STU FIIT

Diakritikovač slovenského textu

Autor: Jakub Gedera
Vedúci práce: Ing. Marián Šimko, PhD.
18.6.2015

Motivácia

- Internet, svet bez diakritiky
- Mnoho textu na internete neobsahuje diakritiku
  - Emaily
  - Diskusné príspevky
  - Blogy
- Strata informácie pri ďalšom spracovaní textu

Motivácia

- Problémy so spracovaním takýchto textov:
  - Kategorizácia textu,
  - Extrakcia metadát, kľúčových slov
- Umožniť používateľovi Webu automaticky rekonštruovať diakritiku
- Obsah s diakritikou vyzerá profesionálnejšie
- Takýto text sa lepšie číta

Identifikovaný problém

- Veta: Nova pracka perie dobre.
- Nova:
  - Nova (eruptívna hviezda)
  - Nová (prídavné meno)
- Pracka:
  - Práčka (prístroj na pranie)
  - Pračka (bitka, ruvačka)
  - Pracka (predmet slúžiaci na spínanie niečoho)
- Perie:
  - Perie (sloveso)
  - Perie (podstatné meno)
- Dobre:
  - Dobre (príslovka)
  - Dobré (prídavné meno)

Naivné riešenia

- Nájsť v slovníku možnosti doplnenia diakritiky pre jednotlivé slová a použiť náhodný výber
- Vybrať slovo s diakritikou, ktoré sa pri vytváraní slovníka vyskytlo najčastejšie
- Slabé výsledky

Využime kontext

- Textová spojitosť
- V bare sa uskutočnila ___?___.
  a) Práčka
  b) Pračka
  c) Pracka
- **Ako využiť kontext?**
- Použiť jazykový model slovenského jazyka

Jazykový model

- Pravdepodobnostný a štatistický model jazyka a postupnosti slov [Jurafsky, Martin, 2000]
- Trénovaný na veľkom korpuse textov
- Využívaný aj pri rozpoznávaní reči
- Analogicky problém k doplneniu diakritiky
- Identifikovať najpravdepodobnejšieho slova na zázname
- Pravdepodobnostný odhad nasledujúceho slova
- Nájdenie najpravdepodobnejšej sekvencie slov

Použitý jazykový model

- Využívame model, ktorý bol vytvorený JÚĽŠ SAV zo Slovenského Národného Korpusu textov
- Je dôležité, na akých typoch textoch je model natrénovaný
- Model použitý v metóde bol natrénovaný na publicistických textoch

Náš algoritmus

1. Tokenizuj vstup a tokeny preveď na malé písmena

   Nova pracka perie dobre.

   nova pracka perie dobre .

2. Pre každý token nájdi v slovníku jeho možnosti diakritikovania

   nova práčka perie dobre .
   nová pračka dobré
   pracka

Náš algoritmus

3. Prechádzaj text po n-ticiach tokenov a vyhodnoť všetky možné n-gramy jazykovým modelom
    - nova pracka perie dobre. -> -24.7071705
    - nová pračka perie dobre. -> -23.2947597
    - ...
    - nová práčka perie dobré. -> -18.7531738
    - nová práčka perie dobre. -> -17.3096389

- (Pre prácu s jazykovým modelom bola použitá knižnica *kenlm* a jej metóda score(), ktorá vráti logaritmus pravdepodobnosti výskytu vyhodnocovaného n-gramu)

Náš algoritmus

4. Vyber n-gram, ktorý bol ohodnotený najlepšie
   - nová práčka perie dobre
5. Rekonštruuj veľkosť písmen
   - Nová práčka perie dobre

Experiment

1. Bez používateľa
   - Sledujeme úspešnosť doplnenia diakritiky
   - Ako dáta boli použité typy textov, ktoré sa na Webe vyskytujú najčastejšie
2. S používateľom
   - Účastníci experimentu vytvorili texty, ktoré sme následne rekonštruovali
   - Účastníci ohodnotili svoju spokojnosť s rekonštruovaným textom

Experiment bez používateľa

- Ako testovacia vzorka boli použité 3 typy textov, aby sme ohodnotili vhodnosť použitého jazykového modelu.
- Typy testovaných textov:
  - Publicistické texty
  - Príspevky na diskusných fórach
  - Umelecké texty

