

Calvin Kranig

COM S 435

TTH 0930 – 1050

PA2

Introduction

This report will detail the methodology behind my work in PA2 as well as report upon the results of various tests.

PageRank

Iterations

	$\beta = 0.85$	$\beta = 0.25$
$\epsilon = 0.05$	4	2
$\epsilon = 0.01$	7	3

Trust Rank

Top 10 Rankings

Trust Rank Order	Regular Rank Order
/wiki/Wayback_Machine : 0.004409	/wiki/Wayback_Machine : 0.004409
/wiki/United_States : 0.003948	/wiki/United_States : 0.003948
/wiki/Association_football : 0.002956	/wiki/Association_football : 0.002956
/wiki/World_War_II : 0.002445	/wiki/World_War_II : 0.002445
/wiki/Basketball : 0.002138	/wiki/Basketball : 0.002138
/wiki/France : 0.002136	/wiki/France : 0.002135
/wiki/Canada : 0.002064	/wiki/Canada : 0.002064
/wiki/National_Diet_Library : 0.002006	/wiki/National_Diet_Library : 0.002005
/wiki/National_Football_League : 0.001971	/wiki/National_Football_League : 0.001971
/wiki/The_New_York_Times : 0.001934	/wiki/The_New_York_Times : 0.001934

Total Difference: 9.647457703856328E-4

Average Difference: 9.647457703856328E-8

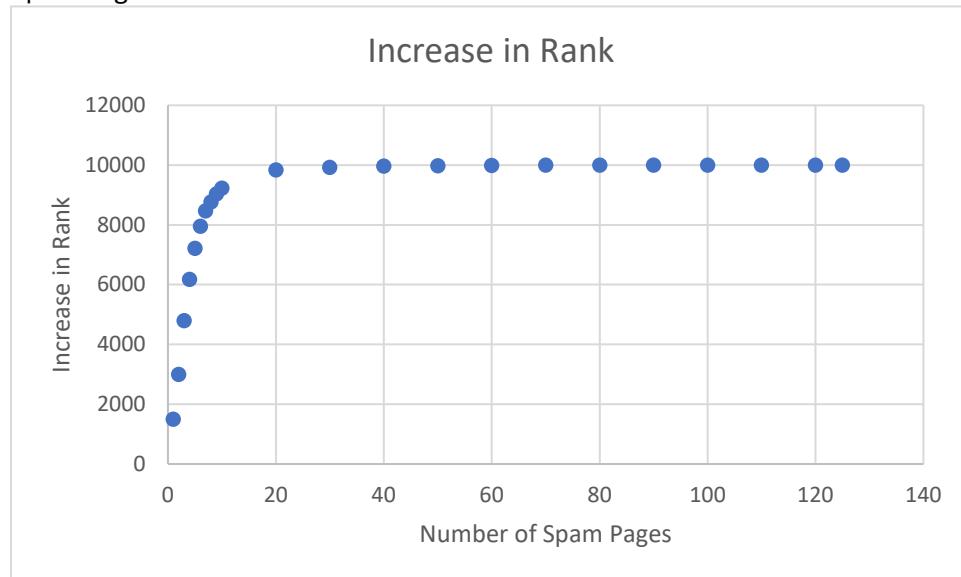
Different at 6995 positions out of 10000

The difference between the top ranks is negligible, and the overall difference is very low. As a result while the two orderings are different from each other at most positions the actual scores of each vertex changed very little, and their ranks only changed minutely.

