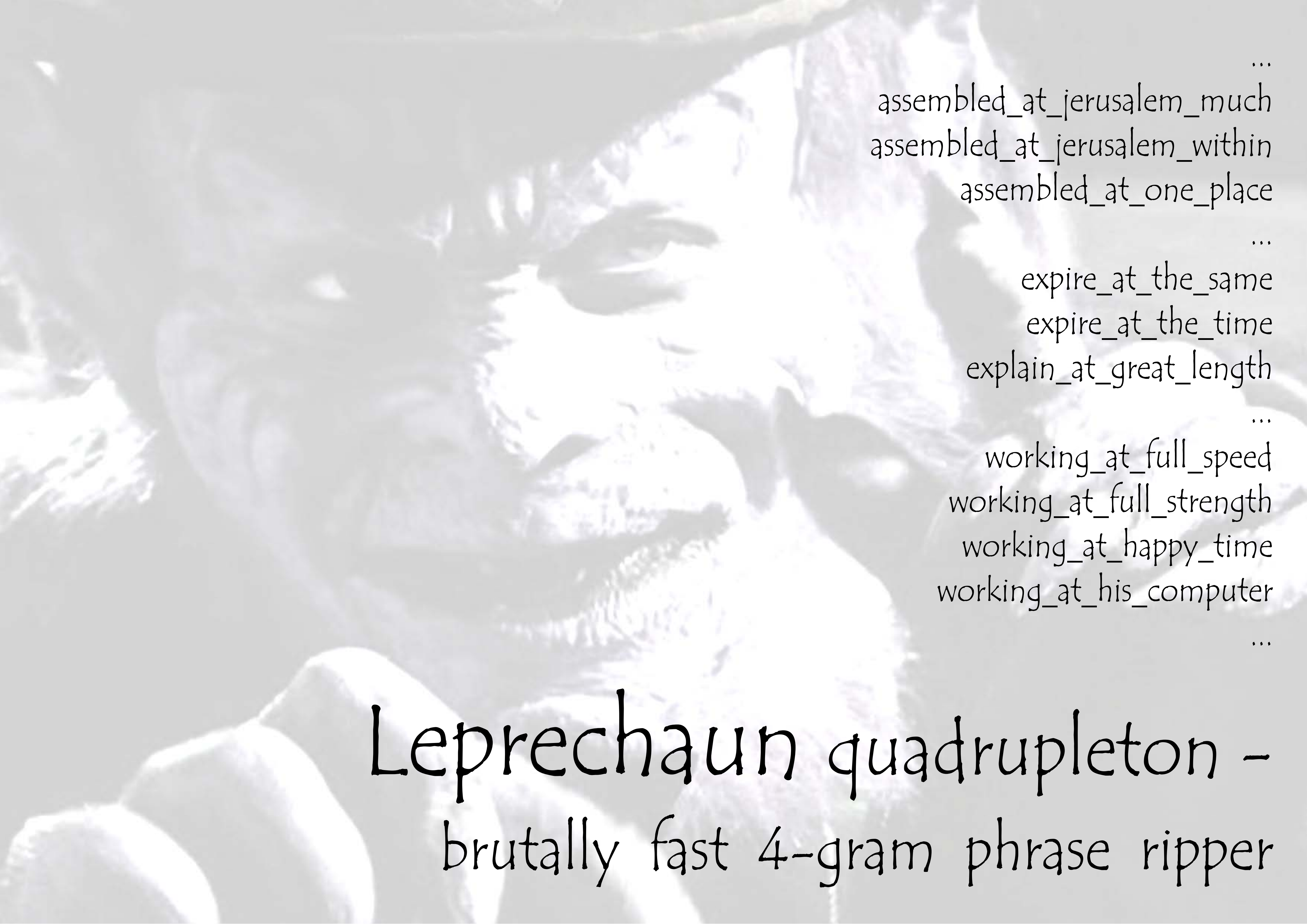

...

assembled_at_jerusalem_much

assembled_at_jerusalem_within

assembled_at_one_place

...

expire_at_the_same

expire_at_the_time

explain_at_great_length

...

working_at_full_speed

working_at_full_strength

working_at_happy_time

working_at_his_computer

...

