# Assignment 2

## Step 1) Data Selection

- **SQL queries:**

CREATE TABLE TX_A2_INITIAL AS

SELECT MESHHEADING, COUNT(DISTINCT TERM) AS NUMBER_OF_TERMS, COUNT(DISTINCT DMUSER.ML330_SODA_MSH.PMID) AS NUMBER_OF_ABSTRACTS

FROM DMUSER.ML330_SODA_MSH, DMUSER.ML330_SODA_NWD

WHERE DMUSER.ML330_SODA_MSH.PMID = DMUSER.ML330_SODA_NWD.PMID

GROUP BY MESHHEADING;

- **Result:**

| | MESHHEADING | NUMBER_OF_TERMS | NUMBER_OF_ABSTRACTS |
|---|---|---|---|
| 1 | Heart Diseases | 67713 | 25000 |
| 2 | Lung Neoplasms | 89205 | 25000 |

For category ‘Heart Diseases’, there are 67713 distinct terms and 25000 abstracts.

For category ‘Lung Neoplasms’, there are 89205 distinct terms and 25000 abstracts.

There are some overlapping terms between the two categories.

## Step 2) Preprocessing

- How many words are there with no vocabulary changes or pruning?

SELECT COUNT(DISTINCT TERM)

FROM DMUSER.ML330_SODA_NWD;

| | COUNT(DISTINCTTERM) |
|---|---|
| 1 | 126848 |

Before any preprocessing, the total number of distinct terms of both categories is 126848.

- How many lower case words are there?

SELECT COUNT(DISTINCT TERM)

FROM DMUSER.ML330_SODA_NWD

WHERE TERM = LOWER(TERM);

| | COUNT(DISTINCTTERM) |
|---|---|
| 1 | 83383 |

There are 83383 terms in lowercase.

SELECT COUNT(DISTINCT LOWER(TERM))

FROM DMUSER.ML330_SODA_NWD;

| | COUNT(DISTINCTLOWER(TERM)) |
|---|---|
| 1 | 114881 |

The total number of distinct terms when converting to lowercase is 114881.

- How many words are there after removing stop words?

I imported stop word list from http://www.lextek.com/manuals/onix/stopwords2.html, and created a table named TX_A2_STOPWORDS. The list contains 571 words.

SELECT COUNT(DISTINCT LOWER(TERM))

FROM DMUSER.ML330_SODA_NWD, TX_A2_STOPWORDS

WHERE LOWER(TERM) = TX_A2_STOPWORDS.STOPWORD;

| | COUNT(DISTINCTLOWER(TERM)) |
|---|---|
| 1 | 490 |

There are 490 words of the terms can be regarded as stop words. So after removing stop words, there are 114881-490=114391 distinct words in lower case.

- Select distinct terms in lowercase without stopwords

CREATE TABLE TX_A2_DISTINCT_LOWER_NOSTOP AS

SELECT DISTINCT LOWER(TERM) AS DISTINCTTERM

FROM DMUSER.ML330_SODA_NWD;

DELETE FROM TX_A2_DISTINCT_LOWER_NOSTOP

WHERE DISTINCTTERM IN (SELECT DISTINCTTERM

FROM TX_A2_DISTINCT_LOWER_NOSTOP, TX_A2_STOPWORDS

WHERE DISTINCTTERM = STOPWORD);

SELECT COUNT(*)

FROM TX_A2_DISTINCT_LOWER_NOSTOP;

| | COUNT(*) |
|---|---|
| 1 | 114391 |

- Identify and remove terms that appear in few abstracts or in most abstracts

(1) Remove terms that appear less than 11 times.

CREATE TABLE TX_A2_DF AS

SELECT DISTINCTTERM, COUNT(DISTINCT DMUSER.ML330_SODA_NWD.PMID) AS ABSTRACTNUMBER

FROM TX_A2_DISTINCT_LOWER_NOSTOP, DMUSER.ML330_SODA_NWD

WHERE TX_A2_DISTINCT_LOWER_NOSTOP.DISTINCTTERM = LOWER(DMUSER.ML330_SODA_NWD.TERM)

GROUP BY DISTINCTTERM

HAVING COUNT(DISTINCT DMUSER.ML330_SODA_NWD.PMID)>10

ORDER BY COUNT(DISTINCT DMUSER.ML330_SODA_NWD.PMID) DESC;

(2) Remove terms that are integers less than 50 and have appeared many times in different abstracts.