Spam Farm

Lowest Ranked Vertex: 5606

Spam Pages: 1 New Rank: 8506 Increase in Rank: 1493
Spam Pages: 2 New Rank: 7006 Increase in Rank: 2993
Spam Pages: 3 New Rank: 5203 Increase in Rank: 4796
Spam Pages: 4 New Rank: 3823 Increase in Rank: 6176
Spam Pages: 5 New Rank: 2781 Increase in Rank: 7218
Spam Pages: 6 New Rank: 2040 Increase in Rank: 7959
Spam Pages: 7 New Rank: 1529 Increase in Rank: 8470
Spam Pages: 8 New Rank: 1230 Increase in Rank: 8769
Spam Pages: 9 New Rank: 962 Increase in Rank: 9037
Spam Pages: 10 New Rank: 774 Increase in Rank: 9225
Spam Pages: 20 New Rank: 166 Increase in Rank: 9833
Spam Pages: 30 New Rank: 74 Increase in Rank: 9925
Spam Pages: 40 New Rank: 30 Increase in Rank: 9969
Spam Pages: 50 New Rank: 19 Increase in Rank: 9980
Spam Pages: 60 New Rank: 10 Increase in Rank: 9989
Spam Pages: 70 New Rank: 4 Increase in Rank: 9995
Spam Pages: 80 New Rank: 3 Increase in Rank: 9996
Spam Pages: 90 New Rank: 2 Increase in Rank: 9997
Spam Pages: 100 New Rank: 2 Increase in Rank: 9997
Spam Pages: 110 New Rank: 2 Increase in Rank: 9997
Spam Pages: 120 New Rank: 1 Increase in Rank: 9998
Spam Pages: 125 New Rank: 0 Increase in Rank: 9999



Spam Pages: 1
New Rank: 8506
Increase in Rank: 1493
Trust Rank Position: 8509
PageRank: 0.000030
TrustRank: 0.000029

Spam Pages: 20
New Rank: 166
Increase in Rank: 9833
Trust Rank Position: 168
PageRank: 0.000494
TrustRank: 0.000493

Spam Pages: 40
New Rank: 30
Increase in Rank: 9969
Trust Rank Position: 31
PageRank: 0.001178
TrustRank: 0.001175

Spam Pages: 60
New Rank: 10
Increase in Rank: 9989
Trust Rank Position: 10
PageRank: 0.001974
TrustRank: 0.001969

Spam Pages: 80
New Rank: 3
Increase in Rank: 9996
Trust Rank Position: 3
PageRank: 0.002754
TrustRank: 0.002747

Spam Pages: 100
New Rank: 2
Increase in Rank: 9997
Trust Rank Position: 2
PageRank: 0.003648
TrustRank: 0.003638

Spam Pages: 125
New Rank: 0
Increase in Rank: 9999
Trust Rank Position: 0
PageRank: 0.004716
TrustRank: 0.004703

Using the trust rank method from the previous test there is no real discernable difference between the trust rank and the page rank. Spam pages can be easy to detect if they follow an optimal strategy of having only one outgoing edge to the target page and one incoming edge from the target page, but just looking at ranks alone makes determining the spam pages difficult.

Positional Index

Query : ball

Document: Fair_ball.txt **TPScore**: 0.000000 **VSScore**: 0.199317 **Relevance**: 0.079727

Document: Batted_ball.txt **TPScore**: 0.000000 **VSScore**: 0.156079 **Relevance**: 0.062432

Document: Dead_ball_(baseball).txt **TPScore**: 0.000000 **VSScore**: 0.152882 **Relevance**: 0.061153

Document: Dead_ball.txt **TPScore**: 0.000000 **VSScore**: 0.152882 **Relevance**: 0.061153

Document: Cincinnati_Browns.txt **TPScore**: 0.000000 **VSScore**: 0.151802 **Relevance**: 0.060721

Document: Ground_out_(baseball).txt **TPScore**: 0.000000 **VSScore**: 0.150784 **Relevance**: 0.060314

Document: Indian_Ball.txt **TPScore**: 0.000000 **VSScore**: 0.150698 **Relevance**: 0.060279

Document: Fly_ball_pitcher.txt **TPScore**: 0.000000 **VSScore**: 0.144912 **Relevance**: 0.057965

Document: Foul_ball.txt **TPScore**: 0.000000 **VSScore**: 0.143091 **Relevance**: 0.057237

Document: Foul_tip.txt **TPScore**: 0.000000 **VSScore**: 0.138544 **Relevance**: 0.055418

Query : pittsburgh pirates

Document: 1909_Major_League_Baseball_season.txt **TPScore**: 2.000000 **VSScore**: 0.545891 **Relevance**: 1.418356

Document: 1965_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.309248 **Relevance**: 1.323699

Document: 1980_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.307982 **Relevance**: 1.323193

Document: 1982_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.306219 **Relevance**: 1.322487

Document: 1968_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.305129 **Relevance**: 1.322052

