# CS347

Lecture 10

May 14, 2001

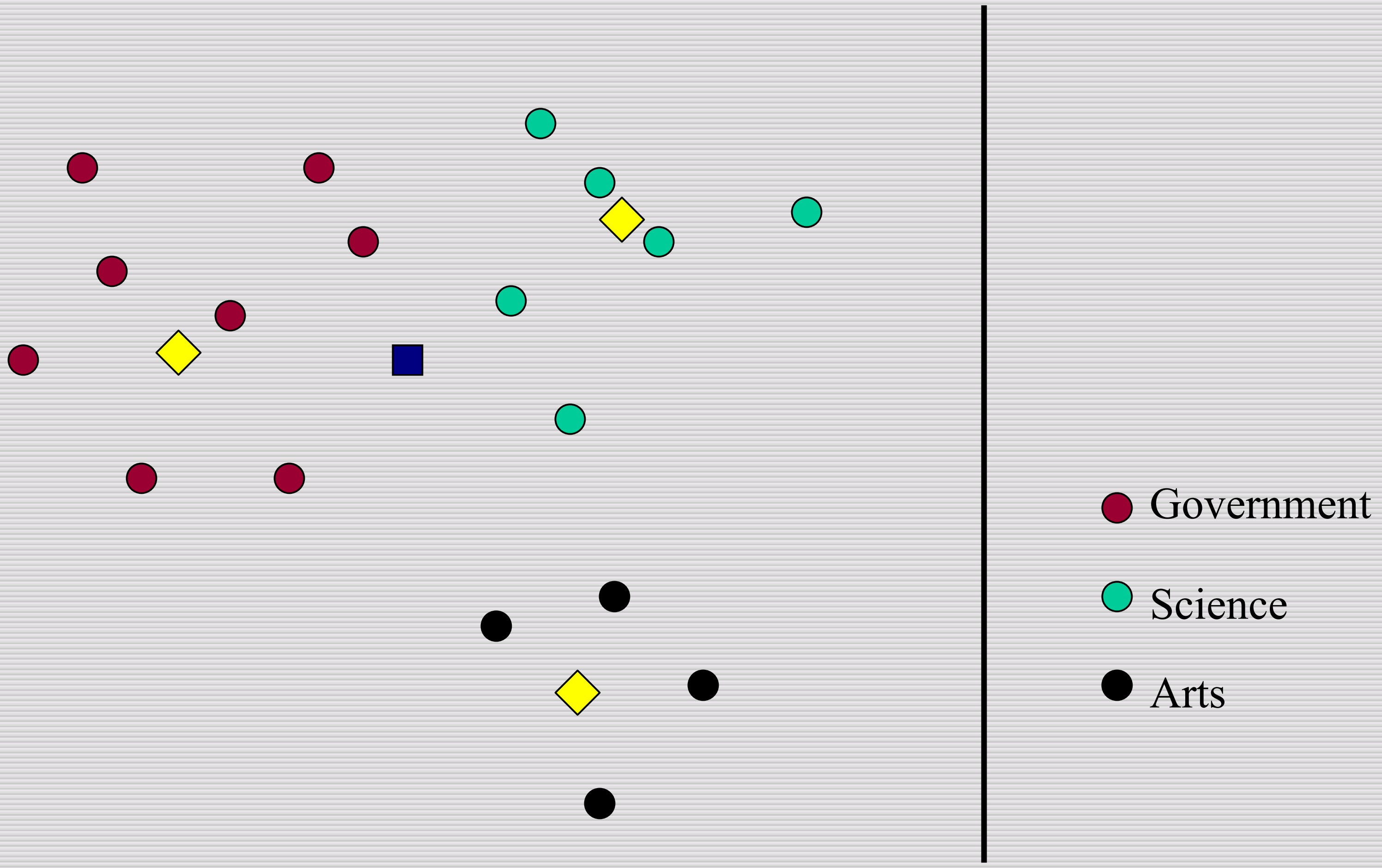
# Topics du jour

- Centroid/nearest-neighbor classification
- Bayesian Classification
- Link-based classification
- Document summarization

# Centroid/NN

- Given training docs for a topic, compute their centroid
- Now have a centroid for each topic
- Given query doc, assign to topic whose centroid is nearest.