Remove some special meaningless words such as "'s".

| | | |
|---|---|---|
| 62 | response | 3127 |
| 63 | 's | 3125 |
| 64 | levels | 3081 |

| | | |
|---|---|---|
| 15 | 2 | 5657 |
| 16 | carcinoma | 5331 |
| 17 | significant | 5270 |
| 18 | cases | 5228 |
| 19 | 1 | 5222 |

(3) The word that appears most among all the abstracts is 'lung'. It appears 17566 times in different abstracts. However, there are 50,000 abstracts in all. So there is no need to remove words which appear so many times.

| | DISTINCTTERM | ABSTRACTNUMBER |
|---|---|---|
| 1 | lung | 17566 |
| 2 | patients | 16473 |
| 3 | cancer | 13690 |
| 4 | results | 11788 |
| 5 | disease | 10648 |

SELECT COUNT(*)

FROM TX_A2_DF;

| | COUNT(*) |
|---|---|
| 1 | 14531 |

The number of terms after step 2 is 14531.

## Step 3) Transformations

### 3.1) Term Frequency (tf) * Inverse Document Frequency (idf)

- SQL queries for computing tf*idf:

CREATE TABLE TX_A2_TF AS

SELECT DISTINCTTERM, PMID, COUNT(LOWER(DMUSER.ML330_SODA_NWD.TERM)) AS TERMNUMBER

FROM DMUSER.ML330_SODA_NWD, TX_A2_DF

WHERE LOWER(DMUSER.ML330_SODA_NWD.TERM) = TX_A2_DF.DISTINCTTERM

GROUP BY DISTINCTTERM, PMID

ORDER BY COUNT(LOWER(DMUSER.ML330_SODA_NWD.TERM)) DESC;

SELECT COUNT(DISTINCT DISTINCTTERM)

FROM TX_A2_TF;

CREATE TABLE TX_A2_TFIDF AS

SELECT TX_A2_TF.DISTINCTTERM, PMID, TERMNUMBER * (LOG(2,50000)-LOG(2,ABSTRACTNUMBER)+1) AS TFIDF

FROM TX_A2_DF, TX_A2_TF

WHERE TX_A2_TF.DISTINCTTERM = TX_A2_DF.DISTINCTTERM

ORDER BY TERMNUMBER * (LOG(2,50000)-LOG(2,ABSTRACTNUMBER)+1) DESC;

SELECT COUNT(*)

FROM TX_A2_TFIDF;

CREATE TABLE TX_A2_TFIDF_TOP AS

SELECT DISTINCT DISTINCTTERM, MAX(TFIDF) AS MAXTFIDF

FROM TX_A2_TFIDF

GROUP BY DISTINCTTERM

ORDER BY MAX(TFIDF) DESC;

- Select top 100/200/500 terms with highest values of tf*idf.

SELECT DISTINCTTERM

FROM TX_A2_TFIDF_TOP

WHERE ROWNUM<=100; or 200 or 500

**3.2) Information Gain**

- Probability of Pr(t)

CREATE TABLE TX_A2_DF_PR1 AS

SELECT DISTINCTTERM, ABSTRACTNUMBER/50000 AS PR1

FROM TX_A2_DF;

- Probability of Pr(Lung Neoplasms|t)

CREATE TABLE TX_A2_DF_LUNGNUMBER AS

SELECT TX_A2_DF.DISTINCTTERM, COUNT(DISTINCT DMUSER.ML330_SODA_NWD.PMID) AS LUNGNUMBER

FROM TX_A2_DF, DMUSER.ML330_SODA_MSH, DMUSER.ML330_SODA_NWD

WHERE TX_A2_DF.DISTINCTTERM = LOWER(DMUSER.ML330_SODA_NWD.TERM) AND DMUSER.ML330_SODA_NWD.PMID = DMUSER.ML330_SODA_MSH.PMID

AND DMUSER.ML330_SODA_MSH.MESHHEADING = 'Lung Neoplasms'

GROUP BY TX_A2_DF.DISTINCTTERM

ORDER BY LUNGNUMBER DESC;