Document: 1983_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.302692 **Relevance**: 1.321077

Document: 1989_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.298365 **Relevance**: 1.319346

Document: 1913_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.297717 **Relevance**: 1.319087

Document: 1961_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.293668 **Relevance**: 1.317467

Document: 1963_Pittsburgh_Pirates_season.txt **TPScore**: 2.000000 **VSScore**: 0.292303 **Relevance**: 1.316921

Query : new york state

Document: Sports_in_New_York.txt **TPScore**: 1.500000 **VSScore**: 0.123914 **Relevance**: 0.949566

Document: List_of_New_York_state_parks.txt **TPScore**: 1.500000 **VSScore**: 0.109147 **Relevance**: 0.943659

Document: Cornelius_Van_Cott.txt **TPScore**: 1.500000 **VSScore**: 0.079749 **Relevance**: 0.931900

Document: Syracuse_Stars_(minor_league_baseball).txt **TPScore**: 1.500000 **VSScore**: 0.075893 **Relevance**: 0.930357

Document: Borough_(New_York_City).txt **TPScore**: 1.500000 **VSScore**: 0.070976 **Relevance**: 0.928390

Document: History_of_New_York.txt **TPScore**: 1.500000 **VSScore**: 0.063570 **Relevance**: 0.925428

Document: Sports_in_New_York_City.txt **TPScore**: 1.500000 **VSScore**: 0.063500 **Relevance**: 0.925400

Document: List_of_museums_in_New_York.txt **TPScore**: 1.500000 **VSScore**: 0.063317 **Relevance**: 0.925327

Document: Administrative_divisions_of_New_York.txt **TPScore**: 1.500000 **VSScore**: 0.059847 **Relevance**: 0.923939

Document: Upstate_New_York.txt **TPScore**: 1.500000 **VSScore**: 0.057999 **Relevance**: 0.923200

Query : the cubs finished fifth

Document: 1922_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.148537 **Relevance**: 0.859415

Document: 1924_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.142667 **Relevance**: 0.857067

Document: 1983_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.135098 **Relevance**: 0.854039

Document: 1916_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.134284 **Relevance**: 0.853714

Document: 1940_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.133620 **Relevance**: 0.853448

Document: 1958_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.132449 **Relevance**: 0.852980

Document: 1975_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.129961 **Relevance**: 0.851984

Document: 1917_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.129033 **Relevance**: 0.851613

Document: 1943_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.128611 **Relevance**: 0.851444

Document: 1952_Chicago_Cubs_season.txt **TPScore**: 1.333333 **VSScore**: 0.128327 **Relevance**: 0.851331

Query : the last cy young winner for the cubs

Document: Chris_Carpenter.txt **TPScore**: 0.333333 **VSScore**: 0.041851 **Relevance**: 0.216741

Document: History_of_the_Detroit_Tigers.txt **TPScore**: 0.333333 **VSScore**: 0.034634 **Relevance**: 0.213853

Document: Detroit_Tigers.txt **TPScore**: 0.333333 **VSScore**: 0.033054 **Relevance**: 0.213222

Document: 2012_Los_Angeles_Dodgers_season.txt **TPScore**: 0.307692 **VSScore**: 0.022316 **Relevance**: 0.193542

Document: History_of_the_Pittsburgh_Pirates.txt **TPScore**: 0.296296 **VSScore**: 0.028464 **Relevance**: 0.189163

Document: Billy_Pierce.txt **TPScore**: 0.296296 **VSScore**: 0.027387 **Relevance**: 0.188733

Document: 2010_Seattle_Mariners_season.txt **TPScore**: 0.285714 **VSScore**: 0.031376 **Relevance**: 0.183979

Document: Barry_Zito.txt **TPScore**: 0.275862 **VSScore**: 0.038026 **Relevance**: 0.180728

Document: Adam_Wainwright.txt **TPScore**: 0.266667 **VSScore**: 0.040985 **Relevance**: 0.176394

Document: Boston_Red_Sox.txt **TPScore**: 0.266667 **VSScore**: 0.028827 **Relevance**: 0.171531

The results were not as satisfactory as I had hoped. There seems to be a need to better highlight the most important words in the query.