جستجوی تقریبی برای زبان فارسی ایران،
پیشنویس ۳

Copyright © 2005 Sharif FarsiWeb, Inc.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

کلیهٔ حقوق برای شرکت فارسی‌وب شریف (سهامی خاص) محفوظ است.

اجازهٔ تکثیر، توزیع، و/یا تغییر این مستند تحت شرایط اجازه‌نامهٔ مستندات آزاد گنو (نسخهٔ 1.2 یا هر نسخهٔ جدیدتری که توسط بنیاد نرم‌افزارهای آزاد منتشر شود) داده می‌شود؛ بدون «قسمتهای بی‌تغییر»، «متن روی جلد»، یا «متن پشت جلد». یک نسخه از اجازه‌نامهٔ مذکور در بخش «اجازه‌نامهٔ مستندات آزاد گنو» آمده است.

۱ مقدمه

جستجوی تقریبی یا نادقیق، در مقابل جستجوی دقیق یا دودویی (binary)، به منظور یافتن عبارتها و زیررشته‌هایی صورت می‌گیرد که تفاوت قابل اغماض با رشتهٔ مورد جستجوی اصلی دارند. یعنی در این نوع جستجو نیازی نیست که عبارت مورد جستجو و عبارت یافته شده نویسه‌به‌نویسه و بایت‌به‌بایت برابر باشند، بلکه کافی است که از نظر خواننده معادل باشند.

جستجوی تقریبی از پردازشهایی است که به زبان متن و ساختار آن وابسته است. مثلاً در زبانهایی که به خط لاتینی نوشته می‌شوند، معروفترین جستجوهای تقریبی، جستجویی است که کوچکی و بزرگی حروف را در نظر نمی‌گیرد و جستجویی که از اَکسانهای روی حروف چشمپوشی می‌کند. جستجوی تقریبی مطلوب و بی‌نقص برای زبان فارسی، مانند هر زبان دیگری، نیاز به اطلاعات زبانی، معنایی، واژگانی، و مانند آنها دارد.

مثلاً اگر بخواهیم برای عبارتی مثل «کُرد» (نام یکی از اقوام ایرانی) در متن جستجو کنیم، که ممکن است بدون اِعراب نیز نوشته شده باشد، تفکیک آن از عبارت «کَرد» (فعل سوم شخص مفرد ماضی)، در مواردی که بدون اِعراب نوشته می‌شود، بدون تحلیل زبانی و دستوری متن ممکن نخواهد بود.

یا اگر بخواهیم به دنبال عبارت «طوطی‌ای» بگردیم، که در بعضی رسم‌الخطها، از جمله رسم‌الخط مرکز نشر دانشگاهی، به شکل «طوطیی» نوشته می‌شود، برای پیدا کردن هر دو حالت املای این عبارت باید از قوانین رسم‌الخطی اطلاع داشته باشیم و دنبالهٔ ‹ی، فاصلهٔ مجازی، الف، ی› را فقط در موارد مشخصی که شامل آن قانون خاص رسم‌الخطی می‌شوند (مثلاً در آخر کلمات در بعضی از مواردی که حرف «ی» صدای [i:] بدهد) با دنبالهٔ ‹ی، ی› معادل بگیریم.

یا ممکن است به دنبال عبارت «رایانه» بگردیم و از آنجا این کلمه به نظر فرهنگستان زبان و ادب فارسی با عبارت «کامپیوتر» هم‌معنی است، بخواهیم عبارت «کامپیوتر» نیز یافته شود. در چنین مواردی باید از روابط معنایی واژه‌ها و عبارات نیز اطلاع داشته باشیم.

در توصیف حاضر، جستجوی تقریبی به شکل بسیار ساده‌شده تعریف شده است. به این معنی که مسائل معنایی، دستوری، رسم‌الخطی، و واژگانی در آن در نظر گرفته نشده‌اند، بلکه فقط شیوه‌هایی که توسط الگوریتمها و جداول ساده قابل پیاده‌سازی هستند در نظر گرفته شده‌اند. به عبارت دیگر، توصیف حاضر فقط به پایه‌ای‌ترین نیازهای جستجوی تقریبی زبان فارسی می‌پردازد. مشابهتهایی که توصیف حاضر به آنها می‌پردازد عبارتند از:

- غلط‌های کدگذاری، از قبیل استفادهٔ نااستاندارد از حرف عربی کاف (ك) به جای حرف کاف فارسی (ک) و حرف عربی ی (ي) به جای حرف ی فارسی (ی)
- غلطهای املایی معروف یا نامحسوس، از قبیل نوشتن کلمهٔ «مؤمن» به شکل «مومن» یا «مسئله» به شکل «مسأله»

- تفاوت در اِعراب‌گذاری عبارات و وجود یا عدم وجود نویسه‌های ضمنی

۲ اصطلاحات و تعاریف

نویسه

نویسه معادل character در استاندارد یونی‌کد است.