# Example



# Bayesian Classification

- As before, train classifier on exemplary docs from classes $c_1, c_2, \ldots c_r$
- Given test doc $d$ estimate

$$\Pr[d \text{ belongs to class } c_j] = \Pr[c_j \mid d]$$

# Apply Bayes' Theorem

$$\Pr[c_j | d] \circ \Pr[d] = \Pr[d | c_j] \circ \Pr[c_j]$$

So $\Pr[c_j | d] = \dfrac{\Pr[d | c_j] \circ \Pr[c_j]}{\Pr[d]}$

Express $\Pr[d]$ as $\sum_{i=1}^{r} \Pr[d | c_i] \circ \Pr[c_i]$

# "Reverse Engineering"

- To compute $\Pr[c_j | d]$ , all we need are $\Pr[d | c_i]$ and $\Pr[c_i]$, for all $i$.
- Will get these from training.

# Training

Given a set of training docs, together with a class label for each training doc.

- e.g., these docs belong to Physics, those others to Astronomy, etc.

# Estimating $\Pr[c_i]$

$\Pr[c_i]$ = Fraction of training docs that are labeled $c_i$.

In practice, use more sophisticated “smoothing” to boost probabilities of classes under-represented in sample.

# Estimating $\Pr[d \mid c_i]$

Basic assumption - each occurrence of each word in each doc is <u>independent</u> of all others.

For a word *w*, (from sample docs)

$\Pr[w \mid c_i]$ = Frequency of word *w* amongst all docs labeled $c_i$ .

$$\Pr[d \mid c_i] = \prod_{w \in d} \Pr[w \mid c_i]$$

# Example

- Thus, the probability of a doc consisting of ***Friends, Romans, Countrymen*** =

$\Pr[\textbf{\textit{Friends}}] \circ \Pr[\textbf{\textit{Romans}}] \circ \Pr[\textbf{\textit{Countrymen}}]$

- In implementations, pay attention to precision/underflow.
- Extract all probabilities from term-doc matrix.

# To summarize

<u>Training</u>

- Use class frequencies in training data for $\Pr[c_i]$ .
- Estimate word frequencies for each word and each class to estimate $\Pr[w \mid c_i]$ .

<u>Test doc *d*</u>

- Use the $\Pr[w \mid c_i]$ values to estimate $\Pr[d \mid c_i]$ for each class $c_i$ .
- Determine class $c_j$ for which $\Pr[c_j \mid d]$ is maximized.

# Abstract features

- So far, have relied on word counts as the "features" to train and classify on.
- In general, could be any statistic.
  - terms in boldface count for more.
  - authors of cited docs.
  - number of equations.
  - square of the number of commas …
- "Abstract features".

# Bayesian in practice

- Many improvements used over “naïve” version discussed above
  - various models for document generation
  - varying emphasis on words in different portions of docs
  - smoothing statistics for infrequent terms
  - classifying into a hierarchy

# Supervised learning deployment issues

- Uniformity of docs in training/test
- Quality of authorship
- Volume of training data

