**Technical Report**

We tried to come up with a strategy by analyzing the patent data extracted in phase 1

We created 3 fields which we thought for our analysis from the patent data(code is attached in the appendix)

* No of forward citations
* No of Backward Citation
* Days for patent Granted

**Regression:**

We performed regression on the data to see if there is relation between Days, Backward and Forward Citation.

We observed that the Days has failed the T-test>0.5%. We have to exclude it from our analysis.

After removing we split the data into Train and Validation . With the model we got from Train data set we ran the Model on Validation data set and observed that there was as slight (0.7%) decrease in MSE for Test data. We then concluded that there is relation between Forward and Backward Citation.

Since there is relation between these variables we Derived metrics out of them to draw a decision tree.

**Decision tree:**

We derived a field named decision with the criterion if (ratio > 2.5 and days <718) = 1 else 0 and Ran decision tree. From the result of the decision tree we observed that the top 10% of the data has the maximum likelihood of getting Response=1. We extracted up to 20% of the data through random Sampling and applied our Decision Tree model on the same. We got the top rated patents according to our metrics.

**Word cloud:**

We ran word cloud on the data obtained from Model and also on the complete data. Both the results match, ie both the top patents and the most concentrated patents are on a similar fields like Desktop/Laptop OS and Display screens.

The details in the market share website confirm the above analysis.

**Insights based on our analysis:**

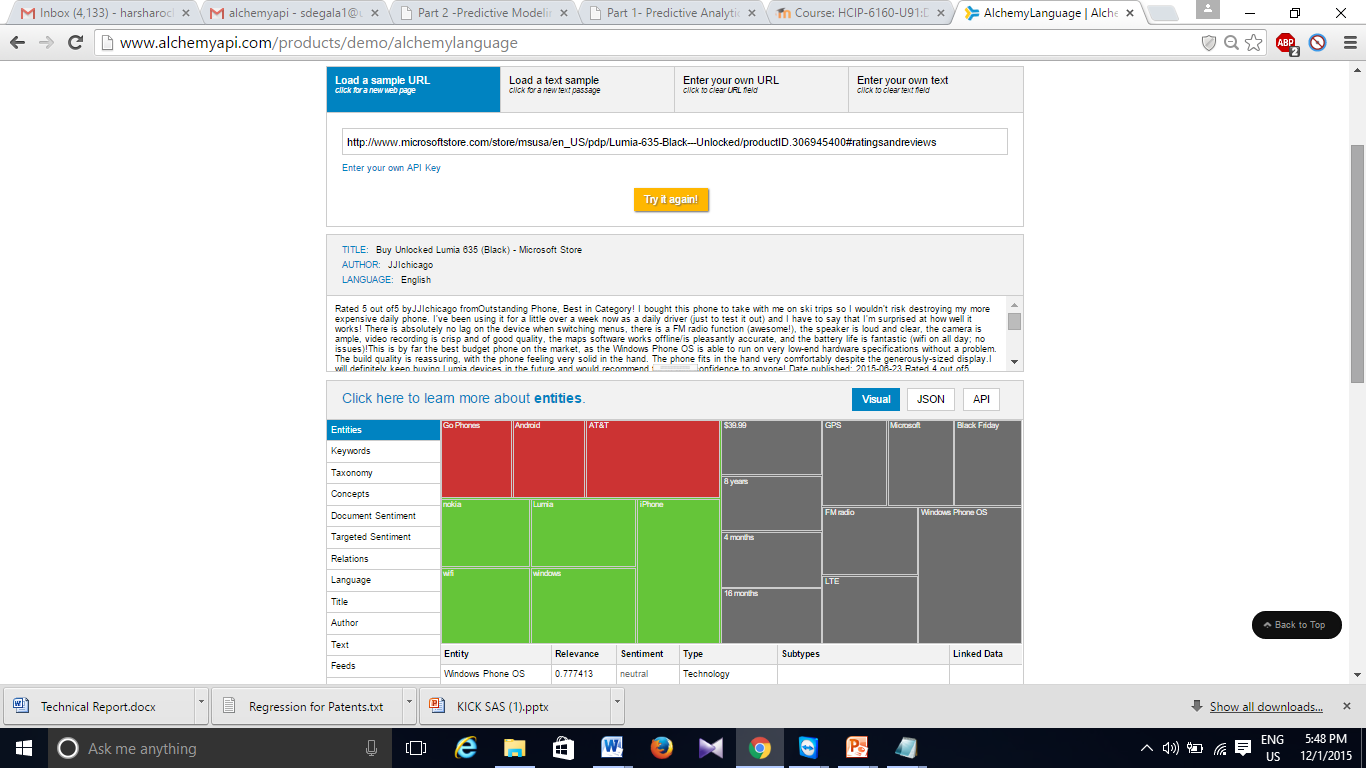
From the data analysis done on the patent data we deduced that Microsoft is predominantly concentrated on Desktop/Laptop OS and Display screens.

