UNIVERSITY OF PORTSMOUTH

Intelligent Data and Text Analytics

Text Mining

Part 2

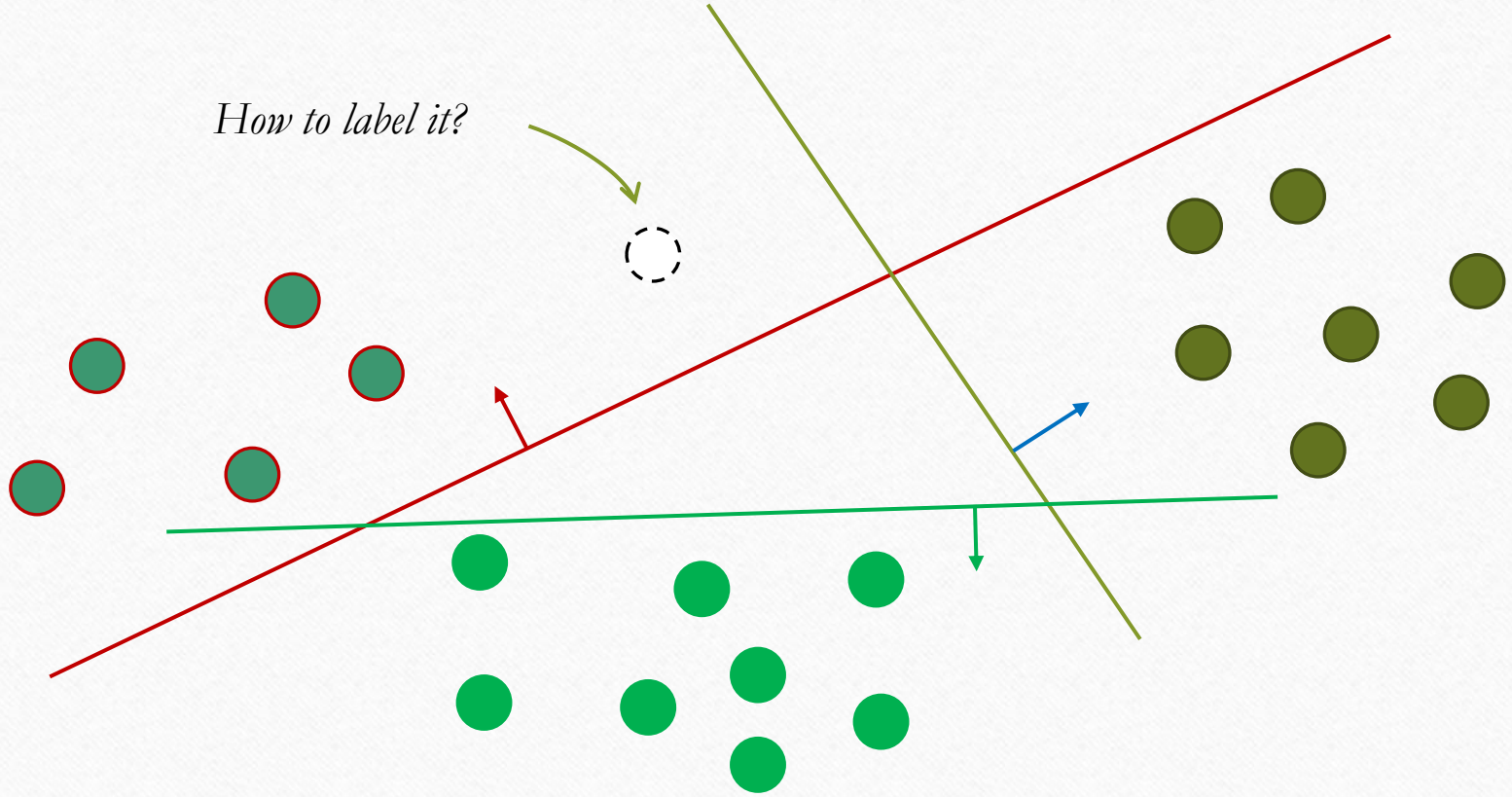
Text Mining – Part 2

- Text clustering
- Text classification
- Topic detection

Text clustering

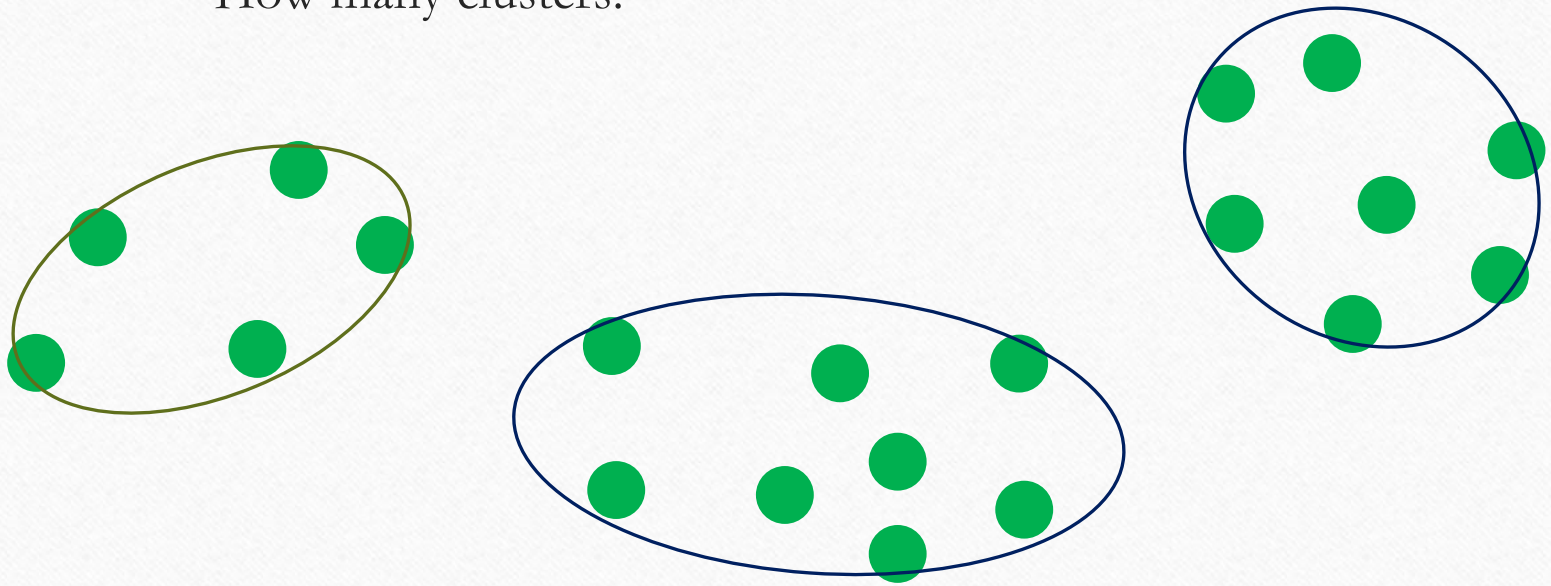
Clustering v.s. Classification

- Assigning documents to their corresponding categories



Clustering problem in general

- Discover "natural structure" of data
  - What is the criterion?
  - How to identify them?
  - How many clusters?



Clustering problem in general

- Clustering - the process of grouping a set of objects into clusters of similar objects
  - Basic criteria
    - high intra-class similarity
    - low inter-class similarity
  - No (little) supervision signal about the underlying clustering structure
  - Need similarity/distance as guidance to form clusters

What is the "natural grouping"?

Clustering is very subjective!

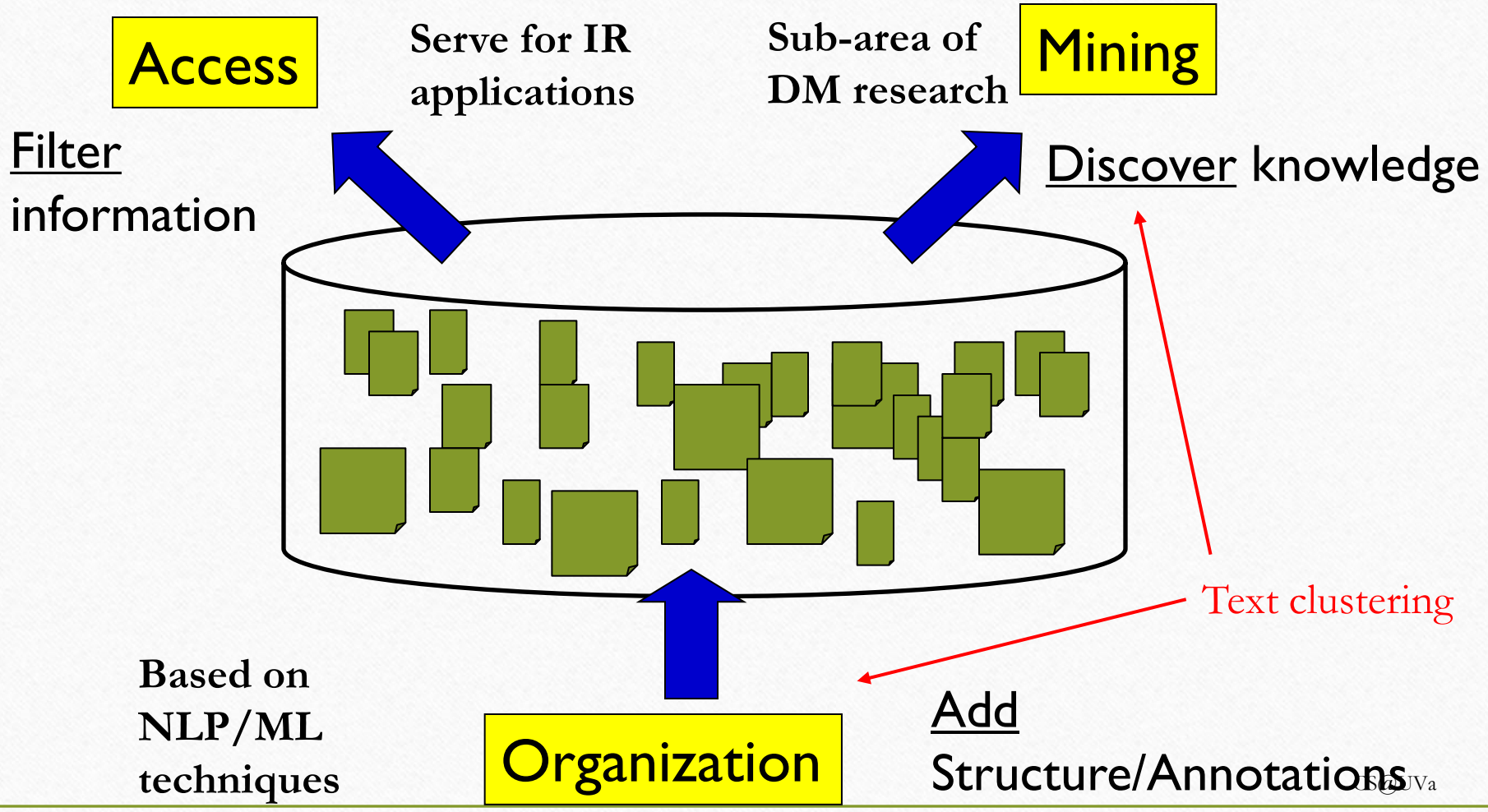
Distance metric is important!

