МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №5
«Поиск коллокаций»
по курсу
«Обработка текстов на естественном языке»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Решение

1) T — test

Предполагается нормальное распределение токенов

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}}$$

2) Chi-square test

Используется распределение хи-квадрат

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

χ^2 = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies

Если значение удовлетворяет заданному уровню значимости alpha = 0.05, то гипотеза H0 о независимости токенов в биграмме отклоняется, следовательно выбирается альтернативная гипотеза H1 о зависимости слов друг от друга, тем самым идентифицируя коллокацию.

Результаты

1) T — test

http www
http org
index php
wikipedia org
org php
org index
wikipedia php
wikipedia index
index title
php title
org title
www com
http wikipedia
http index
diff oldid
oldid diff
diff diff
http com
title diff
diff inks

ser wikipedia
ser org
ser http
should made
title oldid
xternal links
has been
oldid inks
www org
php diff
https org
eferences http
talk page
diff www
xternal http
links http
diff http
the article
lease modify
lease not
not modify
xternal www
links www
the page
discussion the
inks www
further edits
edits made
modify should
modify ubsequent
further should
lease ubsequent
further made
edits should
modify comments
ubsequent comments
comments made
page further
edits this
comments should
ubsequent made
https wikipedia
ubsequent should
have been
https index
not ubsequent
talk further
undoafter undo
edit undo
not comments
should this
page edits

oldid http
talk edits
eferences xternal
should appropriate
appropriate discussion
appropriate such
discussion page
made appropriate
page should
appropriate page
such talk
discussion such
made discussion
eferences links
from the
page such
the first
also eferences
was the
diff ink
for the
oldid wikipedia
nited tates
oldid www
insee home
does not
www home
nbsp nbsp
html www
between and
more than
should page
diff org
undo ink
the same
lass wikipedia
diff wikipedia
article talk
calculated overlap
arch elete
com html
inks com
article further
rev middot
article page
oldid rcid
rcid diff
diff rcid
oldid rev
oldid middot
rev wikipedia
middot wikipedia

middot org
com http
page article
lass https
such article
lass org
with the
made this
www gov
the talk
had been
the appropriate
comments the
preserved archive
index html
above preserved
archive lease
php oldid
above archive
eferences www
following discussion
rev https
middot https
https com
following archived
home asp
discussion preserved
undo www
action edit
www insee
html com
ink www
during the
home page
above discussion
http insee
discussion archive
action undoafter
edit undoafter
action undo
undo undo
oldid https
such the
www net
that are
made result
result was
oldid assessed
oldid ink
this result
eferences insee
did not

the lease
discussion archived
debate modify
debate lease
preserved debate
reason for
archive debate
com news
this page
arch delete
below lease
below modify
inks added
involved accounts
accounts are
added this
debate not
below not
involved are
that involved
made page
made the
arch eep
that accounts
onitored rule
https archive
google com
dded link
reason monitoring
added diff
home home
assessed lass
insee page
archived debate
the discussion
following debate
proposed deletion
archive org
discussion debate
debate deletion
arch arch
part the
archived proposed
debate proposed
arch above
com ser
for monitoring
web web
inks this
http gov
web archive
archive web

inks diff
www news
the nited
com wikipedia
deletion below
rule reason
there are
diff ool
the season
article not
link records
index diff
bgcolor bgcolor
org web
web org
com www
here are
they are
this article
undoafter ink
the century
that would
https web
http home
link com
arch preserved
ser global
web http
archive http
ink com
title action
also http
both and
proposed article
https www
also xternal
deletion article
rcid inks
the end
align center
and the
page result
http net
article lease
preserved the
diff com
article below
but not
the north
was born
assessed https
prev oldid

was released
also links
books books
uefa com
diff prev
html ser
that the
com calculated
com overlap
www uefa
books google
com htm
rcid www
google books
php ser
also www
and other
which was
microsoft com
the tates

2)CHi criteria
bmkx bmkx
bml bml
segs segs
bmkw bmkw
ysr ysr
bmky bmky
qgcs qgcs
rect rect
oinky oinky
poly poly
ioq ioq
kalwini kalwini
ielito ielito
rattstrom rattstrom
leiton ledir
omarqusdepomba azevuoft
chome chome
lii nui
goldengrey kliping
undoafter undo
ional pelor
zipsrch cnty
yran yran
haminade haminade
espnstar pliga
mediatheque floridi
esde asta
boxesetc etrosheet
adius arcmin
alardon uscar

lbrace rbrace
lev levsut
escaladix larticles
affich exte
rarr rarr
trud lepo
enus enus
nui lii
iercurea iuc
asdf asdf
outen roeneveld
yrrh yrrh
fichescom mos
ettaglio tleta
tfrac tfrac
uldhana pincodes
avenna avenna
fnt fny
amg sql
statistiku ilten
ntrez ubmed
ukard riterium
vdots vdots
davidjcorcoran wyroutes
cmer cmer
mungo mungo
raxinus anomala
ortonville ortonville
ungos mungo
comparateur codgeo
articolo sezione
hariyyat hariyyat
oref slogin
osne ilten
tpe lev
aeronautica difesa
blah blah
playerpage ilkid
atvija atvija
ysq ysq
ecuerdo ecuerdo
chra ouari
yzsterious izster
imsky orsakov
arsiv spor
naismith drjamesnaismith
eporting statistics
urowana szmianka
ieczys riterium
hyd gors
eteorologie niei
lexuriserv celex