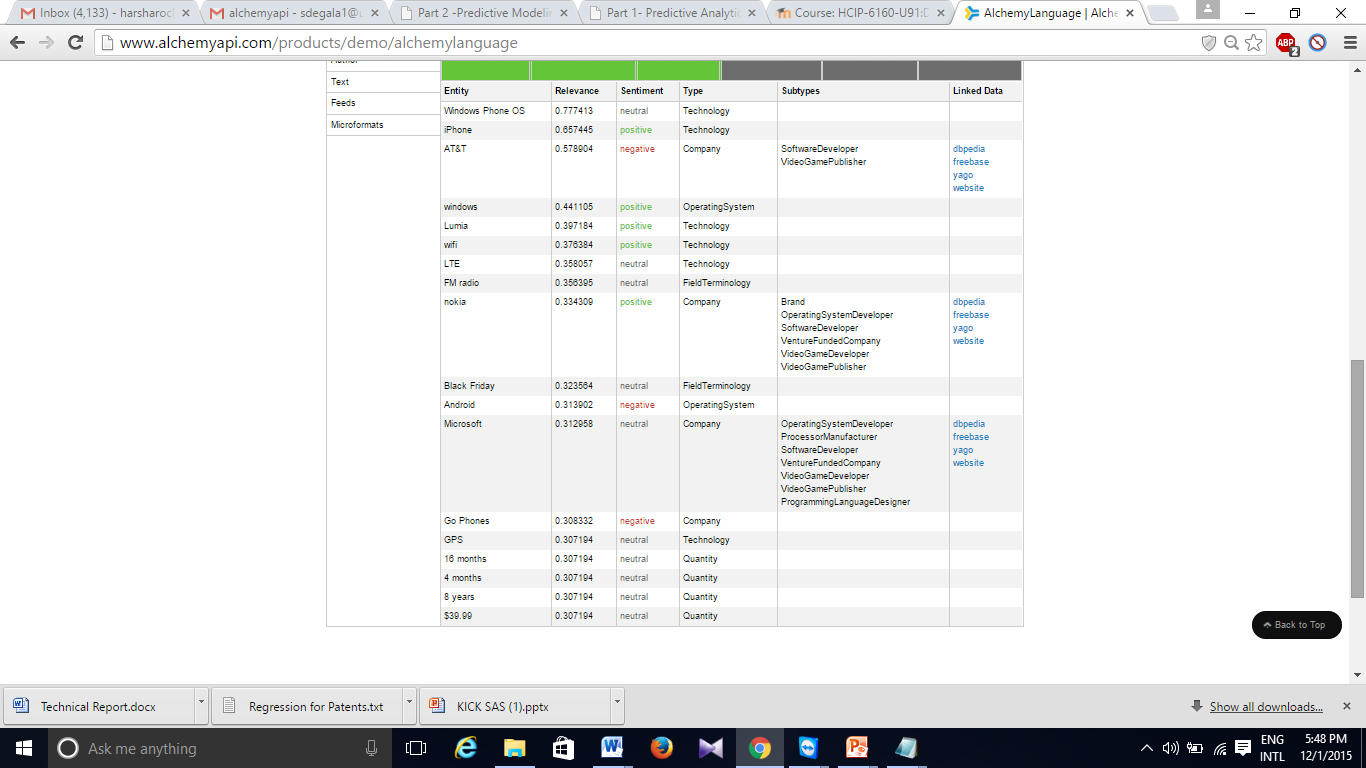
Market share states that Windows are leading in the Desktop/Laptop OS market. But it is far behind in the Mobile market though we see an increase in Market share from 2.07 to 2.45%

**Text analysis using alchemyapi**

We did the text analysis on the reviews provided by Nokia lumia in Microsoft store and amazon.