Leprechaun quadrupleton - brutally fast 4-gram phrase ripper

Leprechaun_quadrupleton_AT_A_GLANCE.txt:

Contents

# Section A: [What is it?]:

Leprechaun_quadrupleton is an extremely fast 32bit console tool for creating 4-gram lists.

# Section B: [Practical usage: an example made in Windows command prompt]:

```
D:\_KA45F~1\_4>dir

12/12/2010  01:37 PM     1,111,609,996 googlebooks-eng-us-all-4gram-20090715-0.csv
01/26/2011  06:46 PM               315 googlebooks-eng-us-all-4gram-20090715-0.csv.EXCERPT
01/26/2011  05:13 AM           514,048 Leprechaun_quadrupleton_Intel_IA-32_11.1.exe

D:\_KA45F~1\_4>type googlebooks-eng-us-all-4gram-20090715-0.csv.EXCERPT
...
It cut me to    2002    4       4       4
It cut me to    2004    4       4       4
It cut me to    2005    6       6       6
It cut me to    2006    2       2       2
It cut me to    2007    1       1       1
It cut me to    2008    1       1       1
It declares that '      1816    1       1       1
It declares that '      1832    2       2       2
It declares that '      1833    1       1       1
It declares that '      1834    1       1       1
It declares that '      1838    1       1       1
...

D:\_KA45F~1\_4>dir *.excerpt/b>test.lst

D:\_KA45F~1\_4>Leprechaun_quadrupleton_Intel_IA-32_11.1.exe test.lst test.wrd
Leprechaun(Fast Greedy Word-Ripper), rev. 13_7pluses quadrupleton_r1, written by Svalqyatchx.
Leprechaun: 'Oh, well, didn't you hear? Bigger is good, but jumbo is dear.'
Kaze: Let's see what a 3-way hash + 6,602,752 Binary-Search-Trees can give us,
      also the performance of a 3-way hash + 6,602,752 B-Trees of order 3.
Size of input file with files for Leprechauning: 53
Allocating memory 424MB ... OK
Size of Input TEXTual file: 315
|; Word count: 39 of them 1 distinct; Done: 64/64
Bytes per second performance: 315B/s
Words per second performance: 39W/s
Flushing unsorted words ...
Time for making unsorted wordlist: 1 second(s)
Deallocated memory in MB: 424
Allocated memory for words in MB: 1
Allocated memory for pointers-to-words in MB: 1
Sorting(with 'MultiKeyQuickSortX26Sort' by J. Bentley and R. Sedgewick) ...
Sort pass 26/26 ...
Flushing sorted words ...
Time for sorting unsorted wordlist: 1 second(s)
Leprechaun: Done.

D:\_KA45F~1\_4>type test.wrd
it_cut_me_to

D:\_KA45F~1\_4>dir

12/12/2010  01:37 PM     1,111,609,996 googlebooks-eng-us-all-4gram-20090715-0.csv
01/26/2011  06:46 PM               315 googlebooks-eng-us-all-4gram-20090715-0.csv.EXCERPT
01/26/2011  06:56 PM               362 Gulliver's-Travels.pdf.txt.EXCERPT
01/26/2011  06:47 PM             4,108 Leprechaun.LOG
01/26/2011  05:13 AM           514,048 Leprechaun_quadrupleton_Intel_IA-32_11.1.exe
01/26/2011  06:47 PM                53 test.lst
01/26/2011  06:47 PM                14 test.wrd

D:\_KA45F~1\_4>dir Gulliver*.excerpt/b>test2.lst

D:\_KA45F~1\_4>type "Gulliver's-Travels.pdf.txt.EXCERPT"
...
And so unmeasureable is the ambition of princes, that he
seemed to think of nothing less than reducing the whole
empire of Blefuscu into a province, and governing it, by
a viceroy; of destroying the Big-endian exiles, and compelling
that people to break the smaller end of their eggs,
by which he would remain the sole monarch of the whole
world.
...
D:\_KA45F~1\_4>Leprechaun_quadrupleton_Intel_IA-32_11.1.exe test2.lst test2.wrd
```

```
Leprechaun(Fast Greedy Word-Ripper), rev. 13_7pluses quadrupleton_r1, written by Svalqyatchx.
Leprechaun: 'Oh, well, didn't you hear? Bigger is good, but jumbo is dear.'
Kaze: Let's see what a 3-way hash + 6,602,752 Binary-Search-Trees can give us,
      also the performance of a 3-way hash + 6,602,752 B-Trees of order 3.
Size of input file with files for Leprechauning: 36
Allocating memory 424MB ... OK
Size of Input TEXTual file: 362
|; Word count: 62 of them 41 distinct; Done: 64/64
Bytes per second performance: 362B/s
Words per second performance: 62W/s
Flushing unsorted words ...
Time for making unsorted wordlist: 1 second(s)
Deallocated memory in MB: 424
Allocated memory for words in MB: 1
Allocated memory for pointers-to-words in MB: 1