CREATE TABLE TX_A2_DF_NUM_PR2 AS

SELECT TX_A2_DF.DISTINCTTERM, TX_A2_DF_LUNGNUMBER.LUNGNUMBER AS LUNGNUMBER

FROM TX_A2_DF_LUNGNUMBER

RIGHT OUTER JOIN TX_A2_DF

ON TX_A2_DF.DISTINCTTERM = TX_A2_DF_LUNGNUMBER.DISTINCTTERM;

- Right outer join is critical with null values for future computation of probability

CREATE TABLE TX_A2_DF_PR2 AS

SELECT TX_A2_DF.DISTINCTTERM, LUNGNUMBER/ABSTRACTNUMBER AS PR2

FROM TX_A2_DF, TX_A2_DF_NUM_PR2

WHERE TX_A2_DF.DISTINCTTERM = TX_A2_DF_NUM_PR2.DISTINCTTERM;

- Probability of Pr(Heart Diseases|t)

CREATE TABLE TX_A2_DF_HEARTNUMBER AS

SELECT TX_A2_DF.DISTINCTTERM, COUNT(DISTINCT DMUSER.ML330_SODA_NWD.PMID) AS HEARTNUMBER

FROM TX_A2_DF, DMUSER.ML330_SODA_MSH, DMUSER.ML330_SODA_NWD

WHERE TX_A2_DF.DISTINCTTERM = LOWER(DMUSER.ML330_SODA_NWD.TERM) AND DMUSER.ML330_SODA_NWD.PMID = DMUSER.ML330_SODA_MSH.PMID

AND DMUSER.ML330_SODA_MSH.MESHHEADING = 'Heart Diseases'

GROUP BY TX_A2_DF.DISTINCTTERM

ORDER BY HEARTNUMBER DESC;

CREATE TABLE TX_A2_DF_NUM_PR3 AS

SELECT TX_A2_DF.DISTINCTTERM, TX_A2_DF_HEARTNUMBER.HEARTNUMBER AS HEARTNUMBER

FROM TX_A2_DF_HEARTNUMBER

RIGHT OUTER JOIN TX_A2_DF

ON TX_A2_DF.DISTINCTTERM = TX_A2_DF_HEARTNUMBER.DISTINCTTERM;

CREATE TABLE TX_A2_DF_PR3 AS

SELECT TX_A2_DF.DISTINCTTERM, HEARTNUMBER/ABSTRACTNUMBER AS PR3

FROM TX_A2_DF, TX_A2_DF_NUM_PR3

WHERE TX_A2_DF.DISTINCTTERM = TX_A2_DF_NUM_PR3.DISTINCTTERM;

- Probability of Pr(not t)

CREATE TABLE TX_A2_DF_PR4 AS

SELECT DISTINCTTERM, 1 - ABSTRACTNUMBER/50000 AS PR4

FROM TX_A2_DF;

- Probability of Pr(Lung Neoplasma|not t)

CREATE TABLE TX_A2_DF_PR5 AS

SELECT TX_A2_DF.DISTINCTTERM, (25000-NVL(LUNGNUMBER,0))/(50000-ABSTRACTNUMBER) AS PR5

FROM TX_A2_DF, TX_A2_DF_NUM_PR2

WHERE TX_A2_DF.DISTINCTTERM = TX_A2_DF_NUM_PR2.DISTINCTTERM;

- Probability of Pr(Heart Diseases|not t)

CREATE TABLE TX_A2_DF_PR6 AS

SELECT TX_A2_DF.DISTINCTTERM, (25000-NVL(HEARTNUMBER,0))/(50000-ABSTRACTNUMBER) AS PR6

FROM TX_A2_DF, TX_A2_DF_NUM_PR3

WHERE TX_A2_DF.DISTINCTTERM = TX_A2_DF_NUM_PR3.DISTINCTTERM;

- Compute IG of each term

CREATE TABLE TX_A2_IG AS

SELECT TX_A2_DF_PR1.DISTINCTTERM, (PR1*(NVL(PR2,0)*LOG(2,NVL(PR2,1)) + NVL(PR3,0)*LOG(2,NVL(PR3,1))) + PR4*(PR5*LOG(2,PR5) + PR6*LOG(2,PR6))) AS IG