دنباله، دنبالهٔ مجموعه‌ای

در این توصیف منظور از دنباله، فهرست متناهی و نمایه‌داری از عددها، نویسه‌ها، رشته‌ها، یا مجموعه‌ها است. دنبالهٔ مجموعه‌ای، دنباله‌ای از مجموعه‌ها است.

رشته

در این توصیف، به هر دنباله‌ای از نویسه‌های یونی‌کدی، رشته گفته می‌شود. در این توصیف فرض می‌شود که رشته‌ها از ترتیب معنایی (در چارچوب الگوریتم دوجهتهٔ یونی‌کد) پیروی می‌کنند.

رشتهٔ خالی

به رشتهٔ دارای طول صفر که هیچ نویسه‌ای ندارد، رشتهٔ خالی گفته می‌شود. در توصیف حاضر این رشته با علامت ε نشان داده می‌شود.

مشابه

در توصیف حاضر، دو نویسه را مشابه می‌نامیم اگر یکسان باشند یا در چارچوب مشابهت‌های در نظر گرفته شده در این توصیف همواره یا در بعضی شرایط به جای نویسهٔ دیگر استفاده شوند. نویسه‌هایی که در این توصیف مشابه در نظر گرفته می‌شوند در بخشهای بعدی مشخص شده‌اند. این رابطهٔ تشابه، بنا به تعاریف نظریهٔ مجموعه‌ها، متعدی/تراگذری نیست، به این معنی که اگر نویسهٔ a با نویسهٔ b مشابه بوده و نویسهٔ b نیز با نویسهٔ c مشابه باشد لزوماً نویسهٔ a با نویسهٔ c مشابه نیست. ولی این رابطه بازتابی/انعکاسی و تقارنی هست، به این معنی که هر نویسهٔ a همیشه با خودش مشابه است و اگر نویسهٔ a با نویسهٔ b مشابه باشد، نویسهٔ b نیز با نویسهٔ a مشابه خواهد بود.

نویسهٔ ضمنی

در این توصیف، بعضی نویسه‌ها که معمولاً در درج آنها در متن اهمال و بی‌دقتی صورت می‌گیرد و بنابراین ممکن است در جای لازم وارد نشده باشند (مثل نویسهٔ اتصال مجازی)، یا حضور آنها هیچ تأثیری در نتیجهٔ جستجوی تقریبی ندارد (مثل نویسهٔ کشیدگی فارسی)، نویسه‌های ضمنی گفته می‌شود. فهرست کامل این نویسه‌ها و روش رفتار الگوریتم جستجوی تقریبی با آنها در متن آمده است.

نویسهٔ فرعی

در این توصیف، به نویسه‌هایی که درج آنها در متن اختیاری است یا معمولاً اختیاری محسوب می‌شود نویسهٔ فرعی می‌گوییم. مثلاً، اِعرابها معمولاً چنین خاصیتی دارند. با وجودی که این نویسه‌ها اختیاری هستند، وجود آنها در هر شرایطی قابل چشمپوشی نیست. مثلاً عبارت «مُلک» نباید هیچگاه در جستجوی تقریبی با «مَلِک» منطبق شود. فهرست کامل این نویسه‌ها و روش رفتار الگوریتم جستجوی تقریبی با آنها در متن آمده است.

ردهٔ تشابه

در این توصیف، به مجموعهٔ نویسه‌هایی که با یک نویسهٔ مشخص مشابه باشند ردهٔ تشابه آن نویسه می‌گوییم. به‌علاوه، اگر نویسه‌ای ضمنی باشد، رشتهٔ خالی نیز به ردهٔ تشابه آن نویسه اضافه می‌شود. مثلاً بنا به تعاریف موجود در بخشهای بعدی این استاندارد، ردهٔ تشابه نویسهٔ «ئ» (ی با همزهٔ بالا) عبارت است از مجموعه‌ای شامل «أ» (الف با همزهٔ بالا)، «ؤ» (واو با همزهٔ بالا)، «ی»، و خود نویسهٔ «ئ». به همین شکل، ردهٔ تشابه نویسهٔ «فاصلهٔ مجازی»، عبارت است از مجموعه‌ای شامل نویسهٔ «فاصله»، رشتهٔ خالی (ε)، و خود نویسهٔ «فاصلهٔ مجازی».