phaeralcea ambigua
idrologie niei
pvl aurentian
sakuracardcaptor activo
achu icchu
esl codgeo
understandingmarket definitionmkting
ublicidaddeunaresidenciaen speleol
streett laytee
speleol usgus
quickfacts qfd
qrio rvlyf
nsrc nloc
nmdxxwyegc rvlyf
nmdxxwyegc qrio
moriyuki abukuma
millesime typgeo
kicsoda letben
janez drnovsek
ettaiyaadu ilaiyaadu
eparhia dvorkin
clarken claytee
cefas homeng
btk ppke
balazs ppke
balazs btk
tamuna sirbiladze
seree ditra
revoluci socialista
miny maxy
kcfixz kldhe
epk epk
emokratikus talakul
dineanu ostolache
diariodigital dinheiro
willvdv chirailfan
tbnh tbnw
padurea craiului
lonee ittlepace
ycanzi dcanzi
urska obota
strateji cukurova
nomasticon oedelicum
manabu ouverneure
informaworld smpp
ebene fiindolo
rumsey geogarage
reporty upotrebleniya
purapelota lvbp
mpressionnistes ymbolistes
litepc ieradicator
geogarage cassinige

bundesverfassungsgericht entscheidungen
rganismo egulador
ousas ushdoony
orsna infoaeropuertos
lchchi aevdharm
itlinksolutions offshoredevelopment
bharti kher
ultraslan gev
sogn fylkesleksikon
sogn fjordane
mairesgenweb resultcommune
ldate hdate
gsyazkampi galatasarayyelken
gstaraftar galatasarayfederasyonu
galatasarayyelken hepsiburada
galatasarayfederasyonu galatasarayfan
galatasarayfan aslanlar
francegenweb resultcommune
francegenweb mairesgenweb
fjordane fylkesleksikon
bigglook biggfootball
biggfootball basini
omira omira
minm maxm
decko pracovi
eknomegisto thehisto
spamr currentsurplusieursarticles
aflhistory linescores
adrianbullock swfc
fenetilaminy polnyj
amfetaminy polnyj
amfetaminy fenetilaminy
pullman lostriver
frigoletto idhmg
cqpolitics wmspage
bcafe pellicola
arodnoe polcheniye
macaudata macauweb
footbel elftallen
biznesa cijas
surveyofwesternp conduoft
zbiory ckcp
wuerzburg topsi
wuerzburg deutop
topsi deutop
jupitermedia jupiterimages
buwcd zbiory
religionswissenschaft mjr
hunterlink maajjs
birdimages birdspecies
boutic dboutic
oyoashihara itsuho

oyoashihara agaioaki
agaioaki itsuho
aircraftinquiry umbertxt
cbll panoid
webgn instituc
cfnavarra webgn
cfnavarra instituc
neilbrown newcastlefans
mooma keshet
iaurif fichescom
calauza afiseaza
activestats psgc
fichecommunale codecom
fichecommunale codedep
codedep codecom
eteorologie idrologie
obswww unige
methodes fichecommunale
nstitutul eteorologie
obswww behrend
npan abcdefghinoq
adastrul nstitutul
nstitutul idrologie
unige behrend
sintesisestadistica cuadros
fichescom ficmos
methodes nomenclatures
dministra adastrul
cuadros oblacion
nomenclatures fichecommunale
tlfq ulaval
esquisar livros
nscb psgc
nscb activestats
artandsocialissues cmaohio
afiseaza trasee
hermandw olymp
unige cou
sbdb sstr
pqasb pqarchiver
retrosheet boxesetc
rusteam permian
behrend cou
nomenclatures cog
cog fichecommunale
webgn normativa
instituc normativa
cog codedep
udleya brevifolia
oez spamsournois
typeprod quelcas
nivgeo typeprod

methodes cog
abcdefghinoq msz
rumious andersnatch
npan msz
bubbleman seifenblasen
omuald azoum
minorplanetobserver pdolc
cdpa nsysu
eteorologie urile
ailnie okruga
ymd vkey
ional adastrul
vkey fext
enrin kyrghiz
bienvenue zugang
idrologie urile
rev middot
aijing aijing
uspicia quaedam
sccwebsite sccwspublications
nstitutul urile
antom arior
ekg odincov
dministra ional
sll sspn
jewishvirtuallibrary jsource
marketshare hitslink
rketinga biznesa
datenbanken dbfoeldeak
bdoubliees journalspirou
hawinigan ataractes
harimaya sengoku
argul rumos
natela nicoli
yaf yma
rticolo ontenuto
tmplrubriche grubrica
maxdocs topdoc
supremecourtus briefoverview
jarnvag rnv
jonathanriley chicsrc
ancdocs mandela
buscar ciudad
poteci muntii
alpinet poteci
efloras florataxon
zkumn pracovi
mpier kleioc
landshold landsholdsdatabasen
decko zkumn
dboutic cdrom
boutic cdrom

pixopale zenfolio
calculated overlap
omoyuki akazawa
mungo ungos
imgres imgurl
emecmua etay
altyapi okul
surreycc sccwebsite
jreast estation
infothek bienvenue