FROM TX_A2_DF_PR1, TX_A2_DF_PR2, TX_A2_DF_PR3, TX_A2_DF_PR4, TX_A2_DF_PR5, TX_A2_DF_PR6

WHERE TX_A2_DF_PR1.DISTINCTTERM = TX_A2_DF_PR2.DISTINCTTERM AND TX_A2_DF_PR2.DISTINCTTERM = TX_A2_DF_PR3.DISTINCTTERM

AND TX_A2_DF_PR3.DISTINCTTERM = TX_A2_DF_PR4.DISTINCTTERM AND TX_A2_DF_PR4.DISTINCTTERM = TX_A2_DF_PR5.DISTINCTTERM

AND TX_A2_DF_PR5.DISTINCTTERM = TX_A2_DF_PR6.DISTINCTTERM

ORDER BY IG DESC;

- Select top 100/200/500 terms with highest values of Information Gain.

SELECT DISTINCTTERM

FROM TX_A2_IG

WHERE ROWNUM<=100; or 200 or 500

**3.3) Create tables for classification**

Create 6 tables with 100, 200 and 500 terms for tf*idf and for information gain respectively.

- Table1: Top 100 terms for tf*idf

CREATE TABLE TX_A2_C1 AS

SELECT TX_A2_FEATURES1.*, DMUSER.ML330_SODA_MSH.MESHHEADING

FROM TX_A2_FEATURES1

JOIN DMUSER.ML330_SODA_MSH

ON TX_A2_FEATURES1.PMID = DMUSER.ML330_SODA_MSH.PMID;

| | PMID | MG_M | ANP | TGF_BETA | TOPOTECAN | BUPIVACAINE | DOT | ANG |
|---|---|---|---|---|---|---|---|---|
| 1 | 35194 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 36578 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 36744 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 37740 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 38426 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- Table2: Top 200 terms for tf*idf

CREATE TABLE TX_A2_C2 AS

SELECT TX_A2_FEATURES2.*, DMUSER.ML330_SODA_MSH.MESHHEADING

FROM TX_A2_FEATURES2

```
JOIN DMUSER.ML330_SODA_MSH
ON TX_A2_FEATURES2.PMID = DMUSER.ML330_SODA_MSH.PMID;
```

- Table3: Top 500 terms for tf*idf

```
CREATE TABLE TX_A2_C3 AS
SELECT TX_A2_FEATURES3.*, DMUSER.ML330_SODA_MSH.MESHHEADING
FROM TX_A2_FEATURES3
JOIN DMUSER.ML330_SODA_MSH
ON TX_A2_FEATURES3.PMID = DMUSER.ML330_SODA_MSH.PMID;
```

- Table4: Top 100 terms for Information Gain

```
CREATE TABLE TX_A2_C4 AS
SELECT TX_A2_FEATURES4.*, DMUSER.ML330_SODA_MSH.MESHHEADING
FROM TX_A2_FEATURES4
JOIN DMUSER.ML330_SODA_MSH
ON TX_A2_FEATURES4.PMID = DMUSER.ML330_SODA_MSH.PMID;
```

| | PMID | LUNG | CANCER | HEART | CARDIAC | CELL | TUMOR |
|---|---|---|---|---|---|---|---|
| 1 | 56439 | 0 | 1 | 0 | 0 | 0 | 3 |
| 2 | 56980 | 2 | 1 | 0 | 0 | 0 | 1 |
| 3 | 56986 | 0 | 0 | 0 | 0 | 0 | 10 |
| 4 | 56987 | 4 | 3 | 0 | 0 | 3 | 0 |
| 5 | 57472 | 0 | 0 | 0 | 0 | 0 | 0 |

- Table5: Top 200 terms for Information Gain

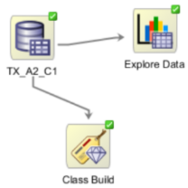
```
CREATE TABLE TX_A2_C5 AS
SELECT TX_A2_FEATURES5.*, DMUSER.ML330_SODA_MSH.MESHHEADING
FROM TX_A2_FEATURES5
JOIN DMUSER.ML330_SODA_MSH
ON TX_A2_FEATURES5.PMID = DMUSER.ML330_SODA_MSH.PMID;
```

- Table6: Top 500 terms for Information Gain

```
CREATE TABLE TX_A2_C6 AS
SELECT TX_A2_FEATURES6.*, DMUSER.ML330_SODA_MSH.MESHHEADING
FROM TX_A2_FEATURES6
JOIN DMUSER.ML330_SODA_MSH
ON TX_A2_FEATURES6.PMID = DMUSER.ML330_SODA_MSH.PMID;
```

## Step 4) Data Mining

Use the oracle data miner to create two classifiers (a decision tree and naïve bayes) for the 6 tables of feature sets. The screenshot below is an example.




