Analysis of unstructered data

Lecture 10 - text classification

Janusz Szwabiński

Outlook:

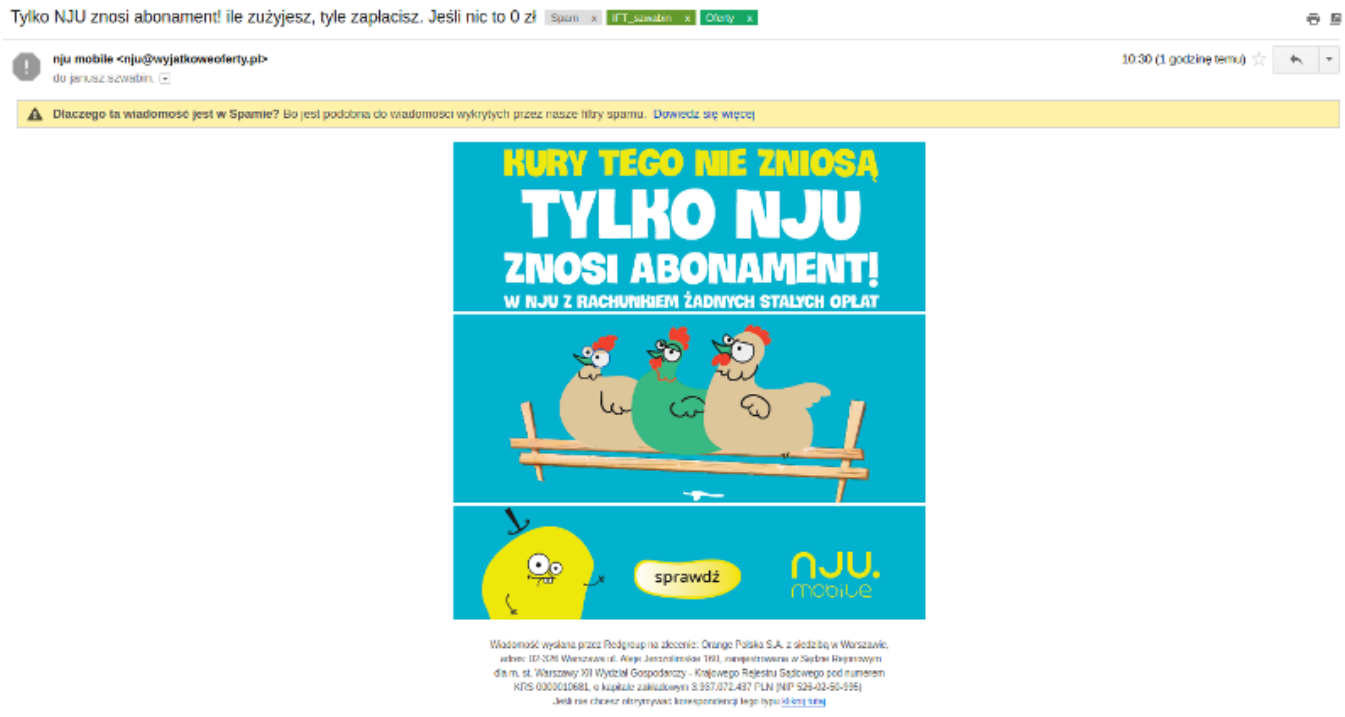
- Motivation
- Text classification
- Naive Bayes classifier
- Decision trees
- Maximum Entropy classifier
- How to improve accuracy of a text classifier?

References:

1. Ch. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*
2. P. Jackson, I. Moulinier, *Natural Language Processing for Online Applications*
3. Dan Jurafsky's presentation, *Text Classification and Naive Bayes*
4. NLTK Book, http://www.nltk.org/book/ (http://www.nltk.org/book/)
5. http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/ (http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/)

Motivation

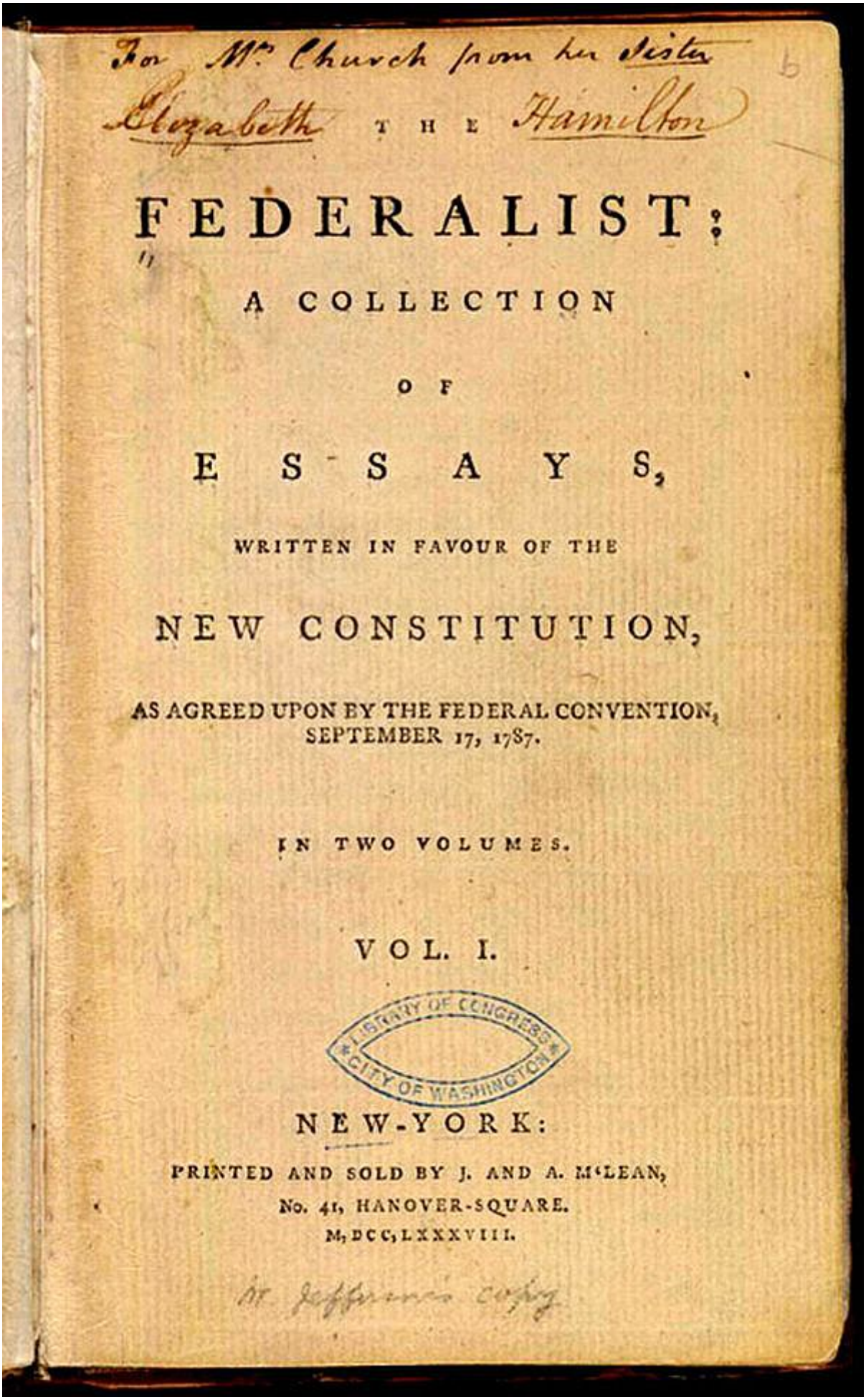
Is that spam?



The Federalist (or The Federalist Papers)

- 1787-1788

- a collection of 85 articles and essays written under the pseudonym "Publius" to promote the ratification of the United States Constitution
- at the time of publication the authorship of the articles was a closely guarded secret
- authors: Alexander Hamilton, James Madison, and John Jay
  - according to Hamilton, he was the author of 63 articles, Jay wrote 5 of them, Madison – 14, 3 were written by Hamilton and Madison
  - in 1818 Madison claimed the authorship of 26 esseys
  - the scholarly detective work of Douglass Adair in 1944 postulated the following assignments of authorship:
    - Alexander Hamilton (51 articles: No. 1, 6–9, 11–13, 15–17, 21–36, 59–61, and 65–85)
    - James Madison (29 articles: No. 10, 14, 18–20,[10] 37–58 and 62–63)
    - John Jay (5 articles: No. 2–5 and 64).
  - in 1963 Mosteller and Wallace corroborated Adair's results by a computer aided text classification


For Mrs Church from her sister
Elizabeth THE Hamilton

FEDERALIST:

A COLLECTION

OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M, DCC, LXXXVIII.

Mr Jefferson's copy

Female or male author?

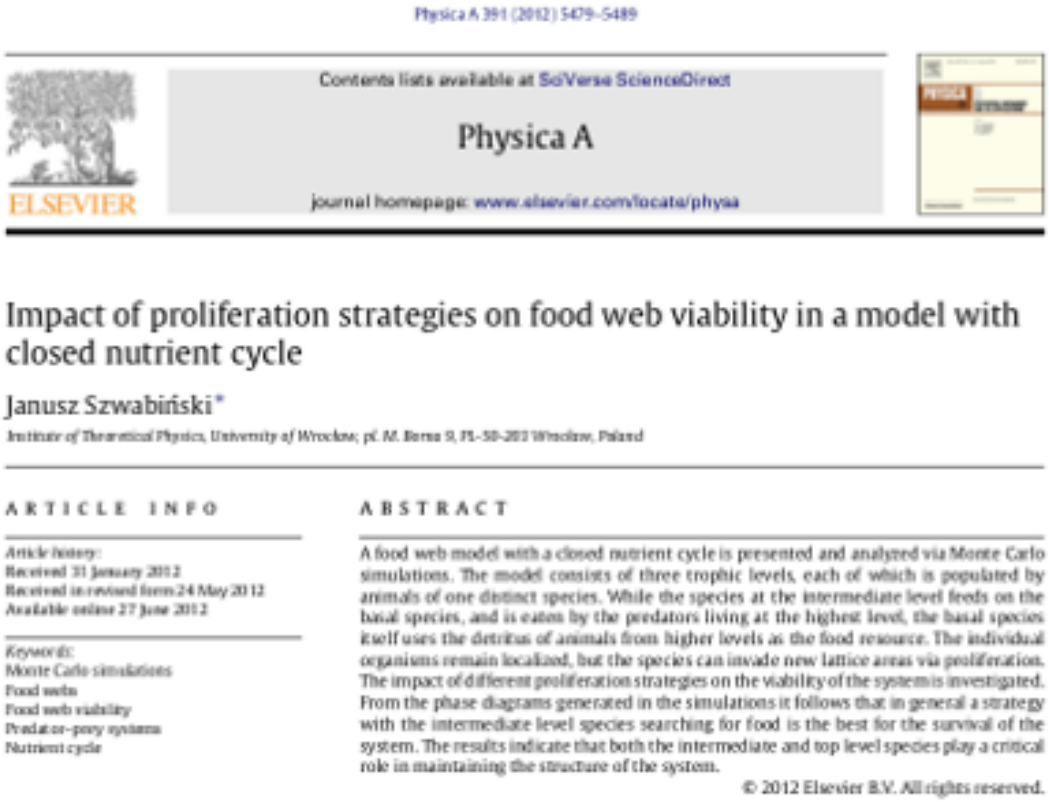
> 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin China; the central area with its imperial capital at Hue was the protectorate of Annam, and the northern region, Tongking, was also a separate ...
>
> Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

- S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts", Text 23, pp. 321–346

Movie reviews

- Unbelievably disappoing
- Full of zany characters and richly applied satire, and some great plot twists
- This is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes

What is this paper about?



Physica A 391 (2012) 5479–5489

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

ELSEVIER

Impact of proliferation strategies on food web viability in a model with closed nutrient cycle

Janusz Szwabiński *

Institute of Theoretical Physics, University of Wrocław, pl. M. Borna 9, PL-50-204 Wrocław, Poland

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form 24 May 2012
Available online 27 June 2012

Keywords:
Monte Carlo simulations
Food webs
Food web viability
Predator–prey systems
Nutrient cycle

ABSTRACT

A food web model with a closed nutrient cycle is presented and analyzed via Monte Carlo simulations. The model consists of three trophic levels, each of which is populated by animals of one distinct species. While the species at the intermediate level feeds on the basal species, and is eaten by the predators living at the highest level, the basal species itself uses the detritus of animals from higher levels as the food resource. The individual organisms remain localized, but the species can invade new lattice areas via proliferation. The impact of different proliferation strategies on the viability of the system is investigated. From the phase diagrams generated in the simulations it follows that in general a strategy with the intermediate level species searching for food is the best for the survival of the system. The results indicate that both the intermediate and top level species play a critical role in maintaining the structure of the system.

'Quantitative Biology' categories on arXiv:

- Quantitative Biology
- Biomolecules
- Cell Behavior
- Genomics
- Molecular Networks
- Neurons and Cognition
- Other Quantitative Biology

- Populations and Evolution
- Quantitative Methods
- Subcellular Processes
- Tissues and Organs

Text classification

From Wikipedia:

> Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done “manually” (or “intellectually”) or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is used mainly in information science and computer science. The problems are overlapping, however, and there is therefore also interdisciplinary research on document classification.

Formal definition

- input data:
  - document d
  - set of labels (classes, categories) $C = \{c_1, c_2, \ldots, c_J\}$
- output data:
  - predicted class c of document d

Applications

- document cataloguing
- message filtering
- spam filtering
- language identification
- sentiment analysis
- genre identification (movies/books/music)
- authorship identification (gender, age)
- ... and many more

Methods

Hand-coded rules

- rules are usually expressed by a combination of words or other features
- example of a spam filtering rule:

```
black-list-address OR("dolars" AND "have been selected")
```

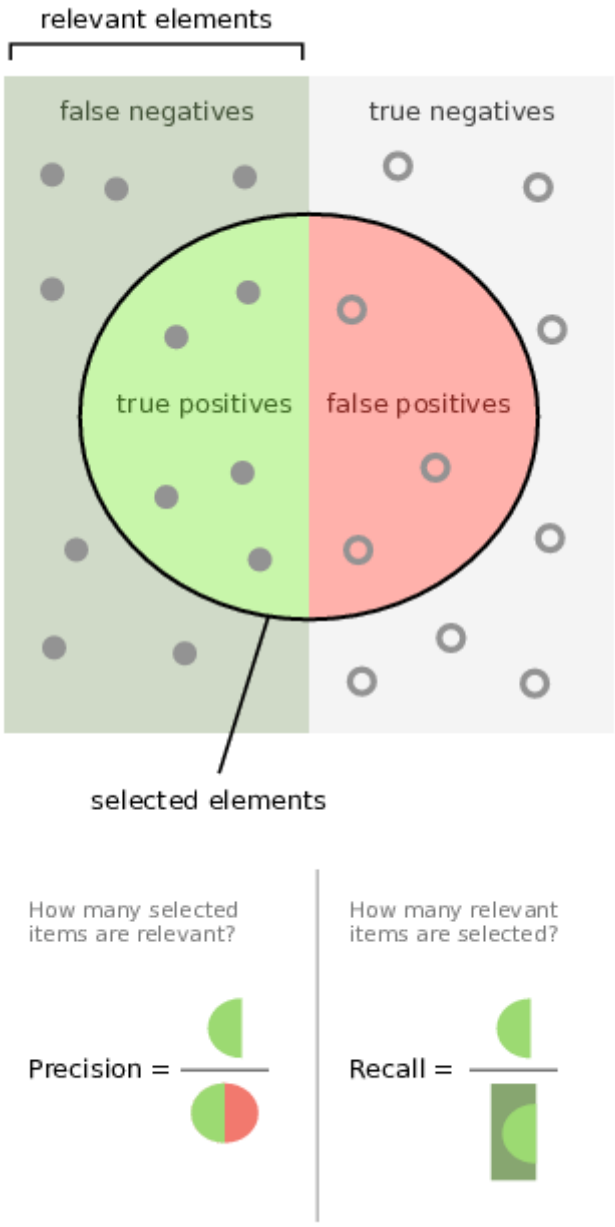
- high accurracy possible
- require continuous updates (by experts)

Supervised machine learning

- input data:
  - document d
  - set of labels (classes) $C = \{c_1, c_2, \ldots, c_J\}$
  - training data, i.e. set of documents labeled manually, $(d_1, c_1), \ldots (d_m, c_m)$
- output data:

  - learned classifier $\gamma: \ d \rightarrow c$
- possible classifiers:
  - naive Bayes
  - logistic regression
  - support vector machine
  - k nearest neighbors
  - maximum entropy
  - decision trees
  - decision rules (lists)
  - . . .

