
NLP

Introduction to NLP

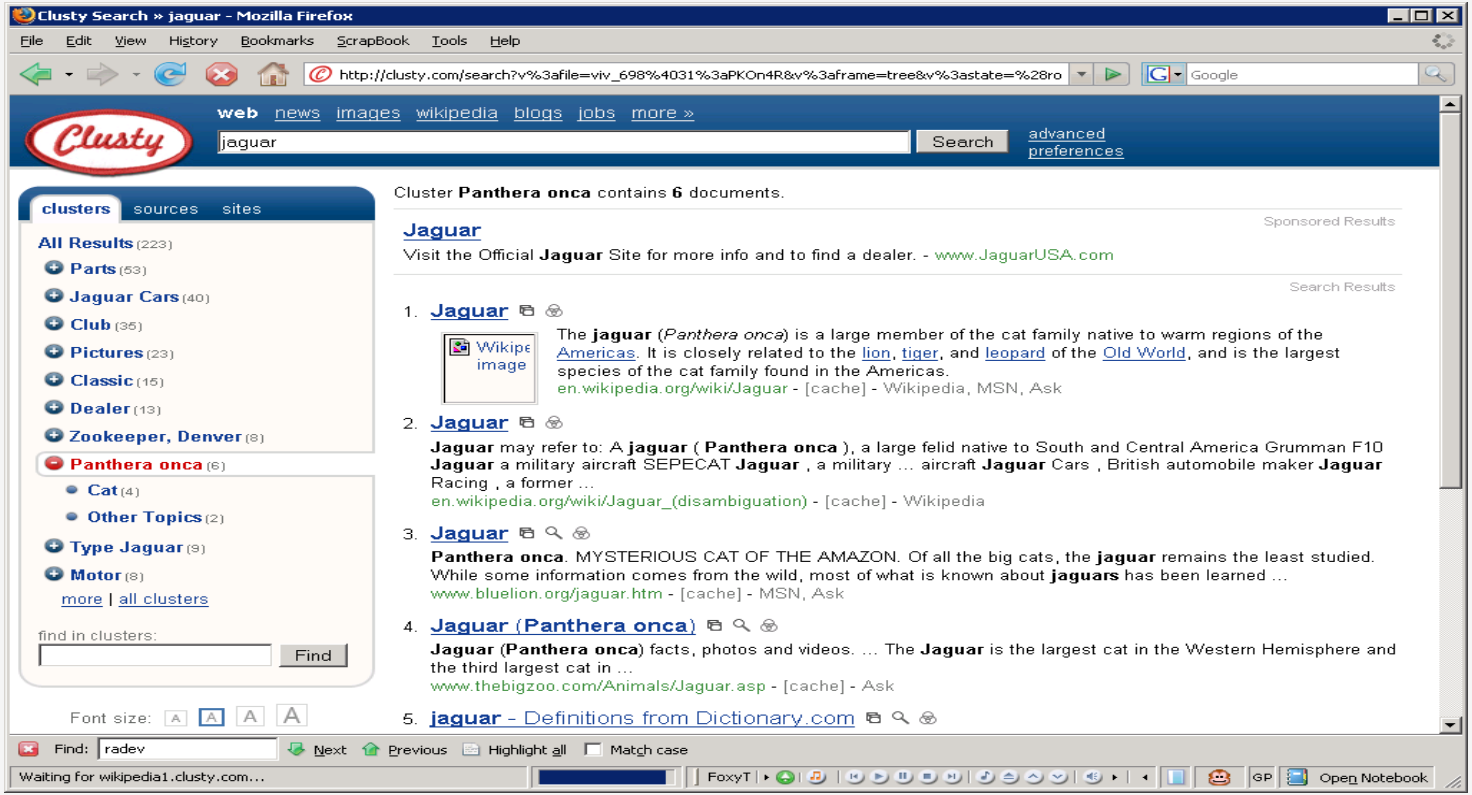
Text Clustering

Clustering

- Exclusive/overlapping clusters
- Hierarchical/flat clusters

- The cluster hypothesis
  - Documents in the same cluster are relevant to the same query
  - How do we use it in practice?

