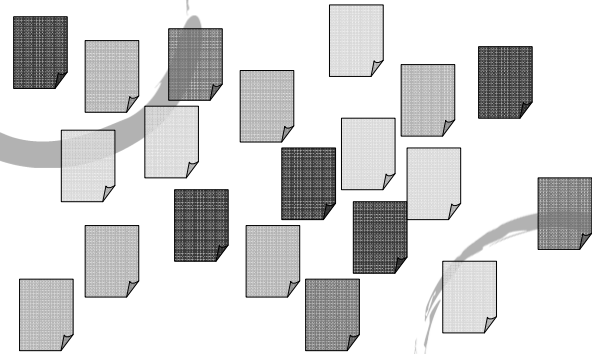


# The Ins and Outs of Clustering

Dawn J. Lawrie  
Loyola College in Maryland

## Goal of Clustering



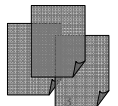
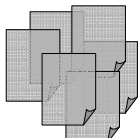
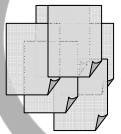
25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

2

## Goal of Clustering

- Create groups of similar documents
- Interested in topical similarity



25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

## Clustering Applications in Information Retrieval

- Cluster-based retrieval
- Search result Clustering
- Collection clustering for browsing
- Language modeling

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

4

## Foundation

- Cluster Hypothesis
  - Documents in the same cluster behave similarly with respect to relevance to information needs.

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

5

## Outline

- Background
- Two popular clustering algorithms
  - K-means
  - Hierarchical Agglomerative Clustering
- Alternative clustering

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

6

## Vector Space Model

- Define each document as a point in n-dimensional space
- Each vocabulary word in a collection represents a single dimension
- Consider "Its nice to be nice to the nice."  
~George Burns

Vector 

1	3	2	1	1
---	---	---	---	---

  
its nice to be the

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

7

## Distance Measures

- Key input into clustering algorithm
- Several ways to compute distance
  - Euclidean distance
  - Cosine Similarity Measure as a distance
- Different distance measures result in different clusterings

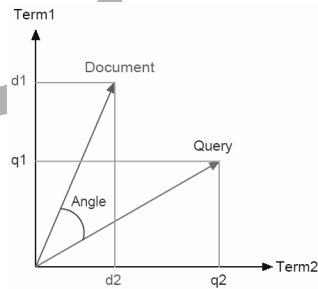
25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

8

## Cosine Similarity

$$\text{Sim}(D, Q) = \cos \theta = \frac{D \cdot Q}{\|D\| \|Q\|} = \frac{d_1 * q_1 + d_2 * q_2}{\sqrt{d_1^2 + d_2^2} \sqrt{q_1^2 + q_2^2}}$$



25/3/08

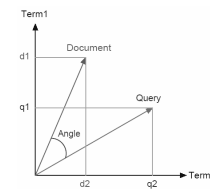
Dawn J. Lawrie  
Loyola College in Maryland

9

## Cosine Similarity

- Formula in N-Dimensional Space

$$\text{Sim}(X, Y) = \frac{\sum_i w_{X_i} w_{Y_i}}{\sqrt{\sum_i w_{X_i}^2} \sqrt{\sum_i w_{Y_i}^2}}$$



25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

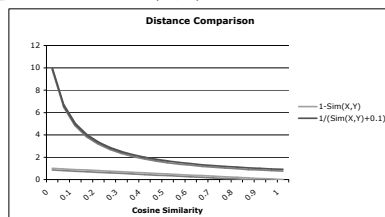
10

## Cosine Similarity

- As a distance formula

$$\text{Dist}(X, Y) = 1.0 - \text{Sim}(X, Y)$$

$$\text{Dist}(X, Y) = \frac{1.0}{\text{Sim}(X, Y) + C}$$



25/3/08

Loyola College in Maryland

11

## Determining Term Weights

- Recall the example vector

Vector	1	3	2	1	1
	its	nice	to	be	the

- Term weights not number of occurrences
  - Difficult to compare short and long documents
  - Long documents to have more occurrences of a term
  - More vs. less interesting terms

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

12

## TF\*IDF

- Term Frequency – Inverse Document Frequency
  - TF is importance of the term within the particular document
  - IDF is the general importance of the term
- Good is common in document, uncommon in collection
- Example Ad-hoc formula

```
tf_idf = (tf / (tf + 0.5 + 1.5 * (dl / avgDocLen))) *  
         (log(0.5 + (collection->num_docs / (double) term->df)) /  
         log(1.0 + log(collection->num_docs)));
```