اَبررشته

در این توصیف، به دنبالهٔ مجموعه‌ای‌ای که از پشت هم گذاشتن رده‌های تشابه نویسه‌های یک رشته حاصل می‌شود، ابررشتهٔ نظیر آن رشته می‌گوییم. مثلاً ابررشتهٔ معادل رشتهٔ «وی» دنباله‌ای به طول دو است که مجموعهٔ اول آن ردهٔ تشابه نویسهٔ «و» و مجموعهٔ دوم آن ردهٔ تشابه نویسهٔ «ی» است.

۳ توصیه‌ها و نیازهای پیاده‌سازی

لزومی ندارد که پیاده‌سازیهای این توصیف دقیقاً همان مراحل مشخص‌شده در این توصیف را انجام دهند. برای سازگاری با این توصیف کافی است خروجی جستجوی هر رشتهٔ فارسی در هر متن فارسی همواره و در همهٔ حالتها همان خروجی‌ای باشد که از این الگوریتم حاصل می‌شود.

این توصیف در مورد جستجوی تقریبی برای زبانهای غیر از فارسی صحبت نمی‌کند و تنها به جستجوی تقریبی عبارتهای فارسی در متنهایی که به زبان فارسی نوشته شده‌اند می‌پردازد.

پیشنهاد می‌شود که پیاده‌سازیها برای آزمایش پیاده‌سازیشان از داده‌های آزمون ضمیمهٔ این توصیف استفاده کنند. در این داده‌های آزمون، یک نمونهٔ متن فارسی و فهرستی از عبارتهای مورد جستجو به همراه نمایهٔ جایشان در جستجوی تقریبی آن متن آمده است. پیاده‌سازیهای دقیق این استاندارد باید همهٔ این عبارتها را در همهٔ آن جاها بیابند و در عین حال آنها را در جاهای دیگر نیابند.

مشابهت‌های مشخص‌شده در توصیف حاضر می‌توانند در پیاده‌سازیهای این توصیف، بسته به نیاز کاربر و امکانات قابل پیاده‌سازی، جستجوهای تقریبی دیگری نیز به این توصیف اضافه کنند یا جستجوهای تقریبی توصیف حاضر را دقیقتر کنند. پیاده‌سازیهایی که ادعای سازگاری با توصیف حاضر را کنند، باید حالتی را نیز پشتیبانی کنند که عبارتها را دقیقاً به شکل مشخص‌شده در توصیف حاضر بیابد، به این معنی که عبارت مورد نظر در هیچ جایی غیر از آن چه توصیف حاضر مشخص می‌کند یافته نشود و در همهٔ جاهایی که توصیف حاضر مشخص می‌کند نیز یافته شود.

۴ نمادها

در این توصیف، نمادها به شکل تعیین شده در زیر به کار می‌روند:

منظور از U+20AC نویسهٔ «علامت یورو» از استاندارد یونی‌کد است که کد شانزده‌شانزدهی آن 20AC است.

منظور از U+06F0..U+06F9، همهٔ نویسه‌هایی از استاندارد یونی‌کد است که کُدشان از 06F0 تا 06F9 است، که شامل ده نویسه می‌شود (این نویسه‌ها ارقام فارسی هستند).

منظور از <م، ـِ ل، ک> رشته‌ای از نویسه‌های «حرف فارسی میم»، «زیر فارسی»، «حرف فارسی لام»، و «حرف فارسی کاف» است. این رشته ممکن است برای خلاصه‌نویسی به شکل «مِلک» نیز نوشته شود. به همین ترتیب، منظور از <ن، ا، م، ه، فاصلهٔ مجازی، ه، ا> رشته‌ای است که به شکل «نامه‌ها» نیز نوشته می‌شود.

کدهای نمایشی در این توصیف به زبان پیتون هستند. این کدها فقط باید برای بهتر فهمیده‌شدن الگوریتم استفاده شوند و نباید به شکل مستقیم در برنامه‌ها استفاده شوند. (مشخصاً، مجوز این کدها، که مجوز مستندات آزاد گنو است، اجازه نمی‌دهد که این کدها بدون اجازهٔ مالک حقوق توصیف حاضر در هیچ برنامه‌ای استفاده شوند.) به‌علاوه، این کدها به منظور کمک به بهتر فهمیده‌شدن، ساده شده‌اند و ممکن است موارد خاص را در نظر نگرفته باشند و/یا برای پیاده‌سازی سریع مناسب نباشند. برای اطمینان از درستی پیاده‌سازیهای این توصیف، نباید از این کدها استفاده کرد، بلکه باید از داده‌های آزمون ضمیمهٔ این توصیف استفاده کرد.

