06H051

web

web

[*1] HTML URI[*2] URI web
HTML HTML[*3]
TOM TOM
TOM
TOM web

[*1]
[*2] Uniform Resource Identifier
[*3] HyperText Markup Language

1

web

1.1

TOM web

1.2

TOM[*4]

1.3

- TOM:

  TOM

- : web MS-Word[*5]

[*4]

[*5] Microsoft Office Word

2

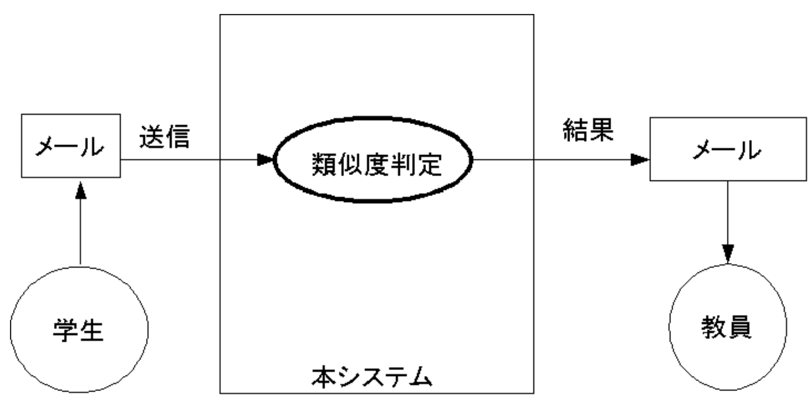
2.1

- (AT)
  - CPU:Intel(R) Celeron(R) CPU 2.80GHz
  - memory:512MB
- OS:Vine Linux 3.2
- :

2.2

web

1



1

2.2.1

[*6] Kakasi

Kakasi

Kakasi GPL[*7]

Kakasi

- Juman: [*8]

  Chasen

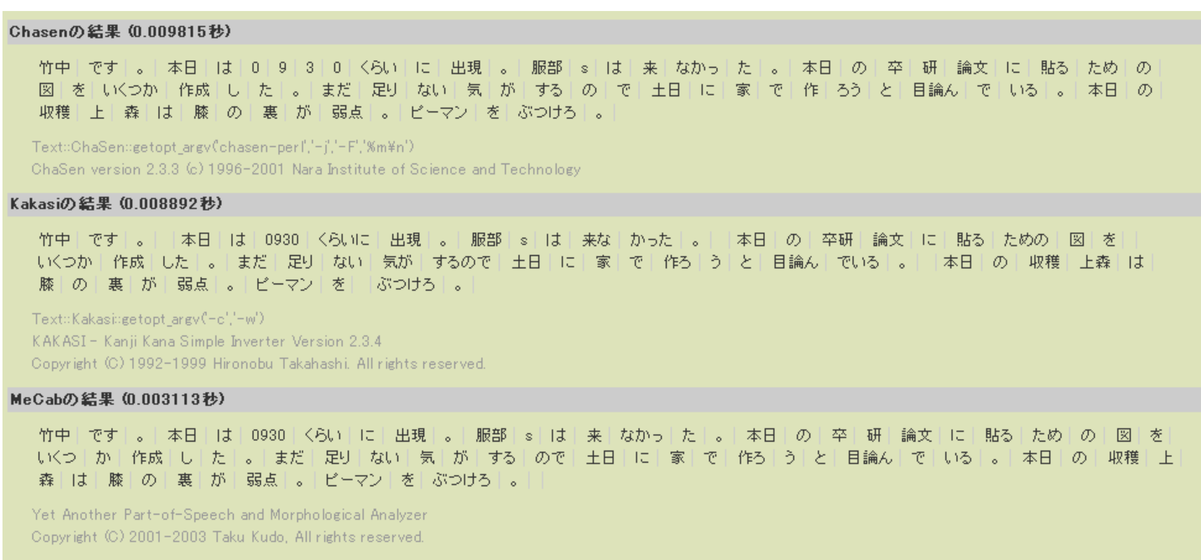
- Chasen:

- MeCab: Chasen

  3-4

2.2.2

2 Kakasi,Mecab,Chasen



2

MeCab Chasen

Kakasi

[*6]

[*7]

[*8]

2.2.3

2.2.4

w3m w3m
/ WWW[*9] kterm rxvt [*10]
Web HTML w3m
w3m HTML [*11]