Assessing a classifier



The performance of the classifier can be also presented in form of a contingency matrix:

	N (predicted)	**P (predicted)**
N (actual)	True negatives, TN	False positives, fp
P (actual)	False negatives, fn	True positives, TP

Then:

$$Precision = \frac{TP}{TP + fp}$$
$$Recall = \frac{TP}{TP + fn}$$

Other measures are possible:

$$Accuracy = \frac{TP + TN}{N}$$
$$Fallout = \frac{fp}{fp + fn}$$
$$NumOfErrors = fp + fn$$

Naive Bayes classifier

Bayes' theorem

Definition - conditional probability

Conditional probability is a measure of the probability of an event given that (by assumption, presumption, assertion or evidence) another event has occurred,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem

Let A_i be a series of events such that $A_i \cap A_j \neq 0$ for $i \neq j$, $\sum_i P(A_i) = 1$ and $P(B) > 0$. Then:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

where

$$P(B) = \sum_j P(B|A_j)P(A_j)$$

Example

Suppose there are two bowls of cookies. Bowl 1 contains 30 vanilla cookies and 10 chocolate cookies. Bowl 2 contains 20 of each.

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

We are looking for conditional probability

$$P(box1|van)$$

It is not obvious how to compute it. On the other hand, the probability of a vanilla cookie given Bowl 1 is much easier to calculate:

$$P(van|box1) = \frac{30}{40} = \frac{3}{4}$$

Now we can use the Bayes' theorem:

$$P(box1|van) = \frac{P(van|box1)P(box1)}{P(van)}$$

In our example we have:

$$P(box1) = \frac{1}{2}$$
$$P(van) = \frac{50}{80} = \frac{5}{8}$$

Thus we get:

$$P(box1|van) = \frac{(1/2)(3/4)}{5/8} = \frac{3}{5}$$

A (slightly) different interpretation

Given a hypothesis H and evidence E , Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence and the probability of the hypothesis after getting the evidence is

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where:

- $P(H)$ - prior (probability that H is true before knowing E)
- $P(H|E)$ - posterior (probability that H is true after knowing E)
- $P(E)$ - normalization constant (probability of getting E for any H)
- $P(E|H)$ - likelihood (probability of getting E if H)

Naive Bayes classifier

- a classifier based on the Bayes' theorem
- appriopriate to multidimensional problems
- "naive" because of the assumption of independence between features
  - very often an oversimplification
  - nevertheless, the performance of the classifier is usually better than expected

Example with structured data

In the following table a training set with customer data of a computer shop is presented:

Id	Age	Income	Graduated	Credit rating	Buyer
1	<30	high	no	good	no
2	<30	high	no	excellent	no
3	[30;40]	high	no	good	yes
4	>40	medium	no	good	yes
5	>40	low	yes	good	yes
6	>40	low	yes	excellent	no
7	[30;40]	low	yes	excellent	yes
8	<30	medium	no	good	no
9	<30	low	yes	good	yes
10	>40	medium	yes	good	yes
11	<30	medium	yes	excellent	yes
12	[30;40]	medium	no	excellent	yes
13	[30;40]	high	yes	good	yes
14	>40	medium	no	excellent	no

Our goal is to check if person X (age under 30, medium income, graduated and with a good credit score) will buy a computer in the shop.

To this end we have to check for which value of i the product

$$P(X|C_i) * P(C_i)$$

has its maximum. Here:

- $C_i, \ \ i = 1, 2$ - two categories: will buy or not
- $P(C_i)$ - prior describing the probability of being a member of class C_i

From the training set we have:

$$P(C_1) = \frac{9}{14} = 0,643$$
$$P(C_2) = \frac{5}{14} = 0,357$$

We calculate the conditional probabilities for each attribute values:

$$P(X|C_1) = P(age =" < 30 " \ |C_1) * P(income =" \ medium " \ |C_1) * P(graduated =" \ ye$$
$$P(X|C_2) = P(age =" < 30 " \ |C_2) * P(income =" \ medium " \ |C_2) * P(graduated =" \ ye$$

Based on the data in the training set we get:

$$P(age =" < 30 " \ |C_1) = \frac{2}{9}, \ \ P(income =" \ medium " \ |C_1) = \frac{4}{9}, \ \ P(graduated =" \ yes$$
$$P(credit =" \ good " \ |C_1) = \frac{6}{9}$$
$$P(age =" < 30 " \ |C_2) = \frac{3}{5}, \ \ P(income =" \ medium " \ |C_2) = \frac{2}{5}, \ \ P(graduated =" \ yes$$
$$P(income =" \ good " \ |C_2) = \frac{2}{5}$$

Hence

$$P(X|C_1) = 0,044, \ \ P(X|C_2) = 0,019$$

and

$$P(X|C_1) * P(C_1) = 0,028$$
$$P(X|C_2) * P(C_2) = 0,007$$

Since the probability is higher for C_1, we classify the person X as a buyer.

Text classification

- bag of word as text representation
- use reduced representation!!!
- additional assumption --> position of a word in a document does not matter

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun**… It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet.

x **love** xxxxxxxxxxxxxxxx **sweet**
xxxxxxx **satirical** xxxxxxxxxx
xxxxxxxxxxx **great** xxxxxxx
xxxxxxxxxxxxxxxxxxxx **fun** xxxx
xxxxxxxxxxxxx **whimsical** xxxx
romantic xxxx **laughing**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx **recommend** xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx **several** xxxxxxxxxxxxxxxxx
xxxxx **happy** xxxxxxxxx **again**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

According to the Bayes' theorem, for a document d and a category c we have

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Classification of a document is nothing but finding the category with highest probability (*maximum a posteriori*, MAP):

$$c_{MAP} = \text{argmax}_{c \in C} P(c|d) = \text{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

For the sake of simplicity let us ommit the normalization constant:

$$c_{MAP} = \text{argmax}_{c \in C} P(d|c)P(c)$$

A document is simply a set of features, thus:

$$c_{MAP} = \text{argmax}_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c)$$

Problems:

- $O(|X|^n \times |C|)$ parameters
- probability can be estimated only for a large training set

"Solution":

- we assume that for every class c the features x_i are independent, i.e.

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) * P(x_2|c) * \cdots * P(x_n|c)$$
$$c_{NB} = \text{argmax}_{c \in C} P(c) \prod_{x \in X} P(x|c)$$

- in the first approximation we can take word frequencies as the probabilities of word occurences

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Is it enough?

No! It may happen that there was no word 'fantastic' in the training set. Then

$$\hat{P}(\text{" } fantastic \text{ "} \mid \text{" } positive \text{ "}) = \frac{count(\text{" } fantastic \text{ "}, \text{" } positive \text{ "})}{\sum_{w \in V} count(w, \text{" } positive \text{ "})} = 0$$

which gives

$$c_{NB} = \text{argmax}_{c \in C} \hat{P}(c) \prod_{x \in X} \hat{P}(x|c) = 0$$

Laplace smoothing

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} (count(w, c_j) + 1)}$$

Naive Bayes in NLTK

In [1]:

```
import nltk
```

Name classification

Let us come back once again to the example from the previous lecture - we want to classify names according to gender. First, we read in the corpus:

In [2]:

```
from nltk.corpus import names
names = ([(name, 'male') for name in names.words('male.txt')] + [(name,
'female') for name in names.words('female.txt')])
```

In [3]:

```
names[:10]
```

Out[3]:

```
[('Aamir', 'male'),
 ('Aaron', 'male'),
 ('Abbey', 'male'),
 ('Abbie', 'male'),
 ('Abbot', 'male'),
 ('Abbott', 'male'),
 ('Abby', 'male'),
 ('Abdel', 'male'),
 ('Abdul', 'male'),
 ('Abdulkarim', 'male')]
```

In [4]:

```
import random
random.shuffle(names)
names[:10]
```

Out[4]:

```
[('Matthew', 'male'),
 ('Burgess', 'male'),
 ('Mendel', 'male'),
 ('Cordy', 'female'),
 ('Katharyn', 'female'),
 ('Faythe', 'female'),
 ('Sigfried', 'male'),
 ('Mindy', 'female'),
 ('Elora', 'female'),
 ('Shay', 'male')]
```

In order to build a classifier, we need a feature extractor. Suppose the last letter of the name will be used as the feature for classification:

In [5]:

```
def gender_features(word):
    return {'last_letter': word[-1]}
gender_features('Shrek')
```

Out[5]:

```
{'last_letter': 'k'}
```

We will use the extractor function to process the input data and train the classifier:

In [6]:

```
featuresets = [(gender_features(n), gender) for (n, gender) in names]
train_set, test_set = featuresets[500:], featuresets[:500]
print(train_set[:10])
```

```
[({'last_letter': 'a'}, 'female'), ({'last_letter': 'a'}, 'femal
e'), ({'last_letter': 'e'}, 'female'), ({'last_letter': 'l'}, 'mal
e'), ({'last_letter': 'd'}, 'male'), ({'last_letter': 'e'}, 'femal
e'), ({'last_letter': 'd'}, 'male'), ({'last_letter': 'a'}, 'femal
e'), ({'last_letter': 'l'}, 'female'), ({'last_letter': 'a'}, 'fema
le')]
```

In [7]:

```
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

Let us check how the classifier works on names that are not in the corpus:

In [8]:

```
classifier.classify(gender_features('Neo'))
```

Out[8]:

```
'male'
```

In [9]:

```
classifier.classify(gender_features('Trinity'))
```

Out[9]:

```
'female'
```

We may use the test set to check the accuracy of the classifier:

In [10]:

```
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.766
```

Checking of the most important features is possible as well:

In [11]:

```
classifier.show_most_informative_features(5)
```

```
Most Informative Features
             last_letter = 'a'            female : male   =     35.
5 : 1.0
             last_letter = 'k'              male : female =     28.
6 : 1.0
             last_letter = 'f'              male : female =     14.
0 : 1.0
             last_letter = 'p'              male : female =     12.
0 : 1.0
             last_letter = 'd'              male : female =     10.
7 : 1.0
```

Let us do some serious testing:

In [12]:

```
from nltk.metrics.scores import precision, recall
```

In [13]:

```
import collections
from nltk.metrics.scores import precision, recall
from nltk.classify import NaiveBayesClassifier

refsets = collections.defaultdict(set)
testsets = collections.defaultdict(set)

for i, (feats, label) in enumerate(test_set):
    refsets[label].add(i)
    observed = classifier.classify(feats)
    testsets[observed].add(i)

print('male precision:', precision(refsets['male'], testsets['male']))
print('male recall:', recall(refsets['male'], testsets['male']))
print('female precision:', precision(refsets['female'], testsets['female']))
print('female recall:', recall(refsets['female'], testsets['female']))
```

```
male precision: 0.7248677248677249
male recall: 0.6782178217821783
female precision: 0.7909967845659164
female recall: 0.825503355704698
```

If we are not happy with the performance, we can extract a different feature set:

In [14]:

```
def gender_features2(name):
    features = {}
    features["first_letter"] = name[0].lower()
    features["last_letter"] = name[-1].lower()
    for letter in 'abcdefghijklmnopqrstuvwxyz':
        features["count({})".format(letter)] = name.lower().count(letter)
        features["has({})".format(letter)] = (letter in name.lower())
    return features
```

In [15]:

```
featuresets = [(gender_features2(n), gender) for (n, gender) in names]
train_set, test_set = featuresets[500:], featuresets[:500]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.746
```

It may happen that it decreases after adding new features, especially in case of small and noisy dataset (*overfitting*). The choice of the feature set is crucial and usually one uses trial and error to find the optimal features.

Document classification

Let us consider now the movie review corpus included in NLTK:

In [16]:

```
from nltk.corpus import movie_reviews
documents = [(list(movie_reviews.words(fileid)), category)
             for category in movie_reviews.categories()
             for fileid in movie_reviews.fileids(category)]
random.shuffle(documents)
print(documents[0])
```

```
(['this', 'is', 'one', 'of', 'the', 'worst', 'big', '-', 'screen',
 'film', 'experiences', 'i', "'", 've', 'had', 'for', 'a', 'while',
'.', 'with', 'this', 'film', ',', 'plus', '`', 'showgirls', "'", 'a
nd', '`', 'basic', 'instinct', "'", ',', 'paul', 'verhoeven', 'ha
s', 'stamped', 'himself', 'as', 'currently', 'one', 'of', 'the', 'w
orst', 'blockbuster', 'directors', '.', 'his', 'celebrated', 'fil
m', '`', 'total', 'recall', "'", 'was', '?', 'i', 'admit', '?', 'su
ccessfully', 'scripted', ',', 'but', 'it', 'nonetheless', 'containe
d', 'directorial', 'flaws', '.', 'obviously', 'nobody', 'wanted',
 'to', 'invest', 'too', 'much', 'money', 'in', 'a', 'production',
 'from', 'someone', 'like', 'verhoeven', ',', 'the', 'result', 'bei
ng', 'that', 'much', 'of', 'the', 'special', 'effects', 'in', '`',
 'starship', "'", 'seemed', 'fake', '.', 'but', 'not', 'everythin
g', 'bad', 'in', 'the', 'film', 'was', 'the', 'director', "'", 's',
'fault', ',', 'even', 'though', 'he', 'was', 'one', 'of', 'the', 'g
uys', 'who', 'employed', 'the', 'actors', '.', 'it', 'is', 'surpris
ing', 'that', 'none', 'of', 'the', 'actors', 'received', 'nominatio
ns', 'for', 'the', 'razzie', 'awards', '(', 'i', 'expected', 'fiv
e', 'for', 'the', 'acting', 'categories', ')', '.', 'casserole', 'v
anity', 'devoid', ',', 'dense', 'ribald', ',', 'dingy', 'miasma',
 ',', 'and', 'jackass', 'bushy', 'are', 'in', 'serious', 'need', 'o
f', 'acting', 'school', '.', 'no', ',', 'they', 'have', 'to', 'pas
s', 'primary', 'school', 'drama', 'classes', 'first', '.', 'while',
'`', 'total', "'", 'was', 'written', 'well', ',', '`', 'starship',
 "'", 'is', 'purely', 'pathetic', '.', 'all', 'right', ',', 'it',
 'is', 'supposed', 'to', 'be', 'a', 'fast', '-', 'paced', 'entertai
nment', 'film', ',', 'and', 'you', "'", 're', 'supposed', 'to', 'tu
rn', 'off', 'your', 'intellect', '(', 'completely', ')', 'and', 'en
joy', 'the', 'action', 'sequences', 'and', 'special', 'effects',
 '(', 'that', 'is', ',', 'guts', 'and', 'gore', ')', '.', 'as',
 'a', 'matter', 'of', 'fact', ',', 'i', 'found', 'the', 'activity',
'incredibly', 'boring', ',', 'a', 'complete', 'waste', 'of', 'mor
e', 'than', 'two', 'hours', '.', 'half', 'of', 'the', 'film', 'wa
s', 'a', 'bad', 'episode', 'of', '`', 'beverly', 'hills', '90210',
 "'", '(', 'dina', 'meyer', 'was', 'in', '"', 'beverly', 'hills',
 '"', ')', ',', 'while', 'another', 'quarter', 'was', 'simply', 'no
thing', '(', 'things', 'like', 'presenting', 'irrelevant', 'informa
tion', 'in', 'an', 'irritating', 'way', 'on', 'the', 'web', ')',
 ',', 'and', 'the', 'rest', 'was', 'a', 'display', 'of', 'humans',
 'fighting', 'computer', '-', 'generated', 'images', '.', 'the', 'b
attles', 'were', 'all', 'the', 'same', '?', 'jumping', 'around',
 ',', 'shoot', 'or', 'get', 'stabbed', '?', 'and', 'on', 'barren',
 'planets', 'that', 'only', 'had', 'giant', 'insects', '.', 'ther
e', 'weren', "'", 't', 'even', 'any', 'stunts', ',', 'which', 'i',
 'consider', 'slightly', 'more', 'exciting', 'than', 'pictures', 'r
unning', 'around', '.', 'i', 'wonder', 'what', 'the', 'insects', 'e
at', ',', 'if', 'there', "'", 's', 'nothing', 'but', 'them', 'on',
 'the', 'planets', '?', 'there', 'is', 'so', 'much', 'laughable',
 'treatment', 'in', 'this', 'film', ',', 'and', 'it', 'is', 'frankl
y', 'not', 'amusing', 'when', 'jokes', 'are', 'intended', '.', 'thi
s', 'type', 'of', 'story', 'is', 'obviously', 'aimed', 'at', '10',
 '-', 'year', '-', 'olds', ',', 'who', 'can', "'", 't', 'see', 'i
t', 'anyway', 'because', 'of', 'the', 'violence', 'and', 'some', 's
exuality', '.', 'but', 'then', ',', 'there', 'are', 'always', '16',
'-', 'year', '-', 'olds', 'who', 'have', 'that', 'frame', 'of', 'mi
nd', '.', 'the', 'pointless', 'plot', 'begins', 'when', 'johnny',
 "'", 's', '(', 'vanity', 'devoid', ')', 'girlfriend', 'carmen',
 '(', 'richards', ')', 'decides', 'that', 'she', 'wants', 'to', 'jo
in', 'the', 'troopers', 'to', 'fight', 'the', 'insects', 'who', 'ar
e', 'throwing', 'asteroids', 'at', 'earth', '.', 'johnny', 'then',
 'signs', 'up', 'as', 'a', 'trooper', 'also', ',', 'after', 'an',
```

