Name: Changhyun, Lee
Andrew ID: changhyl

Machine Learning for Text Mining

Homework 1 - Template

1. Statement of Assurance

I certify that all the materials are original works that was done only by me.

2. Experiments

a) Describe the custom weighing scheme that you have implemented. Explain your motivation for creating this weighting scheme.

$$s_i = \mathrm{w1} * {p_i}^2 + \mathrm{w2} * r_i$$

where s_i is final score and p_i is the PageRank score, r_i is relevance score of document i respectively. The ${p_i}^2$ makes PageRank score well distinguishable. It means that the small one get smaller, the large one get much larger.

b) Report of the performance of the 9 approaches.

I. Metric: MAP

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0457	0.2588	0.2635
QTSPR	0.0434	0.2559	0.2635
PTSPR	0.0455	0.2597	0.2635

II. Metric: Precision at 11 standard recall levels

(Use one table for each recall level, so totally there would be 11 tables.)

Ircl_prn 0.00

Method \ Weighting Scheme	NS	WS	CM
GPR	0.1446	0.8405	0.3963
QTSPR	0.8405	0.8405	0.8405
PTSPR	0.8405	0.8405	0.8405

Ircl_prn 0.10

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0875	0.5926	0.2554
QTSPR	0.5926	0.5926	0.5926
PTSPR	0.5926	0.5926	0.5926

Ircl_prn 0.20

Method \ Weighting Scheme	NS	WS	CM

GPR	0.0786	0.4732	0.2292
QTSPR	0.4732	0.4732	0.4732
PTSPR	0.4732	0.4732	0.4732

Ircl_prn 0.30

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0737	0.3781	0.2021
QTSPR	0.3781	0.3781	0.3781
PTSPR	0.3781	0.3781	0.3781

Ircl_prn 0.40

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0699	0.3145	0.1720
QTSPR	0.3145	0.3145	0.3145
PTSPR	0.3144	0.3145	0.3145

Ircl_prn 0.50

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0653	0.2430	0.1500
QTSPR	0.2430	0.2430	0.2430
PTSPR	0.2432	0.2430	0.2430

Ircl_prn 0.60

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0534	0.1677	0.1107
QTSPR	0.1677	0.1677	0.1677
PTSPR	0.1679	0.1677	0.1677

Ircl_prn 0.70

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0300	0.0914	0.0632
QTSPR	0.0916	0.0915	0.0915
PTSPR	0.0915	0.0915	0.0915

Ircl_prn 0.80

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0115	0.0550	0.0396
QTSPR	0.0550	0.0550	0.0550
PTSPR	0.0550	0.0550	0.0550

Ircl_prn 0.90

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0074	0.0388	0.0287
QTSPR	0.0388	0.0388	0.0388

PTSPR	0.0388	0.0388	0.0388

Ircl_prn 1.00

Method \ Weighting Scheme	NS	WS	CM
GPR	0.0041	0.0101	0.0099
QTSPR	0.0101	0.0101	0.0101
PTSPR	0.0101	0.0101	0.0101

III. Metric: Wall-clock running time in seconds

Method \ Weighting Scheme	NS	WS	CM
GPR	0.2059	0.4180	0.4649
QTSPR	0.3270	0.5460	0.6199
PTSPR	0.3159	0.5360	0.5590

IV. Parameters

GPR : alpha = 0.8,

PTSPR, QTSPR : alpha = 0.8, beta = 0.15, gamma = 0.05,

W1 = 1000, w2 = 1

c) Compare these 9 approaches based on the various metrics described above.
In case of MAP, all of the methods have best score in CM metric with recording 0.2635 score. When using NS metric, worst scores are recorded for all the three methods. Among the methods, GPR recorded best score of 0.0457 in NS metric, and PTSPR recorded best score of 0.2597 in WS metric.

d) Analyze these various algorithms, parameters, and discuss your general observations about using PageRank algorithms.

As for algorithms, GPR has worse result than other methods because it does not take relevance score of each documents into account. The algorithms of both QTSPR and PTSPR are similar but the condition of topic distribution is different. It depends on how we get the distribution when query is given.

Regarding parameters, w1, which is a weight of r vector when we calculate weighted sum of PageRank and relevance score, tends to improve the performance of MAP when it is increased by 1000. It is because the PageRank score is much smaller than the relevance score.

e) 1. What could be some novel ways for search engines to estimate whether a query can benefit from personalization?

Because the query is different from person to person, it is necessary to consider query when we calculate the score. It can be estimated by using topic distribution of Pr(t|q) * Pr(t|u).

2. What could be some novel ways of identifying the user's interests (e.g. the user's topical interest distribution Pr(t|u)) in general?

3. Details of the software implementation

a) Describe your design decisions and high-level software architecture;

This software is composed of total 4 files which are GPR.py, QTSPR.py, PTSPR.py, PR_Modules.py respectively. The GPR.py, QTSPR.py and PTSPR.py files implement calculation of corresponding method. After running those files, we automatically write the result on the project folder. All of the functions that makes the implementation more convenient are in PR_Modules.py file.

b) Describe major data structures and any other data structures you used for speeding up the computation of PageRank;
Transition matrix : transition.txt is read and converted into 81433x81433 sparse matrix.
PageRank vector : It contains PageRank score of 81433 documents.
Weighted sum of PageRank vector

c) Describe any programming tools or libraries and programming environment used;
Tools : Pycharm
Libraries : scipy (for sparse matrix), numpy (matrix calculation), pandas

d) Describe strengths and weaknesses of your design, and any problems that your system encountered
My design has a strength of speed, it is designed to calculate matrix multiplication fast, by using sparse matrix. However, when we try to get each score of GPR, QTSPR, PTSPR, we have to run the files separately.