ANALYSIS AND RANKING OF ONLINE REVIEWS FOR DOCTORS

Harshaneel H. Gokhale
Instructor: Dr. Yu Cheng,
Dr. Huan Liu, Fred Morstatter.

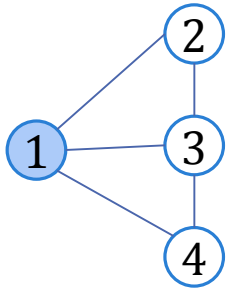
ROADMAP

- Introduction
- Random Walk with Restarts Algorithm
- Sentiment analysis
- Methodology
- Results
- Algorithm
- Scope of the concept and future work

INTRODUCTION

- Problem:
  - We have a bunch of reviews for doctors posted by their patients.
  - Reviews contain meta-information including scores for different parameters for doctor such as, punctuality, knowledge, staff interaction and etc.
  - Reviews also contain the textual review expressing overall feeling of the patient about the doctor.
  - Our task is to come up with a method that helps us to rank doctors with respect to their performance while treating the patients.

RANDOM WALK WITH RESTARTS ALGORITHM



q =

1	0	0	0

A =

0	1	1	1
1	0	1	0
1	1	0	1
1	0	1	0

$\overline{A}$ =

0	0.50	0.33	0.50
0.33	0	0.33	0
0.33	0.50	0	0.50
0.33	0	0.33	0

RWR equation:

$$r = c\overline{A}r + (1 - c)q$$

Closed form solution:

$$r = (1 - c)(I - c\overline{A})^{-1}q$$

Recursive solution:

$$r_t = q$$

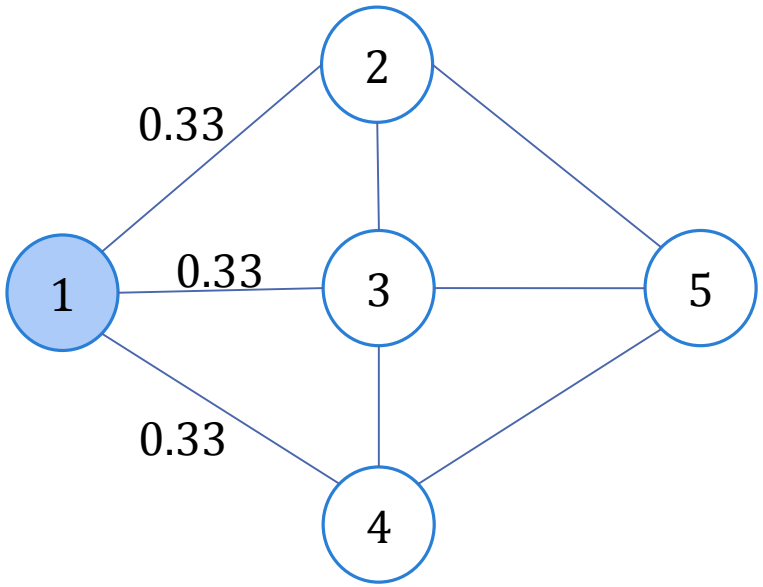
$$r_{t+1} = c\overline{A}r_t + (1 - c)q$$

$$UNTIL\ r_{t+1}\ and\ r_t\ converge:$$

$$r_{t+1} = c\overline{A}r_t + (1 - c)q$$

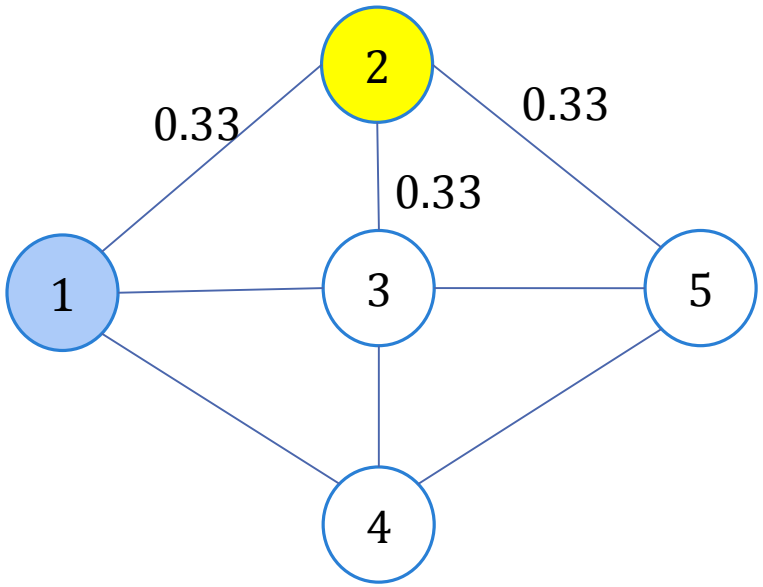
RANDOM WALK WITH RESTARTS ALGORITHM (CONTD..)

