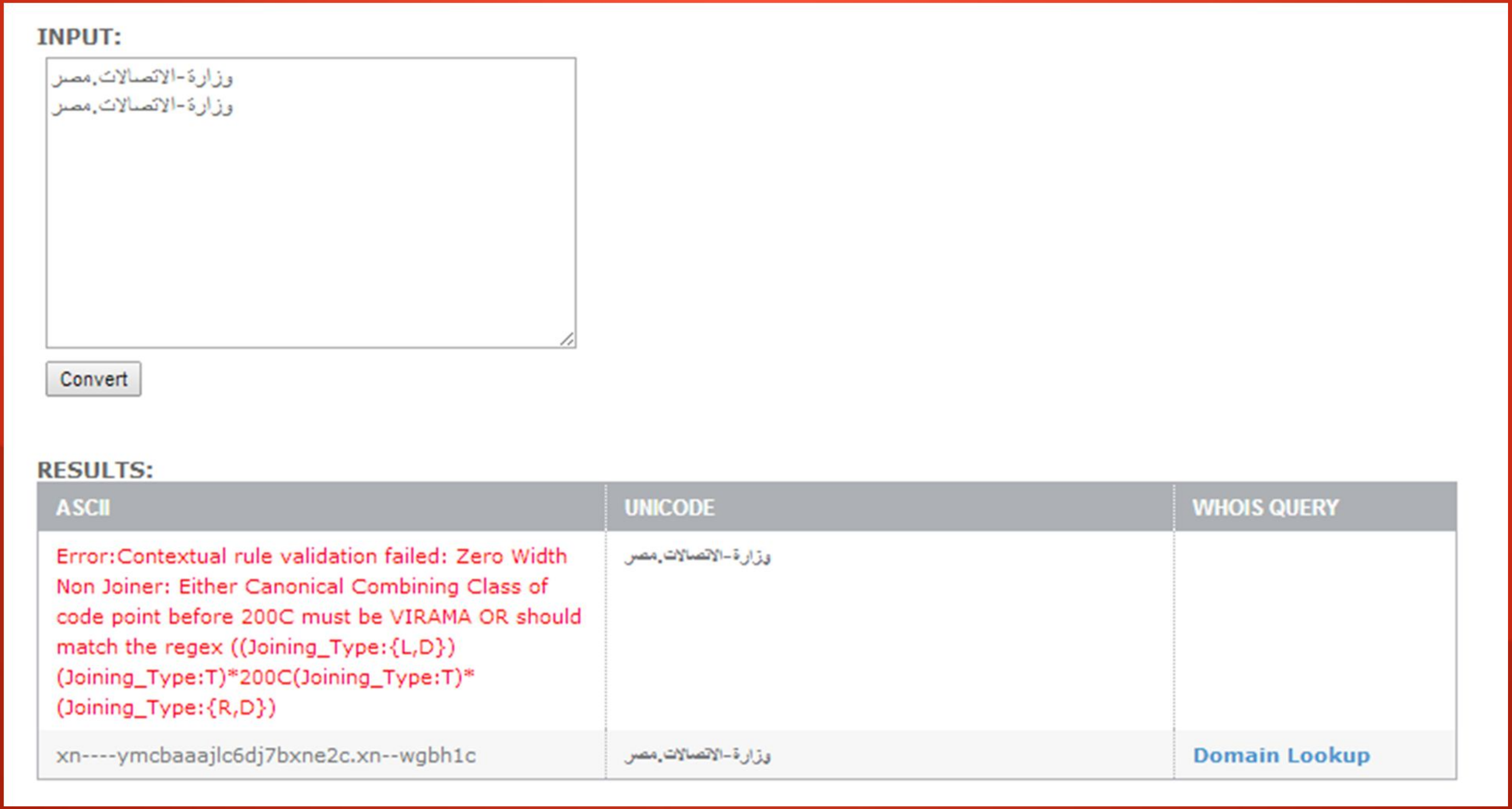
صلاحالدين

صلاح‌الدين

- U+200C
- Non-printable
- When placed between two characters that would otherwise be connected into a ligature, a ZWNJ causes them to be printed in their final and initial forms, this is the effect of space character also.