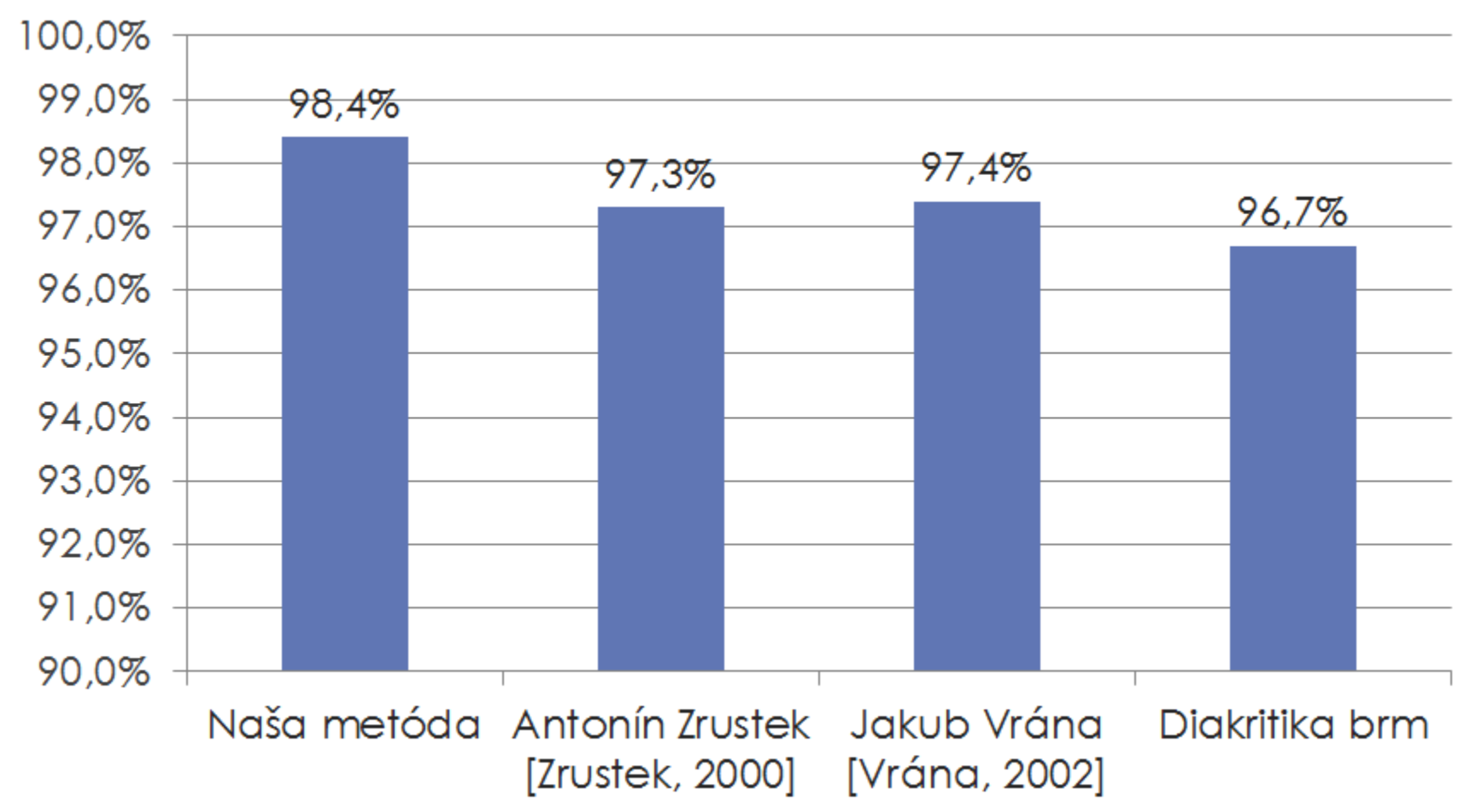
Experiment bez používateľa

Typ textu	Počet slov	Počet chýb	Dosiahnutá úspešnosť
Diskusné fórum	39 123	627	98,4 %
Publicistický štýl	45 321	1 108	97,6 %
Umelecké texty	32 110	3 443	89,3 %