# Typical empirical observations

- Training ~ 1000+ docs/class
- Accuracy

  upto 90% in the very best circumstances

  below 50% in the worst

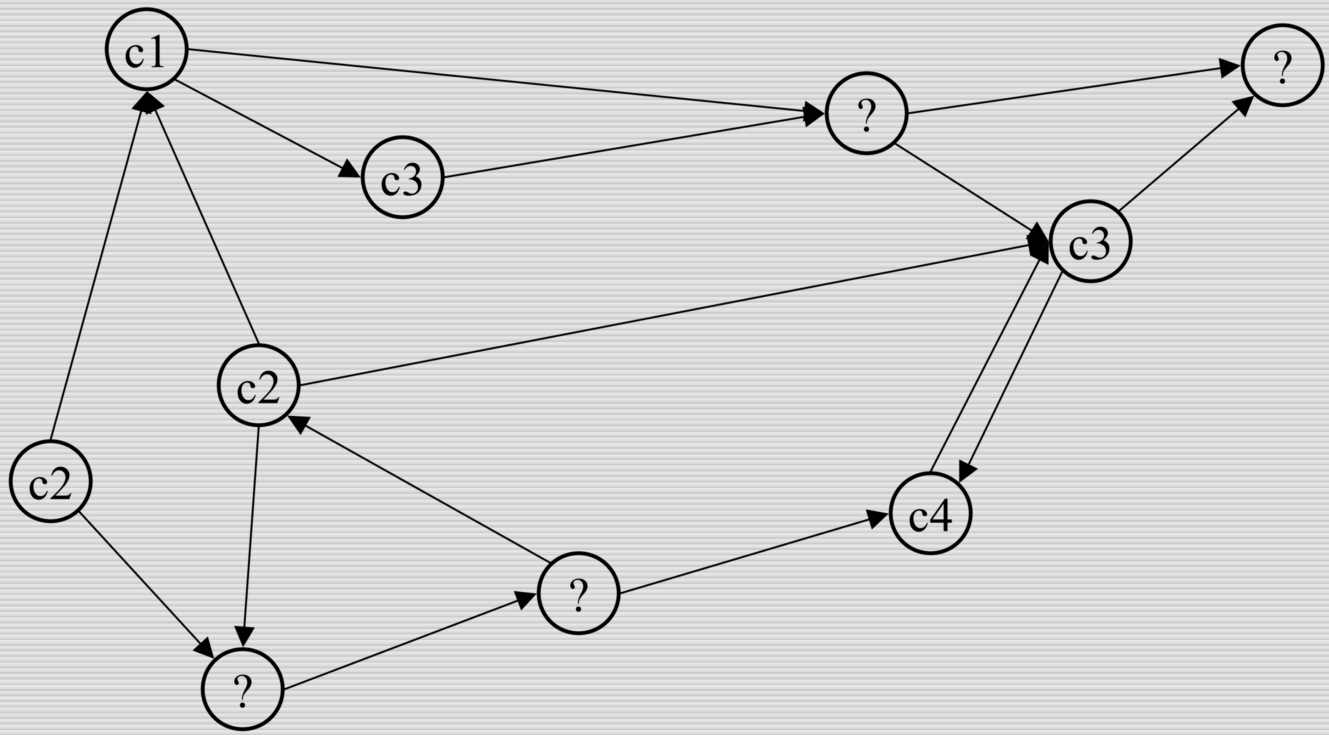
# SVM vs. Bayesian

- SVM appears to beat variety of Bayesian approaches
  - both beat centroid-based methods
- SVM needs quadratic programming
  - more computation than naïve Bayes at training time
  - less at classification time
- Bayesian classifiers also partition vector space, but not using linear decision surfaces

# Classifying hypertext

- Given a set of hyperlinked docs
- Class labels for some docs available
- Figure out class labels for remaining docs

# Example



# Bayesian hypertext classification

- Besides the terms in a doc, derive cues from linked docs to assign a class to test doc.
- Cues could be any abstract features from doc and its neighbors.

# Feature selection

- Attempt 1:
  - use terms in doc + those in its neighbors.
- Generally does worse than terms in doc alone. Why?
- Neighbors' terms diffuse focus of doc's terms.

# Attempt 2

- Use terms in doc, plus tagged terms from neighbors.
- E.g.,
  - ***car*** denotes a term occurring in *d*.
  - ***car@I*** denotes a term occurring in a doc with a link into *d*.
  - ***car@O*** denotes a term occurring in a doc with a link from *d*.
- Generalizations possible: ***car@OIOI***

# Attempt 2 also fails

- Key terms lose density
- e.g., ***car*** gets split into ***car, car@I, car@O***

# Better attempt

- Use class labels of (in- and out-) neighbors as features in classifying *d*.
  - e.g., docs about physics point to docs about physics.
- Setting: some neighbors have pre-assigned labels; need to figure out the rest.

# Content + neighbors' classes

- Naïve Bayes gives Pr[$c_j$|$d$] based on the words in $d$.
- Now consider Pr[$c_j$|$N$] where $N$ is the set of labels of $d$'s neighbors.

  (Can separate $N$ into in- and out-neighbors.)
- Can combine conditional probs for $c_j$ from text- and link-based evidence.

# Training

- As before, use training data to compute $\Pr[N|c_j]$ etc.
- Assume labels of *d*'s neighbors independent (as we did with word occurrences).
- (Also continue to assume word occurrences within *d* are independent.)

# Classification

- Can invert probs using Bayes to derive $\Pr[c_j|N]$.
- Need to know class labels for all of *d*'s neighbors.

# Unknown neighbor labels

- What if all neighbors' class labels are not known?
- First, use word content alone to assign a tentative class label to each unlabelled doc.
- Next, iteratively recompute all tentative labels using word content as well as neighbors' classes (some tentative).