group by gender

group by source of ability

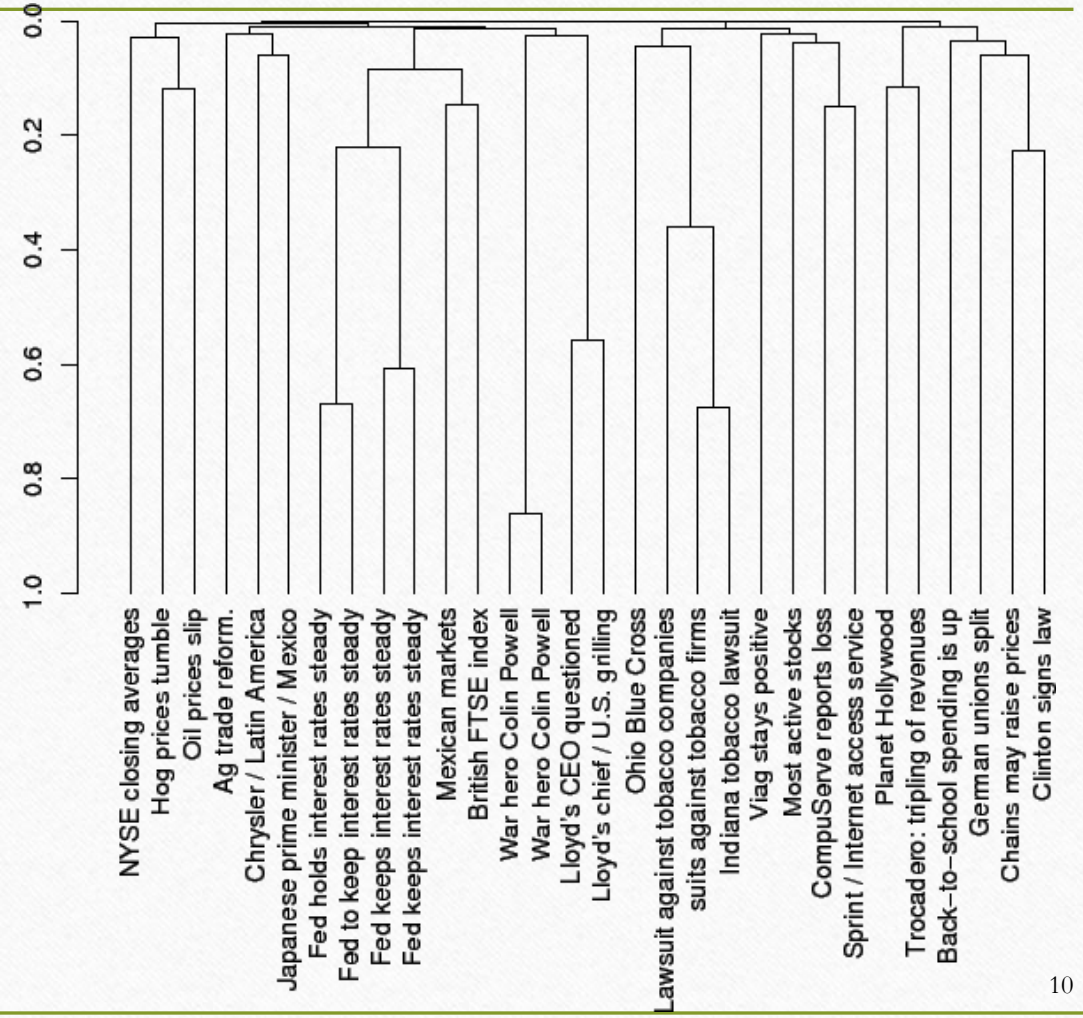
group by costume

Clustering in text mining



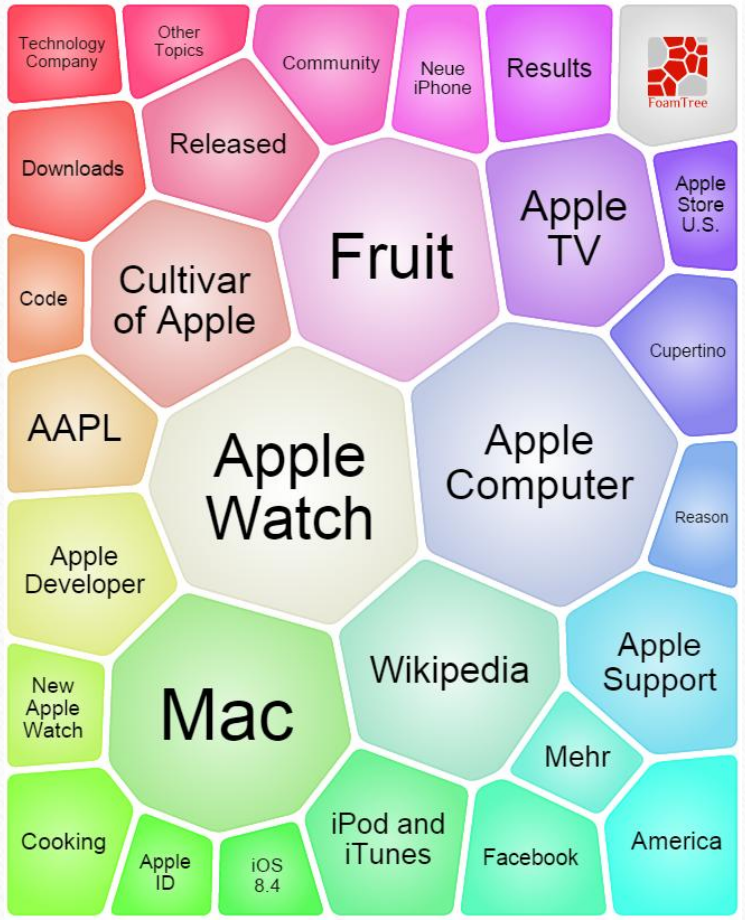
Applications of text clustering

- Organize document collections
  - Automatically identify hierarchical/topical relation among documents



Applications of text clustering

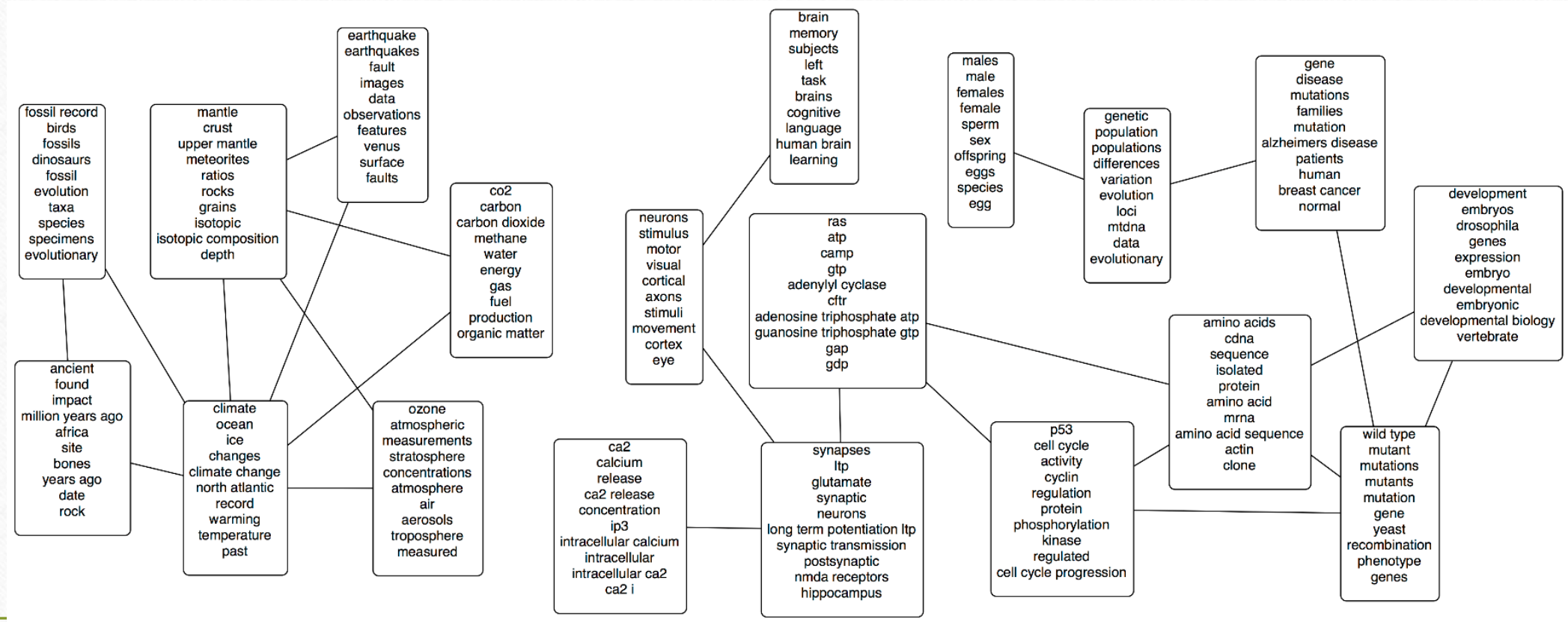
- Grouping search results
  - Organize documents by topics
  - Facilitate user browsing



http://search.carrot2.org/stable/search

Applications of text clustering

- Topic modeling
  - Grouping words into topics



Distance metric

- Basic properties
  - Positive separation
    - $D(x, y) > 0, \forall x \neq y$
    - $D(x, y) = 0$, i.f.f., $x = y$
  - Symmetry
    - $D(x, y) = D(y, x)$
  - Triangle inequality
    - $D(x, y) \leq D(x, z) + D(z, y)$