```
 'overacted', 'argument', 'with', 'his', 'parents', '.', 'but', 'th
ere', 'is', 'another', 'girl', ',', 'dizzy', '(', 'meyer', ')',
 ',', 'who', 'likes', 'johnny', 'and', 'then', 'there', 'is', 'anot
her', 'boy', 'who', 'likes', 'carmen', ',', 'which', 'results', 'i
n', 'a', 'love', 'quadrangle', ',', 'which', 'isn', "'", 't', 'bett
er', ',', 'because', 'it', 'means', 'augmented', 'worse', '-', 'tha
n', '-', 'stereotyped', 'soap', 'opera', ',', 'increased', 'bitchin
ess', ',', 'and', 'more', 'bad', 'beverly', 'hills', '+', 'melros
e', '.', 'and', 'the', 'result', 'of', 'this', 'love', 'quadrangl
e', 'at', 'the', 'end', 'is', 'also', 'rather', 'stupid', '.', 'any
way', ',', 'getting', 'back', 'to', 'the', 'thing', 'you', 'might',
'call', 'plot', ',', 'johnny', 'is', 'too', 'stupid', 'to', 'be',
 'a', 'pilot', 'and', 'has', 'to', 'join', 'the', 'infantry', ',',
 'while', 'his', 'girl', 'and', 'the', 'other', 'dude', 'are', 'i
n', 'the', 'same', 'league', '.', 'dizzy', 'comes', 'chasing', 'joh
nny', 'and', 'joins', 'the', 'infantry', 'also', '.', 'they', 'the
n', 'start', 'training', ',', 'which', 'contains', 'what', 'roger',
'ebert', 'calls', 'ips', '(', 'idiot', 'plot', 'syndrome', ',', 'mo
ments', 'when', 'only', 'an', 'idiot', 'would', 'have', 'made', 'su
ch', 'obvious', 'mistakes', ')', ',', 'then', 'real', 'combat',
 '.', 'and', 'guess', 'what', '?', 'that', "'", 's', 'about', 'as',
'complex', 'as', 'it', 'gets', '.', 'oh', ',', 'and', 'one', 'of',
 'their', 'friends', ',', 'carl', '(', 'neil', 'patrick', 'harris',
'a', '.', 'k', '.', 'a', '.', 'doogie', 'howser', 'm', '.', 'd',
 '.', ')', ',', 'becomes', 'involved', 'in', 'war', 'intelligence',
',', 'and', 'his', 'abilities', 'at', 'the', 'end', 'are', 'reall
y', 'corny', 'and', 'make', 'me', 'want', 'to', 'spray', 'insectici
de', 'on', 'someone', 'for', 'it', '.', 'he', "'", 's', 'my', 'fift
h', 'nom', 'for', 'a', 'razzie', '.', 'the', 'troopers', 'fight',
 ',', 'fall', 'in', 'love', ',', 'die', ',', 'kill', ',', 'and', 't
ry', 'to', 'act', '.', 'naturally', ',', 'they', 'win', ',', 'or',
 'sort', 'of', 'half', '-', 'win', '.', 'of', 'course', ',', 'in',
 'between', '(', 'and', 'at', 'the', 'end', ')', 'there', 'are', 's
oldiers', 'chatting', 'and', 'smiling', 'while', 'carrying', 'griev
ous', 'wounds', 'caused', 'by', 'bug', 'legs', '.', 'the', 'bugs',
 'also', 'suffer', 'from', 'ips', ':', 'why', 'would', 'you', 'rele
ase', 'hold', 'of', 'your', 'captive', 'before', 'killing', 'it',
 '?', 'isn', "'", 't', 'it', 'also', 'amazing', 'that', 'earthling
s', 'haven', "'", 't', 'invented', 'better', 'hand', '-', 'held',
 'weapons', 'by', 'then', '?', 'the', 'only', 'question', 'that',
 'remains', 'is', 'why', 'i', 'gave', 'it', 'one', 'star', 'instea
d', 'of', 'zero', '.', 'well', ',', 'maybe', 'a', '-', 'quarter',
 '(', 'of', 'a', 'star', ')', 'for', 'the', 'originality', 'of', 't
he', 'co', '-', 'sex', 'shower', 'scene', 'and', 'the', '(', 'ver
y', ')', 'brief', 'moments', 'of', 'suspense', ',', 'another', 'qua
rter', 'for', 'copying', '`', 'zulu', "'", 'and', 'letting', 'the',
'good', 'guys', '(', 'the', 'bugs', ')', 'win', ',', 'and', 'half',
'a', 'star', 'for', 'the', 'sucking', '-', 'out', 'of', 'the', 'bra
in', 'of', 'one', 'of', 'those', 'people', 'who', 'call', 'themselv
es', 'actors', '(', 'but', 'there', 'should', 'have', 'been', 'mor
e', ',', 'the', 'troopers', 'deserved', 'to', 'die', ')', '.'], 'ne
g')
```

Again, we define a helper function to extract the features:

In [17]:

```
all_words = nltk.FreqDist(w.lower() for w in movie_reviews.words())
word_features = list(all_words)[:2000]

def document_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    return features
```

Let us check how this function works:

In [18]:

```
print(document_features(movie_reviews.words('pos/cv957_8737.txt')))
```

{'contains(six)': False, 'contains(8034)': False, 'contains(cityscape)': False, 'contains(eloquence)': False, 'contains(alain)': False, 'contains(cheeky)': False, 'contains(oil)': False, 'contains(milla)': False, 'contains(defiency)': False, 'contains(dots)': False, 'contains(gorefest)': False, 'contains(sonorra)': False, 'contains(dusting)': False, 'contains(meaning)': False, 'contains(excitement)': False, 'contains(hides)': False, 'contains(mergers)': False, 'contains(bucket)': False, 'contains(chemo)': False, 'contains(fulfillment)': False, 'contains(separation)': False, 'contains(accomplishment)': False, 'contains(offence)': False, 'contains(wealth)': False, 'contains(shocker)': False, 'contains(mazzello)': False, 'contains(kitschiest)': False, 'contains(woman)': True, 'contains(immortal)': False, 'contains(casserole)': False, 'contains(kettle)': False, 'contains(lavatorial)': False, 'contains(confidence)': False, 'contains(apparent)': False, 'contains(nationalist)': False, 'contains(rake)': False, 'contains(patch)': False, 'contains(edgar)': False, 'contains(prioritized)': False, 'contains(premeditated)': False, 'contains(jaunts)': False, 'contains(unspecified)': False, 'contains(slayers)': False, 'contains(kewpies)': False, 'contains(decomposing)': False, 'contains(evidencing)': False, 'contains(scuffle)': False, 'contains(ploys)': False, 'contains(deane)': False, 'contains(drunks)': False, 'contains(problem)': False, 'contains(endemic)': False, 'contains(douglas)': False, 'contains(coloreds)': False, 'contains(hairline)': False, 'contains(actosta)': False, 'contains(stairwells)': False, 'contains(timm)': False, 'contains(galas)': False, 'contains(doctors)': False, 'contains(misappropriating)': False, 'contains(starship)': False, 'contains(spouses)': False, 'contains(traumatic)': False, 'contains(prescripted)': False, 'contains(shwarzenegger)': False, 'contains(tsi)': False, 'contains(snipe)': False, 'contains(fashioning)': False, 'contains(disliking)': False, 'contains(handwaving)': False, 'contains(deed)': False, 'contains(flotsam)': False, 'contains(jimkendrick)': False, 'contains(penn)': False, 'contains(rave)': False, 'contains(loeb)': False, 'contains(whalberg)': False, 'contains(emannuelle)': False, 'contains(operational)': False, 'contains(motivate)': False, 'contains(initation)': False, 'contains(tripp)': False, 'contains(smartass)': False, 'contains(thereus)': False, 'contains(goggins)': False, 'contains(folded)': False, 'contains(humbled)': False, 'contains(direct)': False, 'contains(godess)': False, 'contains(midland)': False, 'contains(verging)': False, 'contains(dreamworld)': False, 'contains(uncharacteristic)': False, 'contains(reptilian)': False, 'contains(unreal)': False, 'contains(fudd)': False, 'contains(uninterrupted)': False, 'contains(exacted)': False, 'contains(awakens)': False, 'contains(complimented)': False, 'contains(indentity)': False, 'contains(hairbrush)': False, 'contains(naoki)': False, 'contains(enthusiastic)': False, 'contains(shard)': False, 'contains(inanely)': False, 'contains(patronize)': False, 'contains(birdy)': False, 'contains(affordable)': False, 'contains(blurts)': False, 'contains(tragically)': False, 'contains(holly)': False, 'contains(understatement)': False, 'contains(tonya)': False, 'contains(narrowed)': False, 'contains(asked)': False, 'contains(bash)': False, 'contains(dancefloor)': False, 'contains(stooped)': False, 'contains(coutroom)': False, 'contains(vaughns)': False, 'contains(ferraris)': False, 'contains(applesauce)': False, 'contains(mesa)': False, 'contains(classifies)': False, 'contains(elspeth)': False, 'contains(pagniacci)': False, 'contains(cate)': False, 'contains(swampy)': False, 'contains(rhythm)': False, 'contains(hollywood)': False, 'contains(remaining)': False, 'contains(godard)': False, 'contains(solemnly)': False, 'contains(kershner)': False, 'contains(organizing)': False, 'contains(siblings)': False, 'contains(arcs)': False, 'contains(runs)': False, 'contains(career)': False, 'contains(izabella)': False, 'contains(mat)': Fal

se, 'contains(errupts)': False, 'contains(corresponding)': False,
'contains(nero)': False, 'contains(depends)': False, 'contains(mal
kovich)': False, 'contains(sanctimonious)': False, 'contains(guita
r)': False, 'contains(girlfriend)': True, 'contains(staged)': Fals
e, 'contains(ezsterhas)': False, 'contains(etch)': False, 'contains
(counterprogramming)': False, 'contains(distrusts)': False, 'contai
ns(encapsulates)': False, 'contains(parodizing)': False, 'contains
(agent)': False, 'contains(raider)': False, 'contains(faintly)': Fa
lse, 'contains(shoah)': False, 'contains(insurrection)': False, 'co
ntains(affair)': False, 'contains(funky)': False, 'contains(final
s)': False, 'contains(herbie)': False, 'contains(grindhouse)': Fals
e, 'contains(oilrig)': False, 'contains(chemisty)': False, 'contain
s(houseboat)': False, 'contains(belinda)': False, 'contains(disarmi
ng)': False, 'contains(forbes)': False, 'contains(gratuitous)': Fal
se, 'contains(carlotta)': False, 'contains(affront)': False, 'conta
ins(rhyitm)': False, 'contains(skarsgard)': False, 'contains(xero
x)': False, 'contains(synonyms)': False, 'contains(emotional)': Fal
se, 'contains(baseman)': False, 'contains(stirred)': False, 'contai
ns(riot)': False, 'contains(shinohara)': False, 'contains(vagina
l)': False, 'contains(manga)': False, 'contains(newfound)': False,
'contains(ramses)': False, 'contains(blinding)': False, 'contains
(occult)': False, 'contains(jester)': False, 'contains(preteen)': F
alse, 'contains(brigette)': False, 'contains(credit)': False, 'cont
ains(mother)': False, 'contains(flood)': False, 'contains(picturesq
ue)': False, 'contains(thrusts)': False, 'contains(positronic)': Fa
lse, 'contains(crush)': False, 'contains(peculiarities)': False, 'c
ontains(recommendation)': False, 'contains(bathroom)': False, 'cont
ains(giles)': False, 'contains(competent)': False, 'contains(presid
ed)': False, 'contains(tenuous)': False, 'contains(schellenberg)':
False, 'contains(lancaster)': False, 'contains(performances)': Fal
se, 'contains(kilmer)': False, 'contains(dah)': False, 'contains(ch
anning)': False, 'contains(conference)': False, 'contains(1847)': F
alse, 'contains(healy)': False, 'contains(usurper)': False, 'contai
ns(cane)': False, 'contains(circles)': False, 'contains(zanin)': Fa
lse, 'contains(fake)': False, 'contains(burying)': False, 'contains
(reserve)': False, 'contains(harmonizing)': False, 'contains(interv
iewers)': False, 'contains(flabby)': False, 'contains(thoughtfulnes
s)': False, 'contains(efects)': False, 'contains(skeletal)': False,
'contains(animator)': False, 'contains(spineless)': False, 'contain
s(bogus)': False, 'contains(gaius)': False, 'contains(timberland)':
False, 'contains(hunk)': False, 'contains(confrontatory)': False,
'contains(meting)': False, 'contains(smacking)': False, 'contains
(footprints)': False, 'contains(lifeforce)': False, 'contains(prefe
cture)': False, 'contains(ambiguities)': False, 'contains(pyrotechn
ic)': False, 'contains(kewpie)': False, 'contains(zookeepers)': Fal
se, 'contains(wetmore)': False, 'contains(pushy)': False, 'contains
(salami)': False, 'contains(journalists)': False, 'contains(respon
d)': False, 'contains(gently)': False, 'contains(upping)': False,
'contains(&#)': False, 'contains(aspirins)': False, 'contains(seda
ted)': False, 'contains(knott)': False, 'contains(enchanted)': Fals
e, 'contains(ship)': False, 'contains(cotterill)': False, 'contains
(lollipop)': False, 'contains(lockers)': False, 'contains(violentl
y)': False, 'contains(turf)': False, 'contains(wookie)': False, 'co
ntains(summarizing)': False, 'contains(menage)': False, 'contains(p
rague)': False, 'contains(delaurentiis)': False, 'contains(insinuat
ions)': False, 'contains(knees)': False, 'contains(goldberg)': Fals
e, 'contains(interlinked)': False, 'contains(lulu)': False, 'contai
ns(dga)': False, 'contains(speed2)': False, 'contains(ackland)': Fa
lse, 'contains(yawner)': False, 'contains(cking)': False, 'contains
(cyanide)': False, 'contains(ballyhooed)': False, 'contains(moral
e)': False, 'contains(wrongdoings)': False, 'contains(rattigan)': F