# Convergence

- This iterative relabeling will converge provided tentative labels “not too far off”.
- Guarantee requires ideas from Markov random fields, used in computer vision.
- Error rates significantly below text-alone classification.

# End of classification

Move on to document summarization

# Document summarization

- Given a doc, produce a short summary.
- Length of summary a parameter.
  - Application/form factor.
- Very sensitive to doc quality.
- Typically, corpus-independent.

# Summarization

- Simplest algorithm: Output the first 50 (or however many) words of the doc.
- Hard to beat on high-quality docs.
- For very short summaries (e.g., 5 words), drop stop words.

# Summarization

- Slightly more complex heuristics:
- Compute an “importance score” for each sentence.
- Summary output contains sentences from original doc, beginning with the most “important”.

# Example

- **Article from WSJ**
  - 1: Tandon Corp. will introduce a portable hard-disk drive today that will enable personal computer owners to put all of their programs and data on a single transportable cartridge that can be plugged into other computers.
  - 2: Tandon, which has reported big losses in recent quarters as it shifted its product emphasis from disk drives to personal computer systems, asserts that the new device, called Personal Data Pac, could change the way software is sold, and even the way computers are used.
  - 3: The company, based in Moorpark, Calif., also will unveil a new personal computer compatible with International Business Machines Corp.'s PC-AT advanced personal computer that incorporates the new portable hard-disks.
  - 4: "It's an idea we've been working on for several years," said Chuck Peddle, president of Tandon's computer systems division.
  - 5: "As the price of hard disks kept coming down, we realized that if we could make them portable, the floppy disk drive would be a useless accessory.
  - 6: Later, we realized it could change the way people use their computers."
  - 7: Each Data Pac cartridge, which will be priced at about $400, is about the size of a thick paperback book and contains a hard-disk drive that can hold 30 million pieces of information, or the equivalent of about four Bibles.
  - 8: To use the Data Pacs, they must be plugged into a cabinet called the Ad-Pac 2.
  - That device, to be priced at about $500, contains circuitry to connect the cartridge to an IBM-compatible personal computer, and holds the cartridge steadily in place.
  - 9: The cartridges, which weigh about two pounds, are so durable they can be dropped on the floor without being damaged.
  - 10: Tandon developed the portable cartridge in conjunction with Xerox Corp., which supplied much of the research and development funding.
  - 11: Tandon also said it is negotiating with several other personal computer makers, which weren't identified, to incorporate the cartridges into their own designs.
  - 12: Mr. Peddle, who is credited with inventing Commodore International Ltd.'s Pet personal computer in the late 1970s, one of the first personal computers, said the Data Pac will enable personal computer users to carry sensitive data with them or lock it away.
  - ..........

# Example

- **1: Tandon Corp. will introduce a portable hard-disk drive today that will enable personal computer owners to put all of their programs and data on a single transportable cartridge that can be plugged into other computers.**
- **12: Mr. Peddle, who is credited with inventing Commodore International Ltd.'s Pet personal computer in the late 1970s, one of the first personal computers, said the Data Pac will enable personal computer users to carry sensitive data with them or lock it away.**

# Example

- **Article from WSJ**
  - **1: Tandon Corp. will introduce a portable hard-disk drive today that will enable personal computer owners to put all of their programs and data on a single transportable cartridge that can be plugged into other computers.**
  - **7: Each Data Pac cartridge, which will be priced at about $400, is about the size of a thick paperback book and contains a hard-disk drive that can hold 30 million pieces of information, or the equivalent of about four Bibles.**
  - **12: Mr. Peddle, who is credited with inventing Commodore International Ltd.'s Pet personal computer in the late 1970s, one of the first personal computers, said the Data Pac will enable personal computer users to carry sensitive data with them or lock it away.**

# Using Lexical chains

- Score each word based on lexical chains.
- Score sentences.
- Extract hierarchy of key sentences
  - Better (WAP) summarization

# What are Lexical Chains?

- Dependency Relationship between words
  - reiteration:

    e.g. tree - tree
  - superordinate:

    e.g. tree - plant
  - systematic semantic relation

    e.g. tree - bush
  - non- systematic semantic relation

    e.g. tree - tall

# Using lexical chains

- Look for chains of reiterated words
- Score chains
- Use to score sentences
  - determine how important a sentence is to the content of the doc.