Chyby v diakritikovaní

- Dva typy chýb
  1. Pre slovo sa nenašiel správny variant diakritikovania v slovníku (rozšíriť slovník)
  2. Správny variant bol v slovníku, ale vyhodnotenie modelom bolo nesprávne (obohatiť, vytvoriť nový, kvalitnejší model)
- Sú v pomere približne **1:1**

Porovnanie s inými riešeniami



Experiment s používateľmi

- 11 používateľov testovalo spokojnosť s rekonštruovaným textom
- Vytvárali texty vyskytujúce sa bežne na Webe, na diskusných fórach, aplikácia rekonštruovala diakritiku

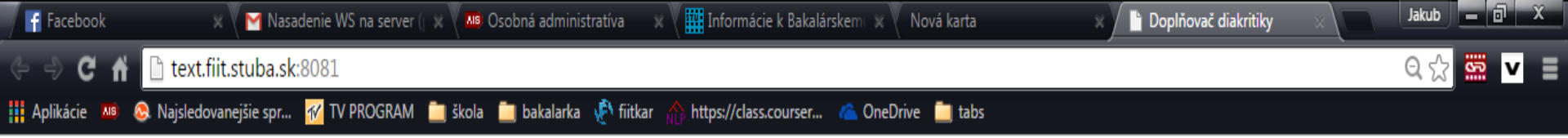
Počet slov	Počet chýb	Dosiahnutá úspešnosť
1 460	36	**97,5 %**

- V priemere jedno slovo zo 40 chybne rekonštruované

Experiment s používateľom

- Hodnotili na stupnici
  - Veľmi nespokojný
  - Nespokojný
  - Spokojný
  - Veľmi spokojný

	Veľmi nespokojný	nespokojný	spokojný	Veľmi spokojný
Počet používateľov	0	0	6	5



Doplňovač diakritiky

Nástroj na doplňovanie diakritiky do textu.

Úvod

V dnešnej dobe sa na internete vyskytuje veľa textov, ktoré neobsahujú diakritiku. Takéto texty sa môžu niekomu ťažšie čítať. Strojové spracovávanie takýchto textov je problematické a v slovenčine to platí dvojnásobne. Preto sme vytvorili tento nástroj, ktorý rekonštruuje diakritiku v texte.

Webová služba

Webová služba umožňuje zrekonštruovať vstupný text bez diakritiky. Komunikovať s webovou službou je možné pomocou GET požiadavky:

GET: 0.0.0.0:8081/diakritika/[text]