```
alse, 'contains(remaking)': False, 'contains(pairing)': False, 'con
tains(bonkers)': False, 'contains(duarte)': False, 'contains(facin
g)': False, 'contains(gifford)': False, 'contains(superflous)': Fal
se, 'contains(renfro)': False, 'contains(sing)': False, 'contains(p
rescribed)': False, 'contains(dell)': False, 'contains(hostile)': F
alse, 'contains(contemplating)': False, 'contains(unearths)': Fals
e, 'contains(unbeknownst)': False, 'contains(befall)': False, 'cont
ains(garlic)': False, 'contains(attending)': False, 'contains(mayb
e)': False, 'contains(phillippe)': False, 'contains(complain)': Fal
se, 'contains(nuptials)': False, 'contains(rewatch)': False, 'conta
ins(blamed)': False, 'contains(unexpected)': False, 'contains(breez
e)': False, 'contains(deception)': False, 'contains(stayin)': Fals
e, 'contains(incongruities)': False, 'contains(fictionalized)': Fal
se, 'contains(hermit)': False, 'contains(giveaways)': False, 'conta
ins(hypnotic)': False, 'contains(hathaway)': False, 'contains(uphol
ds)': False, 'contains(rabal)': False, 'contains(ask)': False, 'con
tains(website)': False, 'contains(surrounding)': False, 'contains(c
onstipated)': False, 'contains(retread)': False, 'contains(dietz)':
False, 'contains(stinks)': False, 'contains(ze)': False, 'contains
(waters)': False, 'contains(gooders)': False, 'contains(talkie)': F
alse, 'contains(smeared)': False, 'contains(octopussy)': False, 'co
ntains(beart)': False, 'contains(cohesiveness)': False, 'contains(w
orded)': False, 'contains(gators)': False, 'contains(robo)': False,
'contains(protruding)': False, 'contains(sire)': False, 'contains(m
outhful)': False, 'contains(infection)': False, 'contains(dong)': F
alse, 'contains(brainard)': False, 'contains(lisbon)': False, 'cont
ains(tally)': False, 'contains(plops)': False, 'contains(predilecti
on)': False, 'contains(costumed)': False, 'contains(gershwin)': Fal
se, 'contains(davidovitch)': False, 'contains(bender)': False, 'con
tains(conceivably)': False, 'contains(arrested)': False, 'contains
(offender)': False, 'contains(glee)': False, 'contains(puffing)': F
alse, 'contains(moviereviews)': False, 'contains(stucker)': False,
 'contains(schmaltzfest)': False, 'contains(booked)': False, 'conta
ins(1773)': False, 'contains(skepticism)': False, 'contains(lowe)':
False, 'contains(recipes)': False, 'contains(juliet)': False, 'cont
ains(evidenced)': False, 'contains(tel)': False, 'contains(callist
o)': False, 'contains(mordantly)': False, 'contains(meets)': False,
'contains(niagara)': False, 'contains(dashing)': False, 'contains(s
treetwalker)': False, 'contains(lagpacan)': False, 'contains(pollst
ers)': False, 'contains(snyder)': False, 'contains(inexpensive)': F
alse, 'contains(ideology)': False, 'contains(connects)': False, 'co
ntains(pausing)': False, 'contains(attainable)': False, 'contains(g
len)': False, 'contains(saim)': False, 'contains(kasa)': False, 'co
ntains(permissiveness)': False, 'contains(punchline)': False, 'cont
ains(file)': False, 'contains(desensitization)': False, 'contains(l
eathers)': False, 'contains(slaughtering)': False, 'contains(tucc
i)': False, 'contains(spraying)': False, 'contains(conjunctions)':
 False, 'contains(ore)': False, 'contains(beltrami)': False, 'conta
ins(collectively)': False, 'contains(jemmons)': False, 'contains(ir
is)': False, 'contains(impregnating)': False, 'contains(vaccinate
d)': False, 'contains(ahh)': False, 'contains(evacuation)': False,
 'contains(copulate)': False, 'contains(archetypal)': False, 'conta
ins(scrawny)': False, 'contains(ballstein)': False, 'contains(gas
p)': True, 'contains(heavy)': False, 'contains(plains)': False, 'co
ntains(diametric)': False, 'contains(gehrig)': False, 'contains(bra
veheart)': False, 'contains(smile)': False, 'contains(kristen)': Fa
lse, 'contains(derelict)': False, 'contains(lee)': False, 'contains
(intellectual)': False, 'contains(twitching)': False, 'contains(tas
ha)': False, 'contains(fighting)': False, 'contains(sadler)': Fals
e, 'contains(snoozing)': False, 'contains(pedophile)': False, 'cont
ains(obtuse)': False, 'contains(dirtier)': False, 'contains(excepti
```

ons)': False, 'contains(curiosities)': False, 'contains(schweethaa
t)': False, 'contains(singular)': False, 'contains(deloreans)': Fal
se, 'contains(penniless)': False, 'contains(normandy)': False, 'con
tains(deadlier)': False, 'contains(singers)': False, 'contains(tod
d)': False, 'contains(vore)': False, 'contains(pollster)': False,
'contains(twirling)': False, 'contains(erupt)': False, 'contains(m
ickelberry)': False, 'contains(reminicent)': False, 'contains(geeki
sh)': False, 'contains(envies)': False, 'contains(doodles)': False,
'contains(investigation)': False, 'contains(manage)': False, 'conta
ins(ditches)': False, 'contains(yong)': False, 'contains(lustre)':
False, 'contains(brighter)': False, 'contains(nhu)': False, 'conta
ins(faze)': False, 'contains(extracurricular)': False, 'contains(ex
perience)': False, 'contains(rejoiced)': False, 'contains(preferabl
y)': False, 'contains(distorts)': False, 'contains(remington)': Fal
se, 'contains(rivals)': False, 'contains(films)': False, 'contains
(funnyman)': False, 'contains(restauranteur)': False, 'contains(pro
jectionist)': False, 'contains(goop)': False, 'contains(jena)': Fal
se, 'contains(redneck)': False, 'contains(writes)': False, 'contain
s(depreciating)': False, 'contains(trenchcoat)': False, 'contains(m
ambos)': False, 'contains(science)': False, 'contains(pysche)': Fal
se, 'contains(lowdown)': False, 'contains(occupied)': False, 'conta
ins(crabs)': False, 'contains(philes)': False, 'contains(crushed)':
False, 'contains(congregation)': False, 'contains(proportions)': Fa
lse, 'contains(odin)': False, 'contains(entitlement)': False, 'cont
ains(tarnish)': False, 'contains(crooning)': False, 'contains(abeya
nce)': False, 'contains(traditionally)': False, 'contains(aristocra
tic)': False, 'contains(striding)': False, 'contains(transvestite
s)': False, 'contains(degree)': False, 'contains(forays)': False,
'contains(retrospectives)': False, 'contains(computerized)': Fals
e, 'contains(failure)': False, 'contains(besson)': False, 'contains
(appreciative)': False, 'contains(misquote)': False, 'contains(vaug
hn)': False, 'contains(minbo)': False, 'contains(oversee)': False,
'contains(thermal)': False, 'contains(dip)': False, 'contains(dozi
er)': False, 'contains(freakish)': False, 'contains(screwball)': Fa
lse, 'contains(grrrl)': False, 'contains(beaulieu)': False, 'contai
ns(glean)': False, 'contains(iconography)': False, 'contains(rela
y)': False, 'contains(manic)': False, 'contains(drab)': False, 'con
tains(specifically)': False, 'contains(bigwigs)': False, 'contains
(souza)': False, 'contains(rainbows)': False, 'contains(mommies)':
False, 'contains(flexibility)': False, 'contains(bedrock)': False,
'contains(intermediary)': False, 'contains(warburton)': False, 'con
tains(sequelitis)': False, 'contains(untold)': False, 'contains(isl
amic)': False, 'contains(demean)': False, 'contains(aiming)': Fals
e, 'contains(melaine)': False, 'contains(transylvania)': False, 'co
ntains(contributions)': False, 'contains(skimming)': False, 'contai
ns(thus)': False, 'contains(academia)': False, 'contains(gascogn
e)': False, 'contains(hackers)': False, 'contains(favreau)': False,
'contains(francois)': False, 'contains(darlene)': False, 'contains
(theft)': False, 'contains(allied)': False, 'contains(h2k)': False,
'contains(amass)': False, 'contains(underlining)': False, 'contains
(sticking)': False, 'contains(fourth)': False, 'contains(atrice)':
False, 'contains(abdomen)': False, 'contains(ariane)': False, 'con
tains(massacre)': False, 'contains(avenging)': False, 'contains(mac
ht)': False, 'contains(tokens)': False, 'contains(_disturbing_behav
ior_)': False, 'contains(boggling)': False, 'contains(go)': False,
'contains(revealing)': False, 'contains(wrights)': False, 'contain
s(biting)': False, 'contains(lovlier)': False, 'contains(renfo)': F
alse, 'contains(unwraps)': False, 'contains(deems)': False, 'contai
ns(motivation)': False, 'contains(benben)': False, 'contains(state
s)': False, 'contains(rancorously)': False, 'contains(febre)': Fals
e, 'contains(prude)': False, 'contains(relieves)': False, 'contains

(diluting)': False, 'contains(dredge)': False, 'contains(sickboy)': False, 'contains(reason)': False, 'contains(allah)': False, 'contains(underscored)': False, 'contains(naunton)': False, 'contains(bangkok)': False, 'contains(pattern)': False, 'contains(antisocial)': False, 'contains(advancements)': False, 'contains(thankless)': False, 'contains(commercialist)': False, 'contains(joel)': False, 'contains(pubescent)': False, 'contains(abbott)': False, 'contains(internalizing)': False, 'contains(neckbraces)': False, 'contains(virtuosity)': False, 'contains(pacing)': False, 'contains(consumerism)': False, 'contains(swann)': False, 'contains(cassette)': False, 'contains(polly)': False, 'contains(blood)': False, 'contains(possum)': False, 'contains(enlightenment)': False, 'contains(wagnerian)': False, 'contains(grainy)': False, 'contains(laconically)': False, 'contains(horseshoes)': False, 'contains(fatass)': False, 'contains(glossed)': False, 'contains(wilson)': False, 'contains(kaaya)': False, 'contains(terminating)': False, 'contains(glo)': False, 'contains(walk)': False, 'contains(suing)': False, 'contains(mil)': False, 'contains(god)': False, 'contains(stroll)': False, 'contains(chronicles)': False, 'contains(94)': False, 'contains(fedoras)': False, 'contains(modernized)': False, 'contains(props)': False, 'contains(concerning)': False, 'contains(gunmen)': False, 'contains(labeling)': False, 'contains(batarangs)': False, 'contains(augusts)': False, 'contains(trained)': False, 'contains(miscalculation)': False, 'contains(posits)': False, 'contains(henslowe)': False, 'contains(suppression)': False, 'contains(blueprints)': False, 'contains(attic)': False, 'contains(uninhabited)': False, 'contains(galactica)': False, 'contains(paperbacks)': False, 'contains(quoc)': False, 'contains(miniscule)': False, 'contains(domestically)': False, 'contains(smolder)': False, 'contains(professors)': False, 'contains(easier)': False, 'contains(protective)': False, 'contains(slot)': False, 'contains(scrubbers)': False, 'contains(earthquakes)': False, 'contains(craig)': False, 'contains(characteristically)': False, 'contains(timidity)': False, 'contains(inadequate)': False, 'contains(pitching)': False, 'contains(pets)': False, 'contains(dedication)': False, 'contains(analyzation)': False, 'contains(gone)': False, 'contains(_huge_)': False, 'contains(mistake)': False, 'contains(slots)': False, 'contains(remix)': False, 'contains(vinegar)': False, 'contains(joker)': False, 'contains(mckee)': False, 'contains(doffs)': False, 'contains(hostility)': False, 'contains(_it)': False, 'contains(vcrs)': False, 'contains(precursor)': False, 'contains(boobies)': False, 'contains(infantry)': False, 'contains(episodes)': False, 'contains(macneill)': False, 'contains(chewed)': False, 'contains(permeated)': False, 'contains(vastly)': False, 'contains(jarring)': False, 'contains(stymied)': False, 'contains(ugly)': False, 'contains(korner)': False, 'contains(bela)': False, 'contains(september)': False, 'contains(vamps)': False, 'contains(trade)': False, 'contains(plex)': False, 'contains(lurks)': False, 'contains(kidlets)': False, 'contains(skyjacked)': False, 'contains(reading)': False, 'contains(zookeeper)': False, 'contains(glimcher)': False, 'contains(uprorously)': False, 'contains(perplexed)': False, 'contains(feigns)': False, 'contains(heartbreak)': False, 'contains(untapped)': False, 'contains(emergency)': False, 'contains(dashiell)': False, 'contains(pertinent)': False, 'contains(aladdin)': False, 'contains(sermon)': False, 'contains(fridge)': False, 'contains(veronica)': False, 'contains(gloat)': False, 'contains(geisha)': False, 'contains(murray)': False, 'contains(interchange)': False, 'contains(apprehensively)': False, 'contains(transpired)': False, 'contains(paltry)': False, 'contains(flabbergastingly)': False, 'contains(excitable)': False, 'contains(recreation)': False, 'contains(settles)': False, 'contains(foredoomed)': False, 'contains(textures)': False, 'contains(equate)': False, 'contains(prototype)': False, 'contains(men_)': Fal

```
se, 'contains(stead)': False, 'contains(masturbator)': False, 'cont
ains(orgasms)': False, 'contains(nba)': False, 'contains(trustworth
y)': False, 'contains(decadenet)': False, 'contains(debatable)': Fa
lse, 'contains(raving)': False, 'contains(eflman)': False, 'contain
s(smug)': False, 'contains(taunts)': False, 'contains(gadfly)': Fal
se, 'contains(unspool)': False, 'contains(seann)': False, 'contains
(mailing)': False, 'contains(faulted)': False, 'contains(gather)':
 False, 'contains(rubbish)': False, 'contains(heartland)': False,
 'contains(escorted)': False, 'contains(insistance)': False, 'conta
ins(gazillionaire)': False, 'contains(pual)': False, 'contains(evel
yn)': False, 'contains(containers)': False, 'contains(ragtag)': Fal
se, 'contains(shannon)': False, 'contains(schmeer)': False, 'contai
ns(pinon)': False, 'contains(levine)': False, 'contains(darryl)': F
alse, 'contains(invaders)': False, 'contains(institutionalized)': F
alse, 'contains(bodyguards)': False, 'contains(cousine)': False, 'c
ontains(254)': False, 'contains(permanant)': False, 'contains(redev
eloped)': False, 'contains(adaptable)': False, 'contains(ditz)': Fa
lse, 'contains(nephew)': False, 'contains(enlightening)': False, 'c
ontains(propose)': False, 'contains(indignity)': False, 'contains(s
todge)': False, 'contains(burkettsville)': False, 'contains(charect
er)': False, 'contains(hideousness)': False, 'contains(dishonor)':
 False, 'contains(flack)': False, 'contains(waging)': False, 'conta
ins(mandated)': False, 'contains(encased)': False, 'contains(blosso
ming)': False, 'contains(tsunami)': False, 'contains(loathed)': Fal
se, 'contains(mute)': False, 'contains(byplay)': False, 'contains(h
olgado)': False, 'contains(mon)': False, 'contains(membership)': Fa
lse, 'contains(ruth)': False, 'contains(persuade)': False, 'contain
s(terrorize)': False, 'contains(realize)': False, 'contains(gilliga
n)': False, 'contains(languishing)': False, 'contains(partition)':
 False, 'contains(motormouth)': False, 'contains(insight)': False,
 'contains(prevue)': False, 'contains(mail)': False, 'contains(mish
andled)': False, 'contains(powered)': False, 'contains(interminglin
g)': False, 'contains(cgi)': False, 'contains(dellaplane)': False,
 'contains(lrighty)': False, 'contains(porcelain)': False, 'contain
s(ultimo)': False, 'contains(hateful)': False, 'contains(bridget)':
False, 'contains(1985)': False, 'contains(overjoyed)': False, 'cont
ains(grouper)': False, 'contains(asswhole)': False, 'contains(overl
ight)': False, 'contains(mesmerize)': False, 'contains(swordsman)':
False, 'contains(loophole)': False, 'contains(weil)': False, 'conta
ins(nationals)': False, 'contains(yuji)': False, 'contains(muddie
s)': False, 'contains(compulsive)': False, 'contains(uh)': False,
 'contains(installs)': False, 'contains(nullified)': False, 'contai
ns(brewing)': False, 'contains(prologue)': False, 'contains(droppe
d)': False, 'contains(guitarist)': False, 'contains(divorcee)': Fal
se, 'contains(rang)': False, 'contains(1990s)': False, 'contains(do
pey)': False, 'contains(frumpy)': False, 'contains(dizzy)': False,
 'contains(stint)': False, 'contains(furrowed)': False, 'contains(t
oll)': False, 'contains(weighty)': False, 'contains(insulin)': Fals
e, 'contains(plow)': False, 'contains(junichiro)': False, 'contains
(knowingly)': False, 'contains(amyls)': False, 'contains(curt)': Fa
lse, 'contains(commoner)': False, 'contains(dishes)': False, 'conta
ins(fortresses)': False, 'contains(dabbling)': False, 'contains(ove
rachieve)': False, 'contains(hallways)': False, 'contains(cycle)':
 False, 'contains(wildly)': False, 'contains(taxes)': False, 'conta
ins(patrolman)': False, 'contains(underworld)': False, 'contains(ne
wsgroup)': False, 'contains(similar)': False, 'contains(garfield)':
False, 'contains(grimmest)': False, 'contains(ads)': False, 'contai
ns(sites)': False, 'contains(dispair)': False, 'contains(designer
s)': False, 'contains(arangements)': False, 'contains(leguizimo)':
 False, 'contains(23rd)': False, 'contains(chord)': False, 'contain
s(disappointed)': False, 'contains(part)': False, 'contains(entrail
```