Microsoft:





Code appendix

**Code used to derive 3 Factual Fields:**

**Code used to Extract Days Information:**

select ((cast(substr(date\_grant,1,4) as int)-cast(substr(date\_publ,1,4) as int))\*365)

+

case when cast(substr(date\_grant,5,2) as int)>=cast(substr(date\_publ,5,2) as int)

then (cast(substr(date\_grant,5,2) as int)-cast(substr(date\_publ,5,2) as int))\*30

else (cast(substr(date\_publ,5,2) as int)-cast(substr(date\_grant,5,2) as int))\*30

end

+

case when cast(substr(date\_grant,7,2) as int)>=cast(substr(date\_publ,7,2) as int)

then (cast(substr(date\_grant,7,2) as int)-cast(substr(date\_publ,7,2) as int))

else (cast(substr(date\_publ,7,2) as int)-cast(substr(date\_grant,7,2) as int))

end

from days

-------------------------------------------------------------------------------------------------------------------------------

**Code used to Extract Backward and Forward Citations:**

select substr(file,(instr(file,'le="'))+4,(instr(file,'.XML'))-(instr(file,'file='))-6) as patentno\_datepubl,

(substr(file,(instr(file,'<doc-number>'))+12,(instr(file,'</doc-number>'))-13)),

case when (substr(file,(instr(file,'<category>'))+10,(instr(file,'</category>'))-11)) like 'cited by%'

then 1 else 0 end as citation

frompatent\_ipgb where file like '%us-patent-grant%' or file like '<category>%'

-------------------------------------------------------------------------------------------------------------------------------

**Regression:**

**Code to run regression on train data with Days Field Included:**

proc import datafile='E:\Hadoop\Analytics\_phase2\hive\_data\_sas.xlsx' DBMS=xlsx out=sample replace;

proc surveyselect data=sample samprate= 0.59 seed=49201 out=test outall method=srs;

run;

ods graphics on;

data patent;set test;

Days1=input(Days,7.);

run;

ods graphics on;

data patent1; set patent;

if(selected=1) then output;

run;

ods graphics on;

data patent2; set patent1;

if(Amber\_ratio>2 && Backward<22 && Days1<720) then decider=1; else decider=0;

run;

Proc univariate data=patent2 plot;

var days1 Forward;

run;

proc reg data=patent2;

model Forward= Backward Days1 ;

plot r.\*p.;

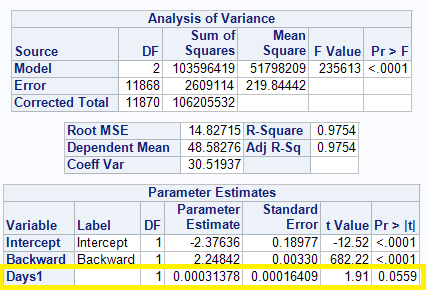
run;

proc sql;

select avg(Days1) as Avg\_days,avg(Forward)as Avg\_forward,avg(Backward) as Avg\_Backward from patent;

run;

**Result:**



-------------------------------------------------------------------------------------------------------------------------------

**Code to run regression on train data without Days:**

proc import datafile='E:\Hadoop\Analytics\_phase2\hive\_data\_sas.xlsx' DBMS=xlsx out=sample replace;

proc surveyselect data=sample samprate= 0.59 seed=49201 out=test outall method=srs;

run;

ods graphics on;

data patent;set test;

Days1=input(Days,7.);

run;

ods graphics on;

data patent1; set patent;

if(selected=1) then output;

run;

ods graphics on;

data patent2; set patent1;

if(Amber\_ratio>2 && Backward<22 && Days1<720) then decider=1; else decider=0;

run;

Proc univariate data=patent2 plot;

var days1 Forward;

run;

proc reg data=patent2;

model Forward= Backward ;

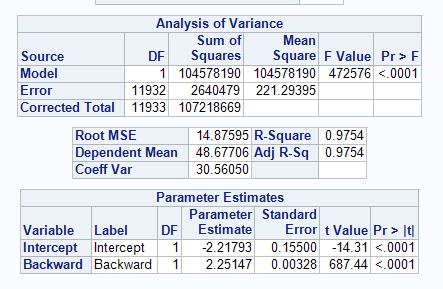
plot r.\*p.;

run;

proc sql;

select avg(Days1) as Avg\_days,avg(Forward)as Avg\_forward,avg(Backward) as Avg\_Backward from patent;

run;



**Code for Test data:**

proc import datafile='E:\Hadoop\Analytics\_phase2\hive\_data\_sas.xlsx' DBMS=xlsx out=sample replace;

procsurveyselect data=sample samprate= 0.59 seed=49201 out=test outall method=srs;

run;

ods graphics on;

datapatent;set test;