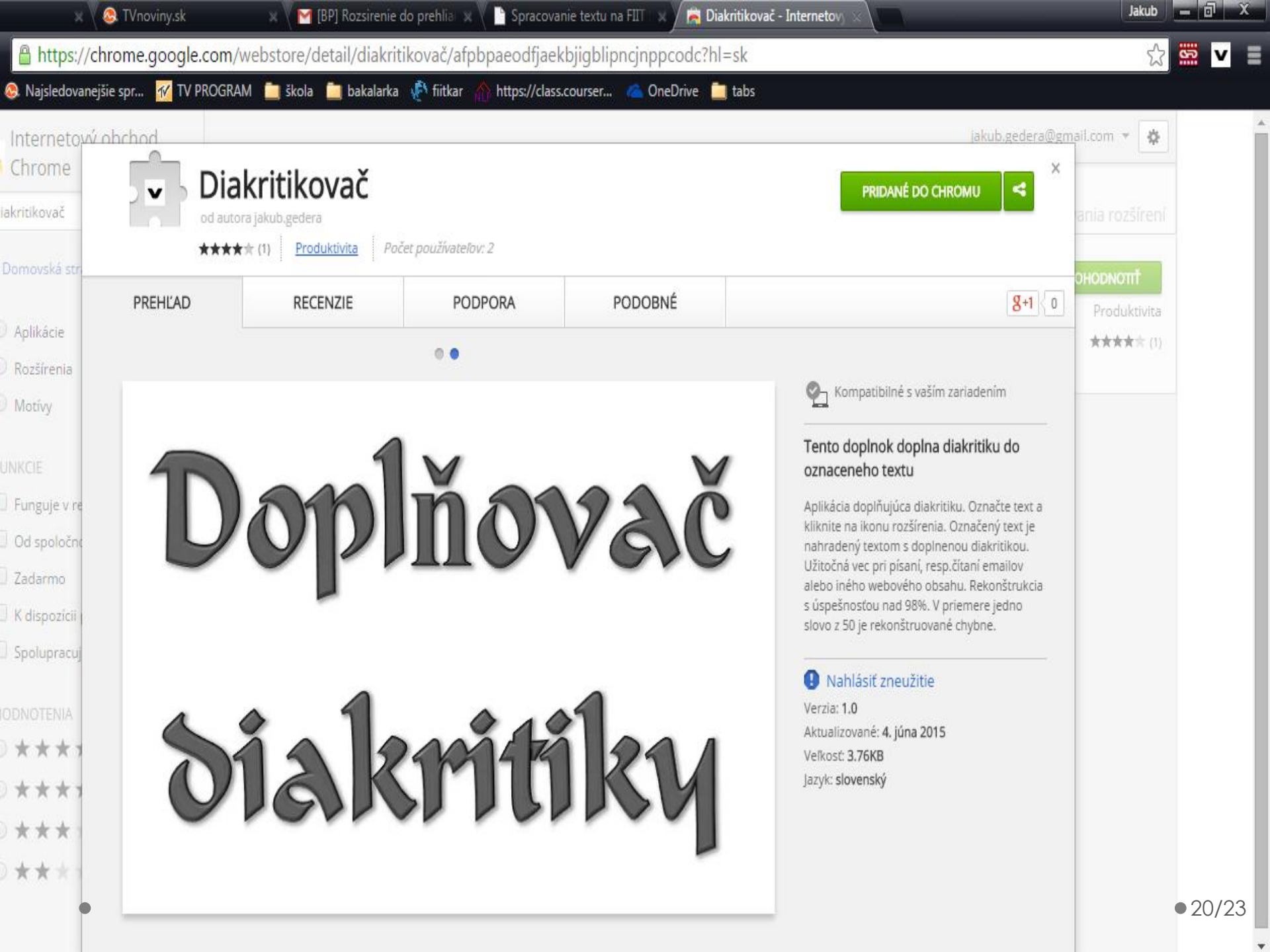
PRÍKLAD: 0.0.0.0:8081/diakritika/Kolko je hodin?