Step 0:



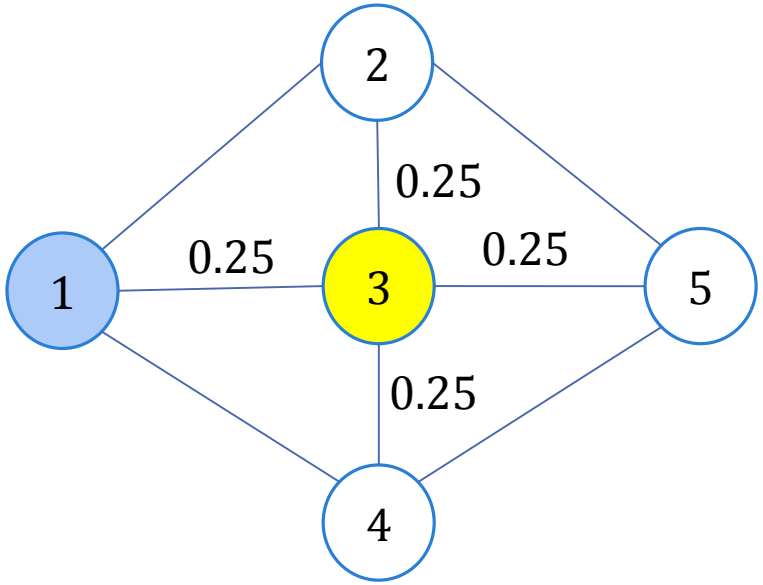
RANDOM WALK WITH RESTARTS ALGORITHM (CONTD..)

Step 1:



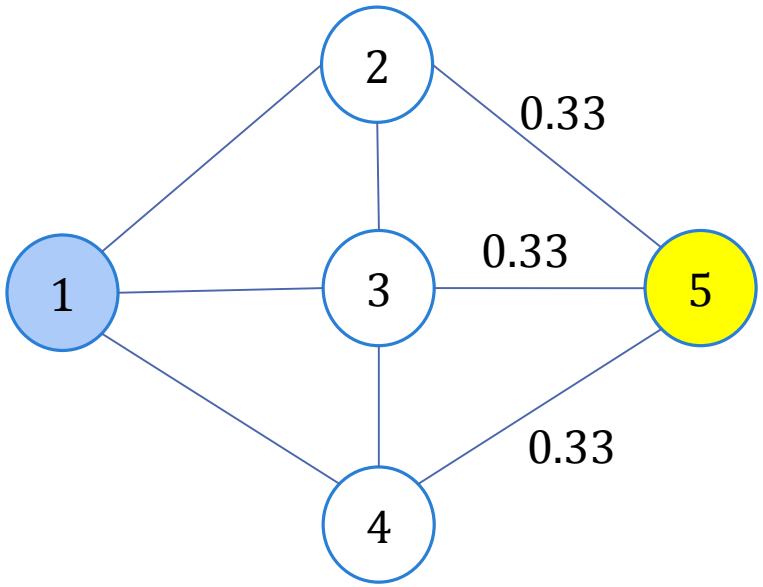
RANDOM WALK WITH RESTARTS ALGORITHM (CONTD..)