2.2.5

yahoo yahoo HTML URI
URI HTML
Google GoogleAPI[*12]
yahoo

2.2.6 URI

web URI

- google
- amazon
- nicovideo
- youtube

---

*9 World Wide Web

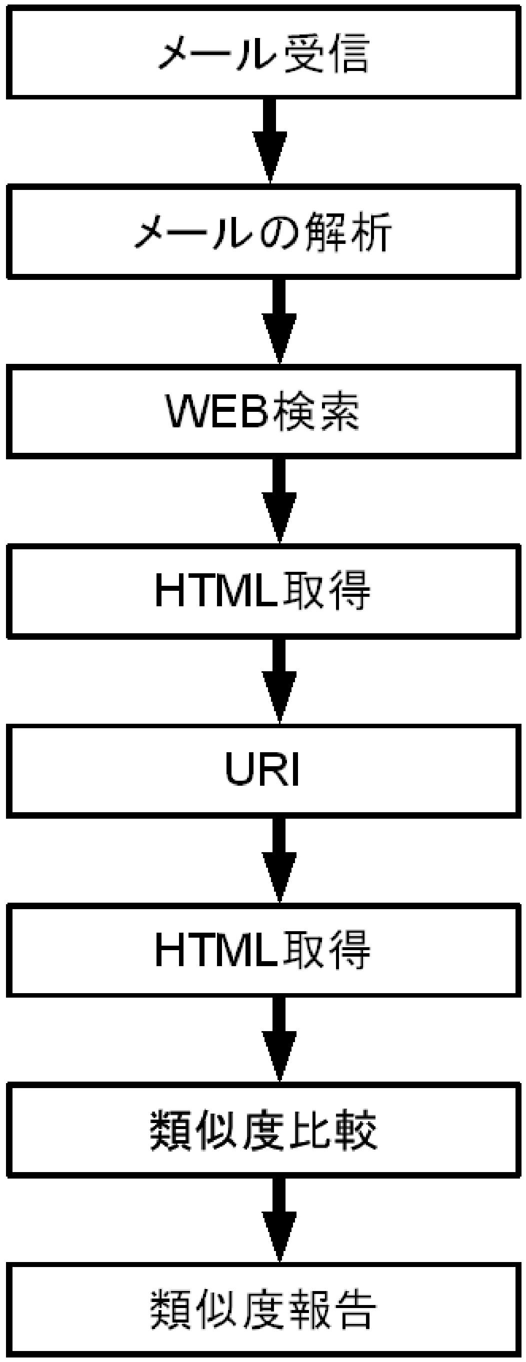
*10

*11

*12 GoogleAPI Google Google
1 1000

2.3

3



3

2.3.1

sed[*13]

2.3.2

kakasi [*14]

2.3.3 web

web w3m

yahoo HTML

2.3.4 URI

HTML URI

2.3.5 web

URI web HTML

2.3.6

HTML

[*13] Linux

[*14]

2.3.7

wikipedia[15]

[16] 4

,

OZ
5 MS 1

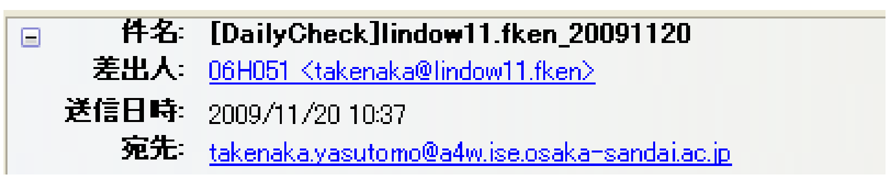
OZ

G

2 1
5 2
G

OZ
MS

OZ
(Deathscythe)
(Death)
(Scythe)
(Deathscythe) Death ( )
Scythe( )



4

[15] Wiki
[16]

3

3.1

ISO-2022.JP[*17]

3.2

HTML web

HTML

[*17] ( ) ISO/IEC 2022
7 JIS

4

wikipedia
75
web
TOM web
web HTML

4.1

- TOM TOM web
- HTML web
- URI web URI
- HTML URI HTML URI URI
-

[1] Shell Script, [http://www.k4.dion.ne.jp/ mms/unix/shellscript/index.html]
[2] Vine Linux, [http://www.vinelinux.org/]
[3] Proc Mail, [http://www.akatsuki.ne.jp/daemon/procmail/setup.html]
[4] w3m, [http://w3m.sourceforge.net/index.ja.html]
[5] , [http://www.lanches.co.jp/search/cache/1219626256.html]