```
s)': False, 'contains(refrains)': False, 'contains(stephane)': Fals
e, 'contains(perry)': False, 'contains(homages)': False, 'contains
(hurts)': False, 'contains(entrusting)': False, 'contains(autumna
l)': False, 'contains(approachment)': False, 'contains(pans)': Fals
e, 'contains(bones)': False, 'contains(paquin)': False, 'contains(s
niggering)': False, 'contains(whereabouts)': False, 'contains(disce
rning)': False, 'contains(gilsig)': False, 'contains(degeneratio
n)': False, 'contains(tricks)': False, 'contains(worms)': False, 'c
ontains(touchstone)': False, 'contains(whut)': False, 'contains(ret
ro)': False, 'contains(donning)': False, 'contains(flabbergasted)':
False, 'contains(ripoff)': False, 'contains(forgiveness)': False,
 'contains(skies)': False, 'contains(irreversible)': False, 'contai
ns(yon)': False, 'contains(differs)': False, 'contains(shandling)':
False, 'contains(cates)': False, 'contains(tm)': False, 'contains(p
otboilers)': False, 'contains(gulp)': False, 'contains(nerd)': Fals
e, 'contains(stivaletti)': False, 'contains(giacomo)': False, 'cont
ains(stempel)': False, 'contains(therese)': False, 'contains(walke
r)': False, 'contains(anabella)': False, 'contains(milo)': False,
 'contains(nada)': False, 'contains(cuddy)': False, 'contains(fift
h)': False, 'contains(planned)': False, 'contains(mm)': False, 'con
tains(outfit)': False, 'contains(omnipresence)': False, 'contains(h
utton)': False, 'contains(weaving)': False, 'contains(wore)': Fals
e, 'contains(humans)': False, 'contains(vaudeville)': False, 'conta
ins(honorary)': False, 'contains(wai)': False, 'contains(traffic)':
False, 'contains(tiering)': False, 'contains(treachery)': False, 'c
ontains(ang)': False, 'contains(hijacker)': False, 'contains(certia
nly)': False, 'contains(whammy)': False, 'contains(rebs)': False,
 'contains(prevalence)': False, 'contains(moviestar)': False, 'cont
ains(oscar)': False, 'contains(san)': False, 'contains(shadyac)': F
alse, 'contains(grunts)': False, 'contains(calahan)': False, 'conta
ins(utmost)': False, 'contains(bombed)': False, 'contains(hagert
y)': False, 'contains(bare)': False, 'contains(assists)': False, 'c
ontains(fleder)': False, 'contains(reform)': False, 'contains(perpe
tually)': False, 'contains(bopper)': False, 'contains(roselyn)': Fa
lse, 'contains(amadeus)': False, 'contains(monroth)': False, 'conta
ins(sabara)': False, 'contains(sayinig)': False, 'contains(frighten
s)': False, 'contains(60s)': False, 'contains(menno)': False, 'cont
ains(h20)': False, 'contains(bishop)': False, 'contains(pumped)': F
alse, 'contains(impressiveness)': False, 'contains(bart)': False,
 'contains(boosts)': False, 'contains(keenan)': False, 'contains(do
cu)': False, 'contains(pit)': False, 'contains(merrill)': False, 'c
ontains(store)': False, 'contains(captor)': False, 'contains(claus
e)': False, 'contains(triumphing)': False, 'contains(terseness)': F
alse, 'contains(loyality)': False, 'contains(abounds)': False, 'con
tains(claustrophobia)': False, 'contains(hanako)': False, 'contains
(provides)': False, 'contains(defuse)': False, 'contains(peeled)':
 False, 'contains(intruding)': False, 'contains(_mafia_)': False,
 'contains(strengthening)': False, 'contains(donnie)': False, 'cont
ains(mcmillan)': False, 'contains(vigilante)': False, 'contains(pic
king)': False, 'contains(tiring)': False, 'contains(mediciney)': Fa
lse, 'contains(sargent)': False, 'contains(windon)': False, 'contai
ns(shone)': False, 'contains(helming)': False, 'contains(peep)': Fa
lse, 'contains(fraternizing)': False, 'contains(transexual)': Fals
e, 'contains(tykes)': False, 'contains(lowell)': False, 'contains(d
itching)': False, 'contains(armstrong)': False, 'contains(enlighten
s)': False, 'contains(spitting)': False, 'contains(prehensile)': Fa
lse, 'contains(pictorial)': False, 'contains(rehabilitation)': Fals
e, 'contains(page)': False, 'contains(robotically)': False, 'contai
ns(accumulate)': False, 'contains(aniston)': False, 'contains(sharp
ly)': False, 'contains(appendages)': False, 'contains(wen)': False,
'contains(patiently)': False, 'contains(slicing)': False, 'contains
```

```
(commiserate)': False, 'contains(whirl)': False, 'contains(tempe
r)': False, 'contains(unsettling)': False, 'contains(ursula)': Fals
e, 'contains(lopsided)': False, 'contains(overtakes)': False, 'cont
ains(slacker)': False, 'contains(glutton)': False, 'contains(foreme
ntioned)': False, 'contains(lightness)': False, 'contains(bogg)': F
alse, 'contains(reprograms)': False, 'contains(splicing)': False,
 'contains(moranis)': False, 'contains(lautrec)': False, 'contains
(toughs)': False, 'contains(bridal)': False, 'contains(151)': Fals
e, 'contains(emotion)': False, 'contains(dumb)': False, 'contains(s
ears)': False, 'contains(blips)': False, 'contains(mongolian)': Fal
se, 'contains(lasers)': False, 'contains(goode)': False, 'contains
(172)': False, 'contains(solos)': False, 'contains(expletitive)': F
alse, 'contains(increased)': False, 'contains(pharaoh)': False, 'co
ntains(lukewarm)': False, 'contains(curing)': False, 'contains(disc
overer)': False, 'contains(18s)': False, 'contains(shoot)': False,
 'contains(superlative)': False, 'contains(escapism)': False, 'cont
ains(barbara)': False, 'contains(unparalled)': False, 'contains(195
0s)': False, 'contains(affiliations)': False, 'contains(fulfillin
g)': False, 'contains(delightfully)': False, 'contains(buzzcocks)':
False, 'contains(carpenters)': False, 'contains(1521)': False, 'con
tains(cozart)': False, 'contains(connundrum)': False, 'contains(cre
ated)': False, 'contains(maimed)': False, 'contains(stolz)': False,
'contains(craftsmanship)': False, 'contains(cough)': False, 'contai
ns(scorsese)': False, 'contains(consultation)': False, 'contains(ab
etting)': False, 'contains(landscapes)': False, 'contains(clockwatc
hers)': False, 'contains(stank)': False, 'contains(palate)': False,
'contains(repertoire)': False, 'contains(foulmouthed)': False, 'con
tains(collars)': False, 'contains(gazing)': False, 'contains(emitte
d)': False, 'contains(decreed)': False, 'contains(carried)': False,
'contains(poon)': False, 'contains(thrash)': False, 'contains(innoc
uously)': False, 'contains(kissed)': False, 'contains(cynically)':
 False, 'contains(adherents)': False, 'contains(laurene)': False,
 'contains(ripley)': False, 'contains(undirected)': False, 'contain
s(deguerin)': False, 'contains(conventional)': False, 'contains(ed
u)': False, 'contains(elderly)': False, 'contains(adjustment)': Fal
se, 'contains(ingnorant)': False, 'contains(trough)': False, 'conta
ins(puppets)': False, 'contains(arc)': False, 'contains(essentiall
y)': False, 'contains(kafka)': False, 'contains(wackier)': False,
 'contains(goldsmith)': False, 'contains(crewman)': False, 'contain
s(nouri)': False, 'contains(rehabilitate)': False, 'contains(arizon
a)': False, 'contains(flaquer)': False, 'contains(filmgoers)': Fals
e, 'contains(schreiber)': False, 'contains(crucifixtion)': False,
 'contains(hipness)': False, 'contains(flyboy)': False, 'contains(s
alt)': False, 'contains(xtdl)': False, 'contains(secure)': False,
 'contains(calvinist)': False, 'contains(photograph)': False, 'cont
ains(webcams)': False, 'contains(revisiting)': False, 'contains(bur
khart)': False, 'contains(flashcards)': False, 'contains(ominousl
y)': False, 'contains(chaired)': False, 'contains(outgrown)': Fals
e, 'contains(machette)': False, 'contains(restrain)': False, 'conta
ins(screenplay)': False, 'contains(dodger)': False, 'contains(wring
ing)': False, 'contains(constrained)': False, 'contains(typist)': F
alse, 'contains(earthward)': False, 'contains(preposterousness)': F
alse, 'contains(bestselling)': False, 'contains(sneeze)': False, 'c
ontains(screen)': True, 'contains(plan)': False, 'contains(opus)':
 False, 'contains(scaffolding)': False, 'contains(evangelical)': Fa
lse, 'contains(gruelling)': False, 'contains(deadline)': False, 'co
ntains(twosomes)': False, 'contains(decipher)': False, 'contains(un
affected)': False, 'contains(okeday)': False, 'contains(productiv
e)': False, 'contains(ricky)': False, 'contains(danced)': False, 'c
ontains(snowfall)': False, 'contains(skarsgaard)': False, 'contains
(leftover)': False, 'contains(fibber)': False, 'contains(pagan)': F
```