Zero-width non-joiner (ZWNJ)

- Security issues arise if permitted in labels, what if used between letters that are not originally joined?



INPUT:

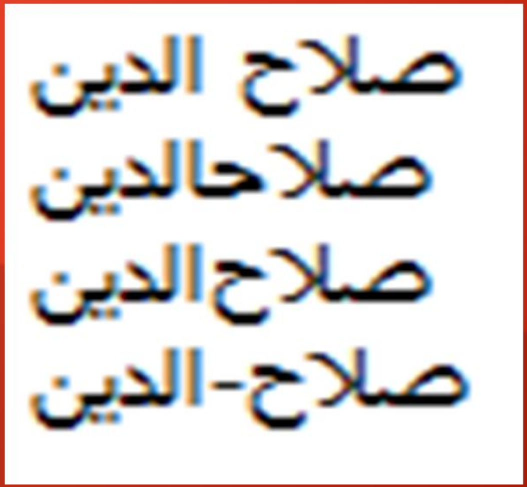
وزارة-الاتصالات.مصر
وزارة-الاتصالات.مصر

RESULTS:

ASCII	UNICODE	WHOIS QUERY
Error:Contextual rule validation failed: Zero Width Non Joiner: Either Canonical Combining Class of code point before 200C must be VIRAMA OR should match the regex ((Joining_Type:{L,D})(Joining_Type:T)*200C(Joining_Type:T)*(Joining_Type:{R,D}))	وزارة-الاتصالات.مصر	
xn----ymcbaaajlc6dj7bxne2c.xn--wgbh1c	وزارة-الاتصالات.مصر	Domain Lookup

Zero-width non-joiner (ZWNJ)

- Each browser deals with ZWNJ differently:
  - Chrome converts it to space
  - MS Explorer removes it.
- It's use is not common in Arabic, but very popular in some other languages like Persian.
- Should not be permitted in labels, a hyphen should be used instead.



صلاح الدين

صلاحالدين

صلاح‌الدين

صلاح-الدين

Character Folding

- Multiple characters that may look alike are folded into one character (shape)

- In some countries some characters may be used interchangeably due to their shape similarity. For example, using ي and ى interchangeably at the end of a word is not allowed in such countries while it is very common in others.

- ـة and ـه at the end of word

- آ,أ,إ, and ا

- و and ؤ

Character Folding (Cont'd)

- It may change the meaning of a word:
  - علي is a noun, but على means “Above”
  - سلمى is a noun, but سلمي means peaceful
- Against some spelling rules:
  - Writing the word العربيه (Arabic language) is against the spelling rule, it should only be written as العربية.

Character Folding (Cont'd)

- Pros:
  - Some types of spelling mistakes will not be an issue, at least technically
  - Will help avoid phishing
- Cons:
  - Against spelling rules, against preserving Arabic language ethics
  - Requires extra levels of processing on registry/registrar side
  - DNS and WEB DEVELOPMENT/HOSTING have to be aware if folding exists for a domain
  - With one folded domain, a registrant can run several different instances with single bought domain name which means a lost income for the registry/registrar.
  - Secure communication like HTTPS and will need special attention.

.masr approach regarding folding

1- Established a permitted code point repository.

- Excluded the Diacritics, Shadda, Tatweel, and non standard Arabic letters.