Typical distance metric

- Minkowski metric
  - $d(x, y) = \sqrt[p]{\sum_{i=1}^{V}(x_i - y_i)^p}$
    - When $p = 2$, it is Euclidean distance
- Cosine metric
  - $d(x, y) = 1 - cosine(x, y)$
    - when $|x|^2 = |y|^2 = 1$, $1 - cosine(x, y) = \frac{r^2}{2}$

Typical distance metric

- Edit distance
  - Count the minimum number of operations required to transform one string into the other
    - Possible operations: insertion, deletion and replacement

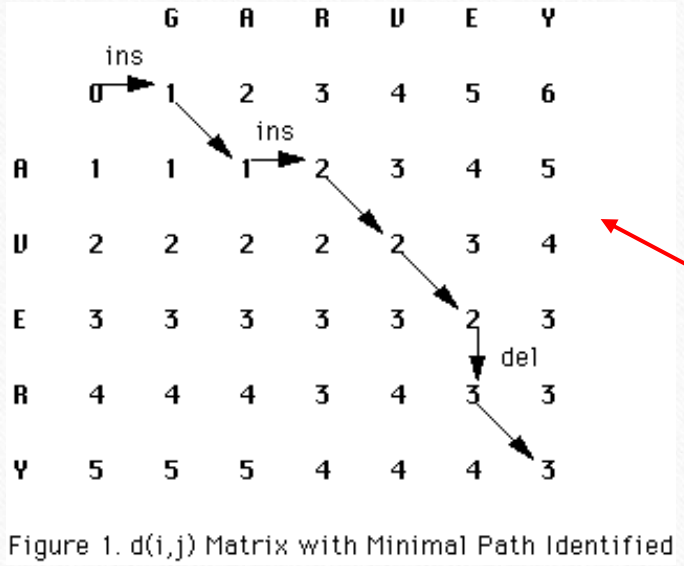


Figure 1. d(i,j) Matrix with Minimal Path Identified

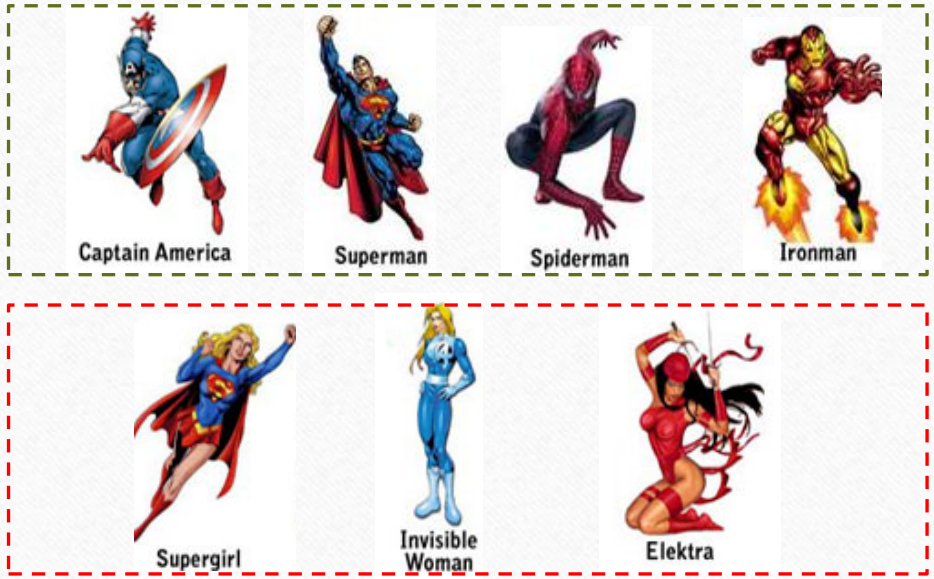
Can be efficiently solved by dynamic programming

Typical distance metric

- Edit distance
  - Count the minimum number of operations required to transform one string into the other
    - Possible operations: insertion, deletion and replacement
  - Extend to distance between sentences

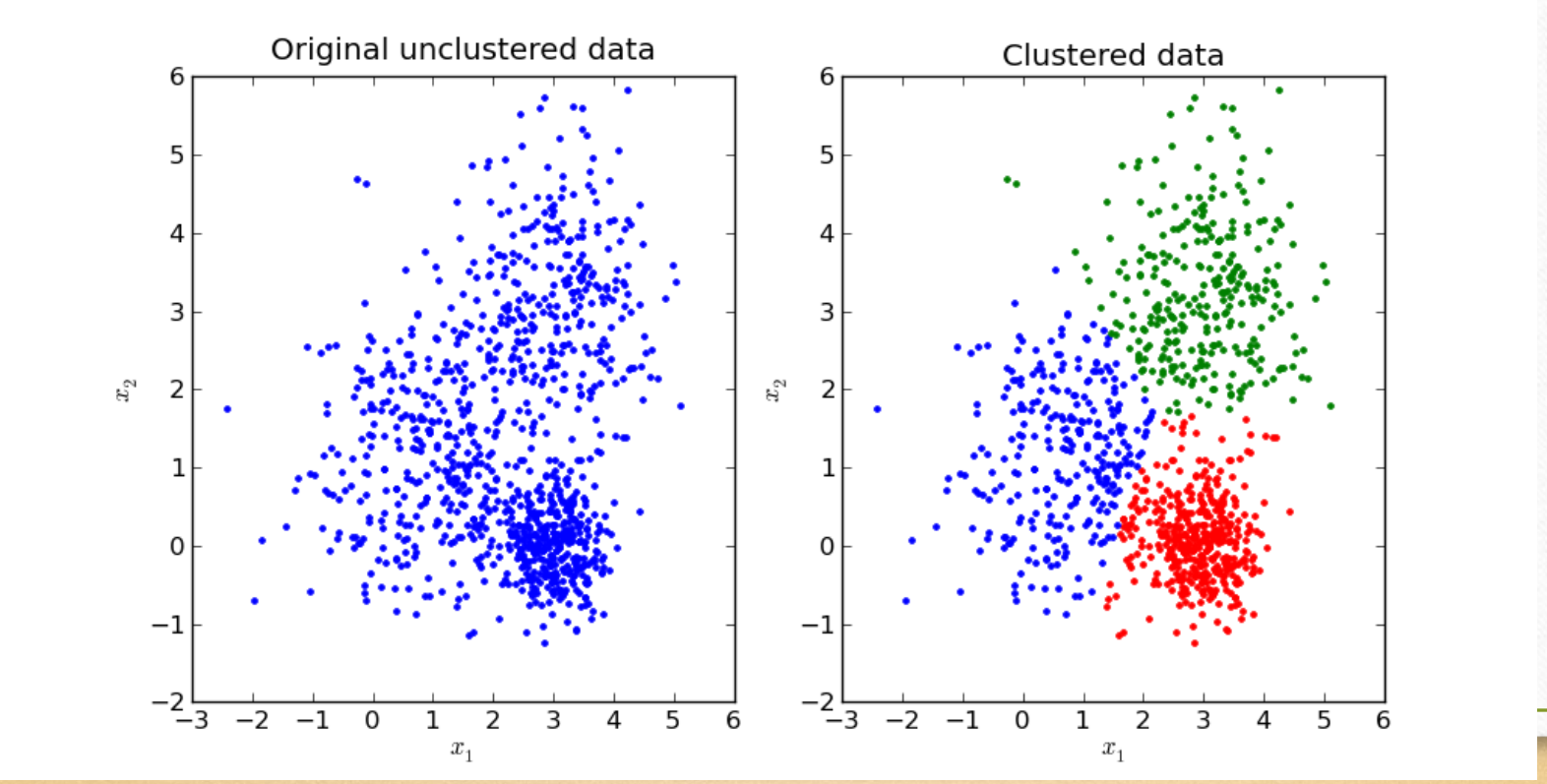
Clustering algorithms

- Partitional clustering algorithms
  - Partition the instances into different groups
  - Flat structure
    - Need to specify the number of classes in advance