# Computing Lexical Chains

- Quite simple if only dealing with reiteration.
  - Issues:
    - How far apart can 2 nodes in a chain be?
    - How do we score a chain?

# Example

- **Article from WSJ**
  - 1: Tandon Corp. will introduce a portable hard-disk drive today that will enable personal computer owners to put all of their programs and data on a single transportable cartridge that can be plugged into other computers.
  - 2: Tandon, which has reported big losses in recent quarters as it shifted its product emphasis from disk drives to personal computer systems, asserts that the new device, called Personal Data Pac, could change the way software is sold, and even the way computers are used.
  - 3: The company, based in Moorpark, Calif., also will unveil a new personal computer compatible with International Business Machines Corp.'s PC-AT advanced personal computer that incorporates the new portable hard-disks.
  - 4: "It's an idea we've been working on for several years," said Chuck Peddle, president of Tandon's computer systems division.
  - 5: "As the price of hard disks kept coming down, we realized that if we could make them portable, the floppy disk drive would be a useless accessory.
  - 6: Later, we realized it could change the way people use their computers."
  - 7: Each Data Pac cartridge, which will be priced at about $400, is about the size of a thick paperback book and contains a hard-disk drive that can hold 30 million pieces of information, or the equivalent of about four Bibles.
  - 8: To use the Data Pacs, they must be plugged into a cabinet called the Ad-Pac 2.
  - That device, to be priced at about $500, contains circuitry to connect the cartridge to an IBM-compatible personal computer, and holds the cartridge steadily in place.
  - 9: The cartridges, which weigh about two pounds, are so durable they can be dropped on the floor without being damaged.
  - 10: Tandon developed the portable cartridge in conjunction with Xerox Corp., which supplied much of the research and development funding.
  - 11: Tandon also said it is negotiating with several other personal computer makers, which weren't identified, to incorporate the cartridges into their own designs.
  - 12: Mr. Peddle, who is credited with inventing Commodore International Ltd.'s Pet personal computer in the late 1970s, one of the first personal computers, said the Data Pac will enable personal computer users to carry sensitive data with them or lock it away.
  - ……….

# Example

- For each word:
  - How far apart can 2 nodes in a chain be?

    Will continue chain if within 2 sentences.
  - How to score a chain?

    Function of word frequency and chain size.

# Chain Structure Analysis

- Compute & Score Chains

| | | | | | | |
|---|---|---|---|---|---|---|
| Chain 1 Score = 3674265 | | | | | | |
| PAC: | 12 | 13 | 14 | 15 | 17 | |
| Chain 2 Score = 2383334 | | | | | | |
| TANDON: | 1 | 2 | 4 | | | |
| Chain 3 Score = 1903214 | | | | | | |
| PORTABLE: | 1 | 3 | 5 | | | |
| Chain 4 Score = 1466674 | | | | | | |
| CARTRIDGE: | 7 | 8 | 9 | 10 | 11 | |
| Chain 5 Score = 959779 | | | | | | |
| COMPUTER: | 1 | 2 | 3 | 4 | 6 | 8 |
| Chain 6 Score = 951902 | | | | | | |
| TANDON: | 15 | 17 | | | | |
| Chain 7 Score = 951902 | | | | | | |
| TANDON: | 10 | 11 | | | | |
| Chain 8 Score = 760142 | | | | | | |
| PEDDLE: 12 | 14 | | | | | |
| Chain 9 Score = 726256 | | | | | | |
| COMPUTER: | 11 | 12 | 13 | 15 | 17 | |
| Chain 10 Score = 476633 | | | | | | |
| DATA: | 12 | 13 | 14 | 15 | 17 | |