۵ پیش‌پردازش رشته‌های فارسی

در این توصیف فرض می‌شود که رشته و متن مورد جستجو به صورت مطلوب جهت پردازش برای جستجوی فارسی هستند. به منظور تبدیل رشته‌ها به صورت مطلوب، باید پیش‌پردازش زیر روی آنها انجام شود. توجه شود که از آنجا که متنی که در آن جستجو صورت می‌گیرد نیز یک رشته است، این پیش‌پردازش باید روی آن متن هم انجام شود. (توجه کنید که لزومی ندارد این پیش‌پردازش به همین شکل روی متن انجام شود، بلکه کافی است کل الگوریتم جستجو طوری رفتار کند که انگار چنین پیش‌پردازشی روی آن انجام شده است.)

ورودی: رشتهٔ s جهت پیش‌پردازش.

خروجی: رشتهٔ پیش‌پردازش‌شدهٔ s.

۱) اگر از نویسه‌های بلوکهای «شکلهای نمایشی عربی» یونی‌کد (با نام انگلیسی Arabic Presentation Forms) یعنی نویسه‌های محدودهٔ U+FB50..U+FDFF (به جز U+FD3E و U+FD3F) و U+FE80 تا U+FEFE در رشتهٔ s استفاده شده بود، آن قسمتهای رشته را با حفظ کامل معنا و در نظر گرفتن احتمال نیاز به درج نویسه‌های فاصلهٔ مجازی و اتصال مجازی، به رشته‌ای که از نویسه‌های عادی فارسی، یعنی U+0600..U+06FF، استفاده می‌کند تبدیل کنید. (مثلاً رشتهٔ <ع آخر، م اول> («عم») به رشتهٔ <اتصال مجازی، ع، فاصلهٔ مجازی، م، اتصال مجازی> و رشتهٔ «لآ» به رشتهٔ <اتصال مجازی، ل، آ> تبدیل می‌شود.)

۲) نویسه‌های فاصلهٔ مجازی و اتصال مجازی موجود در رشتهٔ s را در صورتی که زائد بودند، یعنی حذف آنها در هیچ شرایطی در شیوهٔ نمایش رشته‌ها تغییری ایجاد نمی‌کرد، حذف کنید. (مثلاً رشتهٔ <ل، ک، فاصلهٔ مجازی، فاصلهٔ مجازی، ل، ک> («لک‌لک») به رشتهٔ <ل، ک، فاصلهٔ مجازی، ل، ک> تبدیل می‌شود چون یکی از نویسه‌های فاصلهٔ مجازی زائد است و در هیچ شرایطی تغییری در نمایش رشته ایجاد نمی‌کند.)

۳) رشتهٔ s را به صورت نرمال C (NFC)، مشخص شده در ضمیمهٔ ۱۵ استاندارد یونی‌کد، تبدیل کنید. (مثلاً رشتهٔ <ر، الف، ـٔ، ی> به رشتهٔ <ر، أ، ی> تبدیل می‌شود.)

۶ تشکیل ابررشته از رشتهٔ مورد جستجو

الگوریتم زیر برای تشکیل یک ابررشته از یک رشتهٔ مورد جستجو به کار می‌رود.

ورودی: رشتهٔ پیش‌پردازش‌شدهٔ s به طول l.

خروجی: ابررشتهٔ ss.

۱) ابررشتهٔ ss را خالی کنید.

۲) گام ۳ را برای هر i از ۰ تا $l-$۱، اجرا کنید. سپس الگوریتم را تمام کنید.

۳) اگر s_i یک نویسهٔ ضمنی است و هیچ نویسهٔ دیگر مشابهی ندارد، یا s_i نویسهٔ فرعی است، کاری نکنید. وگرنه ردهٔ تشابه نویسهٔ s_i را به انتهای ss اضافه کنید.

کد نمایشی:

```
ss = []
i = 0
while i < l:
    if not issecondary (s [i]):
        c = simclass [s [i]]
        if len (c) > 2 or not ("" in c):
            ss.append (c)
    i += 1
return ss
```

۷ جستجوی یک ابررشته در یک متن

در این الگوریتم یک ابررشته در یک متن (که در واقع یک رشتهٔ دیگر است) جستجو می‌شود. این الگوریتم مرحلهٔ اول جستجوی رشته است و همهٔ موارد یافته شدن در این الگوریتم لزوماً موارد یافته شدن رشتهٔ مورد جستجو در متن نیستند.