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

13

## K-Means Clustering

- Minimizes the *distance* of documents from their cluster centroid
- Centroid vector is the mean of the document vectors

$$\bar{\mu}(C) = \frac{1}{|C|} \sum_{\vec{d} \in C} \vec{d}$$

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

14

## Worked Example

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

15

## K-Means Algorithm

- Minimizes the *residual sum of squares* or *RSS* over all documents

$$RSS_C = \sum_{\vec{d} \in C_k} |\vec{d} - \bar{\mu}(C_k)|^2$$

$$RSS = \sum_{k=1}^K RSS_k$$

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

16

## K-Means Algorithm

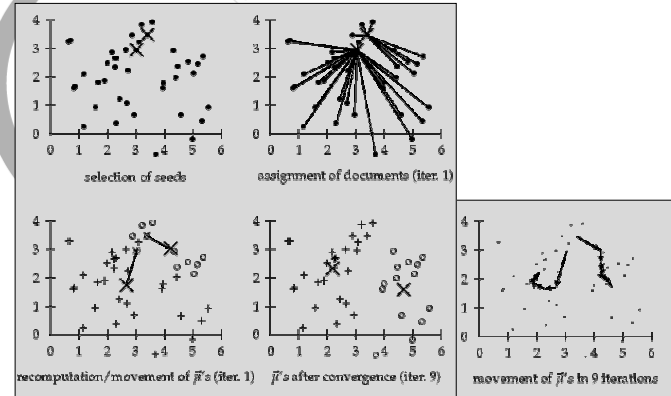
K-Means(D, K)

1. Select K random documents as seeds
2. Initialize centroids to selected documents
3. While not done
  - a) Set clusters to empty set
  - b) Assign each document to the nearest centroid
  - c) Recompute the centroid
4. Return sets of clusters

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

17



25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

18

## Pitfalls of K-Means

- Selecting a outlier document as a seed
  - Methods of avoiding
    - Run algorithm on several sets of seeds - pick lowest cost
    - Exclude outliers
    - Use another method to select seeds

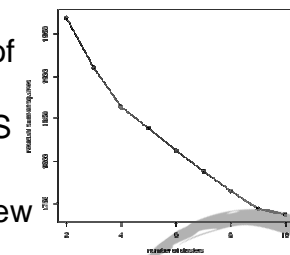
25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

19

## How many clusters?

- Try different numbers of clusters and find point where decrease in RSS becomes smaller
- Impose a penalty for new clusters



25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

20

## K-Means Summary

- Pros
  - Efficient
  - Simple
- Cons
  - cluster relationship?
  - How many clusters?
  - Nondeterministic

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

21

## Hierarchical Clustering

- Pros
  - Outputs a hierarchy
  - Doesn't need number of clusters in advanced
  - Deterministic (ones used in IR)
- Cons
  - Complexity quadratic in number of documents

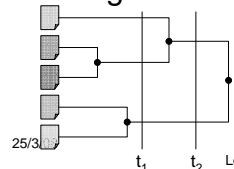
25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

22

## Hierarchical Agglomerative Clustering

- Bottom up approach
- Each document begins as singleton cluster
- Merge most similar clusters until there is a single cluster

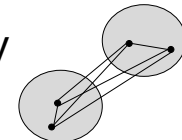


Dawn J. Lawrie  
Loyola College in Maryland

23

## Cluster Similarity

- Single link
  - Based on the most similar members
- Complete link
  - Based on the most dissimilar members
- Group-averaging
  - Based on all similarities between members
- Centroid
  - Based on the similarity of cluster centroids



25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

24

## Cluster Descriptions

- Want to describe topics of the clusters in a few words
- Difficult - any small list is likely to ignore some documents
- Reason - polythetic clusters
  - No document is required to have a particular word

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

25

## Alternative Clustering

- Monothetic clusters
  - Topic is required
- Labeling clusters are simple

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

26

## Monothetic Clusters

- Term Selection Formula

$$T^* = \arg \max_{T \in W} \sum_{i=1}^{|T|} P(\text{Top}(w_i), \text{Pred}(w_i) | w_1 \dots w_{i-1})$$