Step 2:



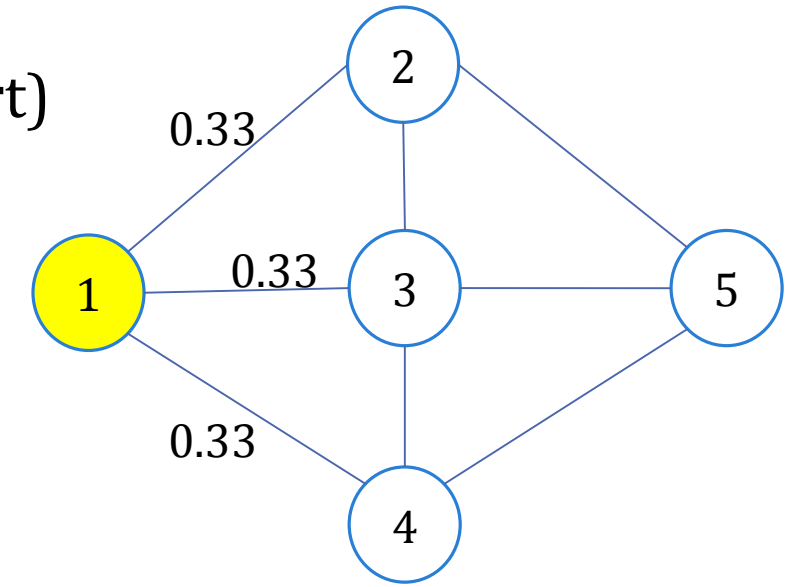
RANDOM WALK WITH RESTARTS ALGORITHM (CONTD..)

Step 3:



RANDOM WALK WITH RESTARTS ALGORITHM (CONTD..)

Step 4:
(Restart)



And so on….

SENTIMENT ANALYSIS

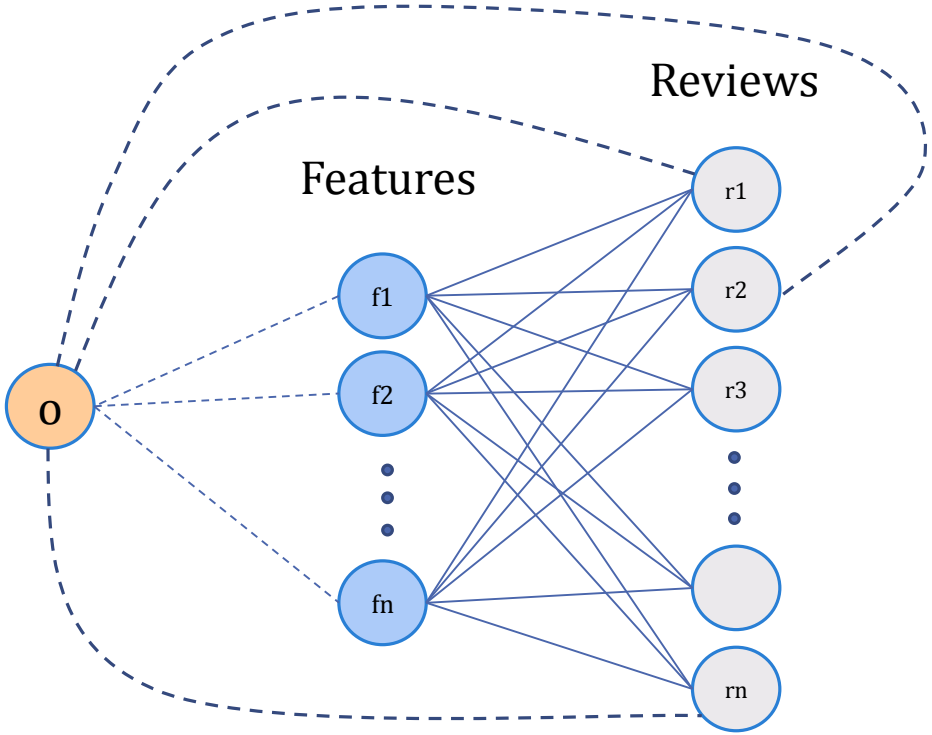
- VADER (Valence Aware Dictionary for sEntiment Reasoning)
- VADER is a simple rule based model for general sentiment analysis.
- VADER takes into consideration, emoticons and slangs to analyze polarity and subjectivity of the text. This makes it useful for our purpose since online reviews are more likely to have colloquial terms and slangs to express sentiment.
- NLTK package has very good implementation for this under package nltk.sentiment.vader.
- Reference: Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

METHODOLOGY

- Feature set:
  - Meta – features:
    - Rating (0-5).
    - Staff (0-5).
    - Punctuality (0-5).
    - Help (0-5).
    - Knowledge (0-5).
  - Sentiment features:
    - Overall polarity of the sentiment from 0 to 1.
    - 1 : highly negative.
    - 0 : highly positive.