Sorting(with 'MultiKeyQuickSortX26Sort' by J. Bentley and R. Sedgewick) ...
Sort pass 26/26 ...
Flushing sorted words ...
Time for sorting unsorted wordlist: 1 second(s)
Leprechaun: Done.

D:\_KA45F~1\_4>type test2.wrd
and_compelling_that_people
and_so_unmeasureable_is
blefuscu_into_a_province
break_the_smaller_end
by_which_he_would
compelling_that_people_to
destroying_the_big_endian
empire_of_blefuscu_into
end_of_their_eggs
he_seemed_to_think
he_would_remain_the
is_the_ambition_of
less_than_reducing_the
monarch_of_the_whole
nothing_less_than_reducing
of_blefuscu_into_a
of_destroying_the_big
of_nothing_less_than
of_the_whole_world
people_to_break_the
reducing_the_whole_empire
remain_the_sole_monarch
seemed_to_think_of
smaller_end_of_their
so_unmeasureable_is_the
sole_monarch_of_the
than_reducing_the_whole
that_he_seemed_to
that_people_to_break
the_ambition_of_princes
the_big_endian_exiles
the_smaller_end_of
the_sole_monarch_of
the_whole_empire_of
think_of_nothing_less
to_break_the_smaller
to_think_of_nothing
unmeasureable_is_the_ambition
which_he_would_remain
whole_empire_of_blefuscu
would_remain_the_sole

D:\_KA45F~1\_4>dir *sword*.excerpt/b>test3.lst

D:\_KA45F~1\_4>type "[2003] When the Last Sword Is Drawn 7.7@imdb CD2.srt.EXCERPT"
...
497
01:02:27,956 --> 01:02:35,089
Morioka, in Nanbu.
It's pretty as a picture!

498
01:02:35,196 --> 01:02:38,723
There's nowhere like it in all Japan!

499
01:02:39,834 --> 01:02:43,827
The Morioka cherry blossom
splits through rock to bloom.

500
01:02:44,506 --> 01:02:48,875
The Morioka magnolia blooms
even facing north.

501
```

```
01:02:49,911 --> 01:02:54,848
So I want you to run ahead
of the times.

502
01:02:55,950 --> 01:03:00,046
Go wild. Bloom.
...
D:\_KA45F~1\_4>Leprechaun_quadrupleton_Intel_IA-32_11.1.exe test3.lst test3.wrd
Leprechaun(Fast Greedy Word-Ripper), rev. 13_7pluses quadrupleton_r1, written by Svalqyatchx.
Leprechaun: 'Oh, well, didn't you hear? Bigger is good, but jumbo is dear.'
Kaze: Let's see what a 3-way hash + 6,602,752 Binary-Search-Trees can give us,
      also the performance of a 3-way hash + 6,602,752 B-Trees of order 3.
Size of input file with files for Leprechauning: 62
Allocating memory 424MB ... OK
Size of Input TEXTual file: 488
|; Word count: 46 of them 25 distinct; Done: 64/64
Bytes per second performance: 488B/s
Words per second performance: 46W/s
Flushing unsorted words ...
Time for making unsorted wordlist: 1 second(s)
Deallocated memory in MB: 424
Allocated memory for words in MB: 1
Allocated memory for pointers-to-words in MB: 1
Sorting(with 'MultiKeyQuickSortX26Sort' by J. Bentley and R. Sedgewick) ...
Sort pass 26/26 ...
Flushing sorted words ...
Time for sorting unsorted wordlist: 1 second(s)
Leprechaun: Done.

D:\_KA45F~1\_4>type test3.wrd
ahead_of_the_times
blooms_even_facing_north
blossom_splits_through_rock
cherry_blossom_splits_through
i_want_you_to
it_in_all_japan
it_s_pretty_as
like_it_in_all
magnolia_blooms_even_facing
morioka_cherry_blossom_splits
morioka_magnolia_blooms_even
nowhere_like_it_in
pretty_as_a_picture
run_ahead_of_the
s_nowhere_like_it
s_pretty_as_a
so_i_want_you
splits_through_rock_to
the_morioka_cherry_blossom
the_morioka_magnolia_blooms
there_s_nowhere_like
through_rock_to_bloom
to_run_ahead_of
want_you_to_run
you_to_run_ahead

D:\_KA45F~1\_4>dir

12/12/2010  01:37 PM     1,111,609,996 googlebooks-eng-us-all-4gram-20090715-0.csv
01/26/2011  06:46 PM               315 googlebooks-eng-us-all-4gram-20090715-0.csv.EXCERPT
01/26/2011  06:56 PM               362 Gulliver's-Travels.pdf.txt.EXCERPT
01/26/2011  07:15 PM            12,354 Leprechaun.LOG
01/26/2011  05:13 AM           514,048 Leprechaun_quadrupleton_Intel_IA-32_11.1.exe
01/26/2011  06:47 PM                53 test.lst
01/26/2011  06:47 PM                14 test.wrd
01/26/2011  06:57 PM                36 test2.lst
01/26/2011  06:58 PM               945 test2.wrd
01/26/2011  07:14 PM                62 test3.lst
01/26/2011  07:15 PM               546 test3.wrd
01/26/2011  07:13 PM               488 [2003] When the Last Sword Is Drawn 7.7@imdb CD2.srt.EXCERPT
```