```
alse, 'contains(distributors)': False, 'contains(vantages)': False,
'contains(religios)': False, 'contains(nyah)': False, 'contains(ber
tolini)': False, 'contains(rectangles)': False, 'contains(fluoro)':
False, 'contains(kathy)': False, 'contains(})': False, 'contains(mu
ltifaceted)': False, 'contains(hunts)': False, 'contains(guccion
e)': False, 'contains(yesssssss)': False, 'contains(expanded)': Fals
e, 'contains(remembers)': False, 'contains(meteorite)': False, 'con
tains(translation)': False, 'contains(copacabana)': False, 'contain
s(swoops)': False, 'contains(outweigh)': False, 'contains(dreams)':
False, 'contains(confusingly)': False, 'contains(scrolled)': False,
'contains(rube)': False, 'contains(deserved)': False, 'contains(ela
stica)': False, 'contains(tape)': False, 'contains(knowledgeable)':
False, 'contains(basil)': False, 'contains(fracture)': False, 'cont
ains(puts)': False, 'contains(overdose)': False, 'contains(overemph
asizing)': False, 'contains(16th)': False, 'contains(overpopulate
d)': False, 'contains(standup)': False, 'contains(analysis)': Fals
e, 'contains(hitting)': False, 'contains(bologna)': False, 'contain
s(wheeler)': False, 'contains(literacy)': False, 'contains(ki)': Fa
lse, 'contains(hypocritical)': False, 'contains(vaults)': False, 'c
ontains(knights)': False, 'contains(fleeting)': False, 'contains(sw
at)': False, 'contains(apprenticeship)': False, 'contains(exam)': F
alse, 'contains(wattage)': False, 'contains(declan)': False, 'conta
ins(rejuvenated)': False, 'contains(respect)': False, 'contains(pre
ening)': False, 'contains(faired)': False, 'contains(hecklers)': Fa
lse, 'contains(overview)': False, 'contains(shave)': False, 'contai
ns(bujold)': False, 'contains(jam)': False, 'contains(betcha)': Fal
se, 'contains(toughened)': False, 'contains(zeffirelli)': False, 'c
ontains(gratuities)': False, 'contains(migraines)': False, 'contain
s(clash)': False, 'contains(rampage)': False, 'contains(brought)':
 False, 'contains(dunn)': False, 'contains(connoisseurs)': False,
 'contains(donned)': False, 'contains(amnesiac)': False, 'contains
(misrepresenting)': False, 'contains(sequoia)': False, 'contains(st
accatto)': False, 'contains(homcoming)': False, 'contains(fully)':
 False, 'contains(dwayne)': False, 'contains(alteration)': False,
 'contains(uncorking)': False, 'contains(maps)': False, 'contains(w
etter)': False, 'contains(imported)': False, 'contains(tossable)':
 False, 'contains(merhi)': False, 'contains(interview)': False, 'co
ntains(jefferson)': False, 'contains(ellie)': False, 'contains(kiss
ing)': False, 'contains(duties)': False, 'contains(fassbinder)': Fa
lse, 'contains(hops)': False, 'contains(imprisoning)': False, 'cont
ains(hatable)': False, 'contains(spit)': False, 'contains(vices)':
 False, 'contains(ancestors)': False, 'contains(demi)': False, 'con
tains(emilio)': False, 'contains(mayoral)': False, 'contains(640)':
False, 'contains(meowth)': False, 'contains(hire)': False, 'contain
s(touring)': False, 'contains(interiors)': False, 'contains(submi
t)': False, 'contains(woodsboro)': False, 'contains(consists)': Fal
se, 'contains(tatooine)': False, 'contains(someone)': False, 'conta
ins(blossom)': False, 'contains(lacking)': False, 'contains(frill
s)': False, 'contains(loafer)': False, 'contains(sinclair)': False,
'contains(unsuspenseful)': False, 'contains(hear)': False, 'contain
s(predictably)': False, 'contains(teaspoons)': False, 'contains(not
ifying)': False, 'contains(merciful)': False, 'contains(sassines
s)': False, 'contains(climactic)': False, 'contains(home)': False,
 'contains(protgaonist)': False, 'contains(decently)': False, 'cont
ains(gaby)': False, 'contains(revolutionized)': False, 'contains(me
ister)': False, 'contains(detection)': False, 'contains(rein)': Fal
se, 'contains(bingo)': False, 'contains(curmudgeons)': False, 'cont
ains(tantric)': False, 'contains(creepers)': False, 'contains(brani
ff)': False, 'contains(payback)': False, 'contains(invests)': Fals
e, 'contains(wrenched)': False, 'contains(layout)': False, 'contain
s(und)': False, 'contains(-`)': False, 'contains(culls)': False, 'c
```

ontains(whooshed)': False, 'contains(field)': True, 'contains(sige
l)': False, 'contains(applicants)': False, 'contains(transistor)':
False, 'contains(cattle)': False, 'contains(flick)': False, 'conta
ins(dustbuster)': False, 'contains(extracted)': False, 'contains(ab
ba)': False, 'contains(horrifying)': False, 'contains(scriptiing)':
False, 'contains(lyon)': False, 'contains(sociology)': False, 'cont
ains(koch)': False, 'contains(pidgeonhole)': False, 'contains(origi
nated)': False, 'contains(stereotypical)': False, 'contains(safel
y)': False, 'contains(sharks)': False, 'contains(reveal)': False,
'contains(reprehensible)': False, 'contains(logistical)': False,
'contains(felix)': False, 'contains(fond)': False, 'contains(dream
quest)': False, 'contains(assuredness)': False, 'contains(boaz)': F
alse, 'contains(hoggett)': False, 'contains(storm)': False, 'contai
ns(brute)': False, 'contains(rochelle)': False, 'contains(daredevil
s)': False, 'contains(survey)': False, 'contains(tagalong)': False,
'contains(exacting)': False, 'contains(appearing)': False, 'contain
s(quake)': False, 'contains(steering)': False, 'contains(farting)':
False, 'contains(muff)': False, 'contains(dum)': False, 'contains(d
ewey)': False, 'contains(min)': False, 'contains(fatalities)': Fals
e, 'contains(macabre)': False, 'contains(courage)': False, 'contain
s(allayah)': False, 'contains(parenthetically)': False, 'contains(f
acade)': False, 'contains(masseur)': False, 'contains(emotionles
s)': False, 'contains(actor)': False, 'contains(majesty)': False,
'contains(alienated)': False, 'contains(confided)': False, 'contai
ns(active)': False, 'contains(shiftlessness)': False, 'contains(win
ningham)': False, 'contains(007)': False, 'contains(weight)': Fals
e, 'contains(undermined)': False, 'contains(rap)': False, 'contains
(ply)': False, 'contains(potente)': False, 'contains(attend)': Fals
e, 'contains(foreman)': False, 'contains(advisor)': False, 'contain
s(vulcan)': False, 'contains(exploded)': False, 'contains(planne
r)': False, 'contains(atreus)': False, 'contains(vitriol)': False,
'contains(preparations)': False, 'contains(neater)': False, 'conta
ins(mannered)': False, 'contains(sculpting)': False, 'contains(perm
anetly)': False, 'contains(gain)': False, 'contains(presnell)': Fal
se, 'contains(copping)': True, 'contains(marley)': False, 'contains
(coca)': False, 'contains(unzipped)': False, 'contains(nighttime)':
False, 'contains(vibrates)': False, 'contains(drink)': False, 'cont
ains(rubbery)': False, 'contains(ethos)': False, 'contains(jovia
l)': False, 'contains(major)': False, 'contains(footloose)': False,
'contains(input)': False, 'contains(bondian)': False, 'contains(kre
mp)': False, 'contains(ferocious)': False, 'contains(assets)': Fals
e, 'contains(sicker)': False, 'contains(francine)': False, 'contain
s(paleontology)': False, 'contains(noraruth)': False, 'contains(sa
b)': False, 'contains(depite)': False, 'contains(randomly)': False,
'contains(wisely)': False, 'contains(inroad)': False, 'contains(lax
atives)': False, 'contains(hydraulic)': False, 'contains(directio
n)': False, 'contains(soul)': False, 'contains(distance)': False,
'contains(relentlessly)': False, 'contains(gq)': False, 'contains
(headfirst)': False, 'contains(procol)': False, 'contains(&)': Fals
e, 'contains(neverland)': False, 'contains(smirk)': False, 'contain
s(ex)': False, 'contains(misogyny)': False, 'contains(seague)': Fal
se, 'contains(excerpts)': False, 'contains(fine)': False, 'contains
(steelworks)': False, 'contains(donation)': False, 'contains(munr
o)': False, 'contains(glamorising)': False, 'contains(washout)': Fa
lse, 'contains(vichy)': False, 'contains(lend)': False, 'contains(e
xecuting)': False, 'contains(thinks)': False, 'contains(whimsica
l)': False, 'contains(garfunkel)': False, 'contains(subtler)': Fals
e, 'contains(onlookers)': False, 'contains(submerged)': False, 'con
tains(sprocket)': False, 'contains(hardy)': False, 'contains(shell
s)': False, 'contains(snuggling)': False, 'contains(_monster_movie
_)': False, 'contains(lemay)': False, 'contains(eloquent)': False,

'contains(burlesque)': False, 'contains(eruting)': False, 'contain
s(articulately)': False, 'contains(horseshit)': False, 'contains(ca
lvin)': False, 'contains(elliott)': False, 'contains(elise)': Fals
e, 'contains(interaction)': False, 'contains(dostoevski)': False,
'contains(snub)': False, 'contains(tunes)': False, 'contains(pranc
ing)': False, 'contains(1995)': False, 'contains(ultraconservativ
e)': False, 'contains(feat)': False, 'contains(indiglo)': False, 'c
ontains(concocts)': False, 'contains(figaro)': False, 'contains(dal
lben)': False, 'contains(lubezki)': False, 'contains(beaker)': Fals
e, 'contains(bribe)': False, 'contains(confusions)': False, 'contai
ns(joplin)': False, 'contains(miscarriage)': False, 'contains(dope
d)': False, 'contains(contented)': False, 'contains(characteristic
s)': False, 'contains(cremer)': False, 'contains(wyatt)': False, 'c
ontains(turrets)': False, 'contains(timeline)': False, 'contains(as
siduously)': False, 'contains(publicist)': False, 'contains(mauna
u)': False, 'contains(keyzer)': False, 'contains(identifiable)': Fa
lse, 'contains(parallel)': False, 'contains(slathers)': False, 'con
tains(garber)': False, 'contains(arsed)': False, 'contains(reinvent
ions)': False, 'contains(malle)': False, 'contains(punished)': Fals
e, 'contains(dignities)': False, 'contains(chili)': False, 'contain
s(stealth)': False, 'contains(pages)': False, 'contains(feelgood)':
False, 'contains(couteney)': False, 'contains(straightest)': False,
'contains(exonerated)': False, 'contains(egos)': False, 'contains(w
itnessing)': False, 'contains(bragging)': False, 'contains(toes)':
False, 'contains(fantasizing)': False, 'contains(chung)': False,
'contains(asleep)': False, 'contains(thierry)': False, 'contains(c
onglomerates)': False, 'contains(gravedigging)': False, 'contains(o
utlet)': False, 'contains(sheryl)': False, 'contains(remakes)': Fal
se, 'contains(partment)': False, 'contains(jacqeline)': False, 'con
tains(stupefied)': False, 'contains(event)': False, 'contains(terre
strials)': False, 'contains(jia)': False, 'contains(bothersome)': F
alse, 'contains(serial)': False, 'contains(twistet)': False, 'conta
ins(richman)': False, 'contains(abundence)': False, 'contains(intui
ts)': False, 'contains(subtitled)': False, 'contains(reach)': Fals
e, 'contains(irvin)': False, 'contains(boorman)': False, 'contains
(misconceptions)': False, 'contains(sheepish)': False, 'contains(cl
eaners)': False, 'contains(loans)': False, 'contains(replete)': Fal
se, 'contains(arqua)': False, 'contains(laconic)': False, 'contains
(penitentiary)': False, 'contains(proposing)': False, 'contains(thi
ng__not)': False, 'contains(zinnia)': False, 'contains(primo)': Fal
se, 'contains(swigert)': False, 'contains(breezy)': False, 'contain
s(moore)': False, 'contains(disoriented)': False, 'contains(charliz
e)': False, 'contains(related)': True, 'contains(uplift)': False,
'contains(adaptions)': False, 'contains(maxim)': False, 'contains
(wresting)': False, 'contains(headlights)': False, 'contains(accoun
ts)': False, 'contains(stretch)': False, 'contains(malibu)': False,
'contains(puppeteering)': False, 'contains(heed)': False, 'contains
(filmming)': False, 'contains(robs)': False, 'contains(rhett)': Fal
se, 'contains(kleiser)': False, 'contains(dreads)': False, 'contain
s(fujioka)': False, 'contains(offended)': False, 'contains(threaten
s)': False, 'contains(boyishly)': False, 'contains(paull)': False,
'contains(dolenz)': False, 'contains(supplying)': False, 'contains
(intolerable)': False, 'contains(nick)': False, 'contains(target)':
False, 'contains(wanderlust)': False, 'contains(drix)': False, 'con
tains(sweden)': False, 'contains(seriously)': False, 'contains(rese
rves)': False, 'contains(counterparts)': False, 'contains(ratted)':
False, 'contains(seng)': False, 'contains(performance)': False, 'co
ntains(tournaments)': False, 'contains(brutish)': False, 'contains
(nominee)': False, 'contains(boatman)': False, 'contains(5000)': Fa
lse, 'contains(penetrate)': False, 'contains(skaro)': False, 'conta
ins(sprayed)': False, 'contains(bresson)': False, 'contains(munchi

```
e)': False, 'contains(thanksgiving)': False, 'contains(horseman)':
 False, 'contains(immeadiately)': False, 'contains(butchers)': Fals
e, 'contains(forty)': False, 'contains(professionalism)': False, 'c
ontains(nighthawks)': False, 'contains(unfetching)': False, 'contai
ns(accelerate)': False, 'contains(icebergs)': False, 'contains(tiar
a)': False, 'contains(spurs)': False, 'contains(nutter)': False, 'c
ontains(scarlett)': False, 'contains(promote)': False, 'contains(da
per)': False, 'contains(hirschfeld)': False, 'contains(purging)': F
alse, 'contains(workout)': False, 'contains(mccracken)': False, 'co
ntains(vertigo)': False, 'contains(examination)': False, 'contains
(graciously)': False, 'contains(stoker)': False, 'contains(daryl)':
False, 'contains(potatohead)': False, 'contains(burglary)': False,
 'contains(piven)': False, 'contains(apart)': False, 'contains(cann
ery)': False, 'contains(rubell)': False, 'contains(apologeticall
y)': False, 'contains(restless)': False, 'contains(bogeyman)': Fals
e, 'contains(derek)': False, 'contains(sequels)': False, 'contains
(turgidson)': False, 'contains(bleaches)': False, 'contains(than)':
False, 'contains(sketching)': False, 'contains(witty)': False, 'con
tains(brauvara)': False, 'contains(_ferris)': False, 'contains(akik
o)': False, 'contains(jing)': False, 'contains(autism)': False, 'co
ntains(goodly)': False, 'contains(soprana)': False, 'contains(publi
city)': False, 'contains(whoopee)': False, 'contains(expect)': Fals
e, 'contains(bistro)': False, 'contains(avalon)': False, 'contains
(_breakfast_)': False, 'contains(necesary)': False, 'contains(feebl
y)': False, 'contains(buttocks)': False, 'contains(farewell)': Fals
e, 'contains(hymn)': False, 'contains(goes)': False, 'contains(mixe
r)': False, 'contains(approve)': False, 'contains(mcdowell)': Fals
e, 'contains(animted)': False, 'contains(duane)': False, 'contains
(natacha)': False, 'contains(seattle)': False, 'contains(brinkfor
d)': False, 'contains(flutters)': False, 'contains(til)': False, 'c
ontains(trick)': False, 'contains(longshot)': False, 'contains(fors
yth)': False, 'contains(martyrs)': False, 'contains(etchings)': Fal
se, 'contains(cunning)': False, 'contains(impossibilities)': False,
'contains(duck)': True, 'contains(mild)': False, 'contains(ultimatu
m)': False, 'contains(gleaming)': False, 'contains(resigning)': Fal
se, 'contains(celebrity)': False, 'contains(vowed)': False, 'contai
ns(evoking)': False, 'contains(recommending)': False, 'contains(vio
lated)': False, 'contains(requirement)': False, 'contains(sicknes
s)': False, 'contains(response)': False, 'contains(slayer)': False,
'contains(bullet)': False, 'contains(ousting)': False, 'contains(wi
shes)': False, 'contains(blunts)': False, 'contains(svelte)': Fals
e, 'contains(blip)': False, 'contains(fertile)': False, 'contains(i
mmuno)': False, 'contains(bawls)': False, 'contains(clarisse)': Fal
se, 'contains(mayersberg)': False, 'contains(1996)': False, 'contai
ns(feasibility)': False, 'contains(medgar)': False, 'contains(coope
ration)': False, 'contains(persuasively)': False, 'contains(regai
n)': False, 'contains(issues)': False, 'contains(immediate)': Fals
e, 'contains(litmus)': False, 'contains(highschool)': False, 'conta
ins(rangoon)': False, 'contains(unpardonable)': False, 'contains(19
88)': False, 'contains(roughly)': False, 'contains(unengaging)': Fa
lse, 'contains(artsier)': False, 'contains(gregarious)': False, 'co
ntains(owl)': False, 'contains(rutger)': False, 'contains(pirate)':
False, 'contains(kinfolk)': False, 'contains(khachaturian)': False,
'contains(conformist)': False, 'contains(hooligans)': False, 'conta
ins(seaside)': False, 'contains(mutiny)': False, 'contains(smother
s)': False, 'contains(wager)': False, 'contains(obscenities)': Fals
e, 'contains(navigator)': False, 'contains(sponge)': False, 'contai
ns(koop)': False, 'contains(pilgrims)': False, 'contains(walters)':
False, 'contains(looker)': False, 'contains(2013)': False, 'contain
s(conspiracy)': False, 'contains(agutter)': False, 'contains(courti
ng)': False, 'contains(mafioso)': False, 'contains(canoeing)': Fals
```