# Sentence scoring

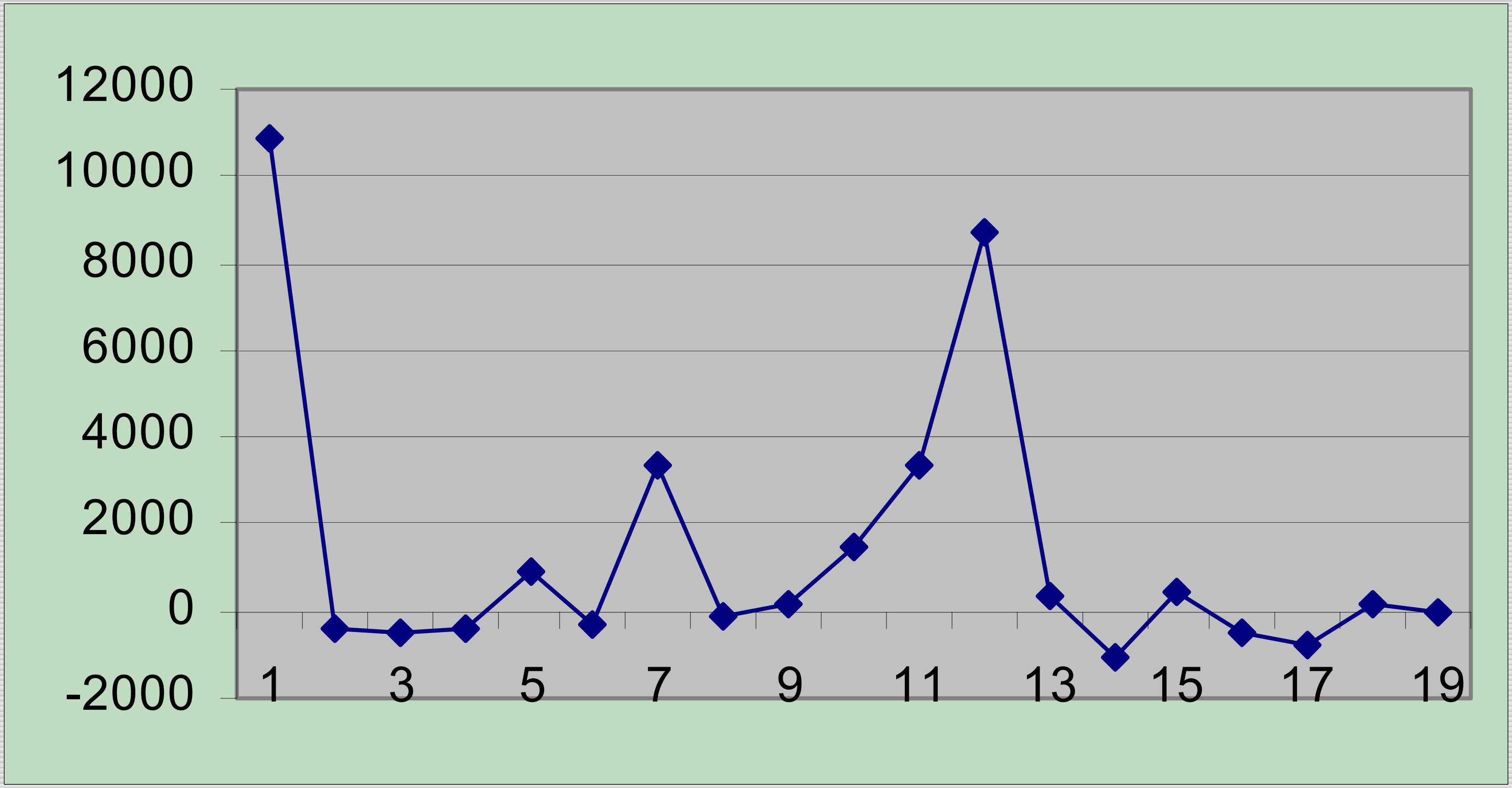
- For each sentence S in the doc.

  f(S) = a*h(S) – b*t(S)

  where h(S) = total score of all chains starting at S

  and t(S) = total score of all chains covering S, but not starting at S

# Semantic Structure Analysis



# Semantic Structure Analysis

- Derive the Document Structure

- 1
  - 1
    - 1
      - 1
      - 2
      - 3
      - 4
    - 5
      - 5
      - 6
  - 7
    - 7
    - 8
    - 9
    - 10
    - 11
  - 12
    - 12
      - 12
      - 13
      - 14
    - 15
      - 15
      - 16
      - 17
    - 18
      - 18
      - 19

# Resources

- S. Chakrabarti, B. Dom, P. Indyk. Enhanced hypertext categorization using hyperlinks.

*http://citeseer.nj.nec.com/chakrabarti98enhanced.html*

- R. Barzilay. Using lexical chains for text summarization.

*http://citeseer.nj.nec.com/barzilay97using.html*