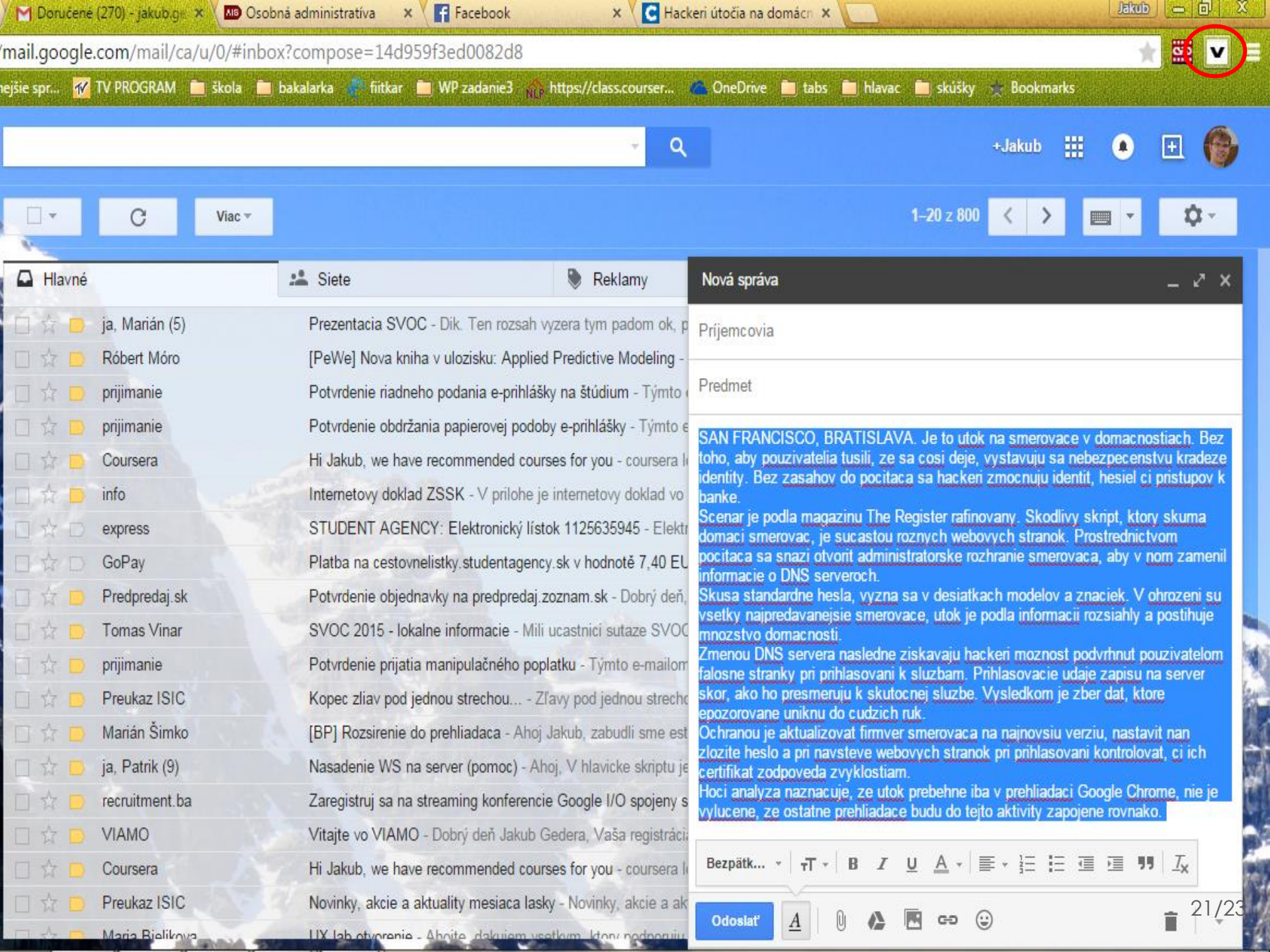
Výsledok dopytu príde vo formáte XML. Tag *in-sentence* obsahuje vetu, ktorú si zadal. Tag *output-sentence* obsahuje vetu, ktorá je po rekonštrukcii.