```
e, 'contains(hubley)': False, 'contains(calhoun)': False, 'contains
(jinn)': False, 'contains(benches)': False, 'contains(vapors)': Fal
se, 'contains(coyle)': False, 'contains(devious)': False, 'contains
(opportunities)': False, 'contains(photogenic)': False, 'contains(h
oary)': False, 'contains(sparing)': False, 'contains(kazantzakis)':
False, 'contains(emotions)': False, 'contains(chipmunk)': False, 'c
ontains(unsworth)': False, 'contains(grapevine)': False, 'contains
(stamp)': False, 'contains(sadist)': False, 'contains(disgruntle
d)': False, 'contains(khe)': False, 'contains(acknowledges)': Fals
e, 'contains(tempers)': False, 'contains(depended)': False, 'contai
ns(revving)': False, 'contains(released)': False, 'contains(telekin
esis)': False, 'contains(makeover)': False, 'contains(drain)': Fals
e, 'contains(cruel)': False, 'contains(glamorized)': False, 'contai
ns(amistad)': False, 'contains(circa)': False, 'contains(cheesines
s)': False, 'contains(homemade)': False, 'contains(enterprising)':
 False, 'contains(cuddly)': False, 'contains(hazel)': False, 'conta
ins(_murder_)': False, 'contains(rothchild)': False, 'contains(spic
er)': False, 'contains(obstructed)': False, 'contains(jost)': Fals
e, 'contains(guessed)': False, 'contains(cultured)': False, 'contai
ns(towards)': False, 'contains(clashed)': False, 'contains(ent)': F
alse, 'contains(politcal)': False, 'contains(gobbling)': False, 'co
ntains(had)': False, 'contains(struts)': False, 'contains(tamala)':
False, 'contains(johner)': False, 'contains(irresponsible)': False,
'contains(shooters)': False, 'contains(subordination)': False, 'con
tains(war)': False, 'contains(orangutan)': False, 'contains(polariz
ation)': False, 'contains(remark)': False, 'contains(alloy)': Fals
e, 'contains(distracted)': False, 'contains(figs)': False, 'contain
s(confounds)': False, 'contains(tentative)': False, 'contains(yann
i)': False, 'contains(descending)': False, 'contains(quirkiness)':
 False, 'contains(quartet)': False, 'contains(supervising)': False,
'contains(request)': False, 'contains(paint)': False, 'contains(cli
ffs)': False, 'contains(sacrifices)': False, 'contains(gunning)': F
alse, 'contains(anand)': False, 'contains(stifling)': False, 'conta
ins(risks)': False, 'contains(spinoff)': False, 'contains(guinee)':
False, 'contains(spiritualistic)': False, 'contains(obligation)': F
alse, 'contains(drew)': False, 'contains(37th)': False, 'contains(f
ungus)': False, 'contains(strikes)': False, 'contains(bobo)': Fals
e, 'contains(declaration)': False, 'contains(topic)': False, 'conta
ins(rin)': False, 'contains(push)': False, 'contains(unwise)': Fals
e, 'contains(heading)': False, 'contains(loquacious)': False, 'cont
ains(markings)': False, 'contains(adlai)': False, 'contains(lela)':
False, 'contains(corresponded)': False, 'contains(1982)': False, 'c
ontains(pedlar)': False, 'contains(sexualized)': False, 'contains(g
ordy)': False, 'contains(pauses)': False, 'contains(balaban)': Fals
e, 'contains(overbearingly)': False, 'contains(fiercer)': False, 'c
ontains(pare)': False, 'contains(articulated)': False, 'contains(de
legates)': False, 'contains(stitch)': False, 'contains(bhatnagar)':
False, 'contains(labors)': False, 'contains(coz)': False, 'contains
(harlin)': False, 'contains(neglecting)': False, 'contains(pesud
o)': False, 'contains(stresses)': False, 'contains(waisted)': Fals
e, 'contains(hurls)': False, 'contains(cellophane)': False, 'contai
ns(fairfax)': False, 'contains(phew)': False, 'contains(rhod)': Fal
se, 'contains(bikers)': False, 'contains(mirren)': False, 'contains
(cling)': False, 'contains(anakins)': False, 'contains(appointed)':
False, 'contains(toiling)': False, 'contains(harvie)': False, 'cont
ains(elegy)': False, 'contains(seething)': False, 'contains(avenu
e)': False, 'contains(desktop)': False, 'contains(climaxing)': Fals
e, 'contains(schuemacher)': False, 'contains(greg)': False, 'contai
ns(letup)': False, 'contains(coffee)': False, 'contains(corral)': F
alse, 'contains(brew)': False, 'contains(brevity)': False, 'contain
s(follows)': False, 'contains(avital)': False, 'contains(thawed)':
```

False, 'contains(lindberg)': False, 'contains(rottingham)': False, 'contains(gymnasium)': False, 'contains(sonar)': False, 'contains(hallucinatory)': False, 'contains(internment)': False, 'contains(safina)': False, 'contains(coagulate)': False, 'contains(columbus)': False, 'contains(aretha)': False, 'contains(golden)': False, 'contains(desparate)': False, 'contains(perlich)': False, 'contains(maintenance)': False, 'contains(putzes)': False, 'contains(anally)': False, 'contains(delacroix)': False, 'contains(stagecoach)': False, 'contains(postlethwaite)': False, 'contains(conferring)': False, 'contains(heighten)': False, 'contains(pain)': False, 'contains(parsley)': False, 'contains(partnered)': False, 'contains(bootstraps)': False, 'contains(geiger)': False, 'contains(squashed)': False, 'contains(meanders)': False, 'contains(tourfilm)': False, 'contains(intellectualizing)': False, 'contains(forbidding)': False, 'contains(suo)': False, 'contains(featuring)': False, 'contains(coax)': False, 'contains(wrapping)': False, 'contains(niftiest)': False, 'contains(rudiments)': False, 'contains(deride)': False, 'contains(offshoot)': False, 'contains(romance)': False, 'contains(groundbreaking)': False, 'contains(dreamt)': False, 'contains(marianne)': False, 'contains(and)': True, 'contains(bolster)': False, 'contains(base)': False, 'contains(nuns)': False, 'contains(brilliant)': False, 'contains(highlighting)': False, 'contains(duplicates)': False, 'contains(trance)': False, "contains('--)": False, 'contains(psychic)': False, 'contains(zit)': False, 'contains(illiteracy)': False, 'contains(commence)': False, 'contains(longshanks)': False, 'contains(beleive)': False, 'contains(swaggering)': False, 'contains(9mm)': False, 'contains(ahem)': False, 'contains(crowned)': False, 'contains(marathon)': False, 'contains(marienbad)': False, 'contains(accountable)': False, 'contains(shadowing)': False, 'contains(rather)': False, 'contains(alzhiemer)': False, 'contains(eligible)': False, 'contains(dislocation)': False, 'contains(voiceovers)': False, 'contains(affay)': False, 'contains(independence)': False, 'contains(mst3k)': False, 'contains(sulking)': False, 'contains(cloying)': False, 'contains(skinheads)': False, 'contains(garp)': False, 'contains(robin)': False, 'contains(suppose)': False, 'contains(confounding)': False, 'contains(priority)': False, 'contains(word)': False, 'contains(silliest)': False, 'contains(gapes)': False, 'contains(pointlessly)': False, 'contains(unflinching)': False, 'contains(intruders)': False, 'contains(deserving)': False, 'contains(quintessential)': False, 'contains(jealously)': False, 'contains(regaled)': False, 'contains(emptying)': False, 'contains(outinen)': False, 'contains(distorted)': False, 'contains(youths)': False, 'contains(tell)': False, 'contains(atheist)': False, 'contains(departs)': False, 'contains(handgun)': False, 'contains(slaughter)': False, 'contains(pays)': False, 'contains(albert)': False, 'contains(damnation)': False, 'contains(hopper)': False, 'contains(recylcled)': False, 'contains(truckloads)': False, 'contains(jnr)': False, 'contains(layouts)': False, 'contains(coffins)': False, 'contains(slate)': False, 'contains(pessimistic)': False, 'contains(agile)': False, 'contains(ravera)': False, 'contains(cheif)': False, 'contains(masur)': False, 'contains(diners)': False, 'contains(lengthy)': False, 'contains(teriffic)': False, 'contains(kilgore)': False, 'contains(sinfully)': False, 'contains(cash)': False, 'contains(tippi)': False, 'contains(blissful)': False, 'contains(sequences)': False, 'contains(mcgavin)': False, 'contains(talmud)': False, 'contains(apprehended)': False, 'contains(batfans)': False, 'contains(families)': False, 'contains(synthesizer)': False, 'contains(macy)': False, 'contains(needles)': False, 'contains(pathos)': False, 'contains(dramatism)': False, 'contains(weasely)': False, 'contains(sterling)': False, 'contains(improvisation)': False, 'contains(rile)': False, 'contains(spinster)': False}

In [19]:

```
len(movie_reviews.words('pos/cv957_8737.txt'))
```

Out[19]:

```
597
```

We are ready to train the classifier:

In [20]:

```
featuresets = [(document_features(d), c) for (d,c) in documents]
train_set, test_set = featuresets[100:], featuresets[:100]
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

Simple check:

In [21]:

```
classifier.show_most_informative_features(5)
```

```
Most Informative Features
       contains(courage) = True              pos : neg    =      6.
3 : 1.0
       contains(cunning) = True              pos : neg    =      6.
2 : 1.0
        contains(stinks) = True              neg : pos    =      5.
9 : 1.0
    contains(dedication) = True              pos : neg    =      5.
5 : 1.0
         contains(elise) = True              pos : neg    =      5.
5 : 1.0
```

In [22]:

```
nltk.classify.accuracy(classifier,test_set)
```

Out[22]:

```
0.59
```

In [23]:

```
doc = "this movie was pathetic".split()
```

In [24]:

```
classifier.classify(document_features(doc))
```

Out[24]:

```
'neg'
```

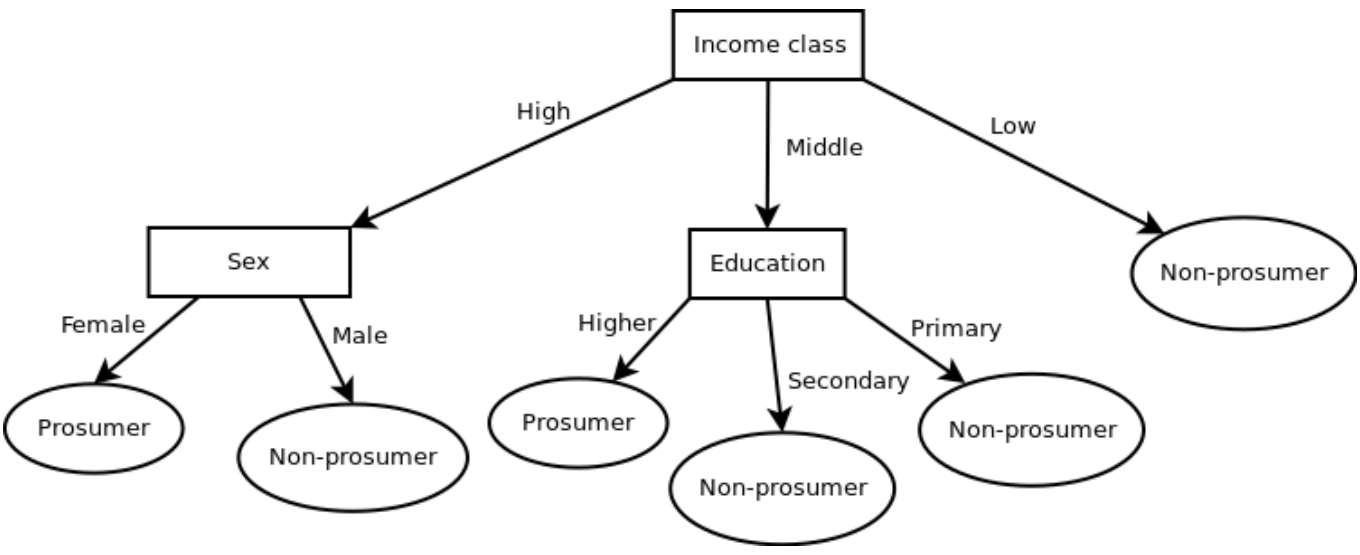
Decision trees

- a data mining method used very often for classification purposes
- other common applications:
  - variable selection, i.e. selection of the most relevant attributes that should be used to form a model
  - relative importance of variables
  - prediction, i.e. forecasting results for future data by using a tree built on historical data
- a decision tree depicts rules for dividing existing data into groups
  - the first rule splits the entire data set into some number of subsets
  - another rules may be applied to each of those subsets forming a next generation of them
  - procedure is repeated until the data in each subset forms a final group
- trees accept several types of variables (nominal, ordinal and interval ones)
- they can handle records with missing data and do not require to drop them

To give an understanding of the concept of decision trees for a non-expert reader let us assume that our goal is to predict if a person is a prosumer taking into consideration features like income, education and sex. A sample data set could look like the one presented in the following table (**training set**):

Income class	Sex	Education	Prosument
High	F	Secondary	YES
High	M	Higher	NO
High	F	Higher	YES
Middle	M	Higher	YES
Middle	F	Secondary	NO
Middle	F	Primary	NO
Low	F	Higher	NO
Low	M	Secondary	NO
Middle	F	Higher	YES
High	F	Higher	YES

A corresponding decision tree is shown below:



- a graph-like structure consisting of non-leaf nodes (depicted by rectangles) and leaves (ovals) connected with each other
- **non-leaf nodes** are attributes characterizing the items in the data set
- **leaves** represent predicted variable
- prediction is easy:
  - a new person - a female with high income and a secondary level education
  - going along appropriate branches will lead us to the conclusion that the person is a prosumer
- interpretation of the decision tree is straightforward and no expert knowledge

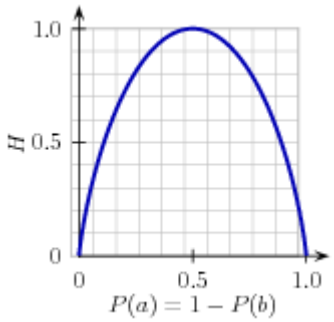
Building a tree

- finding a proper attribute to split the data is not a trivial task
- several algorithms available
- *Iterative Dichotomiser 3* (ID3) - an algorithm invented by Ross Quinlan (1986)
  - it two concepts from the information theory: information entropy and information gain

The entropy H (see C. Shannon, *A Mathematical Theory of Communication*, 1948) is understood as the average information contained in a message or needed to generate it. In our example the message would be simply the `Prosumer` or `Not Prosumer` classes returned by the tree. The entropy of a data set is calculated according to the following formula:

$$H(p,n) = -\frac{p}{p+n}\log\frac{p}{p+n} - \frac{n}{p+n}\log\frac{n}{p+n}$$

Here, p is the number of positive examples in the data set and n - the number of negative ones. Please note that for pure data samples, i.e. all records belonging to the same class, the entropy is equal to zero.



In our example we have 5 prosumers and 5 non-prosumers in the data, thus the initial entropy is

$$H(5,5) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = -\log_2\frac{1}{2} = 1$$

The information gain is the difference between the entropy before and after the split. Let us take the attribute `Income class` as a candidate for breaking down the data. After the split we get following entropies in each subset:

$$H_{High}(3,1) = -0.75 * \log_2(0.75) - 0.25 * \log_2(0.25) = 0.811$$
$$H_{Middle}(2,2) = 1$$
$$H_{Low}(0,2) = 0$$

For the average entropy after the split we have

$$\bar{H}(Income) = 0.4 * 0.811 + 0.4 * 1.0 + 0.2 * 0 = 0.7244.$$

The corresponding information gain is given by

$$I_G(Income) = H(5,5) - \bar{H}(Income) = 0.2756$$

Similarly, for the other attributes we get

$$I_G(Sex) = 0.03$$
$$I_G(Education) = 0.16$$

Since `Income class` yields the largest information gain, we use it to split the data. Then we repeat the same procedure recursively on each subset, considering only attributes not selected before.

In [25]:

```
import math
def entropy(labels):
    freqdist = nltk.FreqDist(labels)
    probs = [freqdist.freq(l) for l in freqdist]
    return -sum(p * math.log(p,2) for p in probs)
```

In [26]:

```
print(entropy(['male', 'male', 'male', 'male']))
```

```
-0.0
```

In [27]:

```
print(entropy(['male', 'female', 'male', 'male']))
```

```
0.8112781244591328
```

In [28]:

```
print(entropy(['female', 'male', 'female', 'male']))
```

```
1.0
```

In [29]:

```
print(entropy(['female', 'female', 'male', 'female']))
```

```
0.8112781244591328
```

In [30]:

```
print(entropy(['female', 'female', 'female', 'female']))
```

```
-0.0
```

Decision trees in NLTK

We will work with the movie reviews from the previous example. Let us define first a couple of helper functions:

In [31]:

```
def bag_of_words(words):
    return dict([(word, True) for word in words])
```

In [32]:

```
def bag_of_words_not_in_set(words, badwords):
    return bag_of_words(set(words) - set(badwords))
```

In [33]:

```
import collections
def label_feats_from_corpus(corp, feature_detector=bag_of_words):
    label_feats = collections.defaultdict(list)
    for label in corp.categories():
        for fileid in corp.fileids(categories=[label]):
            feats = feature_detector(corp.words(fileids=[fileid]))
            label_feats[label].append(feats)
    return label_feats
```

In [34]:

```
def split_label_feats(lfeats, split=0.75):
    train_feats = []
    test_feats = []
    for label, feats in lfeats.items():
        cutoff = int(len(feats) * split)
        train_feats.extend([(feat, label) for feat in feats[:cutoff]])
        test_feats.extend([(feat, label) for feat in feats[cutoff:]])
    return train_feats, test_feats
```

In [35]:

```
from nltk.corpus import movie_reviews
print(movie_reviews.categories())
```

```
['neg', 'pos']
```

In [36]:

```
lfeats = label_feats_from_corpus(movie_reviews)
print(lfeats.keys())
```

```
dict_keys(['neg', 'pos'])
```

In [37]:

```
train_feats, test_feats = split_label_feats(lfeats)
print(len(train_feats))
print(len(test_feats))
```

```
1500
500
```