ورودی: ابررشتهٔ ss به طول l و رشتهٔ پیش‌پردازش‌شدهٔ t به طول m.

خروجی: فهرست I از نمایه‌های مکانهایی در رشتهٔ t که ابررشته یافته می‌شود.

۱) فهرست I را خالی کنید. به ازای هر i از ۰ تا $m-۱$ گام ۳ و ۴ را اجرا کنید. سپس الگوریتم را تمام کنید.

۲) قرار دهید $p = ۰$. تا زمانی که $ss[p]$ شامل رشتهٔ خالی بود p را افزایش دهید. مجموعهٔ pl را برابر همهٔ اعداد از ۰ تا p قرار دهید.

۳) برای هر j از i تا $m-۱$ گام ۵ را اجرا کنید. هرگاه در انتهای گام ۵ مجموعهٔ pl تهی بود یا شامل عدد l بود، شمارش را پایان دهید و اگر pl شامل عدد l بود، i را به فهرست I اضافه کنید.

۴) اگر $t[j]$ نویسهٔ فرعی یا ضمنی بود کاری نکنید، وگرنه گام ۶ را اجرا کنید.

۵) مجموعهٔ nl را برابر تهی قرار دهید. برای هر k که عضو مجموعهٔ pl بود، گام ۷ را اجرا کنید. سپس قرار دهید $pl = nl$.

۶) اگر $t[j]$ عضو $ss[k]$ بود، $k+۱$ را به مجموعهٔ nl اضافه کنید. سپس قرار دهید $p = k+۱$ و تا زمانی که $ss[p]$ شامل رشتهٔ خالی بود p را افزایش دهید و به مجموعهٔ nl اضافه کنید.

کد نمایشی:

```
I = []
i = 0
while i < m:
    p = 0
    while p < l and "" in ss [p]:
        p += 1
    pl = range (p + 1)
    j = i
    while j < m and pl != [] and not (l in pl):
        if not (issecondary (t[j]) or "" in simclass [t[j]]):
            nl = []
            for k in pl:
                if t [j] in ss [k]:
                    p = k + 1
                    nl.append (p)
                    while p < l and "" in ss [p]:
                        p += 1
                        nl.append (p)
            pl = sorted_set (nl)
        j += 1
    if l in pl:
        I.append (i)
    i += 1
return I
```

۸ ایجاد دنبالهٔ فرعی یک رشته

این الگوریتم، از یک رشتهٔ ورودی یک دنباله از رشته‌ها می‌سازد، به این ترتیب که هر یک از این رشته‌ها، اِعرابها و نویسه‌های فرعی دیگری است که روی هر یک از حروف و نویسه‌های پایهٔ رشته ساخته شده است. مثلاً اگر رشتهٔ ورودی <ت، ـَ، م، ـَ، ل، ـُ، ـّ، ک> («تَمَلُّک» باشد، دنبالهٔ خروجی دنباله‌ای است از رشتهٔ خالی، <ـَ>، <ـَ>، <ـُ، ـّ>، و رشتهٔ خالی.

ورودی: رشتهٔ پیش‌پردازش‌شدهٔ s به طول l.

خروجی: دنبالهٔ رشته‌های sl که دنبالهٔ فرعی رشتهٔ s است.

۱) رشتهٔ t و دنبالهٔ sl را خالی کنید و برای هر i از ۰ تا $l-$۱ گام ۲ را اجرا کنید. سپس رشتهٔ t را به انتهای دنبالهٔ sl اضافه کنید و الگوریتم را تمام کنید.

۲) اگر نویسهٔ $s[i]$ ضمنی است، کاری نکنید، وگرنه اگر نویسهٔ $s[i]$ فرعی بود آن را به انتهای رشتهٔ t اضافه کنید ولی اگر فرعی نبود، رشتهٔ t را به انتهای دنبالهٔ sl اضافه کنید و سپس رشتهٔ t را خالی کنید.

کد نمایشی:

```
t = ""
sl = []
i = 0
while i < l:
    if issecondary (s [i]):
        t += s [i]
    elif not ("" in simclass [s [i]]):
        sl.append (t)
        t = ""
    i += 1
sl.append (t)
return sl
```

۹ مطابق بودن دو دنبالهٔ فرعی

این الگوریتم، بررسی می‌کند که دو دنبالهٔ فرعی که از دو رشتهٔ شبیه به هم حاصل شده است قابل تطابق هستند یا نه. مثلاً، دنباله‌های فرعی دو رشتهٔ «ملَّک» و «مُلَک»، از آنجا که ممکن است دو کلمهٔ یکسان با اعراب گذاری متفاوت باشند، مطابق خواهند بود ولی دنباله‌های فرعی دو رشتهٔ «مِلْک» و «مُلْک» از آنجا که امکان یکی بودن کلمات نظیرشان وجود ندارد مطابق نیستند.