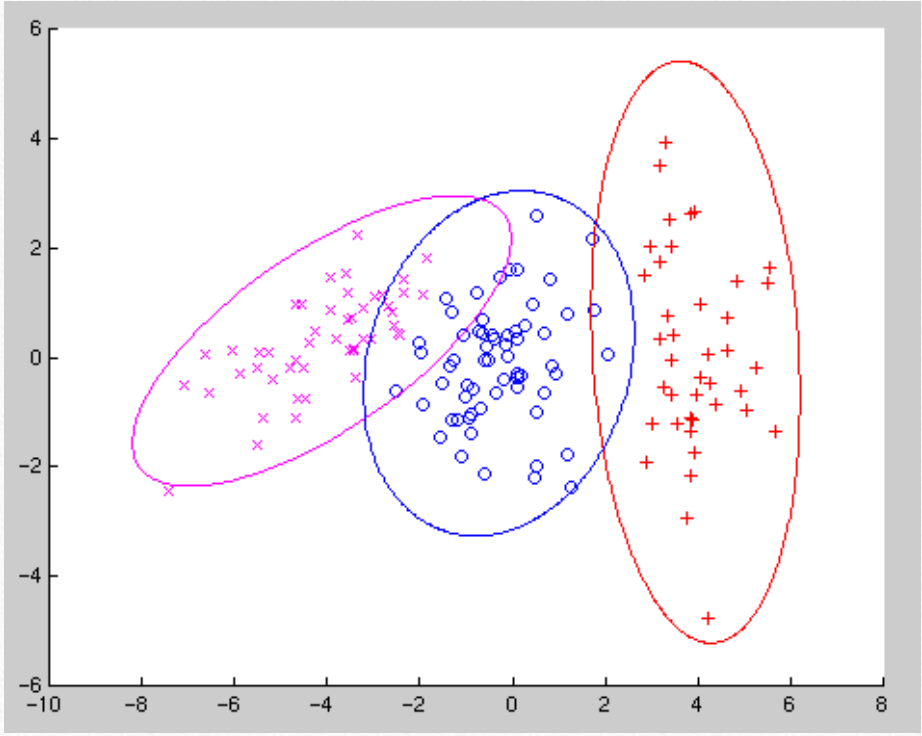
Clustering algorithms

- Typical partitional clustering algorithms
  - *k*-means clustering
    - Partition data by its closest mean



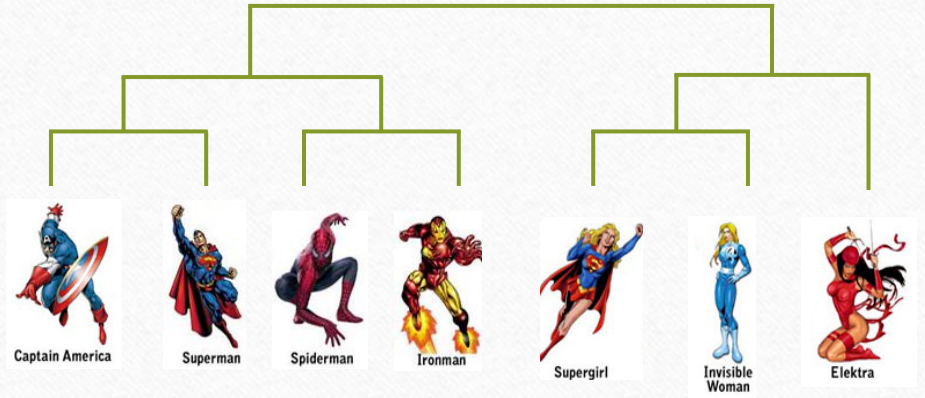
Clustering algorithms

- Typical partitional clustering algorithms
  - *k*-means clustering
    - Partition data by its closest mean
  - Gaussian Mixture Model
    - Consider variance within the cluster as well



Clustering algorithms

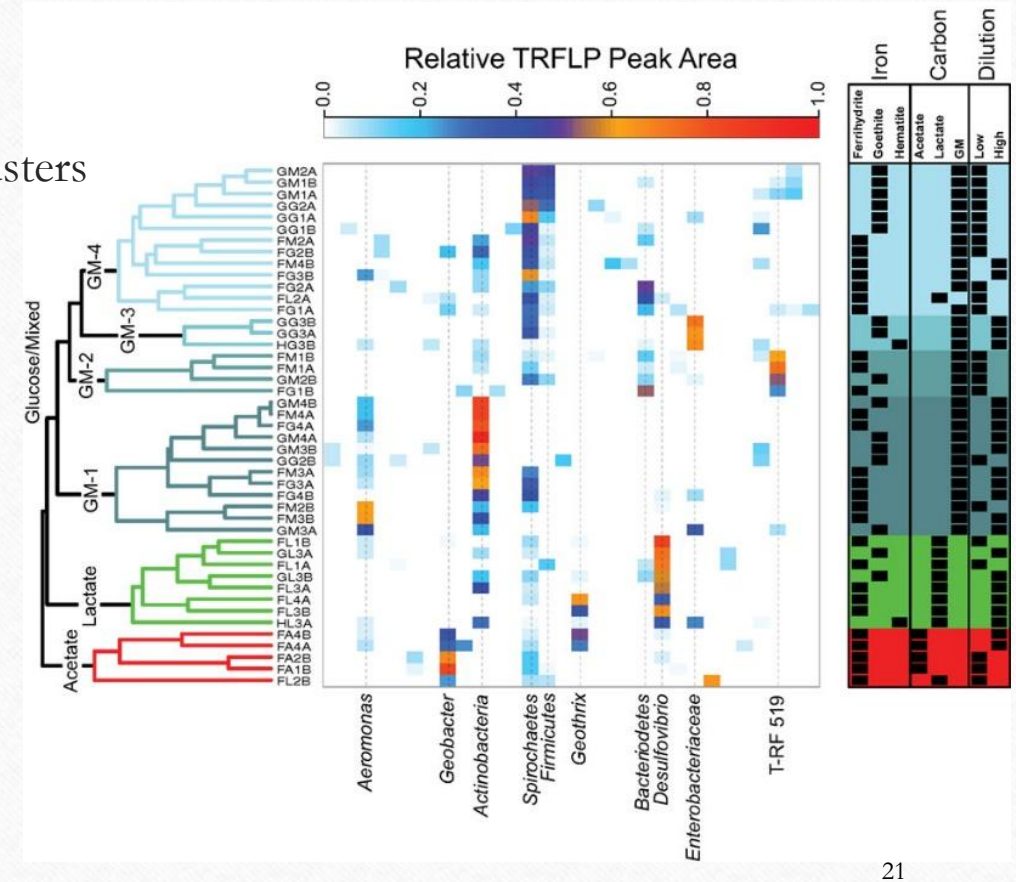
- Hierarchical clustering algorithms
  - Create a hierarchical decomposition of objects
  - Rich internal structure
    - No need to specify the number of clusters
    - Can be used to organize objects



Clustering algorithms

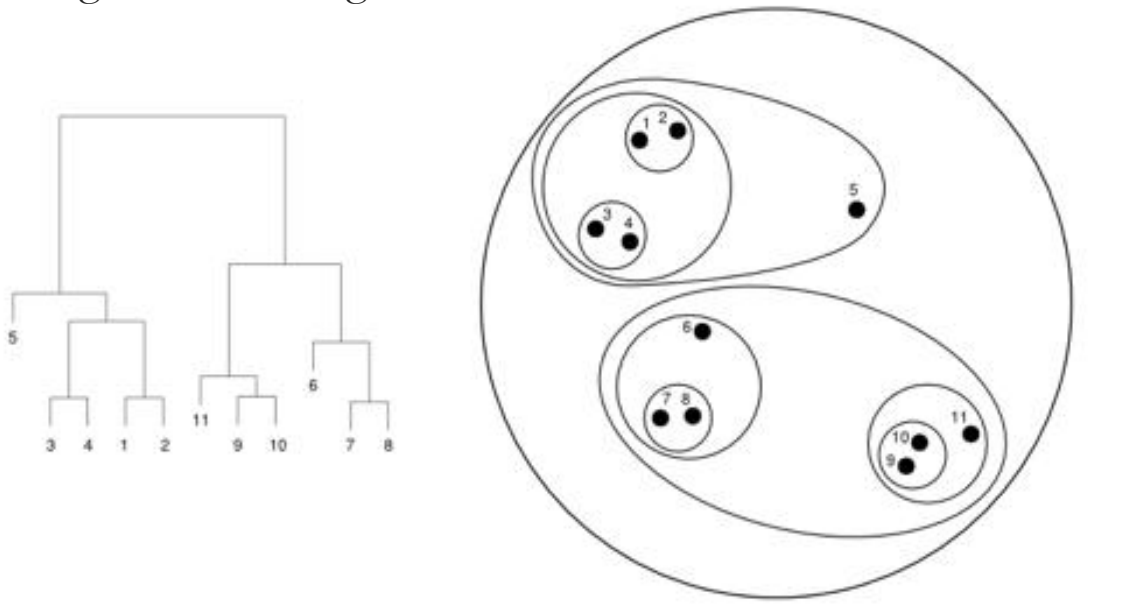
- Typical hierarchical clustering algorithms
  - Bottom-up agglomerative clustering
    - Start with individual objects as separated clusters
    - Repeatedly merge closest pair of clusters

Most typical usage: gene sequence analysis



Clustering algorithms

- Typical hierarchical clustering algorithms
  - Top-down divisive clustering
    - Start with all data as one cluster
    - Repeatedly splitting the remaining clusters into two



Desirable properties of clustering algorithms

- Scalability
  - Both in time and space
- Ability to deal with various types of data
  - No/less assumption about input data
  - Minimal requirement about domain knowledge
- Interpretability and usability

Cluster validation

- Criteria to determine whether the clusters are meaningful
  - Internal validation
    - Stability and coherence
  - External validation
    - Match with known categories

Internal validation

- Coherence
  - Inter-cluster similarity v.s. intra-cluster similarity
  - Davies–Bouldin index
    - $DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$

Evaluate every pair of clusters

- where k is total number of clusters, σ_i is average distance of all elements in cluster i, $d(c_i, c_j)$ is the distance between cluster centroid c_i and c_j.

We prefer smaller DB-index!

Internal validation

- Coherence
  - Inter-cluster similarity v.s. intra-cluster similarity
  - Dunn index
    - $D = \frac{\min\limits_{1 \le i < j \le k} d(c_i, c_j)}{\max\limits_{1 \le i \le k} \sigma_i}$
      - Worst situation analysis

We prefer larger D-index!

