2013 CS276 - PA4 Report

SUNetId: cenk

Task1:

With manual weight from PA3 maximum score obtained was
0.8693

Order of features which are used for learning and testing: url, title, body, header, anchor

Weights: [0.01132102 0.01474673 -0.00174452 0.00673726 0.01298726]
ndcg score from pointwise approach: 0.860943491093

According to weights which are calculated, highest weight should be given to **title**, second is **anchor**, third is **url**, **header** should be low so its not that important and lowest weight should be give to **body**.

Task2:
SVM ndcg score is better than linear regression score.

Order of features which are used for learning and testing: **url, title, body, header, anchor**

Weights: [[0.24849092 0.05652019 0.04435093 0.21833137 0.45965093]]
ndcg score: 0.873787478286

Linear regression output:
query: marguerite
  url: http://transportation.stanford.edu/marguerite/
  url: http://transportation.stanford.edu/marguerite/MargueriteSched.shtml
  url: http://transportation.stanford.edu/pdf/marguerite-map.pdf
  url: http://lbre-apps.stanford.edu/transportation/stanford_ivl/
  url: http://transportation.stanford.edu/marguerite/AboutMarguerite.shtml

SVM output:
query: marguerite
  url: http://transportation.stanford.edu/marguerite/
  url: http://transportation.stanford.edu/marguerite/MargueriteSched.shtml
  url: http://transportation.stanford.edu/pdf/marguerite-map.pdf
  url: http://transportation.stanford.edu/marguerite/AboutMarguerite.shtml
  url: http://lbre-apps.stanford.edu/transportation/stanford_ivl/

So Linear Regression approach puts
url: http://lbre-apps.stanford.edu/transportation/stanford_ivl/

higher than
url: http://transportation.stanford.edu/marguerite/AboutMarguerite.shtml

But SVM approach puts
url: http://transportation.stanford.edu/marguerite/AboutMarguerite.shtml
higher than
url: http://lbre-apps.stanford.edu/transportation/stanford_ivl/

SVM approach weights says that body is more important then header and url, also SVM says body is almost as important as anchor.
Linear Regression weight says that weight of body is almost nothing so body is almost not important at all.
According to url data AboutMarguerite.shtml has 7 body hits and stanford_ivl/ has no body hits.

Thats why because of many body hits SVM approach put AboutMarguerite.shtml higher than stanford_ivl/

And with Linear regression approach stanford_ivl/ has more anchor and linear regression gives high weight to anchor so stanford_ivl has higher place than AboutMarguerite.shtml

Task 3:
Error analysis:
1)For query: marguerite

url: http://mybus.stanford.edu/ has relevance score 2.66

But when i look at Train data file there aren't much information about this url. Body hit is 1 and body length is long , pagerank is low and there isn't anything which we can check for giving higher score for query marguerite.

I think there should be more data, we should be looking at other features also.

2) Because of the weight svm tends to favor urls which has hight number header and urls. i saw many unrelated urls with high ranking at my results. i think somehow i couldnt find correct weights or i should be using more features which can balance importance of urls and headers.

Best Ndcg score i received is 0.875129710331 with these features;**Tfldf with custom weights from PA3 + Page Rank + binaryPdfEnding + SmallWindow**

My suggestion is to consider features related to natural language processing. Because i found urls which has very high relativeness score but not enough data. so there should be other things to consider.

NDCG Scores achieved by tests with different feature sets:

Order of features which are used tfidf learning and testing: url, title, body, header, anchor
I made these test with CalcTFIdfs2.py file PrepareQueryUrlVectorWithExtraFeatures method.

Tfldf + Small Window:

Weights: [[0.23802129 0.06548143 0.08751776 0.2180793   0.45983812 -0.08464347]]

ndcg score: 0.871145815385

Tfldf + Small Window + page rank:

Weights: [[0.26642236 0.12675804 0.07819265 0.2384324   0.40290803 -0.0656504
  0.20747987]]

ndcg score: 0.870695770841

Tfldf + Small Window + page rank + binaryPdfEnding + binaryHtmlending:

Weights: [[0.27682905 0.1386208   0.09206478 0.22999413 0.37677242 -0.08019305
  0.20594618 -0.03638547 -0.05483226]]

ndcg score: 0.869102595465

Tfldf + pageRank:

Weights: [[0.27429459 0.13214166 0.05308744 0.23683108 0.39670525 0.21041693]]
ndcg score: 0.874944106714

Bm25 + Tfldf:

Weights: [[0.80120641 0.25341498 0.05482019 0.28228262 0.33240037 0.70312797
  0.1661316   0.08353606 0.05969748 -0.09915755]]

ndcg score: 0.848059091416

Only Bm25:

Weights: [[-0.02712045 -0.1534903   0.06334145 -0.00251698 -0.31147583]]

ndcg score: 0.857093831897

Tfldf + Bm25 Score:

Weights: [[0.24870344 0.07101332 0.02709478 0.22429749 0.43261579 -0.06604589]]

ndcg score: 0.834931191428

Bm25 + binaryPdfEnding:

Weights: [[-0.04919366 -0.18070319 0.04014055 -0.00063536 -0.34707574 -0.12000507]]

ndcg score: 0.855316483824

Bm25 + binaryHtmlending:
Weights: [[-0.04840736 -0.15707357 0.07307301 -0.00842711 -0.29747003 -0.10556533]]
ndcg score: 0.85543751344

Bm25 + Page Rank:

Weights: [[-8.09614617e-02 -1.85477591e-01   7.79256692e-02   2.17169697e-04
   -2.98821735e-01   1.69720410e-01]]

ndcg score : 0.8555507928394

TfIdf with custom weights from PA3 + Page Rank + binaryPdfEnding:

Weights: [[0.27352225 0.14267708 0.05956987 0.22997251 0.39359345 0.20972791
  -0.02135103]]

ndcg score: 0.874829921929

TfIdf with custom weights from PA3 + Page Rank + binaryPdfEnding + SmallWindow:

Weights: [[0.26778865 0.12842139 0.08647582 0.22953378 0.40016117 -0.06990863
   0.20447635 -0.01934527]]

ndcg score: 0.875129710331