In [38]:

```
print(test_feats[0])
```

```
({'s': True, 'idea': True, 'esteem': True, 'gets': True, 'cute': Tr
ue, 'enjoyable': True, 'twist': True, 'last': True, '-': True, 'i
t': True, 'quite': True, 'own': True, 'when': True, 'funny': True,
 'from': True, 'pet': True, 'both': True, 'predictability': True,
 'zegers': True, "'": True, 'alone': True, 'if': True, 'feat': Tru
e, 'wags': True, 'true': True, 'credits': True, 'not': True, '.': T
rue, '!': True, 'snively': True, ',': True, 'used': True, 'see': Tr
ue, 'actual': True, 'can': True, 'pairs': True, 'yeller': True, 'tw
o': True, 'michael': True, ')': True, 'been': True, 'a': True, 'ar
e': True, 'opens': True, 'movie': True, 'just': True, 'school': Tru
e, 'comedy': True, 'death': True, 'straight': True, 'light': True,
 'marches': True, 'very': True, '(': True, 'norm': True, 'block': T
rue, 'had': True, 'number': True, 'faux': True, 'face': True, 'stup
id': True, 'shots': True, 'surfaces': True, ':': True, 'buried': Tr
ue, 'cope': True, 'big': True, 'lie': True, 'moment': True, 'slapst
ick': True, 'on': True, 'cans': True, 'fernwell': True, 'would': Tr
ue, 'accomplished': True, 'reluctant': True, 'the': True, 'and': Tr
ue, 'comeuppance': True, 'i': True, 'its': True, 'calculated': Tru
e, 'court': True, 'letterman': True, 'matter': True, 'david': True,
'abusive': True, 'dog': True, 'formula': True, 'no': True, 'paint':
True, 'mopey': True, 'chain': True, 'climax': True, 'send': True,
 'may': True, 'least': True, 'well': True, 'once': True, 'heavy': T
rue, 'rockets': True, 'old': True, 'end': True, 'contracts': True,
 'events': True, '--': True, 'animal': True, 'asked': True, 'an': T
rue, 'musical': True, 'himself': True, 'k9': True, '?': True, 'moti
on': True, 'where': True, 'whatever': True, 'style': True, 'sight':
True, 'clown': True, 'have': True, 'clad': True, 'win': True, 'fade
s': True, 'finals': True, 'buddy': True, 'there': True, 'boy': Tru
e, 'one': True, 'else': True, 'although': True, 'is': True, 'visua
l': True, 'friend': True, 'note': True, 'gloom': True, 'all': True,
'for': True, 'game': True, 'proves': True, 'father': True, 'okay':
 True, 'escapes': True, 'exist': True, 'up': True, 'special': True,
'basketball': True, 'seem': True, 'current': True, 'newspapers': Tr
ue, 'name': True, 'with': True, 'approach': True, 'everything': Tru
e, 'absurdity': True, 'picture': True, 'possibly': True, 'moments':
True, 'back': True, 'surprisingly': True, 'move': True, 'off': Tru
e, 'quicker': True, 'trying': True, 'splish': True, 'think': True,
 'tricks': True, 'ends': True, 'his': True, 'this': True, 'or': Tru
e, 'effects': True, 'cleaned': True, 'save': True, 't': True, 'owne
r': True, 'he': True, 'as': True, 'connection': True, 'washington':
True, 'must': True, 'self': True, 'executed': True, 'of': True, 'ev
en': True, 'occasional': True, 'be': True, 'sequences': True, 'min
e': True, 'yeah': True, 'solemn': True, 'begin': True, 'here': Tru
e, 'josh': True, 'we': True, 'courtroom': True, 'prowess': True, 'r
eclaim': True, 'jersey': True, 'appropriate': True, 'actually': Tru
e, 'then': True, 'splash': True, 'in': True, 'at': True, 'tells': T
rue, 'insists': True, 'new': True, 'than': True, 'plays': True, 'ca
nnot': True, 'kevin': True, 'air': True, 'out': True, 'their': Tru
e, 'sneakers': True, 'team': True, 'that': True, 'spilled': True,
 'follows': True, 'player': True, 'realized': True, 'make': True,
 '"': True, 'could': True, 'successful': True, 'sink': True, 'place
s': True, 'to': True, 'hire': True, 'anything': True, 're': True,
 'before': True, 'were': True, 'trick': True, 'tale': True, 'segmen
t': True, 'while': True, 'recent': True, 'interested': True, 'famil
y': True, 'doesn': True, 'forced': True, 'joke': True, 'disney': Tr
ue, 'jeter': True, 'rabies': True, 'montage': True, 'bud': True, 'p
ooch': True, 'mascot': True, 'golden': True, 'saw': True, 'they': T
rue, 'retriever': True, 'kid': True, 'more': True, 'story': True,
 'but': True, 'granted': True, 'better': True, 'cheer': True}, 'ne
g')
```

Now we can train and test the classifier:

In [39]:

```
from nltk.classify import DecisionTreeClassifier
#dt_classifier = DecisionTreeClassifier.train(train_feats, binary=True, entropy_
cutoff=0.8, depth_cutoff=5, support_cutoff=30)
dt_classifier = DecisionTreeClassifier.train(train_feats, depth_cutoff=5)
```

In [40]:

```
print(nltk.classify.accuracy(dt_classifier,test_feats))
```

```
0.682
```

In [41]:

```
for i in test_feats:
    print("Human: ",i[-1],"  |  Machine: ", dt_classifier.classify(i[0]))
```

```
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
```

```
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
```

```
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
```

```
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
```

```
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  neg
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  neg   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
```

```
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
```

```
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
```

```
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
```

```
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  neg
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
Human:  pos   |  Machine:  pos
```

In [44]:

```
doc = bag_of_words("this movie was bad".split())
print(doc)
```

```
{'movie': True, 'was': True, 'this': True, 'bad': True}
```

In [45]:

```
dt_classifier.classify(doc)
```

Out[45]:

```
'neg'
```

In [46]:

```
doc = bag_of_words("this movie was pathetic".split())
print(doc)
```

```
{'movie': True, 'was': True, 'this': True, 'pathetic': True}
```

In [47]:

```
dt_classifier.classify(doc)
```

Out[47]:

```
'pos'
```

Maximum Entropy Classifier

- a probabilistic classifier which belongs to the class of exponential models
- independence of features not required
- based on the principle of Maximum Entropy - from all the models that fit the training data it selects the one which has the largest entropy

> From Wikipedia:
>
> The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with largest entropy, in the context of precisely stated prior data (such as a proposition that expresses testable information).
>
> Another way of stating this: Take precisely stated prior data or testable information about a probability distribution function. Consider the set of all trial probability distributions that would encode the prior data. According to this principle, the distribution with maximal information entropy is the proper one.
>
> The principle was first expounded by E. T. Jaynes in two papers in 1957, where he emphasized a natural correspondence between statistical mechanics and information theory. In particular, Jaynes offered a new and very general rationale why the Gibbsian method of statistical mechanics works. He argued that the entropy of statistical mechanics and the information entropy of information theory are basically the same thing. Consequently, statistical mechanics should be seen just as a particular application of a general tool of logical inference and information theory.

- can be used to solve a large variety of text classification problems:
  - language detection
  - topic classification
  - sentiment analysis
- implementing MaxEnt classifier in a standard programming language is non-trivial primarily due to the numerical optimization problem that one should solve in order to estimate the weights of the model

When to use?

- when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions
- when we can't assume the conditional independence of the features (particularly true in Text Classification problems where the features are usually words which obviously are not independent)

Theoretical Background

Our target is to use the contextual information of the document (unigrams, bigrams, other characteristics within the text) in order to categorize it to a given class (positive/neutral/negative, objective/subjective etc). Following the standard bag-of-words framework that is commonly used in natural language processing and information retrieval, let $\{w1, \ldots, wm\}$ be the m words that can appear in a document. Then each document is represented by a sparse array with 1s and 0s that indicate whether a particular word w_i exists or not in the context of the document.

In order to construct a stochastic model, the first step is to collect a large number of training data which consists of samples in the following format: (x_i, y_i) where the x_i includes the contextual information of the document (the sparse array) and y_i its class.

The second step is to summarize the training sample in terms of its empirical probability distribution:

$$\tilde{p}(x,y) \equiv \frac{1}{N} \times \text{number of times that } (x,y) \text{occurs in the sample}$$

Here, N is the size of the training dataset.

The above empirical distribution will be used to construct the statistical model of the random process which assigns texts to a particular class by taking into account their contextual information.

We introduce the following indicator function, i.e. **feature**:

$$f_j(x,y) = \begin{cases} 1 & \text{if } y = c_i \text{ and x contains } w_k \\ 0 & \text{otherwise} \end{cases}$$

This binary valued indicator function returns 1 only when the class of a particular document is c_i and the document contains the word w_k.

We express any statistic of the training dataset as the expected value of the appropriate binary-valued indicator function f_j. Thus the expected value of feature f_j with respect to the empirical distribution $\tilde{p}(x,y)$ is equal to:

$$\tilde{p}(f_j) \equiv \sum_{x,y} \tilde{p}(x,y) f_j(x,y)$$

If each training sample (x,y) occurs once in training dataset then $\tilde{p}(x,y)$ is equal to $1/N$.

When a particular statistic is useful to our classification, we require our model to accord with it. To do so, we constrain the expected value that the model assigns to the expected value of the feature function f_j, i.e.

$$p(f_j) \equiv \sum_{x,y} \tilde{p}(x) p(y \mid x) f_j(x,y)$$

Here, $\tilde{p}(x)$ is the empirical distribution of x in the training dataset and it is usually set equal to $1/N$.

By constraining the expected value to be the equal to the empirical value and from the above equations we have that:

$$\sum_{x,y} \tilde{p}(x) p(y \mid x) f_j(x,y) = \sum_{x,y} \tilde{p}(x,y) f_j(x,y)$$

The last equation is called **constrain** and we have as many constrains as the number of feature functions j.

The above constrains can be satisfied by an infinite number of models. In order to build our model, we need to select the best candidate based on a specific criterion. According to the principle of **Maximum Entropy**, we should select the model that is as close as possible to uniform. In order words, we should select the model $p*$ with Maximum Entropy:

$$p^* = \arg\max_{p \in C} \left(-\sum_{x,y} \tilde{p}(x) p(y \mid x) \log\ p(y \mid x) \right)$$

We have:

1. $p(y|x) \geq 0$ for all x and y
2. $\sum_y p(y|x) = 1$ for all x
3. $\sum_{x,y} \tilde{p}(x) p(y|x) f_j(x,y) = \sum_{x,y} \tilde{p}(x,y) f_j(x,y)$ for $j \in \{1, 2, \ldots, n\}$

To solve the above optimization problem we introduce the Lagrangian multipliers and estimate them with the maximum likelihood method.

It can be proven that if we find the multipliers $\{\lambda_1, \ldots, \lambda_n\}$ that maximize the dual problem, the probability of a document x to be classified as y is equal to:

$$p^*(y \mid x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{\sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)}$$

Thus, once the lamda parameters are found, all we need to do in order to classify a new document is to use the **maximum a posteriori** decision rule and select the category with the highest probability.

Estimating the lamda parameters requires using an **iterative scaling algorithm** such as the GIS (Generalized Iterative Scaling) or the IIS (Improved Iterative Scaling):

Input: Feature functions $f_1, \ldots, f_n$, empirical distribution $\tilde{p}(x)$
Output: Optimal parameter values λ_i, optimal model p^*
Steps:

1. Start with $\lambda_i = 0$ for all $i \in \{1, 2, \ldots, n\}$
2. For each $i \in \{1, 2, \ldots, n\}$ do:
   A. Let $\Delta\lambda_i$ be the solution to
   $$\sum_{x,y} \tilde{p}(x) p(y|x) f_j(x,y) \exp(\Delta\lambda_i f^s(x,y)) = \tilde{p}(f_i)$$
   with $f^s(x,y) = \sum_{i=1}^n f_i(x,y)$
   B. Update the value of λ_i accordning to $\lambda_i \leftarrow \lambda_i + \Delta\lambda_i$
3. Go to step 2 if not all parameters λ_i have converged

The $f^s(x,y)$ is the total number of features which are active for a particular (x,y) pair. If this number is constant for all documents then $\Delta\lambda_i$ can be calculated in closed-form:

$$\Delta\lambda_i = \frac{1}{C} \log \frac{\tilde{p}(f_i)}{p(f_i)}$$

where $C = f^s(x,y)$. The assumption of $f^s(x,y)$ being constant is rarely realistic in practice. Fortunately, it is only necessary that C is bounded by $f^s(x,y)$ (Goodman 2002, Ratnaparkhi 1997). Thus in some versions of ISS one selects simply

$$C = C_{max} = \arg\max_{x,y} f^s(x,y)$$

MaxEnt classifier in NLTK

- GIS(Generalized Iterative Scaling), IIS(Improved Iterative Scaling), and LM-BFGS (limited memory Broyden-Fletcher-Goldfarb-Shanno) training methods
- GIS and ISS are implemented within NLTK module
  - seem very slow
  - cost large memory

- LM-BFGS supports external libraries like MEGAM (MEGA Model Optimization Package)

In [25]:

```
from nltk import MaxentClassifier
me_classifier = MaxentClassifier.train(train_set,algorithm='gis')
```

```
==> Training (100 iterations)

    Iteration    Log Likelihood    Accuracy
    ----------------------------------------
           1          -0.69315        0.505
           2          -0.69248        0.505
           3          -0.69193        0.505
           4          -0.69138        0.505
           5          -0.69083        0.505
           6          -0.69028        0.507
           7          -0.68973        0.513
           8          -0.68919        0.526
           9          -0.68864        0.543
          10          -0.68810        0.558
          11          -0.68755        0.580
          12          -0.68701        0.599
          13          -0.68647        0.621
          14          -0.68593        0.643
          15          -0.68538        0.662
          16          -0.68485        0.673
          17          -0.68431        0.691
          18          -0.68377        0.702
          19          -0.68323        0.713
          20          -0.68270        0.725
          21          -0.68216        0.738
          22          -0.68163        0.750
          23          -0.68110        0.761
          24          -0.68057        0.767
          25          -0.68004        0.773
          26          -0.67951        0.777
          27          -0.67898        0.783
          28          -0.67845        0.782
          29          -0.67793        0.790
          30          -0.67740        0.796
          31          -0.67688        0.801
          32          -0.67635        0.804
          33          -0.67583        0.807
          34          -0.67531        0.808
          35          -0.67479        0.812
          36          -0.67427        0.813
          37          -0.67375        0.814
          38          -0.67323        0.815
          39          -0.67272        0.815
          40          -0.67220        0.814
          41          -0.67168        0.817
          42          -0.67117        0.819
          43          -0.67066        0.819
          44          -0.67015        0.821
          45          -0.66963        0.822
          46          -0.66912        0.824
          47          -0.66862        0.827
          48          -0.66811        0.828
          49          -0.66760        0.828
          50          -0.66709        0.829
          51          -0.66659        0.831
          52          -0.66608        0.832
          53          -0.66558        0.832
          54          -0.66508        0.832
          55          -0.66458        0.831
          56          -0.66408        0.829
          57          -0.66358        0.830
```

```
           58          -0.66308        0.831
           59          -0.66258        0.831
           60          -0.66208        0.831
           61          -0.66159        0.831
           62          -0.66109        0.831
           63          -0.66060        0.830
           64          -0.66010        0.830
           65          -0.65961        0.831
           66          -0.65912        0.832
           67          -0.65863        0.832
           68          -0.65814        0.832
           69          -0.65765        0.832
           70          -0.65716        0.832
           71          -0.65667        0.833
           72          -0.65619        0.833
           73          -0.65570        0.834
           74          -0.65522        0.834
           75          -0.65474        0.834
           76          -0.65425        0.834
           77          -0.65377        0.834
           78          -0.65329        0.835
           79          -0.65281        0.835
           80          -0.65233        0.835
           81          -0.65186        0.835
           82          -0.65138        0.835
           83          -0.65090        0.835
           84          -0.65043        0.835
           85          -0.64995        0.836
           86          -0.64948        0.835
           87          -0.64901        0.835
           88          -0.64853        0.835
           89          -0.64806        0.835
           90          -0.64759        0.836
           91          -0.64712        0.836
           92          -0.64666        0.836
           93          -0.64619        0.836
           94          -0.64572        0.836
           95          -0.64526        0.837
           96          -0.64479        0.837
           97          -0.64433        0.838
           98          -0.64387        0.838
           99          -0.64340        0.839
        Final          -0.64294        0.839
```

In [26]:

```
me_classifier.show_most_informative_features(5)
```

```
  -0.110 contains(cunning)==True and label is 'neg'
  -0.102 contains(skarsgard)==True and label is 'neg'
  -0.102 contains(dedication)==True and label is 'neg'
  -0.101 contains(tatooine)==True and label is 'neg'
  -0.100 contains(elise)==True and label is 'neg'
```

In [27]:

```
nltk.classify.accuracy(classifier,test_set)
```

Out[27]:

0.59

How to improve accuracy of a text classifier?

- eliminate low quality features (words)
  - L1 regularization
- recursively grow your stopword list (by analyzing the top features):
  - frequently used words
  - countries
  - cities
  - adjectives
  - temporal words (Tuesday, tomorrow, January)
- look beyond unigrams
  - considering bigrams often boosts performance
- diversify your corpus
  - it helps dilute word features that are specific to one particular corpus
  - context of the classification problem is important while selecting the corpus
  - example: social media texts classifier - it is not enough to include only Twitter corpus
- tweak precision and recall
  - tweak the system so when it fails, it does so in a manner that is more tolerable
  - shifting False Positive (or Precision) under-performance to False Negative (or recall) under-performance
  - or viceversa
- eliminate low quality predictions
  - instead, return "class unknown" if necessary
- canonicalize words through lemma reduction
  - reduced probability space of the algorithm
- normalize exaggerations
  - haaaaapppyyyy --> happy
- eliminate numerals, punctuation and corpus specific text
  - if you are dealing with a dataset from Twitter, you might want to eliminate (using a RegEx perhaps) any usernames (of the format @user) because they do not contribute towards the classification problem you have at hand
- try a different classification algorithm
- lowering all text is not always smart
  - if you are classifying **Intensity** or **Mood**, then capitalization might be an important feature that contributes positively to the accuracy of the predictions
  - if you are trying to classify text into topics or categories, then lowering all text might have a very healthy impact on overall accuracy
- manual curation of corpus data
  - of particular importance if the training set involves human entry, or people trying to game the system for their own benefit