Example


Clusty Search » jaguar - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
http://clusty.com/search?v%3afile=viv_698%4031%3aPkOn4R&v%3aframe=tree&v%3astate=%28ro
Google
Clusty
web news images wikipedia blogs jobs more »
jaguar
Search
advanced preferences
clusters sources sites
All Results (223)
Parts (53)
Jaguar Cars (40)
Club (35)
Pictures (23)
Classic (15)
Dealer (13)
Zookeeper, Denver (8)
Panthera onca (6)
Cat (4)
Other Topics (2)
Type Jaguar (9)
Motor (8)
more | all clusters
find in clusters:
Find
Font size:
Cluster Panthera onca contains 6 documents.
Sponsored Results
Jaguar
Visit the Official Jaguar Site for more info and to find a dealer. - www.JaguarUSA.com
Search Results
1. Jaguar
Wikipe image
The jaguar (Panthera onca) is a large member of the cat family native to warm regions of the Americas. It is closely related to the lion, tiger, and leopard of the Old World, and is the largest species of the cat family found in the Americas.
en.wikipedia.org/wiki/Jaguar - [cache] - Wikipedia, MSN, Ask
2. Jaguar
Jaguar may refer to: A jaguar (Panthera onca), a large felid native to South and Central America Grumman F10 Jaguar a military aircraft SEPECAT Jaguar , a military ... aircraft Jaguar Cars , British automobile maker Jaguar Racing , a former ...
en.wikipedia.org/wiki/Jaguar_(disambiguation) - [cache] - Wikipedia
3. Jaguar
Panthera onca. MYSTERIOUS CAT OF THE AMAZON. Of all the big cats, the jaguar remains the least studied. While some information comes from the wild, most of what is known about jaguars has been learned ...
www.bluelion.org/jaguar.htm - [cache] - MSN, Ask
4. Jaguar (Panthera onca)
Jaguar (Panthera onca) facts, photos and videos. ... The Jaguar is the largest cat in the Western Hemisphere and the third largest cat in ...
www.thebigzoo.com/Animals/Jaguar.asp - [cache] - Ask
5. jaguar - Definitions from Dictionary.com
Find: radev
Next Previous Highlight all Match case
Waiting for wikipedia1.clusty.com...
FoxyT
GP
Open Notebook

k-means

- Iteratively determine which cluster a point belongs to, then adjust the cluster centroid, then repeat
- Needed: small number k of desired clusters
- hard decisions

k-means

1 initialize cluster centroids to arbitrary vectors
2 **while** further improvement is possible **do**
3 **for** each document *d* **do**
4 find the cluster *c* whose centroid is ***closest*** to *d*
5 assign *d* to cluster *c*
6 **end for**
7 **for** each cluster *c* **do**
8 recompute the centroid of cluster *c* based on its documents
9 **end for**
10 **end while**

Example

- Cluster the following vectors into two groups:
  - A = <1,6>
  - B = <2,2>
  - C = <4,0>
  - D = <3,3>
  - E = <2,5>
  - F = <2,1>

Demos

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- http://cgm.cs.mcgill.ca/~godfried/student_projects/bonnef_k-means
- http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster
- http://www.cc.gatech.edu/~dellaert/FrankDellaert/Software.html
- http://www-2.cs.cmu.edu/~awm/tutorials/kmeans11.pdf
- http://web.archive.org/web/20110223234358/http://www.ece.neu.edu/groups/rpl/projects/kmeans/

Evaluation of Clustering

- Purity
  - considering the majority class in each cluster
- RAND index
  - See next slide

Purity

- Three clusters

  XXXOO

  OOOX%

  %%%%XX

- Purity:
  - (3+3+4)/16=62.5%

Rand Index

- Accuracy when preserving object-object relationships.
- RI=(TP+TN)/(TP+FP+FN+TN)
- In the example:

$$TP + FP = \binom{5}{2} + \binom{5}{2} + \binom{6}{2} = 35$$

$$TP = \binom{3}{2} + \binom{3}{2} + \binom{4}{2} + \binom{2}{2} = 13$$

$$FP = 35 - 13 = 22$$

Rand Index

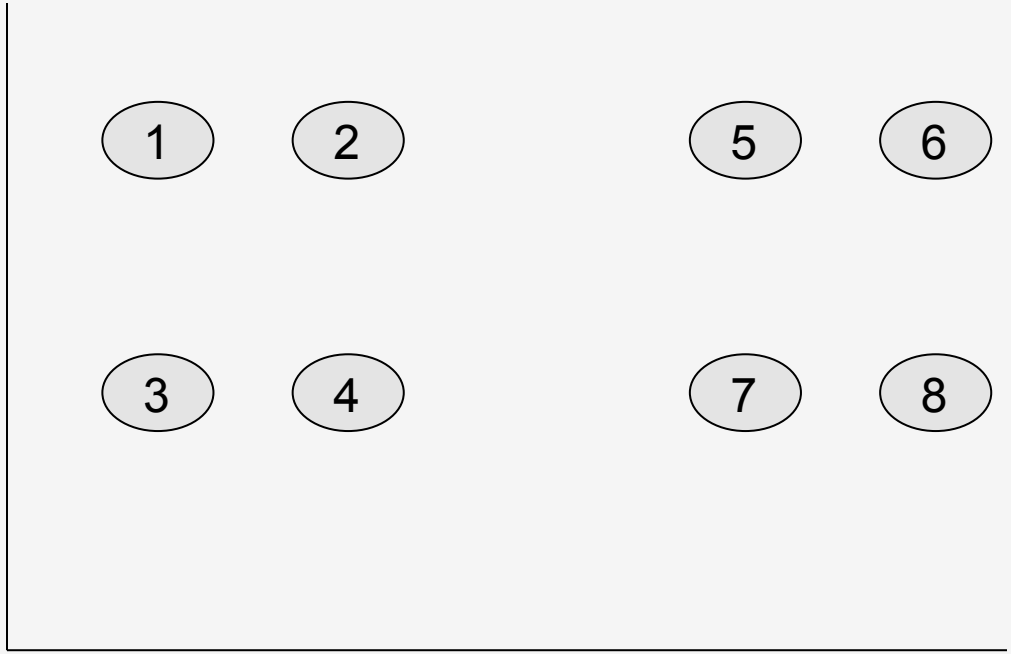
	Same cluster	
Same class	TP=13	FN=21
	FP=22	TN=64

RI = (TP+TN)/(TP+TN+FP+FN)=(13+64)/(13+64+22+21)=0.64

Hierarchical Clustering Methods

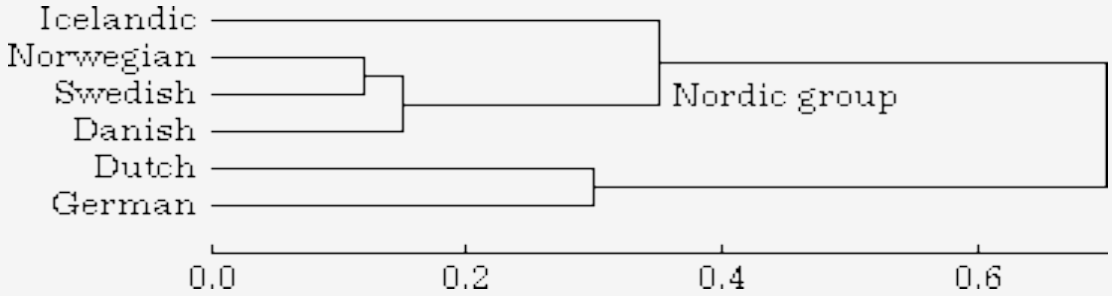
- Single-linkage
  - One common pair is sufficient
  - disadvantages: long chains
- Complete-linkage
  - All pairs have to match
  - Disadvantages: too conservative
- Average-linkage

Hierarchical Clustering



Hierarchical Agglomerative Clustering Dendrograms

E.g., language similarity:



http://odur.let.rug.nl/~kleiweg/clustering/clustering.html

Clustering Using Dendrograms

Example: cluster the following sentences:

A B C B A
A D C C A D E
C D E F C D A
E F G F D A
A C D A B A

REPEAT
- Compute pairwise similarities
- Identify closest pair
- Merge pair into single node

UNTIL only one node left

Q: what is the equivalent Venn diagram representation?


NLP