# Section C: [Benchmarking googlebooks-eng-us-all-4gram 1GB chunk]:

```
D:\_KA45F~1\_4>dir *.csv

12/12/2010  01:37 PM     1,111,609,996 googlebooks-eng-us-all-4gram-20090715-0.csv

D:\_KA45F~1\_4>dir *.csv/b>test_speed.lst

D:\_KA45F~1\_4>Leprechaun_quadrupleton_Intel_IA-32_11.1.exe test_speed.lst test_speed.wrd 4700
Leprechaun(Fast Greedy Word-Ripper), rev. 13_7pluses quadrupleton_r1, written by Svalqyatchx.
Leprechaun: 'Oh, well, didn't you hear? Bigger is good, but jumbo is dear.'
Kaze: Let's see what a 3-way hash + 6,602,752 Binary-Search-Trees can give us,
      also the performance of a 3-way hash + 6,602,752 B-Trees of order 3.
Size of input file with files for Leprechauning: 45
Allocating memory 1948MB ... OK
Size of Input TEXTual file: 1,111,609,996
/; Word count: 114,671,215 of them 381,294 distinct; Done: 64/64
```

```
Bytes per second performance: 55,580,499B/s
Words per second performance: 5,733,560W/s
Flushing unsorted words ...
Time for making unsorted wordlist: 20 second(s)
Deallocated memory in MB: 1948
Allocated memory for words in MB: 9
Allocated memory for pointers-to-words in MB: 2
Sorting(with 'MultiKeyQuickSortX26Sort' by J. Bentley and R. Sedgewick) ...
Sort pass 26/26 ...
Flushing sorted words ...
Time for sorting unsorted wordlist: 1 second(s)
Leprechaun: Done.

D:\_KA45F~1\_4>dir test_speed.wrd

01/26/2011  08:58 PM         8,754,241 test_speed.wrd

D:\_KA45F~1\_4>type test_speed.wrd|more
a_a_matter_of
a_account_of_the
a_and_b_above
a_and_the_range
a_are_equal_to
a_as_the_standard
a_as_the_sum
a_ba_in_speech
a_babe_when_he
a_baby_due_to
a_baby_s_interest
a_baby_was_in
a_baby_who_cries
a_bachelor_at_the
a_backdrop_of_green
a_background_of_night
a_background_of_substantial
a_backhanded_sort_of
a_backyard_full_of
a_bacteriological_diagnosis_of
a_bacterium_may_be
a_bad_character_for
a_bad_feeling_between
a_bad_fix_and
a_bad_hangover_from
a_bad_moral_character
a_bad_sunburn_and
a_bad_time_is
a_bad_way_as
a_bad_week_of
a_bag_marked_with
a_bag_of_fleas
a_bag_of_hot
a_bag_of_nerves
a_balance_between_births
a_balance_of_forces
a_balance_of_four
a_balance_of_tone
a_balance_sheet_is
a_balance_were_struck
a_balcony_behind_the
a_bald_catalogue_of
a_ball_as_you
a_ball_coming_at
a_ball_game_over
a_ball_shot_from
a_ballad_about_it
a_ballistic_galvanometer_as
a_balm_for_our
a_ban_on_such
a_band_leader_and
^C
D:\_KA45F~1\_4>
```