METHODOLOGY (CONTD..)

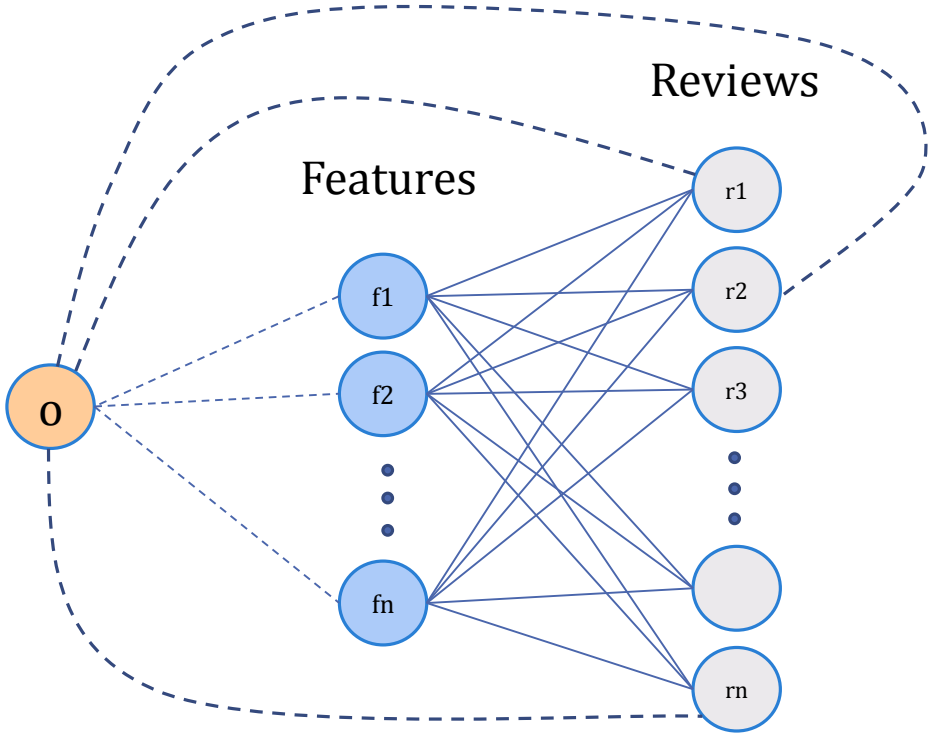
- Graph construction:



$F \equiv set(Features)$
$R \equiv set(Reviews)$
$e(f, r) \equiv value\ A[f, r]$

That means, edge between f_i and r_i indicates value of feature f_i for review r_i.

METHODOLOGY (CONTD..)