Days1=input(Days,7.);

run;

data patent1; set patent;

if(selected=0) then output;

run;

ods graphics on;

data patent2; set patent1;

if(Amber\_ratio>2 && Backward<22 && Days1<720) then decider=1; else decider=0;

run;

Procunivariate data=patent2 plot;

var days1 Forward;

run;

data patent3; set patent2;

fwd=-2.21793+2.25147\*Backward;

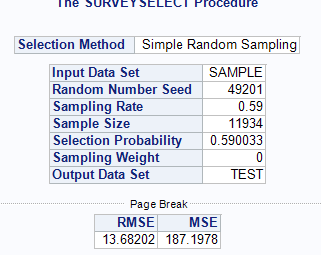
e=(Forward-Fwd)\*(Forward-Fwd);

run;

proc sql;

select sqrt(avg(e)) as RMSE,avg(e) as MSE from patent3;

run;



--------------------------------------------------------------------------------------------------------------------------------

**Code to get 20% of data as per Decision tree**

proc import datafile='E:\Hadoop\hive\_data\_sas.xlsx' DBMS=xlsx out=sample replace;

procsurveyselect data=sample samprate= 0.20 seed=49201 out=test outall method=srs;

run;

ods graphics on;

datapatent;set test;

if(selected=1) then output;

run;

data patent1;set patent;

Days1=input(Days,7.);

run;

data patent2; set patent1;

if(Amber\_ratio>2 && Backward<22 && Days1<720) then output;

run;

