Homework 4

Nicholas Antonov & Pat Grasso

October 27, 2016

Contents

1 Overview

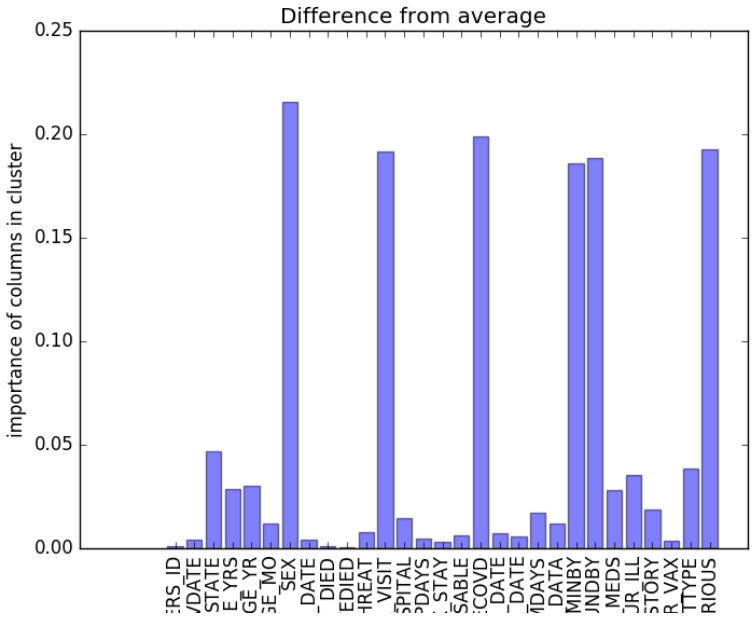
What we did was first cluster the documents. Then, for each group of documents in a cluster, we did a summation of the number of times each unique column value appeared. We then repeated this over the entire document set.

After doing this, we computed the squared distance of each cluster's average values for a column, verses the average values over the whole document set. These values show how different any cluster's makeup is from the document as a whole. We then graphed these values too see patterns, and found that the clustering does appear to be working, as different clusters have different characteristics, such as in the example submitted, cluster 2 has a very unique SPLTTYPE compared to all the other clusters