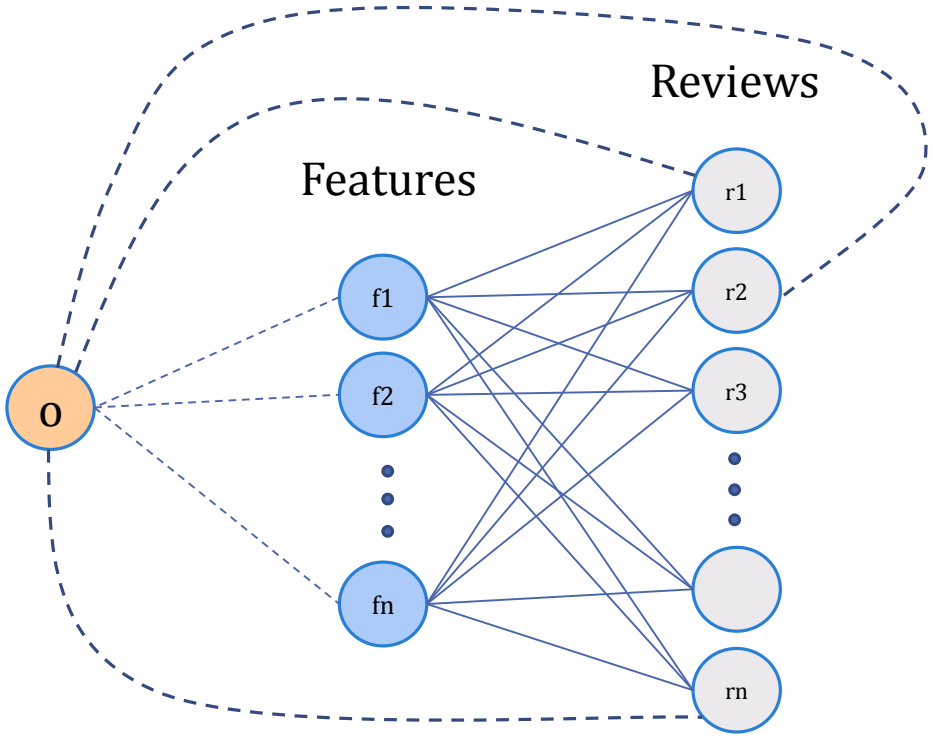


Why new node O?

Answer is simple, we want to query on all the features simultaneously. Instead, we can query node O which is connected to all the features. Hence in steady state, a random particle taking walk on this graph would have travelled over all the features.

Additionally, it allows us to incorporate feature importance. More important feature should have higher probability of visiting which we can specify by giving higher weight to $e(o, f_i)$.

METHODOLOGY (CONTD..)

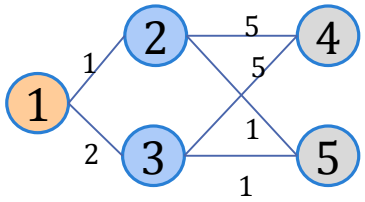


Why new edges from Reviews to O?

Because, we don't want to lose information while normalizing Adjacency matrix. Adding very weak links does have a strong effect on ranking. Adding these edges does not change rank of the matrix but helps us keeping information stored in feature values. Since we do column normalization of adjacency matrix before using random walk algorithm, all the information stored in features is lost.

More on this on the next slide…..

METHODOLOGY (CONTD..)



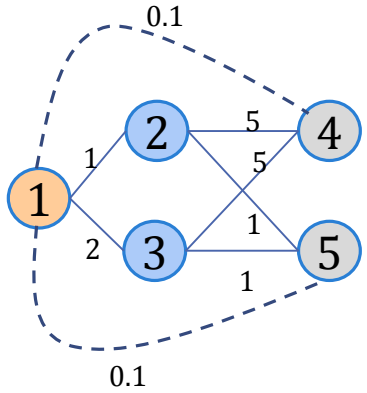
0	1	2	0	0
1	0	0	5	1
2	0	0	5	1
0	5	5	0	0
0	1	1	0	0

0	Feat.	Reviews

0	0.14	0.25	0	0
0.33	0	0	0.5	0.5
0.66	0	0	0.5	0.5
0	0.71	0.62	0	0
0	0.14	0.12	0	0

0	Features	Reviews

Column 4 and 5 are same after normalization. Hence, the information stored in feature values is lost! ☹



0	1	2	0.1	0.1
1	0	0	5	1
2	0	0	5	1
0.1	5	5	0	0
0.1	1	1	0	0

0	0.14	0.25	0.009	0.06
0.31	0	0	0.49	0.47
0.62	0	0	0.49	0.47
0.03	0.71	0.62	0	0
0.03	0.14	0.12	0	0

Column 4 and 5 are not same! Higher weight is assigned for features in column 4 than 5. ☺