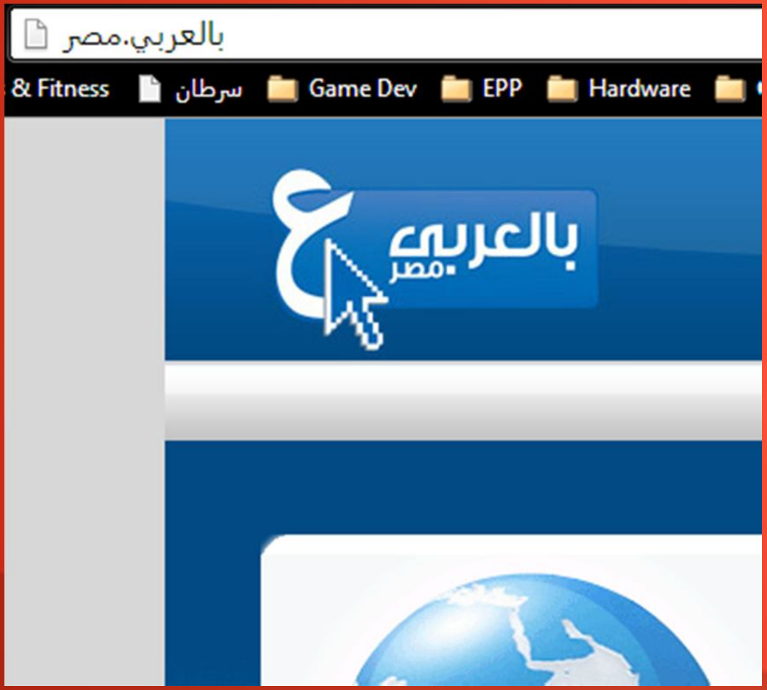
2- No numbers allowed at beginning of the label (RegEx)

3- No hyphens allowed at the end of beginning of a label

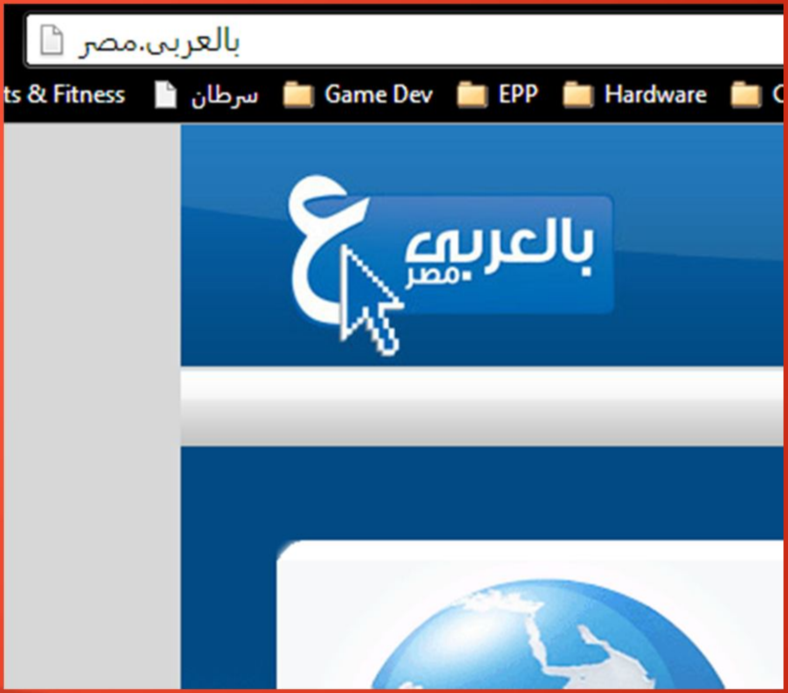
4- Established a folding rules:

- Folding ي and ى at the end of a label.
- Folding Eastern (Indic) and Western Arabic numbers
- No folding between ـه and ـة at the end of label
- No folding between different Alef (ا أ آ إ)

.masr approach regarding folding (Cont'd)



بالعربي.مصر



بالعربى.مصر

Folding in .emarat

- Folding all Alef (أ ا إ اّ)

