Project Farseer

News Popularity Prediction

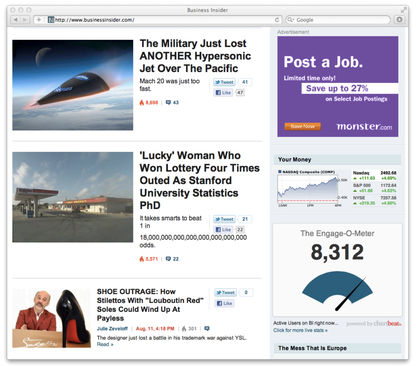
Vladimir Zaytsev
Pablo Paredes
Eduard Tuchfeld
UC Berkeley

1. Goal and Opportunity

- Predict the popularity (# page views) of a news article given its content
- .Create an analysis tool to improve content creation.

2. Observations

Businessinsider.com



download
100,000 news

News

Documents

Title: the 10 most import featur in max os x lion

Pageviews: 143529

Tags: apps mac mac os x features

Content: apple final launch it newest oper system ...

Parse & Stem

Build term dictionary & Convert docs into vectors

Document

Title: the 10 most import featur in max os x lion

Tags: apps mac mac os x features

Content: apple final launch it newest oper system ...

Dictionary

61	the
...	...
117	featur
...	...
671	os
...	...
981	mac
...	...
1027	in
...	...
1271	most
...	...
1398	10

Vector

61	+1
...	...
117	+1
...	...
671	+1
...	...
981	+1
...	...
1027	+1
...	...
1271	+1
...	...
1398	+1

3. Reduction

Calculate score - average number of page views

total page views	frequency	score	term
1800552	15	120037	valley
1626632	16	101665	silicon
1136660	12	94722	present
1306454	18	72581	woman
794235	11	72203	inspire
1405330	20	70267	tour
3438812	52	66131	20
1815864	29	62616	hot
1324035	22	60183	differ
1699716	31	54830	truth
1038872	22	47221	fact
1208126	26	46466	half
1172500	26	45096	men
656446	15	43763	epic
511907	13	39377	download
841427	22	38247	crush
978864	27	36254	pictur
526572	15	35105	shock
1023056	30	34102	os

Select the terms which are the most correlated with number of pageviews

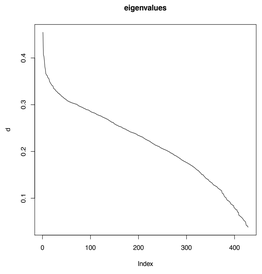
Take top **a** and bottom **b** percents of terms and remove others from the dictionary

	total page views	frequency	score	term
TOP TERMS a%	valley silicon inspire 20 woman tour fact epic download picture shock os 14 21 photo highest incredible future fake lion x comic beautiful most hollywood			
BOTTOM TERMS b%	forecast tool dax trend outlook weak forex armageddon finance letterman monuments content fantacy hunt alert oil fuel cloud politics direct conan yanke radio tv summit			

Positive influence on the number of page views

Negative influence on the number of page views

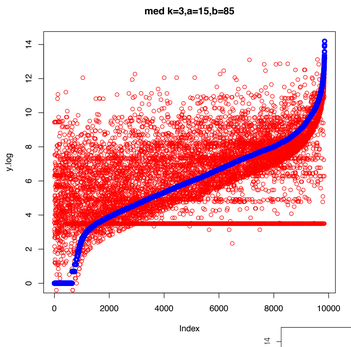
Apply PCA



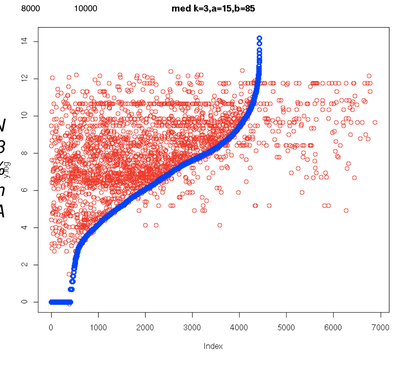
Some Data Facts

- Used different parameters a and b
- Used different datasets - with all the content of news or just only with titles
- With PCA or without PCA
- Tried to use medians to calculate scores
- Produced about 250 possible datasets with100,000 documents and from 400 to 1500 terms

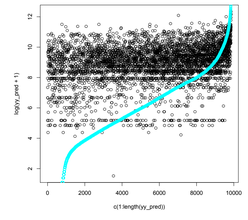
4. Results



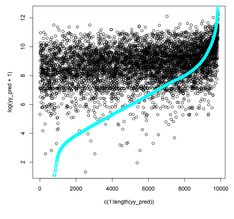
KNN
k=3
Top & Bottom =20%
Score = Median
No PCA



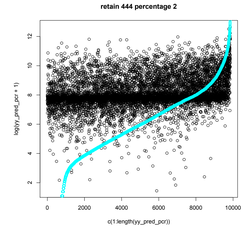
KNN
k=3
Top & Bottom =20%
Score = Median
PCA



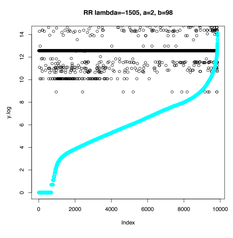
Linear Reg. 20%



Lasso 20% λ=0.1



PCR 20% #PC=444



Ridge 20% λ-1505

Summary

1) Preparation and selection of data is important.
2) We can only predict lower bounds of page views.

Ideas for Future Work

1) Use larger data sets (from various sources)
2) Add new components (emotions, time, pictures)
3) Define a multidimensional metric of popularity (pg. views, tweets, FB likes, others)
4) Use other classification techniques (SVM, Random Trees, GBM, etc.)