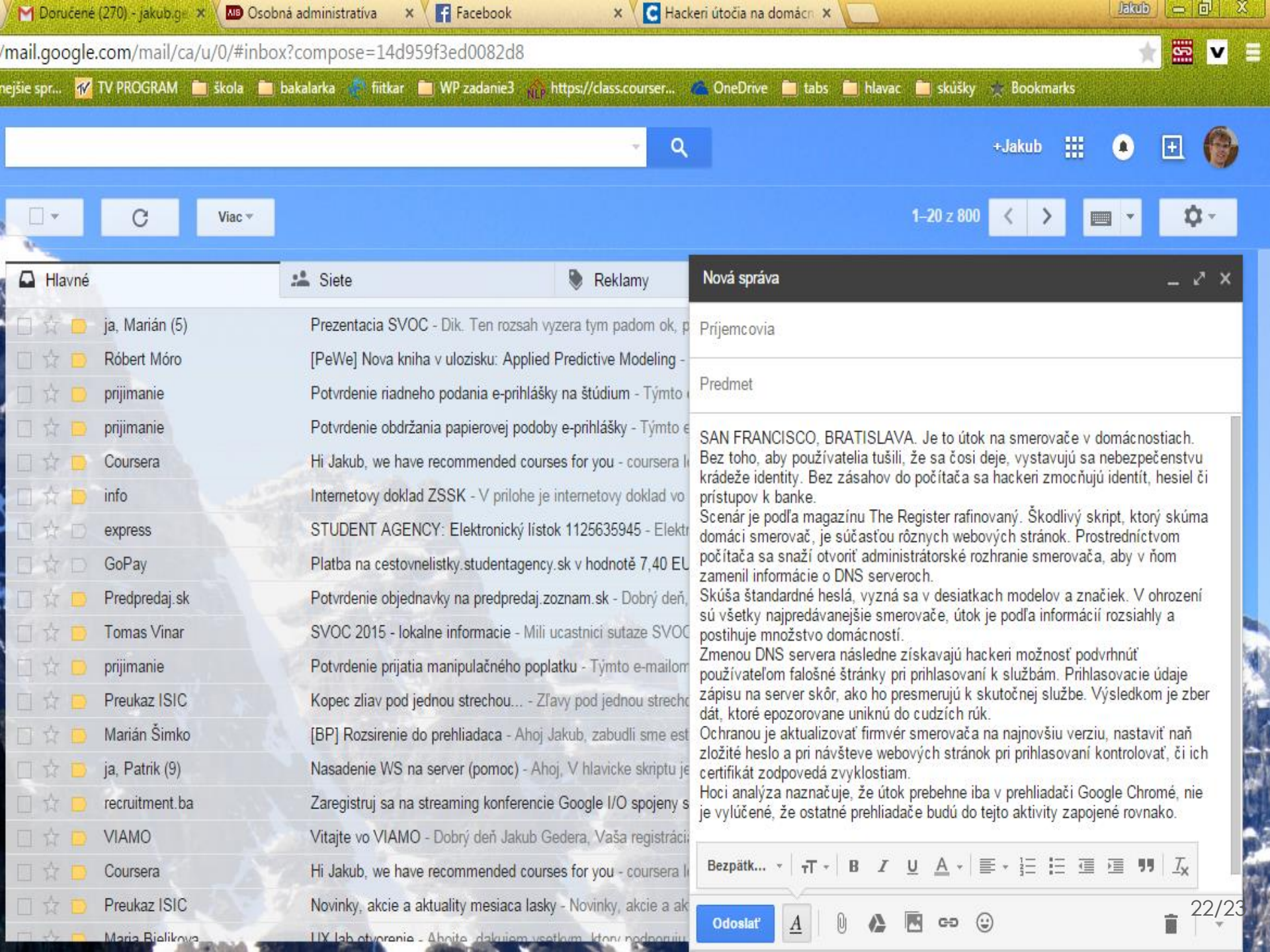
Príklad

Aplikáciu si môžeš vyskúšať aj priamo, zadaním textu nižšie.

Diakritikuj


TVnoviny.sk
[BP] Rozsirenie do prehli
Spracovanie textu na FIIT
Diakritikovač - Internetov
Jakub
https://chrome.google.com/webstore/detail/diakritikovač/afpbpaeodfjaekbjigblipncjnppcodc?hl=sk
Najsledovanejšie spr...
TV PROGRAM
škola
bakalarka
fiitkar
https://class.courser...
OneDrive
tabs
Internetový obchod Chrome
jakub.gedera@gmail.com
Diakritikovač
od autora jakub.gedera
PRIDANÉ DO CHROMU
(1)
Produktivita
Počet používateľov: 2
PREHĽAD
RECENZIE
PODPORA
PODOBNÉ
+1
0
Doplňovač diakritiky
Kompatibilné s vaším zariadením
Tento doplnok doplna diakritiku do oznaceneho textu
Aplikácia doplňujúca diakritiku. Označte text a kliknite na ikonu rozšírenia. Označený text je nahradený textom s doplnenou diakritikou. Užitočná vec pri písaní, resp.čítaní emailov alebo iného webového obsahu. Rekonštrukcia s úspešnosťou nad 98%. V priemere jedno slovo z 50 je rekonštruované chybne.
Nahlásiť zneužitie
Verzia: 1.0
Aktualizované: 4. júna 2015
Veľkosť: 3.76KB
Jazyk: slovenský
Aplikácie
Rozšírenia
Motívy
Produktivita
(1)