## Step 5) Interpretation

### 1. Table of top 100 terms for tf*idf

- Accuracy and confusion matrix of decision trees

| | |
|---|---|
| Average Accuracy: | 53.4149 |
| Overall Accuracy: | 53.4149 |
| Total Cost: | 18,580 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,876 | 95 | 9,971 | 99.0472 | 190 |
| Lung Neoplasms | 9,195 | 776 | 9,971 | 7.7826 | 18,390 |
| Total | 19,071 | 871 | 19,942 | | |
| Correct % | 51.7854 | 89.093 | | | |
| Cost | 18,390 | 190 | | | |

- Accuracy and confusion matrix of naïve bayes

| | |
|---|---|
| Average Accuracy: | 54.3927 |
| Overall Accuracy: | 54.3927 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,867 | 104 | 9,971 | 98.957 | |
| Lung Neoplasms | 8,991 | 980 | 9,971 | 9.8285 | |
| Total | 18,858 | 1,084 | 19,942 | | |
| Correct % | 52.3226 | 90.4059 | | | |
| Cost | | | | | |

### 2. Table of top 200 terms for tf*idf

- Accuracy and confusion matrix of decision trees

| | |
|---|---|
| Average Accuracy: | 53.9214 |
| Overall Accuracy: | 53.9214 |
| Total Cost: | 18,378 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,917 | 54 | 9,971 | 99.4584 | 108 |
| Lung Neoplasms | 9,135 | 836 | 9,971 | 8.3843 | 18,270 |
| Total | 19,052 | 890 | 19,942 | | |
| Correct % | 52.0523 | 93.9326 | | | |
| Cost | 18,270 | 108 | | | |

- Accuracy and confusion matrix of naïve bayes

| | |
|---|---|
| Average Accuracy: | 57.4967 |
| Overall Accuracy: | 57.4967 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,753 | 218 | 9,971 | 97.8137 | |
| Lung Neoplasms | 8,258 | 1,713 | 9,971 | 17.1798 | |
| Total | 18,011 | 1,931 | 19,942 | | |
| Correct % | 54.1502 | 88.7105 | | | |
| Cost | | | | | |

**3. Table of top 500 terms for tf*idf**

- Accuracy and confusion matrix of decision trees

| | |
|---|---|
| Average Accuracy: | 58.7403 |
| Overall Accuracy: | 58.7403 |
| Total Cost: | 16,456 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 1,795 | 8,176 | 9,971 | 18.0022 | 16,352 |
| Lung Neoplasms | 52 | 9,919 | 9,971 | 99.4785 | 104 |
| Total | 1,847 | 18,095 | 19,942 | | |
| Correct % | 97.1846 | 54.8162 | | | |
| Cost | 104 | 16,352 | | | |

- Accuracy and confusion matrix of naïve bayes

| | |
|---|---|
| Average Accuracy: | 62.9024 |
| Overall Accuracy: | 62.9024 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,666 | 305 | 9,971 | 96.9411 | |
| Lung Neoplasms | 7,093 | 2,878 | 9,971 | 28.8637 | |
| Total | 16,759 | 3,183 | 19,942 | | |
| Correct % | 57.6765 | 90.4178 | | | |
| Cost | | | | | |

**4. Table of top 100 terms for Information Gain**

- Accuracy and confusion matrix of decision trees

| | |
|---|---|
| Average Accuracy: | 90.7582 |
| Overall Accuracy: | 90.7582 |
| Total Cost: | 3,686 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,764 | 207 | 9,971 | 97.924 | 414 |
| Lung Neoplasms | 1,636 | 8,335 | 9,971 | 83.5924 | 3,272 |
| Total | 11,400 | 8,542 | 19,942 | | |
| Correct % | 85.6491 | 97.5767 | | | |
| Cost | 3,272 | 414 | | | |