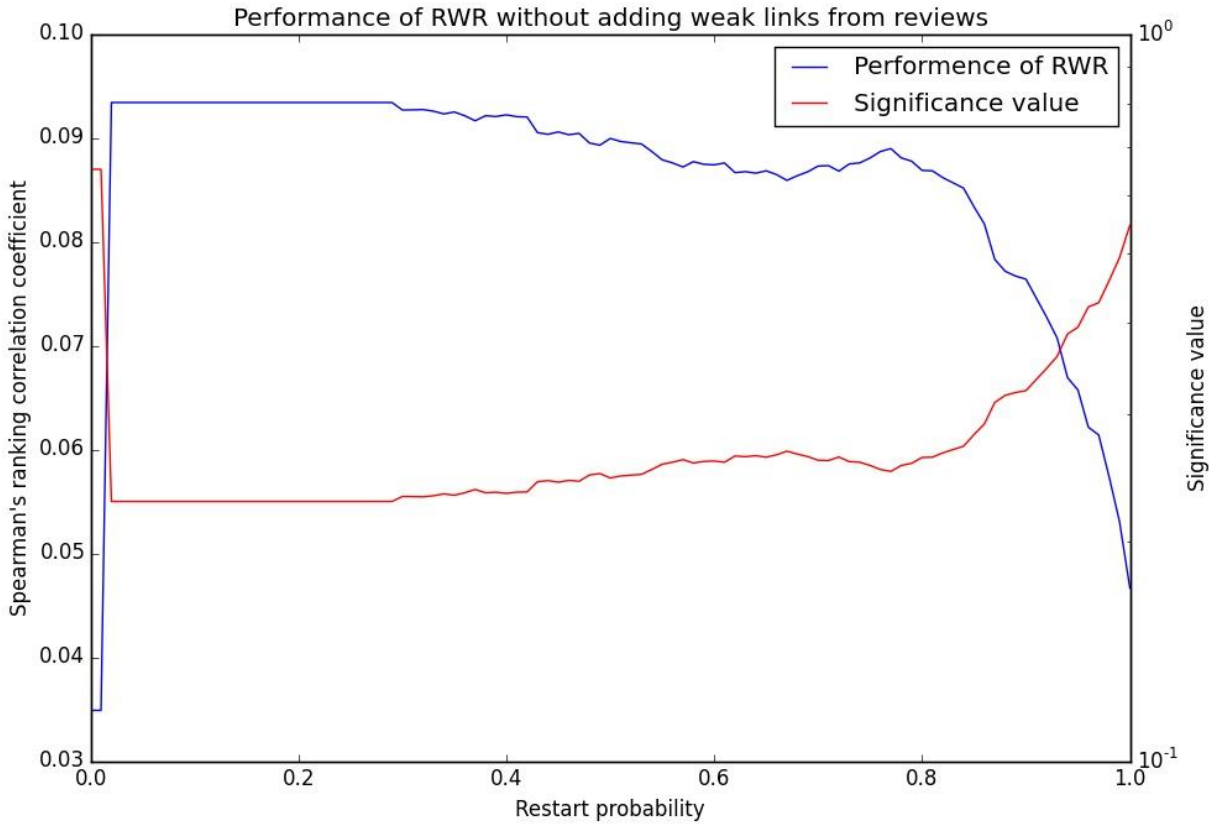
METHODOLOGY (CONTD..)

- Feature importance.
- How can we measure importance of feature for ranking?? (Can we use Entropy?)

F1:	1	1	1	1	1	2	2	2	2	2
F2:	1	2	3	4	5	6	7	8	9	10

- Which one do you think is a better feature for ranking?
- Is there any way of finding best one among first five elements using F1?
- Is there any way of finding best one among first five elements using F2?
- *Hint:*
  Entropy(F1) = 0.3, Entropy(F2) = 1.0

RESULTS (CONTD..)



Maximum value of correlation coefficient = 0.09 at c = 0.01

p-value = 0.22 (two sided)

H_0= Proposed ranking is correlating by chance.
H_a= Proposed ranking is significant.

Since p-value is much more than 0.01 we can accept null hypothesis because there no sufficient evidence for alternative hypothesis

This serves as another evidence for adding weak links.