- Topicality
- Predictiveness

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

27

## Estimating Topicality

- Term's contribution to relative entropy
  - Compares document model to general English Model

$$\text{KL contribution}(w) = P_D(w) \log_2 \frac{P_D(w)}{P_{GE}(w)}$$

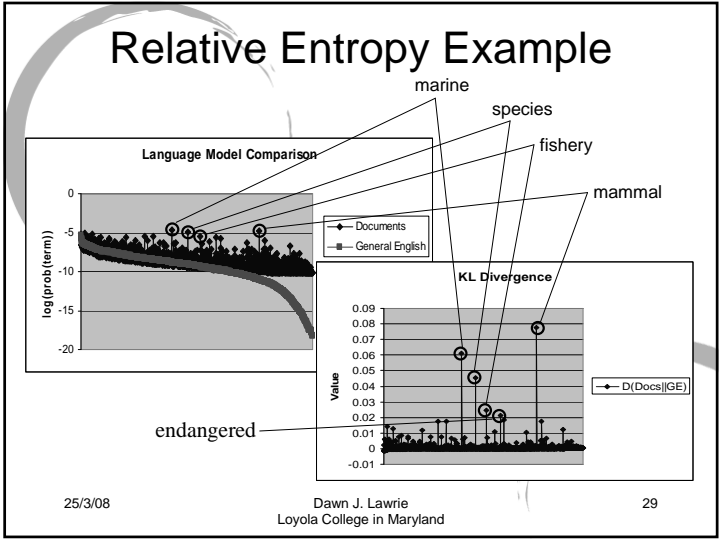
- Estimate unigram language model

$$P_D(w) = \frac{\text{occurrences}(w)}{\text{total words}}$$

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

28



## Identifying Predictive Words

- Topics co-occur with a distinct set of words (subtopics)
- Estimate
 
$$P(w_i | w_j) = \frac{\text{segments}_k(w_i \cap w_j)}{\text{occurrences}(w_j)}$$
  - $k$  is the maximum distance between  $w_i$  and  $w_j$

25/3/08 Dawn J. Lawrie Loyola College in Maryland 30

## Estimating Predictiveness

- Use heuristic to Dominating Set Problem

$$P(\text{Pred}(w_i | w_1 \dots w_{i-1})) = \frac{1}{|W_i|} \sum_{w \in W_i} P(w_i | w)$$

$$W_i = W_{i-1} - S_{i-1}$$

$$W_0 = W$$

The graph shows a network of vertices connected by edges. Some vertices are highlighted with circles, representing dominating sets. A legend below the graph explains the symbols: a circle with a dot for 'Vertex  $v_i$  and dominated vertices', a circle with a square for 'Vertex  $v_i$  and dominated vertices', and a circle with a triangle for 'Vertex  $v_i$  and dominated vertices'.

25/3/08 Dawn J. Lawrie Loyola College in Maryland 31

## Example Clustering

The screenshot shows a web interface for a topic hierarchy. The search query is 'endangered species mammals'. The results are displayed as a tree structure of topics. The 'ALL > mamma Mammals' path is highlighted. The resulting menu includes topics like 'Endangered Species Act - 56', 'Endangered Species Program - 6', 'Endangered Species Reports - 3', 'Endangered Mammal Species - 1', 'different species - 1', 'rare mammals - 1', 'World's Rarest Mammals - 1', 'Texas Threatened - 1', 'Society - 2', 'minors - 2', and 'primarily - 2'. Other related topics include 'Endangered Species - 454', 'species - 482', 'mammals - 620', '<ALL> - 620', 'endangered - 405', '<ALL> - 405', 'endangered species list - 23', 'Threatened Species - 36', 'Endangered Mammals - 21', 'threatened - 136', 'Mammals species - 4', 'birds - 146', 'extinct species - 7', 'ENDANGERED SPECIES PROTECTION - 3', 'SchoolWorld Endangered Species Project - 2', and 'Endangered Marine Mammals - 1'.

25/3/08 Dawn J. Lawrie Loyola College in Maryland 32

## Summary

- Myriad of approaches to clustering in IR
- Most popular
  - K-Means
  - Hierarchical Agglomerative Clusterings
- Polythetic vs Monothetic clusters
- Best technique will depend on the problem needs

25/3/08

Dawn J. Lawrie  
Loyola College in Maryland

33