( )URI 2009 11 25

A

```
#! /bin/sh
TMP1=/var/tmp/TMP1.$$
TMP2=/var/tmp/TMP2.$$

#MAIL=/home/2009/takenaka/sotuken/test/mail.txt

ENEMY=/home/2009/takenaka/sotuken/sign2.txt
cat > $OMAIL
MAIL=$OMAIL

#
cat $MAIL | \
sed -e '1,/^$/d' | sed -n -e '1,/^-- $/p' | \
grep -v "^>" | grep -v "^#" | nkf -Z1 | \
tr -s '\t' ' ' |  tr -d '\n' | \
kakasi -w | kakasi -JH | tr -s ' ' '\n' > $MAIL
cat $MAIL > /home/2009/takenaka/sotuken/body.txt

#
nkf -e /home/2009/takenaka/sotuken/body.txt > /home/2009/takenaka/sotuken/ebody.txt

#w3m
for WORD in `cat /home/2009/takenaka/sotuken/ebody.txt | sed 's/,/ /g' | awk '{print $2,$3,$4}'`
do

#
w3m -dump_source http://search.yahoo.co.jp/search?p=$WORD >> /home/2009/takenaka/sotuken/search.txt
#w3m  -dump 'http://search.yahoo.co.jp/search?p='$WORD' ie=euc-jp' >> /home/2009/takenaka/sotuken/search.txt
#w3m -dump 'http://www.google.co.jp/search?q='$WORD' ie=euc-jp'
#w3m -dump  http://search.yahoo.co.jp/search?p=$WORD > /home/2009/takenaka/sotuken/search.txt

#cat  /home/2009/takenaka/sotuken/search.txt |  sed 's/<em class="yschurl">/\n/g' | sed 's/<wbr>/\n/g' > /home/2009/takenaka/sotuken/URI.txt

# for URI in `cat /home/2009/takenaka/sotuken/search.txt | grep 'div class="abs"' \
#              | sed 's/.*"yschurl">//g' | sed 's/<[^>]*>//g' | head -n 3`
#  do
    #echo -n ,$URI >> /home/2009/takenaka/sotuken/URI.txt
#done
done

URI=/home/2009/takenaka/sotuken/search.txt

#URI
cat $URI | grep 'div class="abs"' \
 | sed 's/.*"yschurl">//g' | sed 's/<[^>]*>//g' | head -n 3 >> $URI

cat $URI | grep -v google | grep -v amazon | grep -v nicovideo | grep -v youtube  > $URI
cat $URI > /home/2009/takenaka/sotuken/test/test2.txt


wget -np  --restrict-file-names=nocontrol --directory-prefix=/home/2009/takenaka/sotuken/html/ -i  /home/2009/takenaka/sotuken/test/test2.txt

cd /home/2009/takenaka/sotuken/html/

i=1
#
for name in *.html ; do
   mv $name $name.txt
done

#
for name2    in ?* ; do
   mv $name2 $i.txt
let i=${i}+1
done
# find . -type f | xargs -n 10 nkf -e --overwrite

#
OBDY=/home/2009/takenaka/sotuken/ebody.txt
for rname in *.txt ; do
    nkf -e $rname >>  /home/2009/takenaka/sotuken/test/test3.txt
```

```
    WGET=/home/2009/takenaka/sotuken/test/test3.txt
#echo $WGET

    tail +3 $OBDY | grep -v "^>" >$TMP1
    tail +3 $WGET | grep -v "^>" >$TMP2

    CO=`cat $TMP1 | wc -l`
   echo $CO
    CC=`/usr/bin/diff -b -B $TMP1 $TMP2 | wc -l`
    HAN=`expr 100 - $CC \* 100 / $CO - 1`
#    cat $HAN
if [ $HAN -gt 75 ]
then
echo  "${HAN}" | cat >  /home/2009/takenaka/sotuken/p3.txt
fi


#cat  $HAN >> /home/2009/takenaka/sotuken/p3.txt
done

adad=takenaka.yasutomo@a4w.ise.osaka-sandai.ac.jp

cat <<END | /usr/lib/sendmail -t
To: takenaka.yasutomo@a4w.ise.osaka-sandai.ac.jp
Subject: [DailyCheck]`hostname`_20`date +%y%m%d`
Mime-Version: 1.0
Content-Type: text/plain; charset=ISO-2022-JP

Diese E-Mail aus dem Internet konnten {$HAN}% dupliziert haben
END

exit 0
```

A.1 1.1