باید توجه شود که نمی‌توان در مواردی که قبلاً عبارتهای پایهٔ مقایسه نشده و مطابق تشخیص داده نشده‌اند، از این الگوریتم استفادهٔ مفیدی کرد.

ورودی: دو دنبالهٔ فرعی و همطول a و b به طول l جهت مقایسه از نظر تطابق.

خروجی: «درست» در صورتی که دو دنباله مطابق باشند و «نادرست» در صورتی که مطابق نباشند.

۱) برای هر i از ۰ تا $l-$۱، گام ۲ را اجرا کنید. سپس به گام ۳ بروید.

۲) اگر همهٔ نویسه‌های رشتهٔ $a[i]$ در رشتهٔ $b[i]$ هم وجود داشت، یا اگر همهٔ نویسه‌های رشتهٔ $b[i]$ در رشتهٔ $a[i]$ هم وجود داشت، کاری نکنید. وگرنه خروجی الگوریتم «نادرست» است و الگوریتم را تمام کنید.

۳) خروجی الگوریتم «درست» است. الگوریتم را تمام کنید.

کد نمایشی:

```
i = 0
while i < l:
    ok = True
    for c in a[i]:
        if not (c in b[i]):
            ok = False
            break
    if not ok:
        ok = True
        for c in b[i]:
            if not (c in a[i]):
            ok = False
            break
        if not ok:
            return False
    i += 1
return True
```

۱۰ جستجوی یک عبارت در یک متن

این الگوریتم اصلی توصیف حاضر است که به جستجوی یک عبارت در یک متن می‌پردازد. این الگوریتم ابتدا از عبارت مورد جستجو یک ابررشته می‌سازد و آن را در متن جستجو می‌کند. سپس در تک‌تک موارد یافته شدن ابررشته، عبارت اولیه از نظر نویسه‌های فرعی از قبیل اِعرابها با محل یافته شدنش در متن مقایسه می‌شود تا فقط در صورتی که مطابقت دارند، آن مکان در خروجی در نظر گرفته شود.

ورودی: رشتهٔ s به طول l به‌عنوان عبارت و رشتهٔ t به طول m به عنوان متن.

خروجی: فهرست I از نمایه‌های مکانهایی در رشتهٔ t که رشتهٔ s یافته می‌شود.

۱) رشته‌های s و t را بر اساس الگوریتم «پیش‌پردازش رشته‌های فارسی» پردازش کنید. اجرای این گام در صورتی که بتوان فرض کرد رشته‌های ورودی از قبل برای پردازش توسط الگوریتم حاضر مناسبند، لزومی ندارد.

۲) با استفاده از الگوریتم تشکیل ابررشته از رشتهٔ مورد جستجو، از رشتهٔ s، ابررشتهٔ ss را بسازید.

۳) الگوریتم جستجوی ابررشته در متن را برای ابررشتهٔ ss و رشتهٔ t انجام دهید و فهرست J را برابر خروجی آن قرار دهید.

۴) الگوریتم ایجاد دنبالهٔ فرعی را روی رشتهٔ s اجرا کنید و sl را برابر خروجی آن قرار دهید.

۵) فهرست I را خالی کنید. برای هر i که عضو J بود، گام ۶ و ۷ را اجرا کنید. سپس الگوریتم را تمام کنید.

۶) رشتهٔ u را زیررشته‌ای از رشتهٔ t قرار دهید که از نمایهٔ i شروع شده و تا انتهای رشته ادامه می‌یابد. الگوریتم ایجاد دنبالهٔ فرعی را روی رشتهٔ u اجرا کنید و ul را برابر خروجی آن قرار دهید.

۷) اگر تعداد رشته‌های دنبالهٔ ul از تعداد رشته‌های دنبالهٔ sl کمتر بود، کاری نکنید. وگرنه گام ۸ و ۹ را اجرا کنید.

۸) اگر تعداد رشته‌های دنبالهٔ ul از تعداد رشته‌های دنبالهٔ sl بیشتر بود، از انتهای دنبالهٔ ul آن قدر رشته حذف کنید تا تعداد رشته‌های دو دنباله برابر شوند.