Doručené (270) - jakub.ge
Osobná administratíva
Facebook
Hackeri útočia na domácn
Jakub
mail.google.com/mail/ca/u/0/#inbox?compose=14d959f3ed0082d8
TV PROGRAM
škola
bakalarka
fitkar
WP zadanie3
https://class.courser...
OneDrive
tabs
hlavac
skúšky
Bookmarks
+Jakub
Viac
1–20 z 800
Hlavné
Siete
Reklamy
ja, Marián (5)
Prezentacia SVOC - Dik. Ten rozsah vyzera tym padom ok,
Róbert Móro
[PeWe] Nova kniha v ulozisku: Applied Predictive Modeling -
prijimanie
Potvrdenie riadneho podania e-prihlášky na štúdium - Týmto
prijimanie
Potvrdenie obdržania papierovej podoby e-prihlášky - Týmto
Coursera
Hi Jakub, we have recommended courses for you - coursera
info
Internetovy doklad ZSSK - V prilohe je internetovy doklad vo
express
STUDENT AGENCY: Elektronický lístok 1125635945 - Elektr
GoPay
Platba na cestovnelistky.studentagency.sk v hodnotě 7,40 EU
Predpredaj.sk
Potvrdenie objednavky na predpredaj.zoznam.sk - Dobrý deň,
Tomas Vinar
SVOC 2015 - lokalne informacie - Mili ucastnici sutaze SVOC
prijimanie
Potvrdenie prijatia manipulačného poplatku - Týmto e-mailom
Preukaz ISIC
Kopec zliav pod jednou strechou... - Zľavy pod jednou strecho
Marián Šimko
[BP] Rozsirenie do prehliadaca - Ahoj Jakub, zabudli sme est
ja, Patrik (9)
Nasadenie WS na server (pomoc) - Ahoj, V hlavicke skriptu je
recruitment.ba
Zaregistruj sa na streaming konferencie Google I/O spojeny s
VIAMO
Vitajte vo VIAMO - Dobrý deň Jakub Gedera, Vaša registráci
Coursera
Hi Jakub, we have recommended courses for you - coursera
Preukaz ISIC
Novinky, akcie a aktuality mesiaca lasky - Novinky, akcie a ak
Nová správa
Príjemcovia
Predmet
SAN FRANCISCO, BRATISLAVA. Je to utok na smerovace v domacnostiach. Bez toho, aby pouzivatelia tusili, ze sa cosi deje, vystavuju sa nebezpecenstvu kradeze identity. Bez zasahov do pocitaca sa hackeri zmocnuju identit, hesiel ci pristupov k banke.
Scenar je podla magazinu The Register rafinovany. Skodlivy skript, ktory skuma domaci smerovac, je sucastou roznych webovych stranok. Prostrednictvom pocitaca sa snazi otvorit administratorske rozhranie smerovaca, aby v nom zamenil informacie o DNS serveroch.
Skusa standardne hesla, vyzna sa v desiatkach modelov a znaciek. V ohrozeni su vsetky najpredavanejsie smerovace, utok je podla informacii rozsiahly a postihuje mnozstvo domacnosti.
Zmenou DNS servera nasledne ziskavaju hackeri moznost podvrhnut pouzivatelom falosne stranky pri prihlasovani k sluzbam. Prihlasovacie udaje zapisu na server skor, ako ho presmeruju k skutocnej sluzbe. Vysledkom je zber dat, ktore epozorovane uniknu do cudzich ruk.
Ochranou je aktualizovat firmver smerovaca na najnovsiu verziu, nastavit nan zlozite heslo a pri navsteve webovych stranok pri prihlasovani kontrolovat, ci ich certifikat zodpoveda zvyklostiam.
Hoci analyza naznacuje, ze utok prebehne iba v prehliadaci Google Chrome, nie je vylucene, ze ostatne prehliadace budu do tejto aktivity zapojene rovnako.
Bezpätk...
Odoslať
21/23