ALGORITHM

$Given: Training\ data\ X(n \times m)$

$Output: Ranking\ list\ of\ all\ the\ doctors$

$A = preprocess(X);$

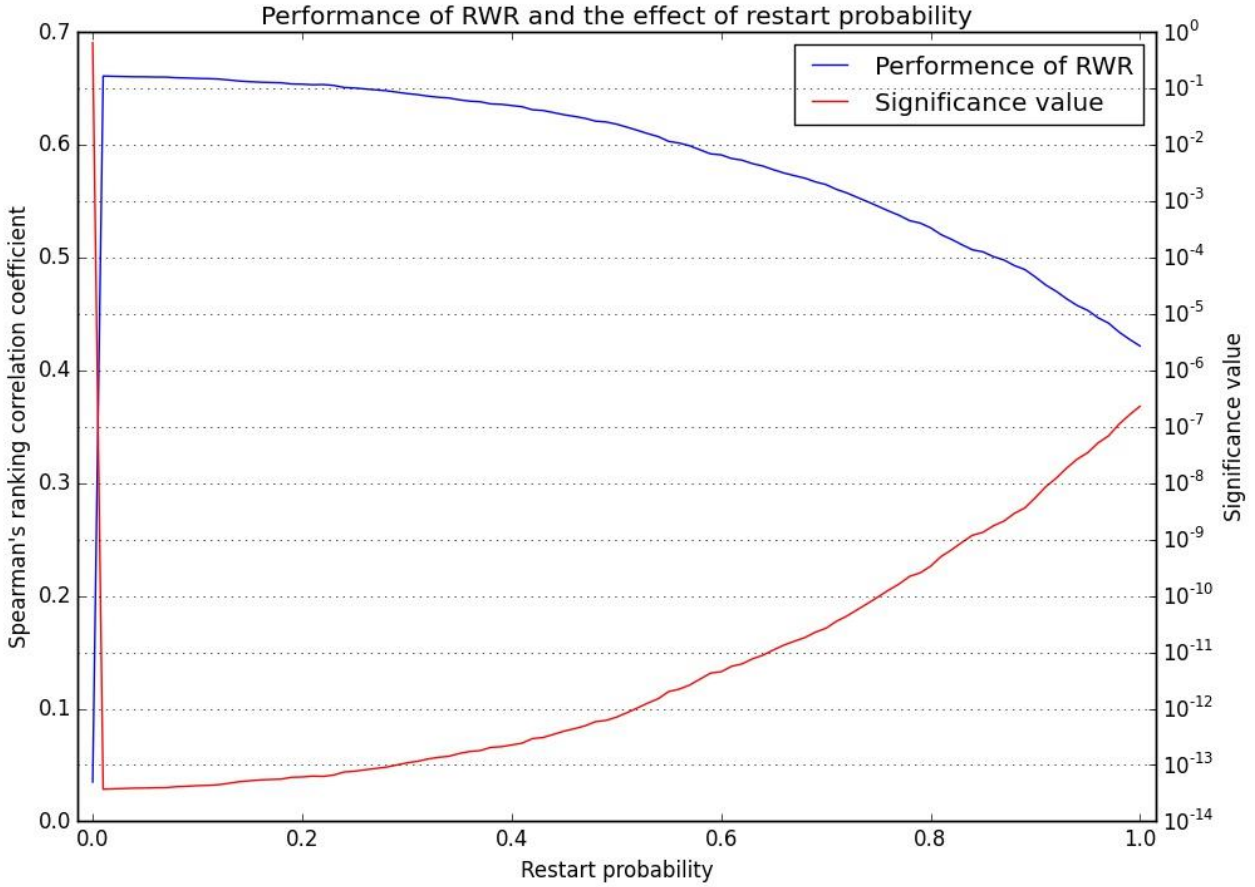
$q = construct\_query(0); c \leftarrow 0.01;$

$r = rwr(A, q, c);$

$doctors\_ranks = extract\_and\_group(r);$

$return\ doctors\_ranks;$

RESULTS



Maximum value of correlation coefficient = 0.67 at c = 0.01

p-value = 3.71e-14 (two sided)

H_0= Proposed ranking is correlating by chance.
H_a= Proposed ranking is significant.

Since p-value is far less than 0.01 we can say that we have insufficient confidence to accept null hypothesis and we have sufficient confidence to accept alternative hypothesis.

THANK YOU!

You can refer to my work at- http://github.com/harshaneelhg/Online-Doctor-Reviews-Analysis