- Accuracy and confusion matrix of naïve bayes

| | |
|---|---|
| Average Accuracy: | 91.1192 |
| Overall Accuracy: | 91.1192 |

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,837 | 134 | 9,971 | 98.6561 | |
| Lung Neoplasms | 1,637 | 8,334 | 9,971 | 83.5824 | |
| Total | 11,474 | 8,468 | 19,942 | | |
| Correct % | 85.733 | 98.4176 | | | |
| Cost | | | | | |

**5. Table of top 200 terms for Information Gain**

- Accuracy and confusion matrix of decision trees

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,764 | 207 | 9,971 | 97.924 | 414 |
| Lung Neoplasms | 1,636 | 8,335 | 9,971 | 83.5924 | 3,272 |
| Total | 11,400 | 8,542 | 19,942 | | |
| Correct % | 85.6491 | 97.5767 | | | |
| Cost | 3,272 | 414 | | | |

| | |
|---|---|
| Average Accuracy: | 90.7582 |
| Overall Accuracy: | 90.7582 |
| Total Cost: | 3,686 |

- Accuracy and confusion matrix of naïve bayes

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,829 | 142 | 9,971 | 98.5759 | |
| Lung Neoplasms | 1,746 | 8,225 | 9,971 | 82.4892 | |
| Total | 11,575 | 8,367 | 19,942 | | |
| Correct % | 84.9158 | 98.3029 | | | |
| Cost | | | | | |

| | |
|---|---|
| Average Accuracy: | 90.5325 |
| Overall Accuracy: | 90.5325 |

**6. Table of top 500 terms for Information Gain**

- Accuracy and confusion matrix of decision trees

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,764 | 207 | 9,971 | 97.924 | 414 |
| Lung Neoplasms | 1,636 | 8,335 | 9,971 | 83.5924 | 3,272 |
| Total | 11,400 | 8,542 | 19,942 | | |
| Correct % | 85.6491 | 97.5767 | | | |
| Cost | 3,272 | 414 | | | |

| | |
|---|---|
| Average Accuracy: | 90.7582 |
| Overall Accuracy: | 90.7582 |
| Total Cost: | 3,686 |

- Accuracy and confusion matrix of naïve bayes

| | Heart Diseases | Lung Neoplasms | Total | Correct % | Cost |
|---|---|---|---|---|---|
| Heart Diseases | 9,786 | 185 | 9,971 | 98.1446 | |
| Lung Neoplasms | 2,275 | 7,696 | 9,971 | 77.1838 | |
| Total | 12,061 | 7,881 | 19,942 | | |
| Correct % | 81.1376 | 97.6526 | | | |
| Cost | | | | | |

| | |
|---|---|
| Average Accuracy: | 87.6642 |
| Overall Accuracy: | 87.6642 |

**7. Findings from the above results**

In general, performances of text classification of feature sets based on tf*idf selection method are much worse than those of feature sets based on Information Gain. The overall accuracy of tf*idf feature sets is about 50% to 65%, while the overall accuracy of Information Gain feature sets is about 90%. So information gain feature selection method is more effective for this text classification project.

Discrepancies of accuracy between decision trees and naïve bayes are not obvious for the same table (source data). For top 100/200/500 terms based on tf*idf and top 100 terms based on information gain, the performances of naïve bayes are a little bit better than decision trees.

For tf*idf feature sets, as the number of terms increases, the accuracy of both decision trees and naïve bayes increase. As can be seen in the confusion matrix, the accuracy of ‘Heart Diseases’ is much higher than ‘Lung Neoplasms’. However, there is a reverse for the accuracy of decision trees with 500 terms based on tf*idf, which shows that the performance of identification of ‘Lung Neoplasms’ is much better than the other.

For information gain feature sets, as the number of terms increases, there is no difference between the accuracy of decision trees. Decision trees have the same performance on all the three tables based on information gain regardless of the number of terms. On the other hand, however, as the number of terms rises, the accuracy of naïve bayes declines for feature sets based on information gain.

As the number of terms increase, the model of decision trees become more sophisticated with more braches. After taking a look at the selected terms that played an important role for text classification, I find that top terms selected by information gain look more reasonable and anticipated, while top terms selected by tf*idf seem surprised to me.