Doručené (270) - jakub.g
Osobná administratíva
Facebook
Hackeri útočia na domácn
Jakub
/mail.google.com/mail/ca/u/0/#inbox?compose=14d959f3ed0082d8
nejšie spr...
TV PROGRAM
škola
bakalarka
fiitkar
WP zadanie3
https://class.courser...
OneDrive
tabs
hlavac
skúšky
Bookmarks
+Jakub
Viac
1–20 z 800
Hlavné
Siete
Reklamy
ja, Marián (5) Prezentacia SVOC - Dik. Ten rozsah vyzera tym padom ok, p
Róbert Móro [PeWe] Nova kniha v ulozisku: Applied Predictive Modeling -
prijimanie Potvrdenie riadneho podania e-prihlášky na štúdium - Týmto
prijimanie Potvrdenie obdržania papierovej podoby e-prihlášky - Týmto e
Coursera Hi Jakub, we have recommended courses for you - coursera l
info Internetovy doklad ZSSK - V prilohe je internetovy doklad vo
express STUDENT AGENCY: Elektronický lístok 1125635945 - Elektr
GoPay Platba na cestovnelistky.studentagency.sk v hodnotě 7,40 EU
Predpredaj.sk Potvrdenie objednavky na predpredaj.zoznam.sk - Dobrý deň,
Tomas Vinar SVOC 2015 - lokalne informacie - Mili ucastnici sutaze SVOC
prijimanie Potvrdenie prijatia manipulačného poplatku - Týmto e-mailom
Preukaz ISIC Kopec zliav pod jednou strechou... - Zľavy pod jednou strecho
Marián Šimko [BP] Rozsirenie do prehliadaca - Ahoj Jakub, zabudli sme est
ja, Patrik (9) Nasadenie WS na server (pomoc) - Ahoj, V hlavicke skriptu je
recruitment.ba Zaregistruj sa na streaming konferencie Google I/O spojeny s
VIAMO Vitajte vo VIAMO - Dobrý deň Jakub Gedera, Vaša registráci
Coursera Hi Jakub, we have recommended courses for you - coursera l
Preukaz ISIC Novinky, akcie a aktuality mesiaca lasky - Novinky, akcie a ak
Maria Bielikova UX lab otvorenie - Ahojte, dakujem vsetkym, ktory podporuju
Nová správa
Príjemcovia
Predmet
SAN FRANCISCO, BRATISLAVA. Je to útok na smerovače v domácnostiach. Bez toho, aby používatelia tušili, že sa čosi deje, vystavujú sa nebezpečenstvu krádeže identity. Bez zásahov do počítača sa hackeri zmocňujú identít, hesiel či prístupov k banke.
Scenár je podľa magazínu The Register rafinovaný. Škodlivý skript, ktorý skúma domáci smerovač, je súčasťou rôznych webových stránok. Prostredníctvom počítača sa snaží otvoriť administrátorské rozhranie smerovača, aby v ňom zamenil informácie o DNS serveroch.
Skúša štandardné heslá, vyzná sa v desiatkach modelov a značiek. V ohrození sú všetky najpredávanejšie smerovače, útok je podľa informácií rozsiahly a postihuje množstvo domácností.
Zmenou DNS servera následne získavajú hackeri možnosť podvrhnúť používateľom falošné štránky pri prihlasovaní k službám. Prihlasovacie údaje zápisu na server skôr, ako ho presmerujú k skutočnej službe. Výsledkom je zber dát, ktoré epozorovane uniknú do cudzích rúk.
Ochranou je aktualizovať firmvér smerovača na najnovšiu verziu, nastaviť naň zložité heslo a pri návšteve webových stránok pri prihlasovaní kontrolovať, či ich certifikát zodpovedá zvyklostiam.
Hoci analýza naznačuje, že útok prebehne iba v prehliadači Google Chromé, nie je vylúčené, že ostatné prehliadače budú do tejto aktivity zapojené rovnako.
Bezpätk...
Odoslať
22/23

Záver

- Pri dopĺňaní diakritiky sme **využili kontext** pre nájdenie správneho variantu slova s diakritikou
- Priblížili sme aplikáciu diakritikovania bežným používateľom Webu formou **rozšírenia pre webový prehliadač**
- **Overili** sme **úspešnosť** diakritikovania
- Aplikácia bola zo strany používateľov ohodnotená **pozitívne** a vedia si ju predstaviť v **praxi** (emaily, blogy, neformálne dokumenty)
- Nie len pri písaní textov ale aj pri čítaní