# Section D: [Defining what type of 4-gram becomes a Leprechaun-quadruple-phrase]:

- A quadruple-phrase has exactly 4 words;
- Only alpha ASCII chars form our words;
- A quadruple-phrase is lowercased i.e. contains only small letters 'a'..'z';
- A quadruple-phrase has length between 12 and 31 chars inclusive;
- Four words concatenated with '_' to form one quadruple-phrase;
- Symbols not allowed between 4 words forming a quadruple-phrase: '.', '!', '?', ':', ';', ',', '\t'.

# Section E: [The C source segment deciding what 4-gram enters our list of quadruple phrases]:

```
// Quadruple! [
// Sliding window for 'wrd': The incoming string 'a lot of things must' becomes 'a_lot_of_things' and 'lot_of_things_must':

// ain_t_that_a
// didn_t_feel_a
// i_didn_t_feel
// t_feel_a_thing
// t_that_a_cake

// 316
// 00:17:55,859 --> 00:17:58,447
// Ain't that a cake ? I didn't feel a thing !

                           if ( PLE_words_INITflag == 0 && ( (PLE_words != 0) || (PLE_words == 0 && wrdlen != 0) ) )
                           if ( workbyte == '.' || workbyte == '!' || workbyte == '?' || workbyte == ':' || workbyte == ';' || workbyte == ',' ||
     workbyte == '\t' ) {
                              PLE_words_INITflag = 1;
                           }
// Quadruple! ]

...

// Quadruple! [
PLE_words++;
if (PLE_words == 1)
       strcpy( wrd1st, wrd );
else if (PLE_words == 2)
       strcpy( wrd2nd, wrd );
else if (PLE_words == 3)
       strcpy( wrd3rd, wrd );
else if (PLE_words == 4) {
       strcpy( wrd4th, wrd );
       wrdlen = strlen(wrd1st)+strlen(wrd2nd)+strlen(wrd3rd)+strlen(wrd4th)+1+1+1; // '_''_''_'
       //wrdlen = strlen(wrd);
       if ( wrdlen <= 31 ) {
       strcpy(wrd, wrd1st);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd2nd);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd3rd);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd4th);
       }
}
else {
       PLE_words = 4;
       strcpy( wrd1st, wrd2nd );
       strcpy( wrd2nd, wrd3rd );
       strcpy( wrd3rd, wrd4th );
       strcpy( wrd4th, wrd );
       wrdlen = strlen(wrd1st)+strlen(wrd2nd)+strlen(wrd3rd)+strlen(wrd4th)+1+1+1; // '_''_''_'
       //wrdlen = strlen(wrd);
       if ( wrdlen <= 31 ) {
       strcpy(wrd, wrd1st);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd2nd);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd3rd);
       strcat(wrd, DelimiterUnderscore);
       strcat(wrd, wrd4th);
       }
}
// Quadruple! ]

                        if ( ( PLE_words == 4 ) && ( 12 <= wrdlen ) && ( wrdlen <= 31 ) ) {
...
```

# Section F: [Homepage]:

www.sanmayce.com/Downloads

Enjoy!
2011 Jan 26,
Kaze