- Limitation
  - No indication of actual application's performance
  - Bias towards a specific type of clustering algorithm if that algorithm is designed to optimize similar metric

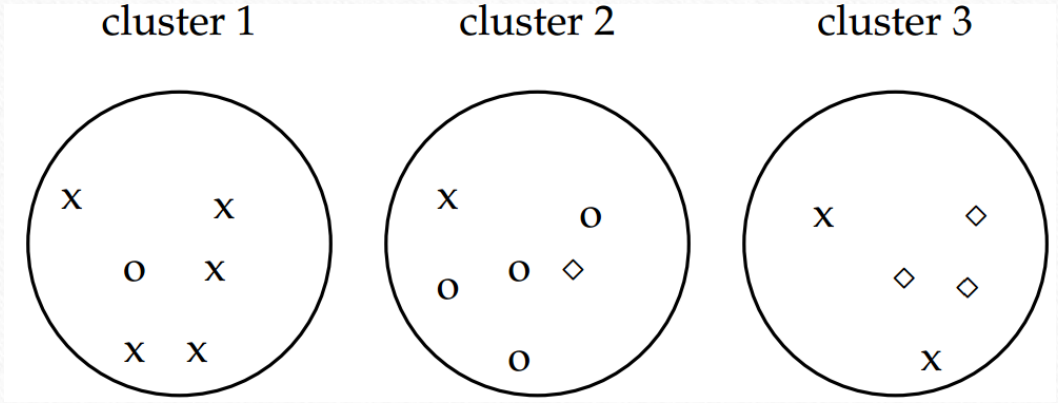
External validation

Required, might need extra cost

- Given class label Ω on each instance
  - Purity: correctly clustered documents in each cluster
    - $purity(\Omega, C) = \frac{1}{N}\sum_{i=1}^{k} \max_j |c_i \cap w_j|$
      - where c_i is a set of documents in cluster i, and w_j is a set of documents in class j

Not a good metric if we assign each document into a single cluster

$$purity(\Omega, C) = \frac{1}{17}(5 + 4 + 3)$$



Text classification

Applications of text classification

- Automatically classify politic news from sports news

political

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email
Share
Tweet
Save
More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

sports

PRO FOOTBALL | ANALYSIS

Super Bowl 2015: Patriots' Red-Hot Offense Faces Seahawks' Dominant Defense

By CHASE STUART JAN. 20, 2015

Email
Share
Tweet
Pin
Save
More

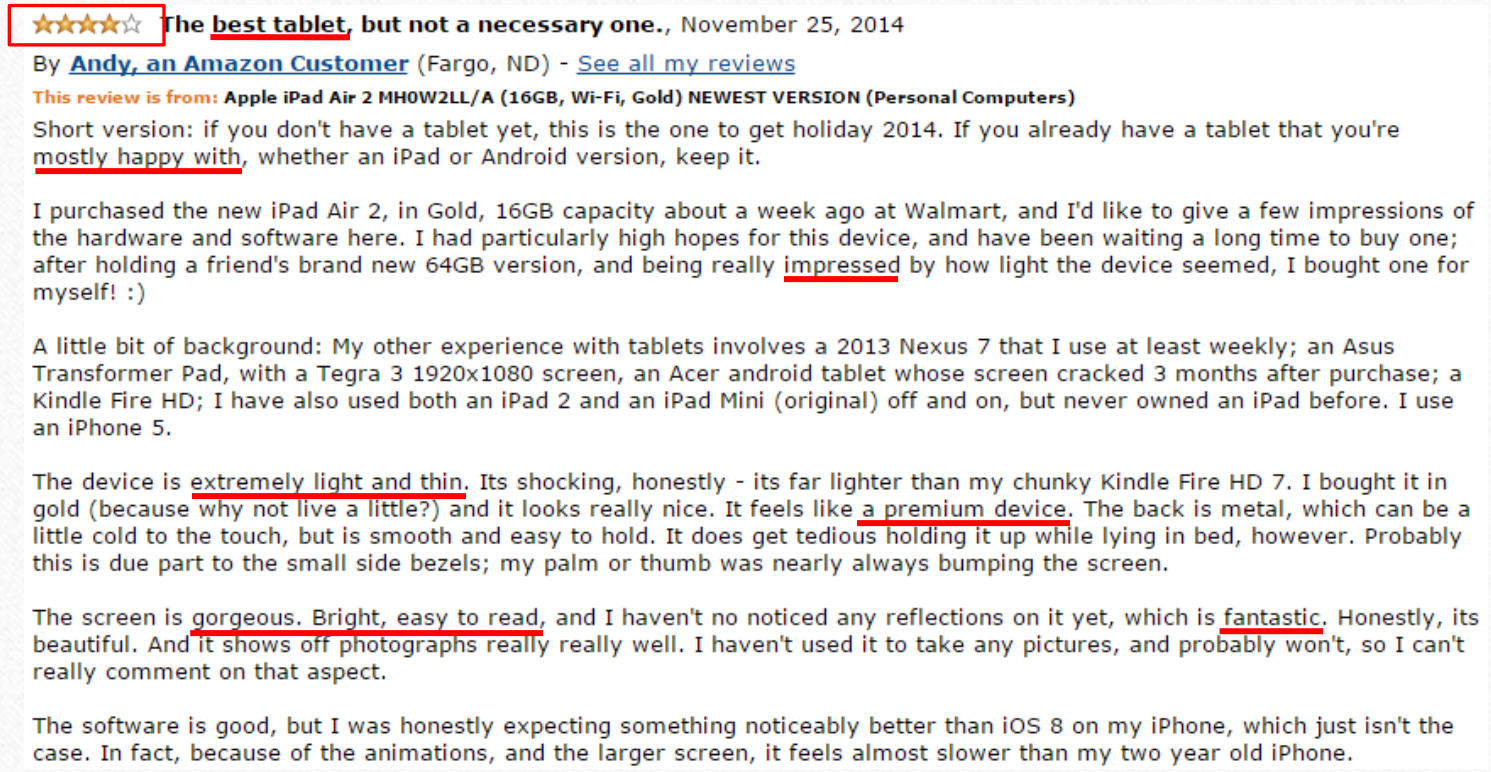
Last year's Super Bowl pitted one of the greatest single-season offenses in N.F.L. history against one of the greatest single-season defenses. Using slightly different time frames, this year's Super Bowl can boast similar claims.

The Seahawks' Richard Sherman intercepting a pass against the Packers in the N.F.C. title game.
David J. Phillip/Associated Press

Both the New England Patriots and the Seattle Seahawks had slow starts in 2014. After New England's 41-14 loss to the Kansas City Chiefs in Week 4, pundits wondered if we were witnessing the end of the Tom Brady/Bill Belichick-era Patriots. But since that game, the offensive line emerged as a cohesive unit, Rob Gronkowski's health improved and Brady became red-hot. Since that

Applications of text classification

- Sentiment analysis



★★★★☆ **The best tablet, but not a necessary one.**, November 25, 2014

By **Andy, an Amazon Customer** (Fargo, ND) - See all my reviews

This review is from: Apple iPad Air 2 MH0W2LL/A (16GB, Wi-Fi, Gold) NEWEST VERSION (Personal Computers)

Short version: if you don't have a tablet yet, this is the one to get holiday 2014. If you already have a tablet that you're mostly happy with, whether an iPad or Android version, keep it.

I purchased the new iPad Air 2, in Gold, 16GB capacity about a week ago at Walmart, and I'd like to give a few impressions of the hardware and software here. I had particularly high hopes for this device, and have been waiting a long time to buy one; after holding a friend's brand new 64GB version, and being really impressed by how light the device seemed, I bought one for myself! :)

A little bit of background: My other experience with tablets involves a 2013 Nexus 7 that I use at least weekly; an Asus Transformer Pad, with a Tegra 3 1920x1080 screen, an Acer android tablet whose screen cracked 3 months after purchase; a Kindle Fire HD; I have also used both an iPad 2 and an iPad Mini (original) off and on, but never owned an iPad before. I use an iPhone 5.

The device is extremely light and thin. Its shocking, honestly - its far lighter than my chunky Kindle Fire HD 7. I bought it in gold (because why not live a little?) and it looks really nice. It feels like a premium device. The back is metal, which can be a little cold to the touch, but is smooth and easy to hold. It does get tedious holding it up while lying in bed, however. Probably this is due part to the small side bezels; my palm or thumb was nearly always bumping the screen.

The screen is gorgeous. Bright, easy to read, and I haven't no noticed any reflections on it yet, which is fantastic. Honestly, its beautiful. And it shows off photographs really really well. I haven't used it to take any pictures, and probably won't, so I can't really comment on that aspect.

The software is good, but I was honestly expecting something noticeably better than iOS 8 on my iPhone, which just isn't the case. In fact, because of the animations, and the larger screen, it feels almost slower than my two year old iPhone.