PROC EXPORT

DATA=WORK.patent2

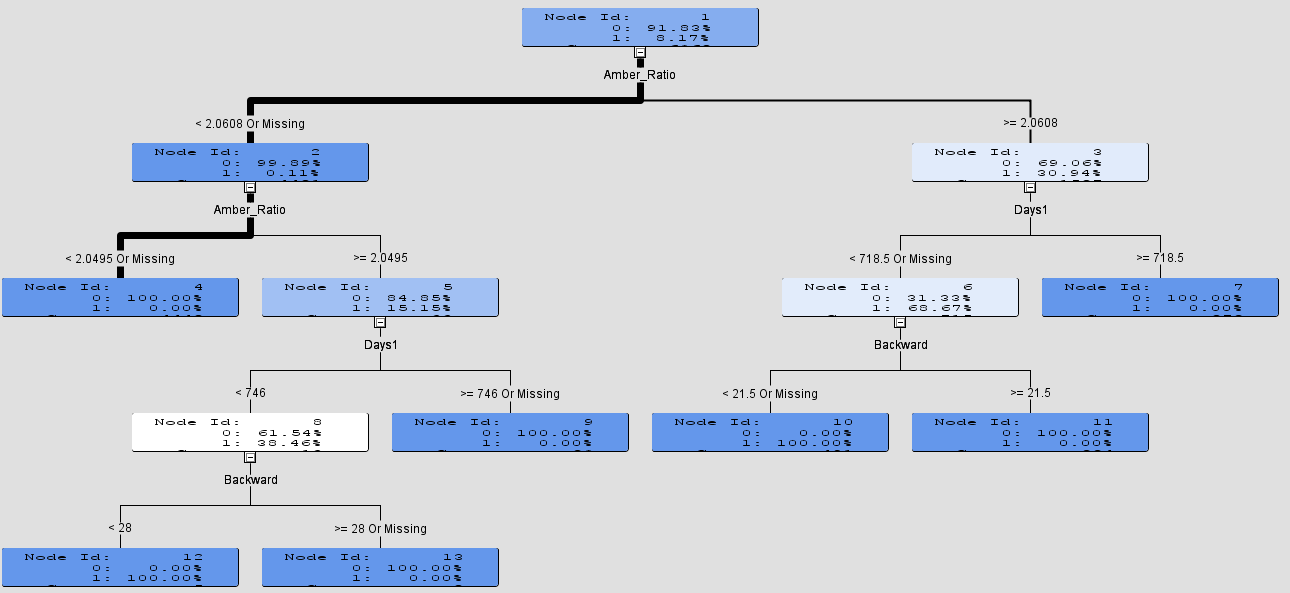
DBMS=xls

LABEL

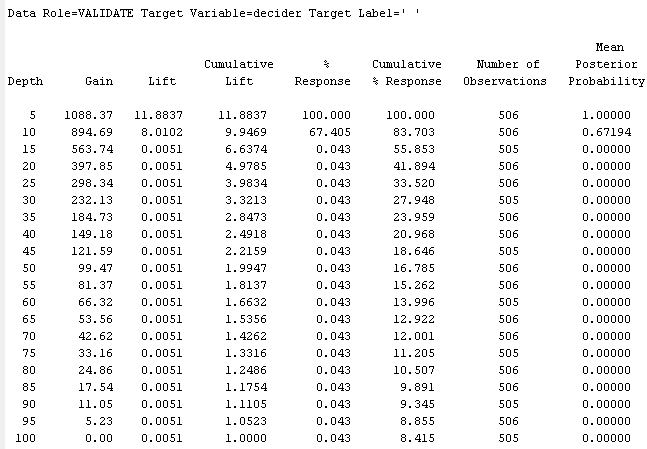
OUTFILE="E:\Hadoop\Analytics\_phase2\patent.xls"

REPLACE;

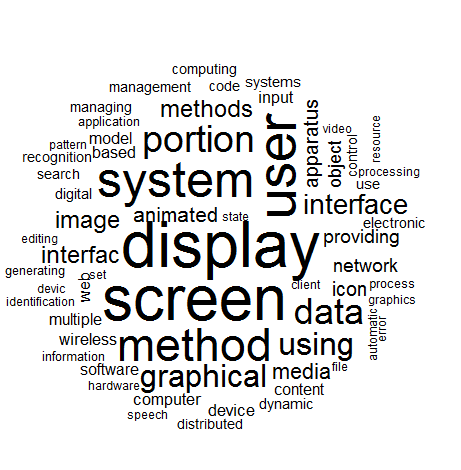
-------------------------------------------------------------------------------------------------------------------------------

****

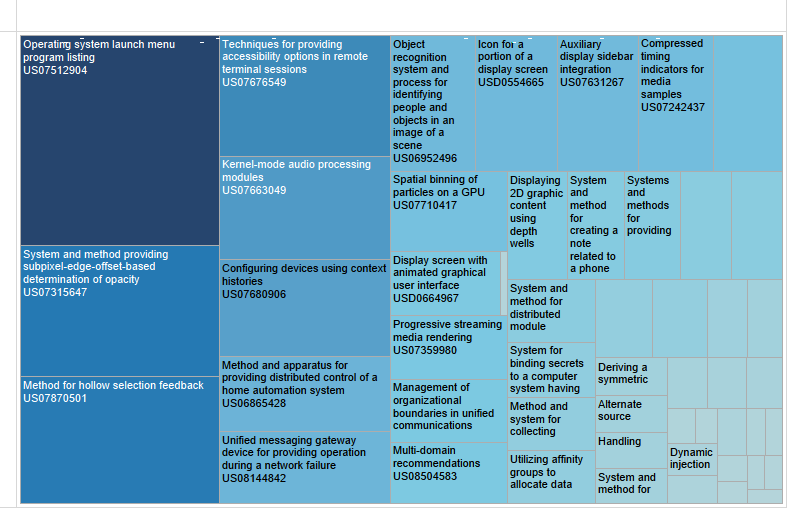
**output:**

****

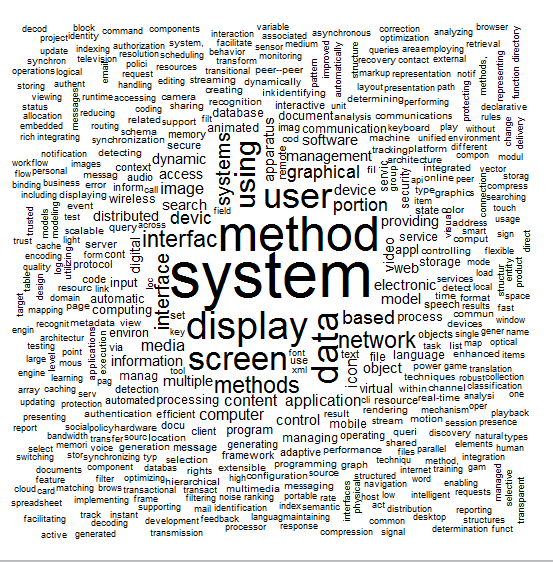
**World cloud of Microsoft patents with high Amber ratio: *(With Model)***



**Tree map for the top patents based on weight of patents:**



**Microsoft Patents highly concentrated area based on word cloud of complete data*:(Without Model)***

****