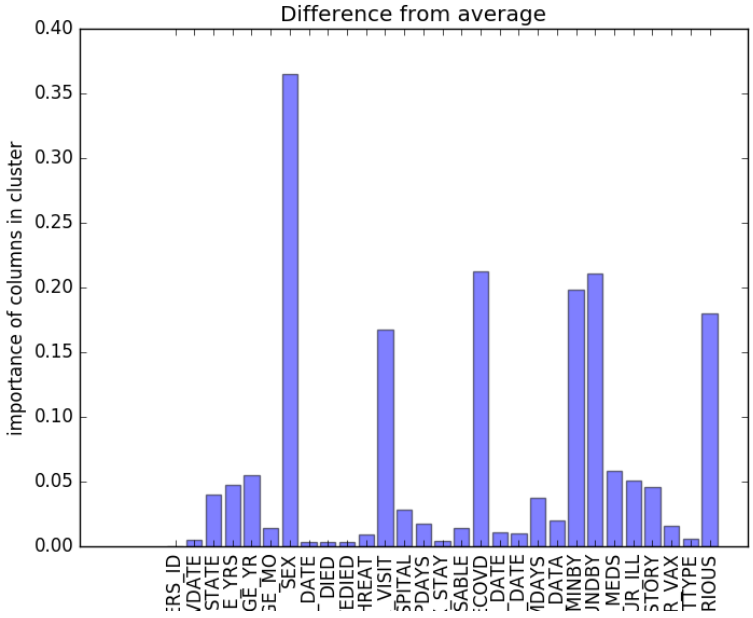
2 Output

2.1 Graphs

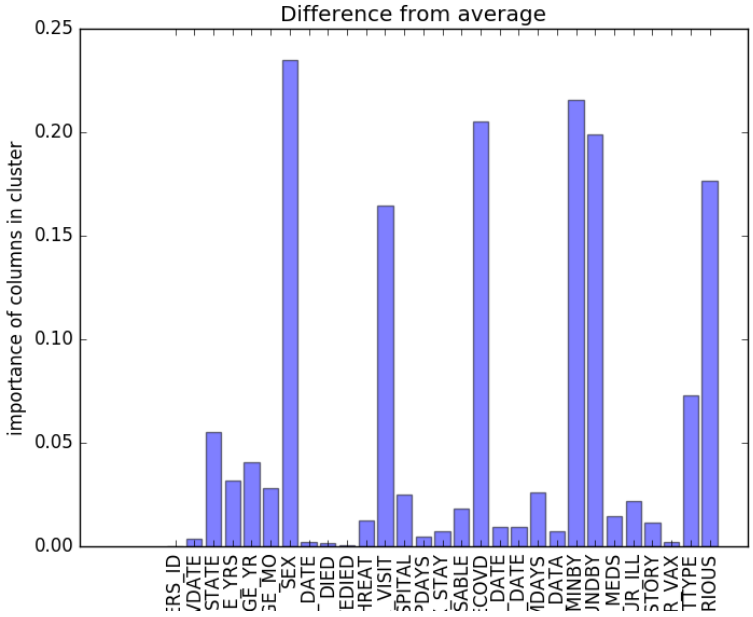
2.1.1 0



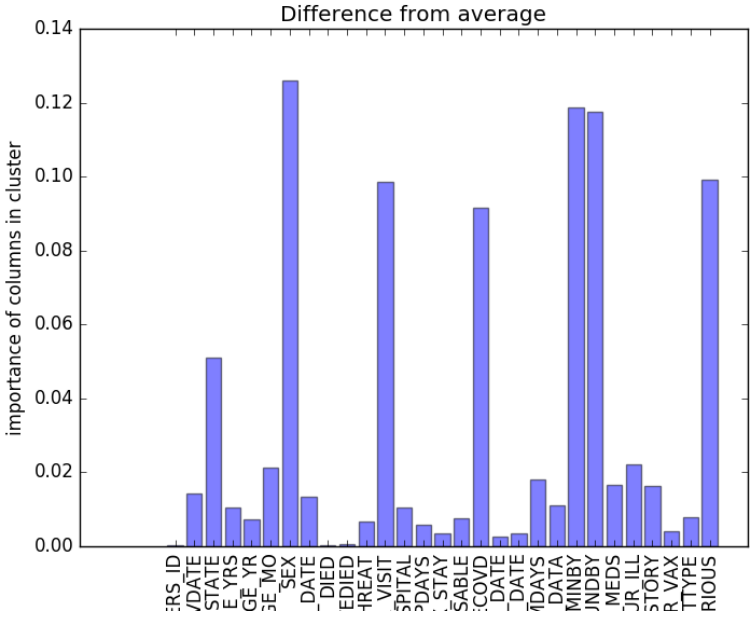
2.1.2 1


Difference from average
importance of columns in cluster
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
SEX
DATE
DIED
VISIT
STAY
DATE
DATE
DATA
MEDS

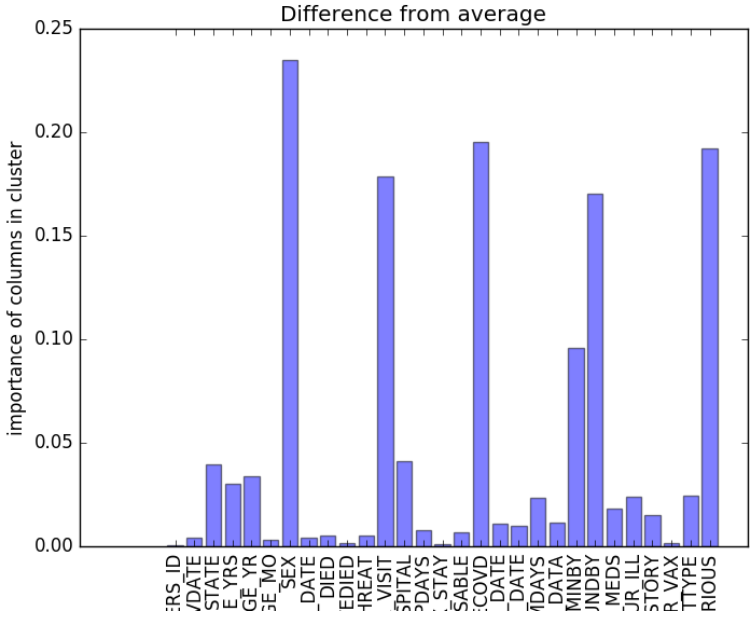
2.1.3 2


Difference from average
importance of columns in cluster
0.25
0.20
0.15
0.10
0.05
0.00
SEX
DATE
DIED
VISIT
STAY
DATE
DATE
DATA
MEDS

2.1.4 3


Difference from average
importance of columns in cluster
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
SEX
VISIT
COVID
MINBY
JNDBY
MEDS
STORY
RIOUS

2.1.5 4


Difference from average
importance of columns in cluster
0.25
0.20
0.15
0.10
0.05
0.00
SEX
VISIT
COVID
MINBY
JNDBY
MEDS
STORY
RIOUS

2.2 Textual

```
tfidf matrix shape:
(13829, 51)
Cluster 0
Size: 1037
# serious: 423
Cluster 1
Size: 4896
# serious: 170
Cluster 2
Size: 3988
# serious: 1561
Cluster 3
Size: 2648
# serious: 307
Cluster 4
Size: 1260
# serious: 513
Segment Profile:
cluster 0
SEX 0.21582183788869094
RECOVD 0.19884222144923286
SERIOUS 0.19285210673167152
ER_VISIT 0.191547565445499
V_FUNDBY 0.18849893018113967
V_ADMINBY 0.18619691417928239
cluster 1
SEX 0.3652979722338193
RECOVD 0.21215650333589559
V_FUNDBY 0.2105062352917194
V_ADMINBY 0.19841400708750392
SERIOUS 0.18033309631129432
ER_VISIT 0.16754784867643394
cluster 2
SEX 0.23501963898679087
V_ADMINBY 0.2156288075482672
RECOVD 0.20536075236375181
V_FUNDBY 0.19879649441440214
SERIOUS 0.17636895428493882
```

ER_VISIT 0.16483942753127417
cluster 3
SEX 0.12599394380693987
V_ADMINBY 0.11882411816667653
V_FUNDBY 0.11745204517150949
SERIOUS 0.09911876264227788
ER_VISIT 0.09854892483700894
RECOVD 0.09144331558006305
cluster 4
SEX 0.23505336662980583
RECOVD 0.19507734197438256
SERIOUS 0.19208753860934058
ER_VISIT 0.17877568073533076
V_FUNDBY 0.1704460651124881
V_ADMINBY 0.09599558487209404