- Folding ي and ى

- Folding ه and ة

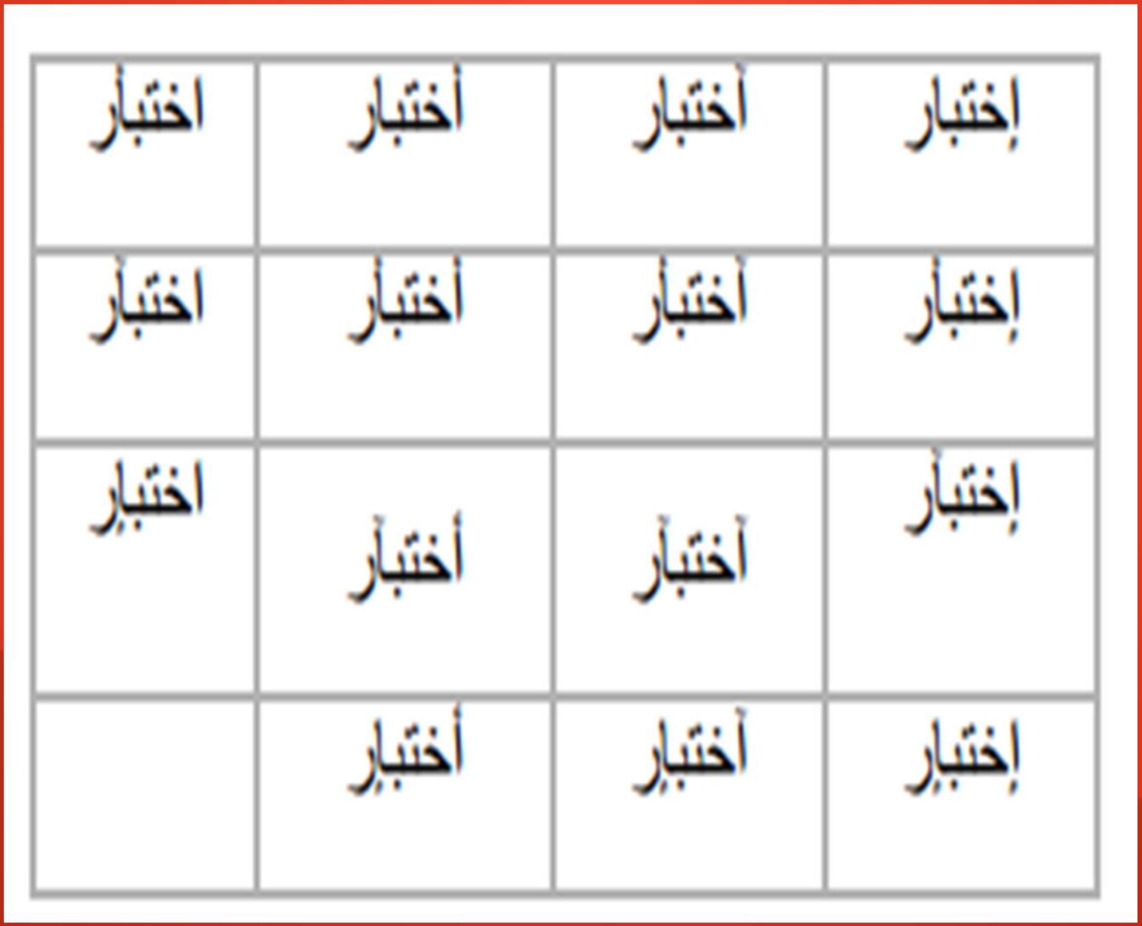
- At time of registration the registrant is given only the requested variation of a domain

- All other variations are blocked.

- The registrant is allowed activate any of the blocked variation, and the remaining will stay blocked.

Folding in .emarat (Cont'd)

- Example: in case of registering ┘ | امارات.اختبار



اختبأر	أختبأر	آختبأر	إختبأر
اختبآر	أختبآر	آختبآر	إختبآر
اختبإر	أختبإر	آختبإر	إختبإر
	أختبار	آختبار	إختبار

Folding in .qatar

- The registrant is given only the requested variation of a domain

- All other variations are blocked.

- The registrant is allowed to later request up to five (5) variation or a domain, any remaining variations will be blocked.

شكراً

شکریه

با سپاس

تشكر

Thank You