۹) اگر دو دنبالهٔ فرعی sl و ul مطابق بودند، i را به فهرست I اضافه کنید.

کد نمایشی:

```
s = prep (s)
t = prep (t)
ss = superstring (s)
J = findsuperstring (ss, t)

sl = secondarysequence (s)
I = []
for i in J:
    u = t[i:]
    ul = secondarysequence (u)
    if len (ul) < len (sl):
        continue
    if len (ul) > len (sl):
        ul = ul [:len (sl)]
    if secondarymatch (sl, ul):
        I.append (i)
return I
```

۱۱ فهرست مشابهت‌های نویسه‌ها

جدول زیر فهرست مشابهت‌های نویسه‌ها است. در این جدول اگر دو نویسه مشابه باشند، این مشابهت تنها یک بار ذکر شده است. به‌علاوه، مشابهت هر نویسه‌ای با خودش در این جدول ذکر نشده است.

این فهرست می‌تواند بسته به نیاز کاربرد گسترش یابد:

جدول ۱: فهرست مشابهت‌های نویسه‌ها

نام نویسهٔ اول	شکل نویسهٔ اول	کد یونی‌کد	نام نویسهٔ دوم	شکل نویسهٔ دوم	کد یونی‌کد
حرف فارسی آ	آ	U+0622	حرف فارسی الف	ا	U+0627
حرف الف وصل	ٱ	U+0671	حرف فارسی الف	ا	U+0627
حرف فارسی الف با همزه‌ی بالا	أ	U+0623	حرف فارسی الف	ا	U+0627
حرف فارسی الف با همزه‌ی بالا	أ	U+0623	حرف فارسی ی با همزه‌ی بالا	ئ	U+0626
حرف الف با همزه‌ی پایین	إ	U+0625	حرف فارسی الف	ا	U+0627
حرف فارسی واو با همزه‌ی بالا	ؤ	U+0624	حرف فارسی واو	و	U+0648
حرف فارسی واو با همزه‌ی بالا	ؤ	U+0624	حرف فارسی ی با همزه‌ی بالا	ئ	U+0626
حرف فارسی ی با همزه‌ی بالا	ئ	U+0626	حرف فارسی ی	ی	U+06CC
حرف فارسی ت	ت	U+062A	حرف ت گرد	ة	U+0629

جدول ۱: فهرست مشابهت‌های نویسه‌ها					
نام نویسهٔ اول	شکل نویسهٔ اول	کد یونی‌کد	نام نویسهٔ دوم	شکل نویسهٔ دوم	کد یونی‌کد
حرف فارسی کاف	ک	U+06A9	حرف کاف عربی	ك	U+0643
حرف فارسی ه	ه	U+0647	حرف ت گرد	ة	U+0629
حرف فارسی ی	ی	U+06CC	حرف ی عربی بی‌نقطه	د	U+0649
حرف فارسی ی	ی	U+06CC	حرف ی عربی نقطه‌دار	ي	U+064A
رقم فارسی صفر	۰	U+06F0	رقم صفر عربی	٠	U+0660
رقم فارسی یک	۱	U+06F1	رقم یک عربی	١	U+0661
رقم فارسی دو	۲	U+06F2	رقم دو عربی	٢	U+0662
رقم فارسی سه	۳	U+06F3	رقم سه عربی	٣	U+0663
رقم فارسی چهار	۴	U+06F4	رقم چهار عربی	٤	U+0664
رقم فارسی پنج	۵	U+06F5	رقم پنج عربی	٥	U+0665
رقم فارسی شش	۶	U+06F6	رقم شش عربی	٦	U+0666
رقم فارسی هفت	۷	U+06F7	رقم هفت عربی	٧	U+0667
رقم فارسی هشت	۸	U+06F8	رقم هشت عربی	٨	U+0668

جدول ۱: فهرست مشابهت‌های نویسه‌ها					
نام نویسهٔ اول	**شکل نویسهٔ اول**	**کد یونی‌کد**	**نام نویسهٔ دوم**	**شکل نویسهٔ دوم**	**کد یونی‌کد**
رقم فارسی نه	۹	U+06F9	رقم نه عربی	۹	U+0669
فاصله		U+0020	فاصلهٔ مجازی		U+200C

۱۲ فهرست نویسه‌های ضمنی

جدول زیر فهرست نویسه‌های ضمنی است. این فهرست می‌تواند بسته به نیاز کاربرد گسترش یابد:

جدول ۲: فهرست نویسه‌های ضمنی		
نام نویسه	**شکل نویسه**	**کد یونی‌کد**
کشیدگی فارسی	ـ	U+0640
فاصلهٔ مجازی		U+200C
اتصال مجازی		U+200D
نشانه‌ی چپ‌به‌راست		U+200E
نشانه‌ی راست‌به‌چپ		U+200F
زیرمتن چپ‌به‌راست		U+202A
زیرمتن راست‌به‌چپ		U+202B
پایان زیرمتن		U+202C
زیرمتن اکیداً چپ‌به‌راست		U+202D
زیرمتن اکیداً چپ‌به‌راست		U+202E
نشانه‌ی ترتیب بایت‌ها		U+FEFF

۱۳ فهرست نویسه‌های فرعی

جدول زیر فهرست نویسه‌های فرعی است. باید توجه شود که همهٔ نویسه‌های این فهرست اِعراب نیستند. این فهرست می‌تواند بسته به نیاز کاربرد گسترش یابد:

جدول ۳: فهرست نویسه‌های فرعی		
نام نویسه	**شکل نویسه**	**کد یونی‌کد**
حرف فارسی همزه	ء	U+0628
زبر فارسی (فتحه)	ـَ	U+064E
زیر فارسی (کسره)	ـِ	U+0650
پیش فارسی (ضمه)	ـُ	U+064F
دوزبر فارسی (تنوین نصب)	ـً	U+064B
دوزیر فارسی (تنوین جر)	ـٍ	U+064D
دوپیش فارسی (تنوین رفع)	ـٌ	U+064C
تشدید فارسی	ـّ	U+0651
ساکن فارسی	ـْ	U+0652
مد فارسی	ـٓ	U+0653
همزه‌ی فارسی بالا	ـٔ	U+0654

جدول ۳: فهرست نویسه‌های فرعی		
نام نویسه	شکل نویسه	کد یونی‌کد
همزه‌ی فارسی پایین	ـٕ	U+0655
الف مقصوره‌ی فارسی	ـٰ	U+0670
الف مقصوره‌ی فارسی پایین	ٰ	U+0647

۱۴ مراجع

[1] ISO/IEC 10646:2003, “Information technology — Universal Multiple-Octet Coded Character Set (UCS)”.

[2] The Unicode Consortium, *The Unicode Standard, Version 4.0*, Reading, Massachusetts, Addison-Wesley, 2003. ISBN 0-321-18578-1.

[3] Mark Davis, Unicode Standard Annex #9, “The Bidirectional Algorithm”, Version 4.0.1, 2004.

[4] Mark Davis and Ken Whistler, Unicode Technical Standard #10, “Unicode Collation Algorithm”, Version 4.0, 2004.

[5] Mark Davis and Martin Dürst, Unicode Standard Annex #15, “Unicode Normalization Forms”, Version 4.0.0, 2003.

[6] *Wikipedia: The Free Encyclopedia*, http://en.wikipedia.org/.

[۷] استاندارد ملی شمارهٔ ۳۳۴۲ سال ۱۳۷۲، «استاندارد کد تبادل اطلاعات ۸ بیتی فارسی».

[۸] استاندارد ملی شمارهٔ ۶۲۱۹ سال ۱۳۸۱، «فنّاوریِ اطلاعات ــ تبادل و شیوه‌ی نمایشِ اطلاعاتِ فارسی بر اساسِ یونی‌کُد».

[۹] *دستور خط فارسی، مصوب فرهنگستان زبان و ادب فارسی*، فرهنگستان زبان و ادب فارسی (نشر آثار)، تهران، ۱۳۸۱. شابک ۳-۱۳-۷۵۳۱-۹۶۴.

[۱۰] *شیوه‌نامه*، مرکز نشر دانشگاهی، ویرایش دوم، تهران، ۱۳۷۲. شابک ۷-۸۱۲۷-۰۱-۹۶۴.

۱۵ اجازه‌نامهٔ مستندات آزاد گنو (GNU Free Documentation License)

Version 1.2, November 2002

Copyright © 2000,2001,2002 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the

subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to

be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- **A.** Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- **B.** List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- **C.** State on the Title page the name of the publisher of the Modified Version, as the publisher.
- **D.** Preserve all the copyright notices of the Document.
- **E.** Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- **F.** Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- **G.** Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- **H.** Include an unaltered copy of this License.
- **I.** Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- **J.** Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- **K.** For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- **L.** Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- **M.** Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

- **N.** Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- **O.** Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your

rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

```
Copyright (c)  YEAR  YOUR NAME.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.2
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover
Texts.  A copy of the license is included in the section entitled
"GNU Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being
LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.