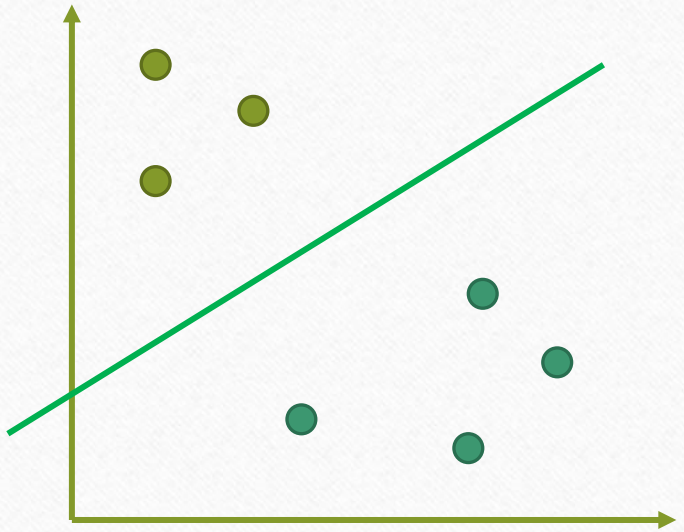
Basic notions about classification

- Data points/Instances
  - X: an m-dimensional feature vector

vector space representation

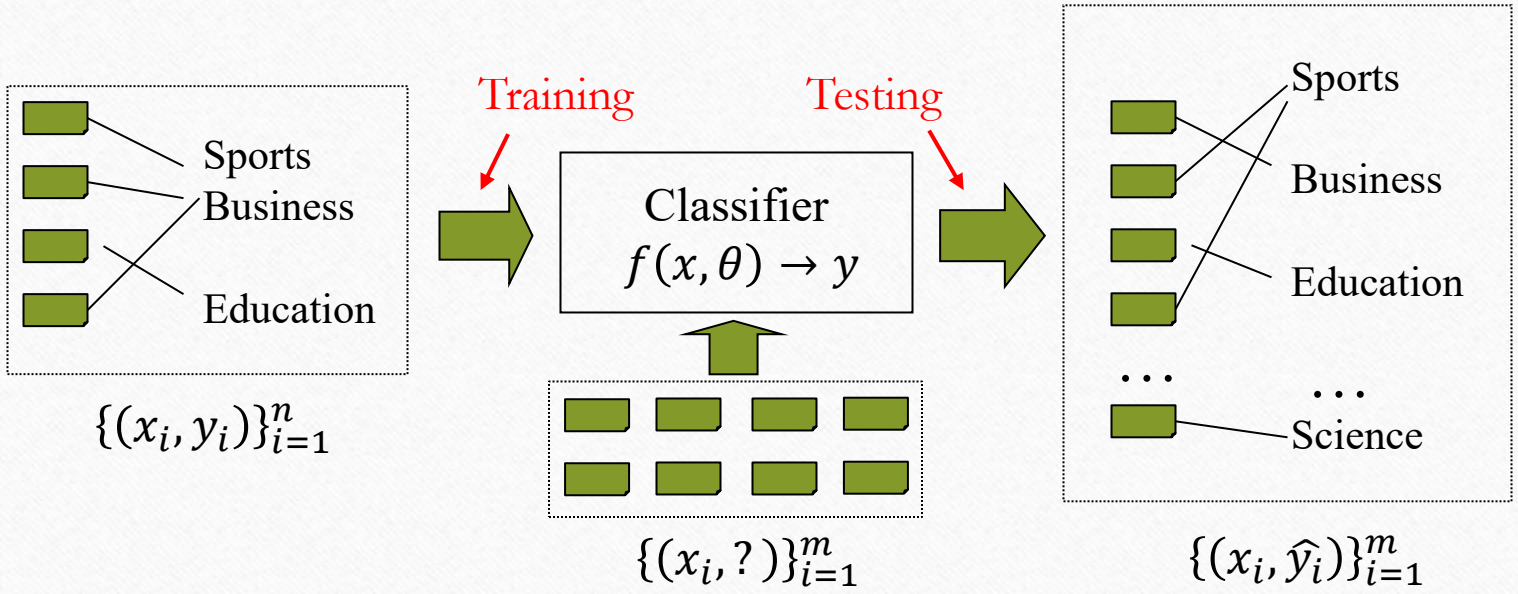
- Labels
  - y: a categorical value from $\{0, \ldots, k-1\}$
- Classification hyper-plane
  - $f(X) \rightarrow y$

Key question: how to find such a mapping?



Text classification

- Supervised learning
  - Estimate a model/method from labeled data
  - It can then be used to determine the labels of the unobserved samples



General steps for text classification

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email
Share
Tweet
Save
More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

Political News

Sports News

Entertainment News

1. Feature construction
2. Model specification
3. Model estimation and selection
4. Evaluation

General steps for text classification

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email
Share
Tweet
Save
More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

PLAY VIDEO

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

1. Feature construction
2. Model specification
3. Model estimation and selection
4. Evaluation

Political News

Sports News

Entertainment News

Consider:

1.1 How to represent the text documents?

1.2 Do we need all those features?

Feature construction for text categorization

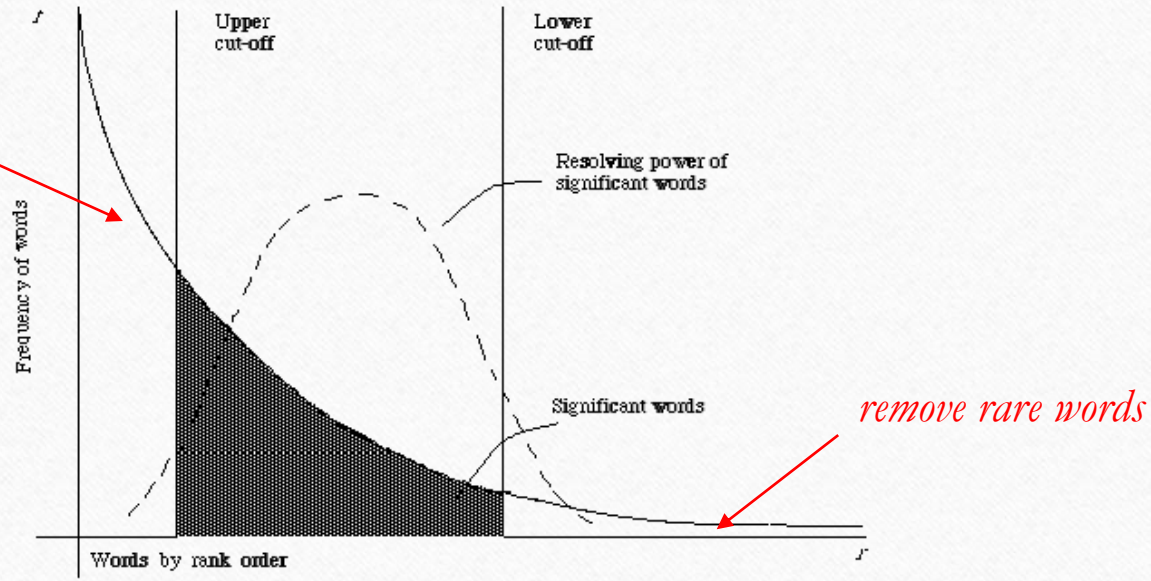
- Vector space representation
  - Standard procedure in document representation
  - Features
    - N-gram, POS tags, named entities, topics
  - Feature value
    - Binary (presence/absence)
    - TF-IDF (many variants)

Feature selection for text categorization

- Select the most informative features for model training
  - Reduce noise in feature representation
    - Improve final classification performance
  - Improve training/testing efficiency
    - Less time complexity
    - Fewer training data

Feature scoring metrics

- Document frequency
  - Rare words: non-influential for global prediction, reduce vocabulary size



remove head words

remove rare words

Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email
Share
Tweet
Save
More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

1. Feature construction
2. Model specification
3. Model estimation and selection
4. Evaluation

Political News

Sports News

Entertainment News

Consider:

2.1 What is the unique property of this problem?

2.2 What type of classifier we should use?

Model specification

- Specify dependency assumptions
  - Linear relation between x and y
    - $w^T x \rightarrow y$
    - Features are <u>independent</u> among each other
      - Naïve Bayes, linear SVM
  - Non-linear relation between x and y
    - $f(x) \rightarrow y$, where $f(\cdot)$ is a non-linear function of x
    - Features are <u>not independent</u> among each other
      - Decision tree, kernel SVM, mixture model
- Choose based on our domain knowledge of the problem

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email Share Tweet Save More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

Political News

Sports News

Entertainment News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

3.1 How to estimate the parameters in the selected model?

3.2 How to control the complexity of the estimated model?

Model estimation and selection

- General philosophy
  - Loss minimization

$$E[L] = L_{1,0}p(y=1)\int_{R_0} p(x)dx + L_{0,1}p(y=0)\int_{R_1} p(x)dx$$

Penalty when misclassifying c_1 to c_0

Penalty when misclassifying c_0 to c_1

Empirically estimated from training set

Empirical loss!

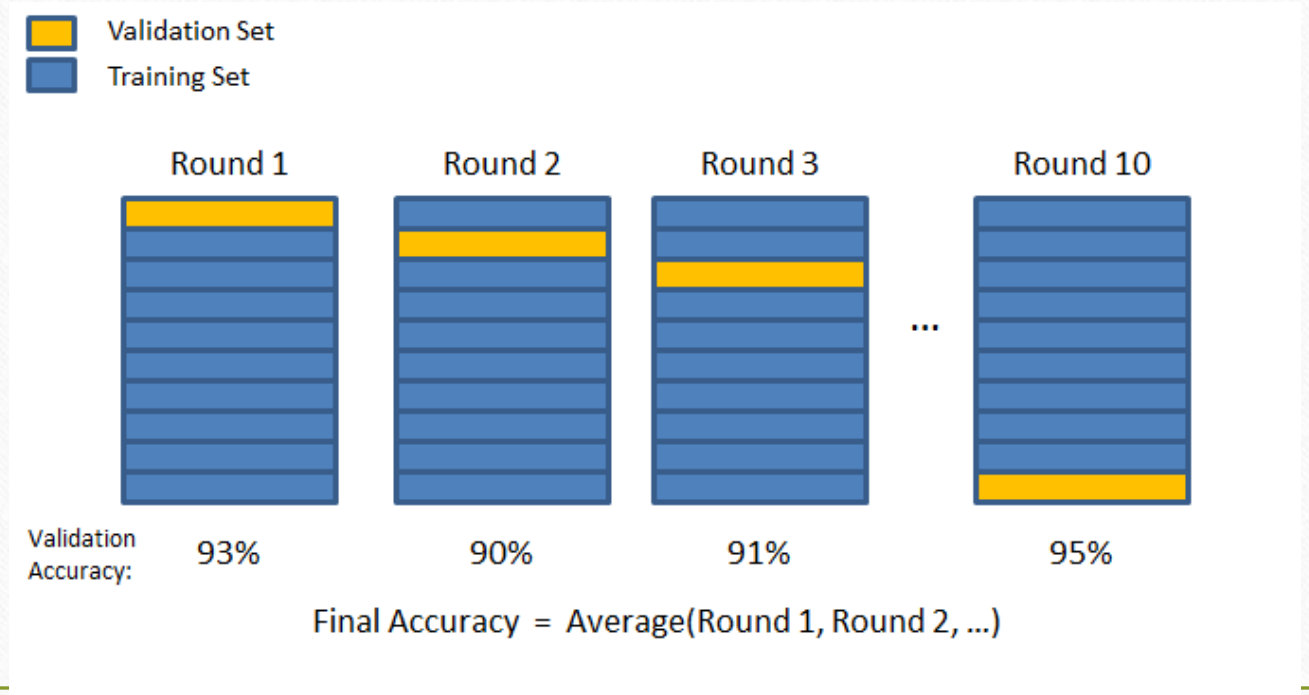
Key assumption: Independent and Identically Distributed!

Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation
    1. Partition all training data into *k* equal size disjoint subsets;
    2. Leave one subset for validation and the other *k*-1 for training;
    3. Repeat step (2) *k* times with each of the *k* subsets used exactly once as the validation data.

Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation



Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation
    - Choose the model (among different models or same model with different settings) that has the best average performance on the test sets

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email
Share
Tweet
Save
More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

Political News

Sports News

Entertainment News

1. Feature construction
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

4.1 How to judge the quality of learned model?

4.2 How can you further improve the performance?

Classification evaluation

- Accuracy
  - Percentage of correct prediction over all predictions, i.e., $p(y^* = y)$
  - Limitation
    - Highly skewed class distribution
      - $p(y^* = 1) = 0.99$
        - Trivial solution: all testing cases are positive
      - Classifiers' capability is only differentiated by 1% testing cases

Evaluation of binary classification

- Precision
  - Fraction of predicted positive documents that are indeed positive, i.e., $p(y^* = 1|y = 1)$
- Recall
  - Fraction of positive documents that are predicted to be positive, i.e., $p(y = 1|y^* = 1)$

	$y^* = 1$	$y^* = 0$
$y = 1$	true positive (TP)	false positive (FP)
$y = 0$	false negative (FN)	true negative (TN)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision and recall trade off

- Precision decreases as the number of documents predicted to be positive increases (unless in perfect classification), while recall keeps increasing
- These two metrics emphasize different perspectives of a classifier
  - Precision: prefers a classifier to recognize fewer documents, but highly accurate
  - Recall: prefers a classifier to recognize more documents

No.	Approach	Precision AVG	Precision STD	Recall AVG	Recall STD
1	Triple-S	0.31	0.19	0.36	0.26
2	BP Graph Matching	**0.60**	0.45	0.19	0.30
3	RefMod-Mine/NSCM	0.37	0.22	0.39	0.27
4	RefMod-Mine/ESGM	0.16	0.26	0.12	0.21
5	Bag-of-Words Similarity	0.56	0.23	0.32	0.28
6	PMLM	0.12	0.05	**0.58**	0.20
7	ICoP	0.36	0.24	0.37	0.26

Summarizing precision and recall

- With a single value
  - In order to compare different classifiers
  - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad \left(F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}\right)$$

Equal weight between precision and recall

  - Why harmonic mean?
    - Classifier1: P:0.53, R:0.36
    - Classifier2: P:0.01, R:0.99

F	Accuracy
0.429	0.445
0.019	0.500

Summarizing precision and recall

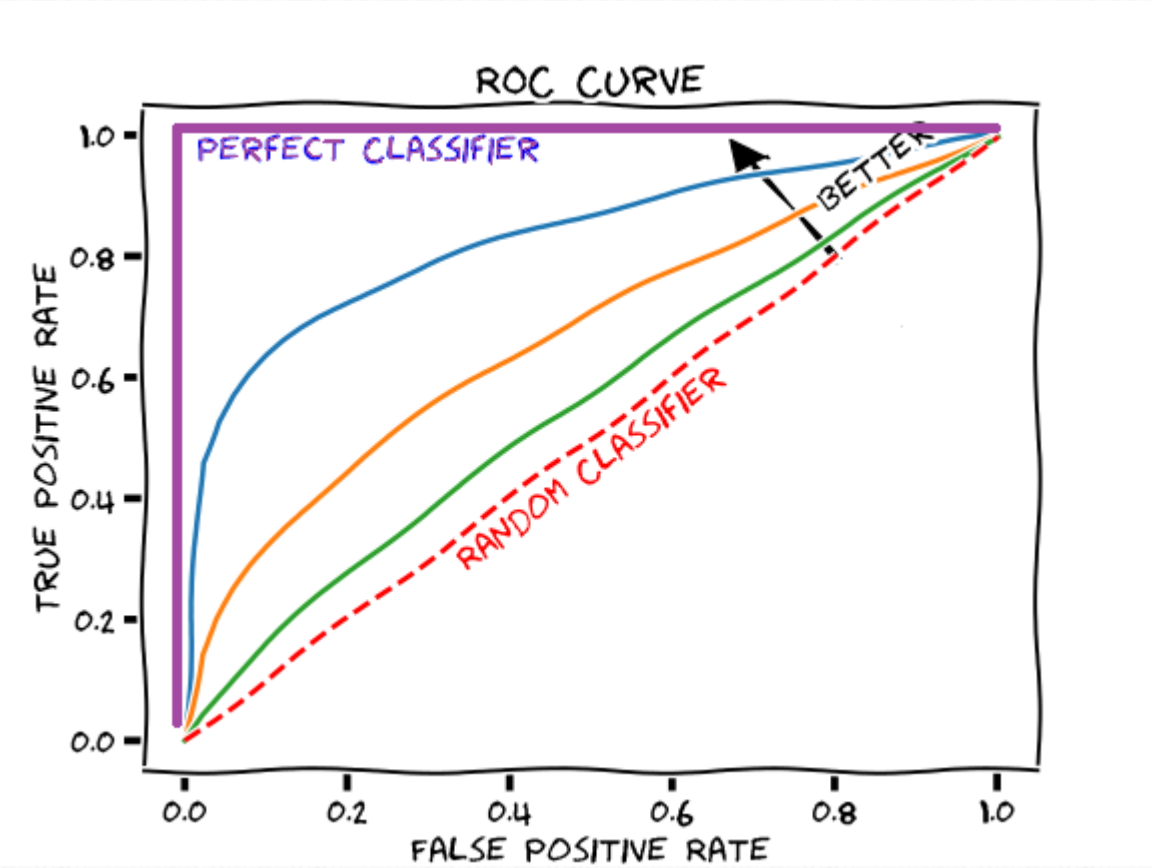
- With a curve – ROC and AUC
  - ROC curve (receiver operating characteristic curve)
  - True Positive rate (same as recall):

$$TPR = \frac{TP}{TP + FN}$$

  - False Positive rate:

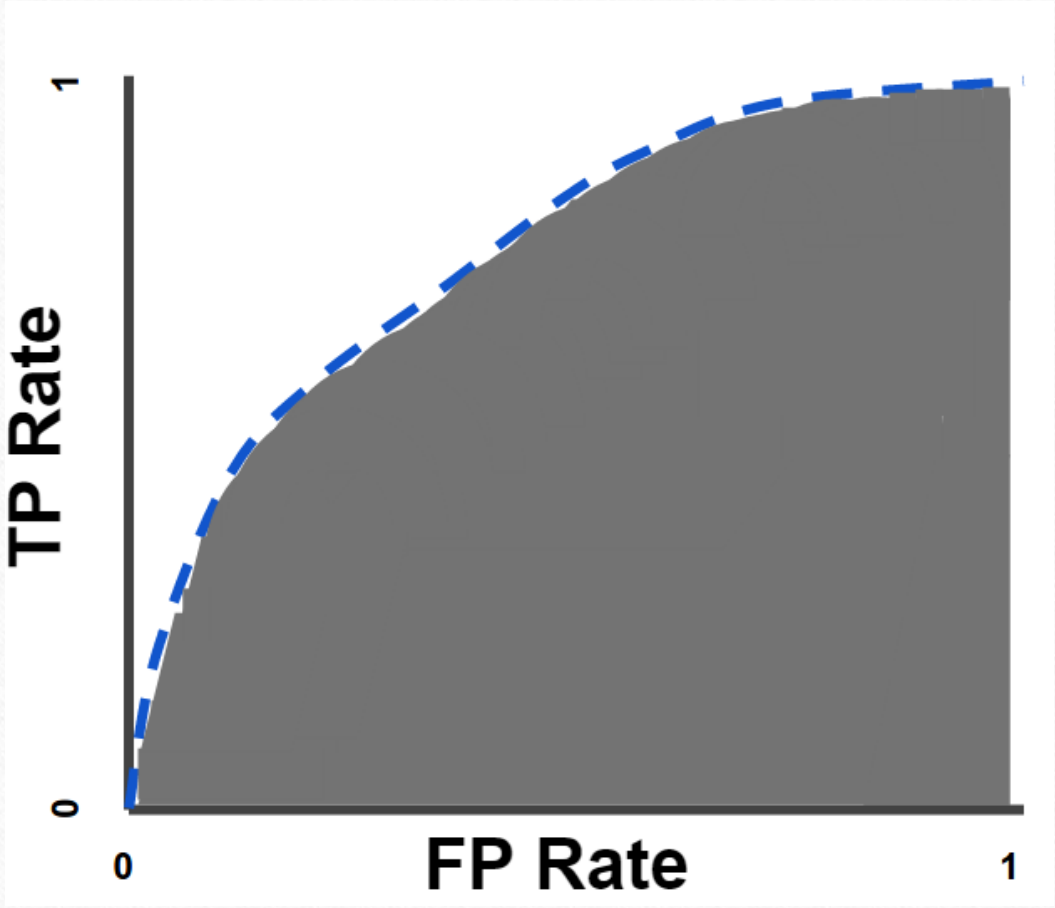
$$FPR = \frac{FP}{FP + TN}$$

	$y^* = 1$	$y^* = 0$
$y = 1$	TP	FP
$y = 0$	FN	TN



Summarizing precision and recall

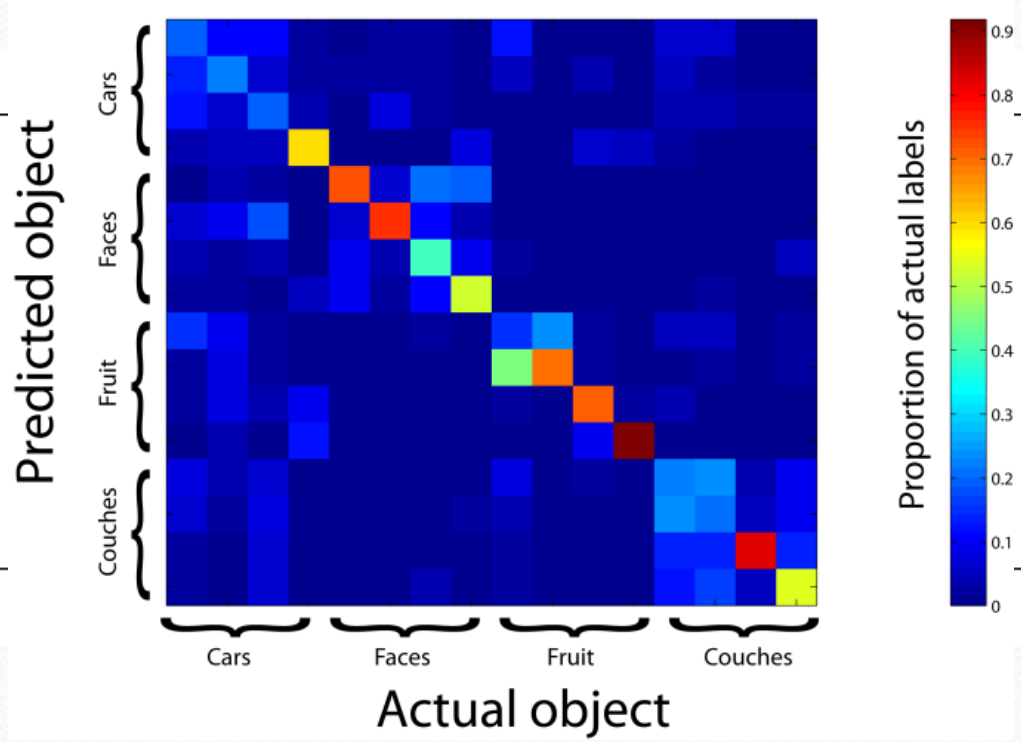
- With a curve – ROC and AUC
  - Area Under Curve (AUC)



Multi-class classification

- Confusion matrix
  - A generalized contingency table for precision and recall

Actual/ Predicted	Class 1
Class 1	**9.06**
Class 2	
Class 3	0.03
Class 4	0.01
Class 5	
Class 6	
Class 7	0.02
Class 8	0.20
Class 9	
Class 10	
Total	9.32
Precision	97.21

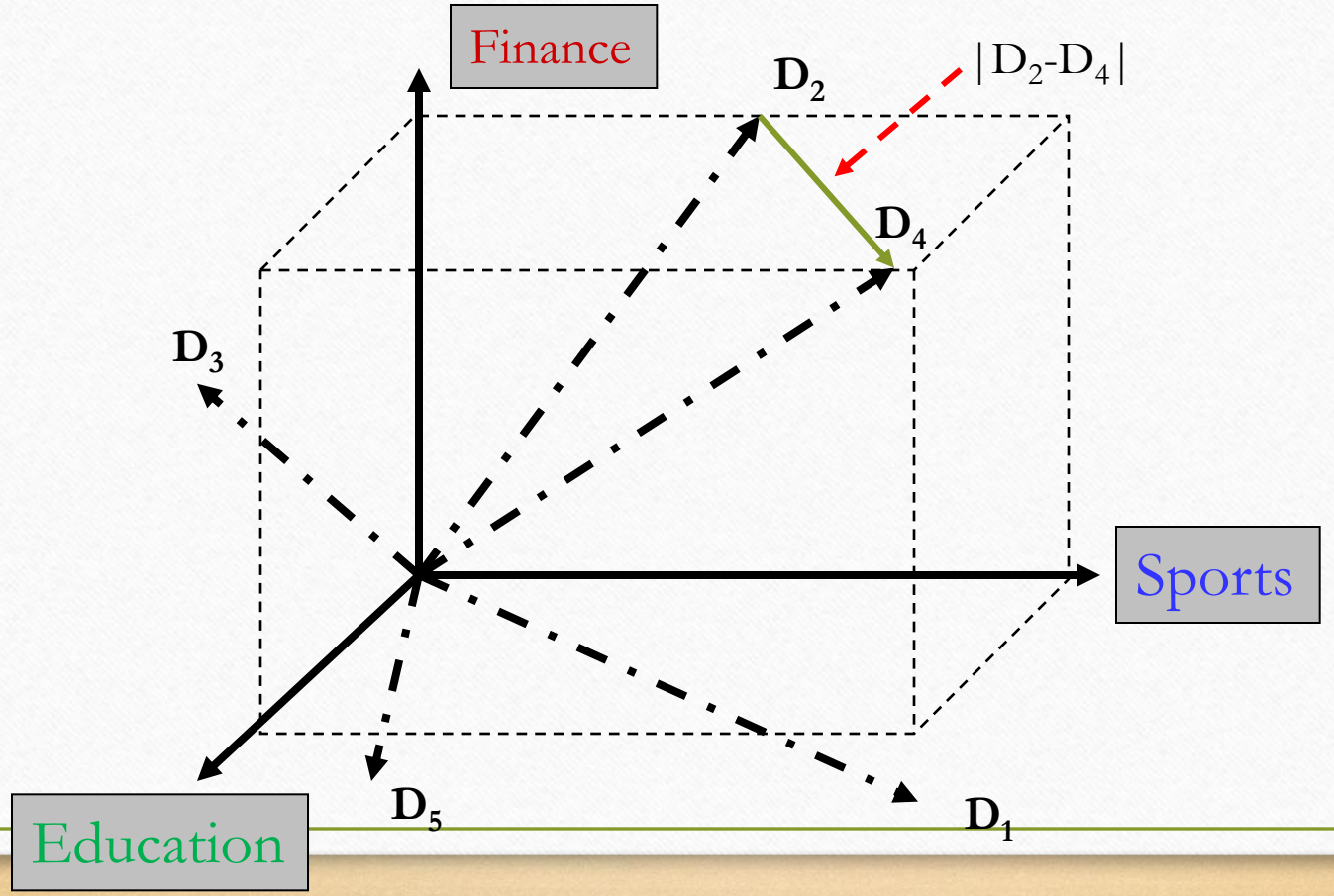


Total	Recall
10	90.60
10	82.00
10	95.20
10	90.10
10	26.70
10	77.50
10	36.60
10	97.00
10	68.40
10	53.70
100	

Topic detection

Vector space model

- All documents are projected into this concept space



Topic modelling - LDA

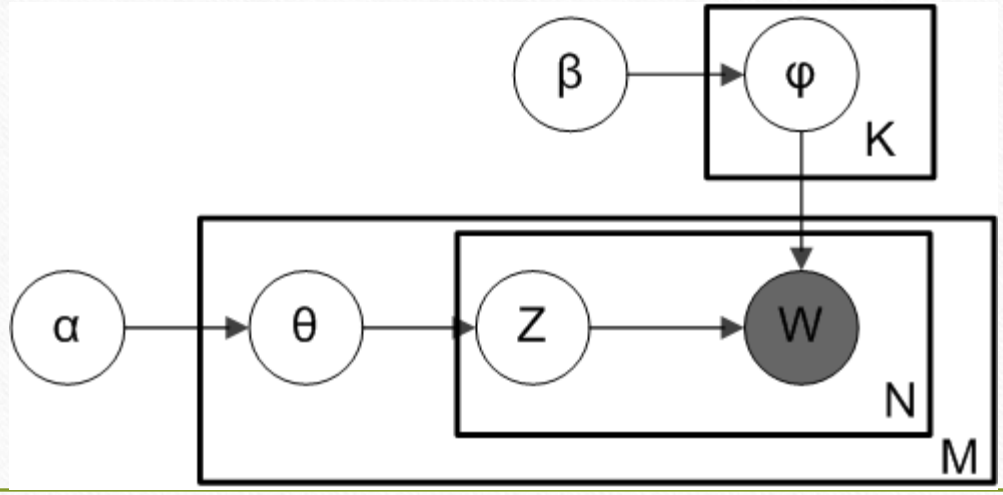
- LDA = Latent Dirichlet Allocation
- LDA = a form of unsupervised learning that views documents as bags of words (i.e. order does not matter).
- Key assumption:
  - the way a document was generated was by picking a set of topics, and then
  - for each topic picking a set of words.
- How does it find topics?
  - it reverse engineers this process.

Topic modelling - LDA

- For each document *m*:
  - Assume there are *k* topics across all of the documents
  - Distribute these *k* topics across document *m* (this distribution is known as α) by assigning each word a topic.
  - For each word *w* in document *m*, assume its topic is wrong but every other word is assigned the correct topic.
  - Probabilistically assign word *w* a topic based on two things:
    - what topics are in document *m*
    - how many times word *w* has been assigned a particular topic across all of the documents (this distribution is called *β)*
- Repeat this process a number of times for each document

Topic modelling - LDA

- Plate diagram of an LDA model where:
  α is the per-document topic distributions,
  β is the per-topic word distribution,
  θ is the topic distribution for document *m*,
  φ is the word distribution for topic *k*,
  z is the topic for the *n*-th word in document *m*, and
  w is the specific word



Acknowledgments

- Slides have been compiled from several sources:
- Hongning Wang, Lecture slides on Text Mining, University of Virginia, USA
- Jiawei Han, Micheline Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University (Data Mining: Concepts and Techniques 3rd ed.)
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd