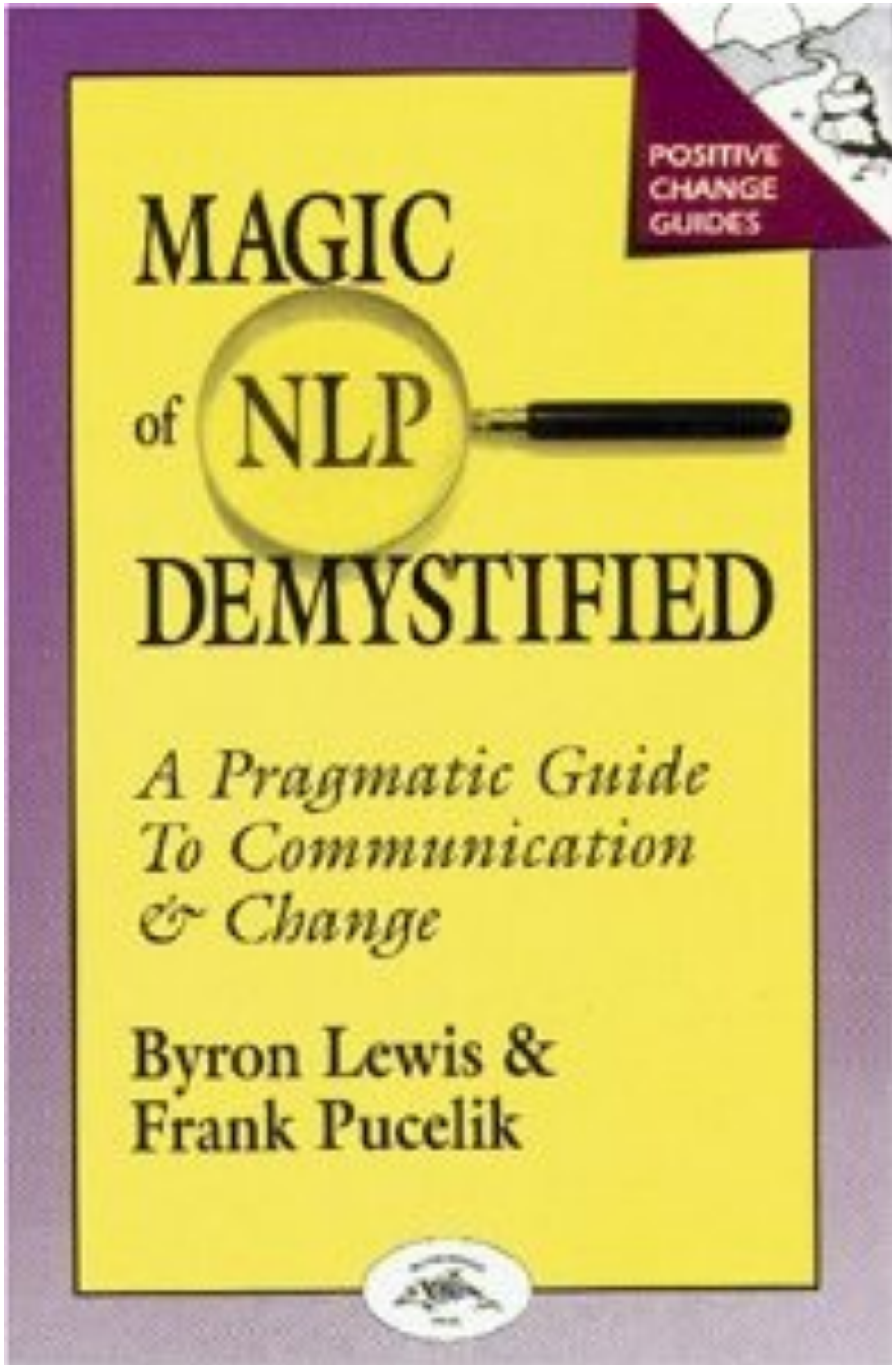
Text in the Data Pipeline

Noah Smith
Language Technologies Institute
Machine Learning Department
School of Computer Science
Carnegie Mellon University
nasmith@cs.cmu.edu


POSITIVE
CHANGE
GUIDES
MAGIC
of NLP
DEMYSTIFIED
A Pragmatic Guide
To Communication
& Change
Byron Lewis &
Frank Pucelik

NLP?

Outline

1. Automatically categorizing documents
2. Decoding sequences of words
3. Clustering documents and/or words

Categorizing Documents: Examples

- Mosteller and Wallace (1964): authorship of the *Federalist* papers
- News categories: U.S., world, sports, religion, business, technology, entertainment, ...
- How positive or negative is a review of a film or restaurant?
- Is a given email message spam?
- What is the reading level of a piece of text?
- How influential will a research paper be?
- Will a congressional bill pass committee?

The Vision

- Human experts label some data
- Feed the data to a learning algorithm that constructs an automatic labeling function
- Apply that function to as much data as you want!

Basic Recipe for Document Categorization

1. Obtain a pool of correctly categorized documents *D*.
2. Define a function $\boldsymbol{f}$ from documents to feature vectors.
3. Define a parameterized function $h_{\mathbf{w}}$ from feature vectors to categories.
4. Select *h*'s parameters **w** using a training sample from *D*.
5. Estimate performance on a held-out sample from *D*.

1. Obtain Categorized Documents

Spinoza, 17^{th} century rationalist

2. Define the Feature Vector Function

- Simplest choice: one dimension per word, and let $[\,\boldsymbol{f}(d)\,]_j$ be the count of w_j in d.
- Twists:
  - Monotonic transforms, like dividing by the length of d or taking a log.
  - Increase the weights of words that occur in fewer documents ("inverse document frequency")
  - n-grams
  - Count specially defined groupings of words 
  - Statistical tests to select words likely to be informative

Basic Recipe for Document Categorization

1. Obtain a pool of correctly categorized documents *D*.
2. Define a function ***f*** from documents to feature vectors.
3. Define a parameteriz , from feature vectors to cat
4. Select *h*'s parameters ining sample from *D*.
5. Estimate performanc ut sample from *D*.

3. Define a Function from Feature Vectors to Categories

- Simplest choice: linear model

$$h_{\mathbf{w}}(d) = \arg\max_{c} \mathbf{w}_c^\top \boldsymbol{f}(d) + w_c^{bias}$$

w$_c$ is the vector of coefficients associating each feature with class *c* (can be positive or negative).

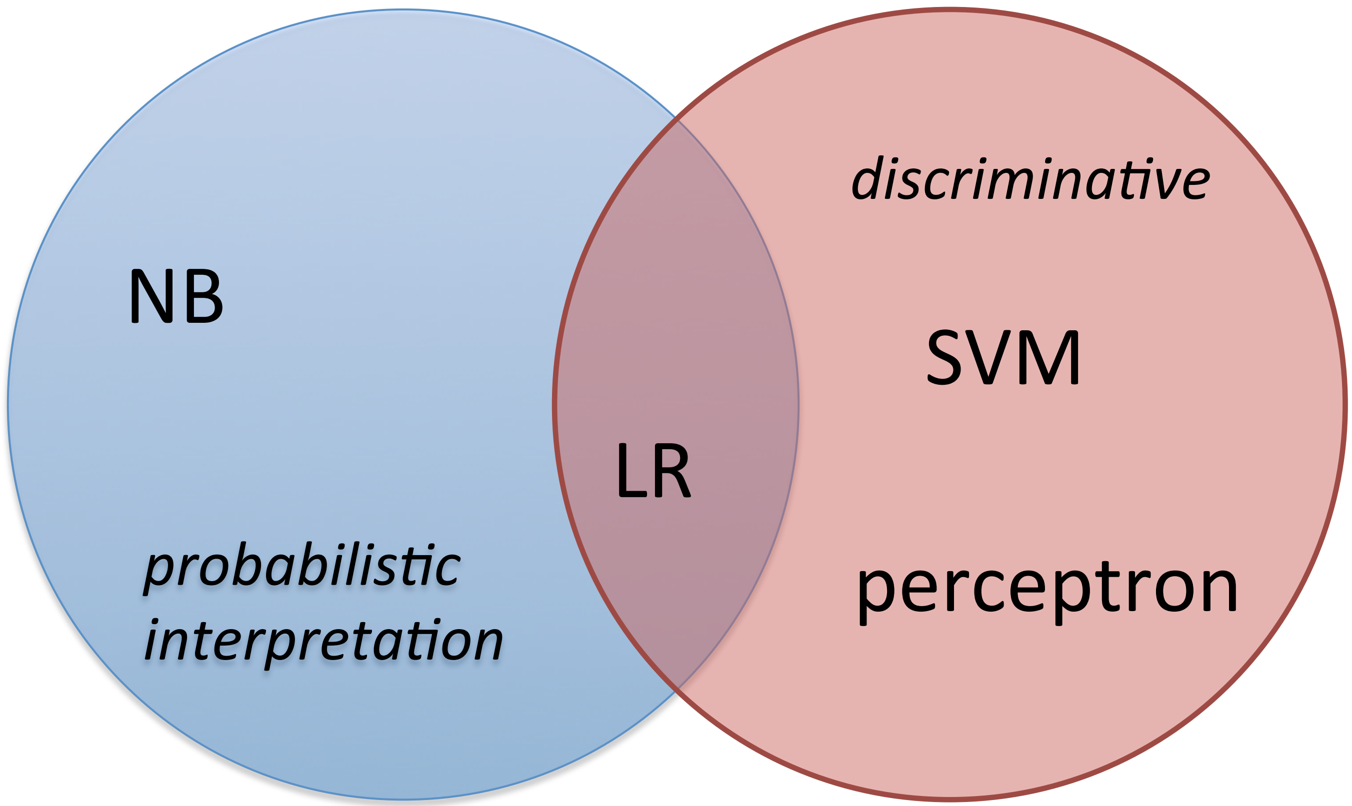
  – Advantage: interpretability

  – Advantage: computational efficiency

- Some alternatives: k-nearest neighbors, decision trees, neural networks, ...

4. Select Parameters using Data

- Also known as "machine learning."
- Many learning options for linear classifiers!



4. Select Parameters using Data

Optimization view of learning:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} R(\mathbf{w}) + \frac{1}{|D_{train}|} \sum_{d \in D_{train}} L(d; \mathbf{w})$$

"regularization" to avoid overfitting

"empirical risk" = average loss over training data

Typical loss functions for linear models are **convex** and can be efficiently optimized using online or batch iterative algorithms with convergence guarantees.

4. Select Parameters using Data

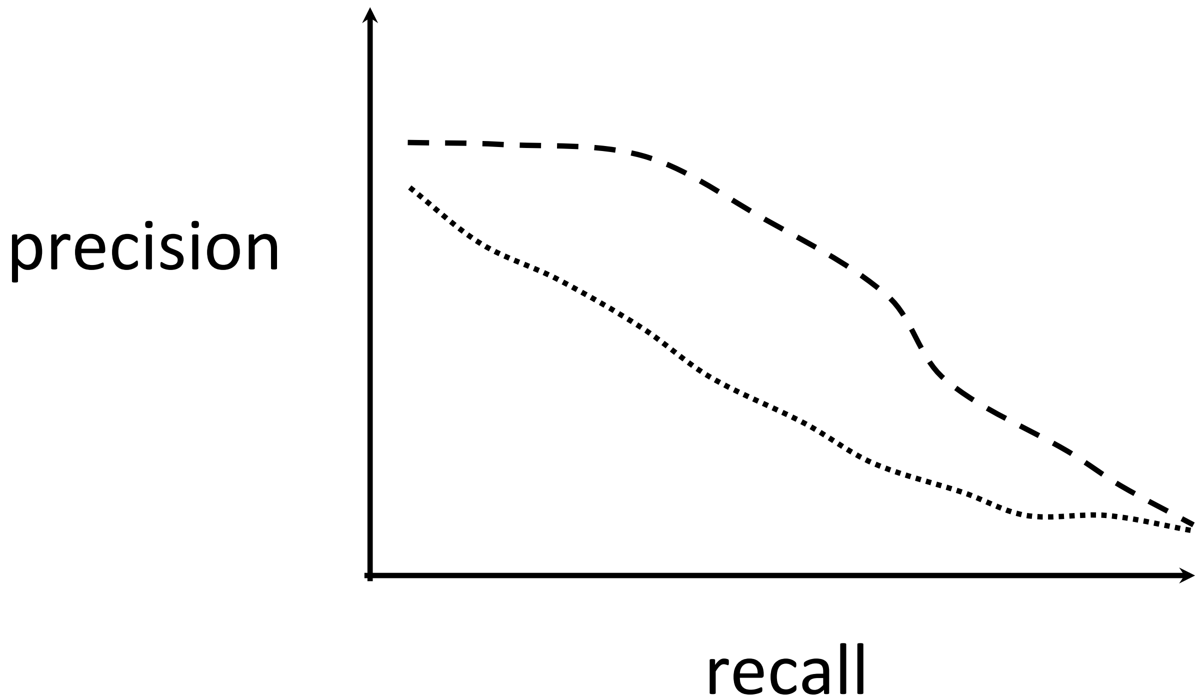
Considerations:

- Do you want posterior probabilities, or just labels?
- What methods do you understand well enough to explain in your paper?
- What methods will your "customers" understand?
- What implementations are available?
  - Cost, scalability, programming language, compatibility with your workflow, ...
- How well does it work (on held-out data)?

5. Estimate Performance

- Always, always, always use **held-out** data.
  - Multiple rounds of tests? Fresh testing data!
- Consider the "most frequent class" baseline.
- Consider inter-annotator agreement.
- What to measure?
  - Accuracy
  - When one class is special: precision/recall

5. Estimate Performance



$$h_{\mathbf{w}}(d) = \arg\max_{c} \mathbf{w}_c^\top \boldsymbol{f}(d) + w_c^{bias}$$

Basic Recipe for Document Categorization

1. Obtain a pool of correctly categorized documents *D*.
2. Define a function ***f*** from documents to feature vectors.
3. Define a parameterized function $h_{\mathbf{w}}$ from feature vectors to categories.
4. Select *h*'s parameters **w** using a training sample from *D*.
5. Estimate performance on a held-out sample from *D*.

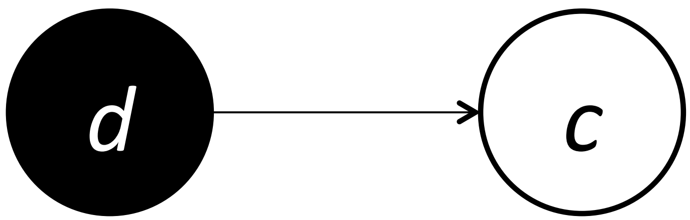
Outline

✓ Automatically categorizing documents
2. Decoding sequences of words
3. Clustering documents and/or words

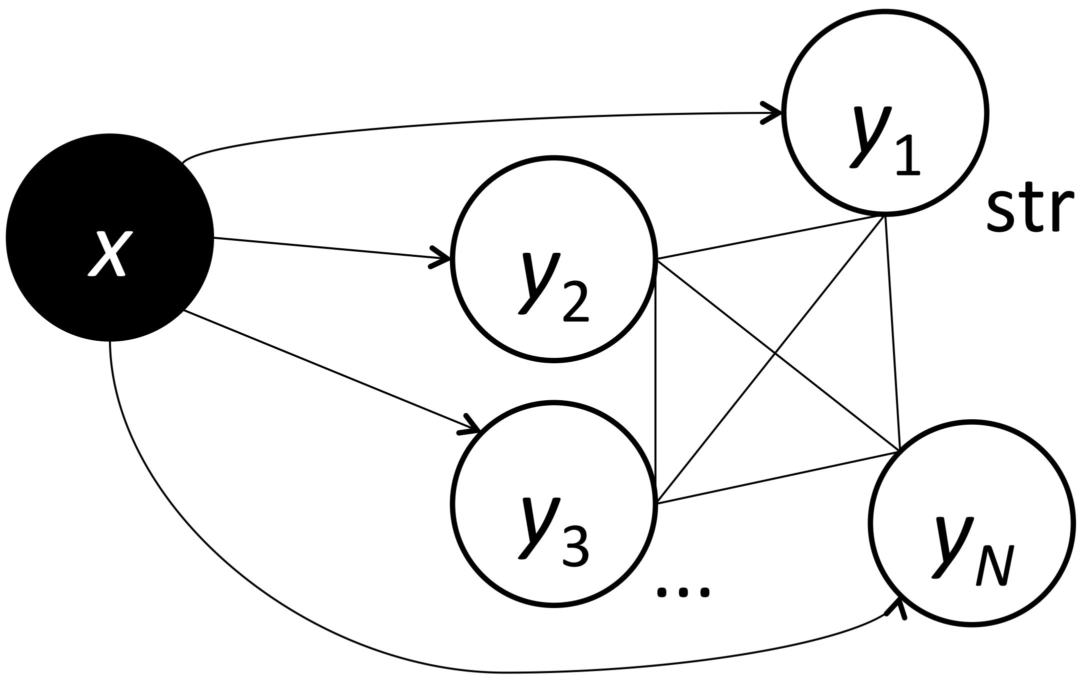
Decoding Word Sequences: Examples

- Categorizing each word by its part-of-speech or semantic class
- Recognizing mentions of named entities
- Segmenting a document into parts
- Parsing a sentence into a grammatical or semantic structure

High-Level View



classification



structured prediction

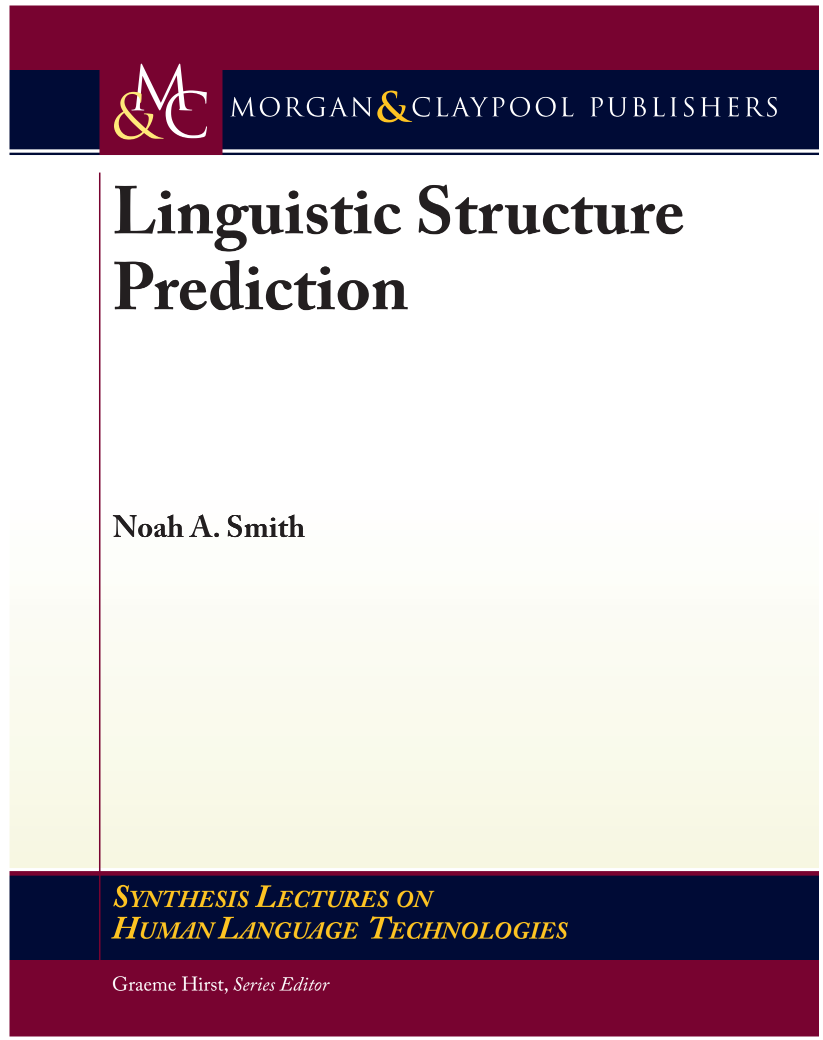
Possible Lines of Attack

1. Transform into a sequence of classification problems (see part 1).
2. Transform into a sequence labeling problem and use a variant of the Viterbi algorithm.
3. Design a representation, prediction algorithm, and learning algorithm for your particular problem.

Shameless Self-Promotion

$56.02 on amazon.com

free in electronic form, through CMU's library



Lines of Attack

1. Reduce to a sequence of classification problems (see part 1).
2. Reduce to a sequence labeling problem and use a variant of the Viterbi algorithm.
3. Design a representation, prediction algorithm, and learning algorithm for your problem.

Sequence Labeling

- Input: sequence of symbols $x_1\ x_2 \ldots x_L$
- Output: sequence of labels $y_1\ y_2 \ldots y_L$ each $\in \Lambda$

Prediction rule:

$$h_{\mathbf{w}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathbf{w}^\top \boldsymbol{f}(x_1 \ldots x_L, y_1 \ldots y_L)$$

Problem: there are $O(|\Lambda|^L)$ choices for $y_1\ y_2 \ldots y_L$!

Sequence Labeling with Local Features

A key assumption about $\boldsymbol{f}$ allows us to solve the problem exactly, in O(|Λ|^2 L) time and O(|Λ|L) space.

$$
\begin{aligned}
h_{\mathbf{w}}(\boldsymbol{x}) &= \arg\max_{\boldsymbol{y}} \mathbf{w}^\top \boldsymbol{f}(x_1 \ldots x_L, y_1 \ldots y_L) \\
&= \arg\max_{\boldsymbol{y}} \mathbf{w}^\top \left(\sum_{\ell=1}^{L-1} \boldsymbol{f}_{local}(x_1 \ldots x_L, y_\ell y_{\ell+1}) \right)
\end{aligned}
$$

If I knew the best label sequence for $x_1 \ldots x_{L-1}$, then y_L would be easy.

That decision would depend only on state $L - 1$.

$$y_L^* = \arg\max_{y_L \in \Lambda} \mathbf{w}^\top \left(\sum_{\ell=1}^{L-2} \boldsymbol{f}_{local}(x_1 \ldots x_L, y_\ell^* y_{\ell+1}^*) \right) + \mathbf{w}^\top \boldsymbol{f}_{local}(x_1 \ldots x_L, y_{L-1}^* y_L)$$
$$= \mathbf{w}^\top \left(\sum_{\ell=1}^{L-2} \boldsymbol{f}_{local}(x_1 \ldots x_L, y_\ell^* y_{\ell+1}^*) \right) + \arg\max_{y_L \in \Lambda} \mathbf{w}^\top \boldsymbol{f}_{local}(x_1 \ldots x_L, y_{L-1}^* y_L)$$

I don't know that best sequence, but there are only $|\Lambda|$ options at $L - 1$.

So I only need the score of the best sequence up to $L - 1$, for each possible label at $L - 1$. Call this $V[L - 1, y]$ for $y \in \Lambda$. From this, I can score each label at L, for each hypothetical label at $L - 1$.

Score of the best sequences up to $L - 1$ relies similarly on score of the best sequences up to $L - 2$. Ditto, at every other timestep $L - 2$, $L - 3$, ... 1.

(Featurized) Viterbi Algorithm

- Precompute V[*, *] from left to right. V[1, *] = 0. For ℓ = 2 to L, for each y in Λ:

$$V[\ell, y] = \max_{y' \in \Lambda} V[\ell - 1, y'] + \mathbf{w}^\top \boldsymbol{f}_{local}(x_1 \ldots x_L, y'y)$$

$$B[\ell, y] = \arg\max_{y' \in \Lambda} V[\ell - 1, y'] + \mathbf{w}^\top \boldsymbol{f}_{local}(x_1 \ldots x_L, y'y)$$

- Backtrack and select the labels from right to left.

$$y_L^* = \arg\max_y V[L, y]$$

For ℓ = L - 1 to 1:

$$y_\ell^* = B[\ell + 1, y_{\ell+1}^*]$$

Part of Speech Tagging

After paying the medical bills , Frances was nearly broke .

RB	VBG	DT	JJ	NNS	,	NNP	VBZ	RB	JJ	.

- Adverb (RB)
- Verb (VBG, VBZ, and others)
- Determiner (DT)
- Adjective (JJ)
- Noun (NN, NNS, NNP, and others)
- Punctuation (., ,, and others)

Named Entity Recognition

With Commander Chris Ferguson at the helm ,

Atlantis touched down at Kennedy Space Center .

Named Entity Recognition

O B-person I-person I-person O O O O

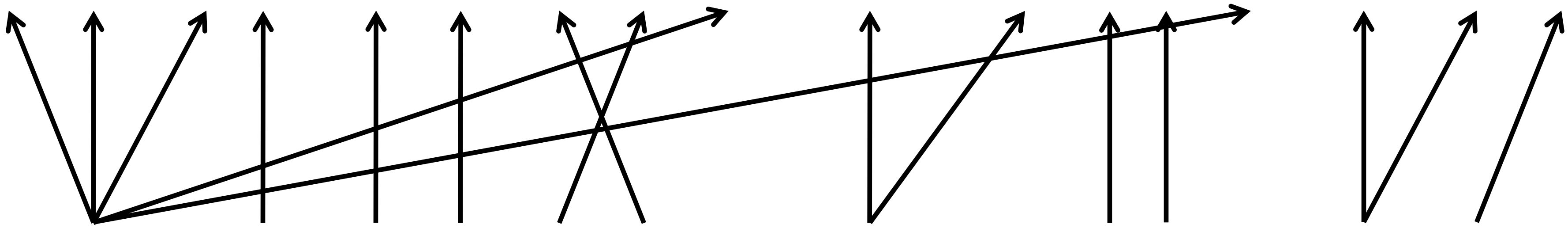
With Commander Chris Ferguson at the helm ,

B-space-shuttle O O O B-place I-place I-place O

Atlantis touched down at Kennedy Space Center .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.



NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah's ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah's ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah's ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

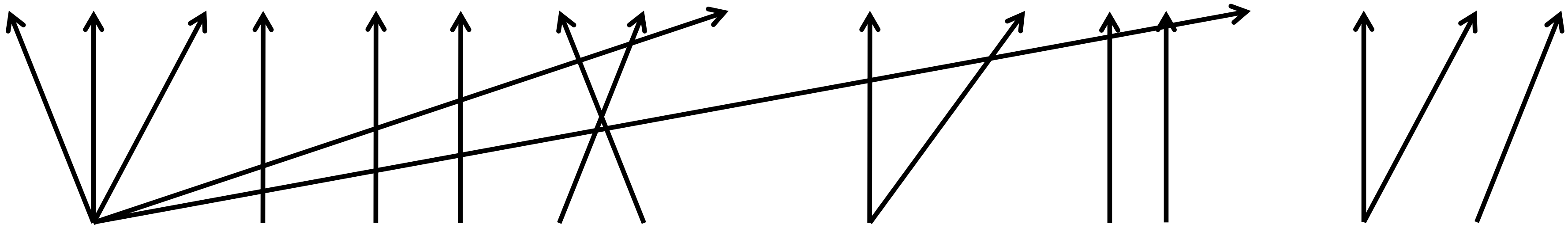
Word Alignment

Mr. President , Noah's ark was filled not with production factors , but with living creatures.

NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Word Alignment

Mr. President , Noah’s ark was filled not with production factors , but with living creatures.



NULL Noahs Arche war nicht voller Productionsfactoren , sondern Geschöpfe .

Basic Recipe for ~~Document Categorization~~ Sequence Labeling

1. Obtain a pool of correctly labeled sequences *D*.
2. Define a locally factored function ***f*** from sequences and labelings to feature vectors.

3. Define a parameterized function $h_{\mathbf{w}}$ from feature vectors to labelings.
4. Select *h*'s parameters **w** using a training sample from *D*.
5. Estimate performance on a held-out sample from *D*.

Structured Learners Generalize Linear Classification Learners!

- hidden Markov models ⟵ naïve Bayes
- conditional random fields ⟵ logistic regression
- structured perceptron ⟵ perceptron
- structured SVM ⟵ support vector machine

Additional Notes

- Outputs that are trees, graphs, logical forms, other strings ...

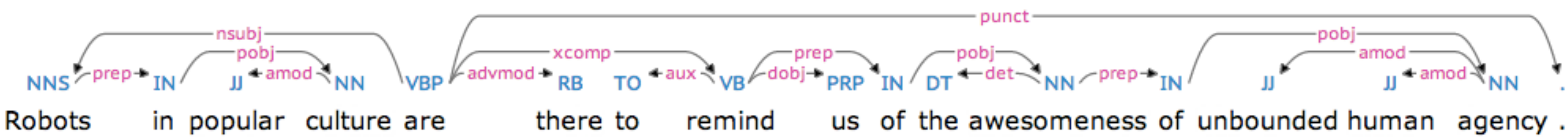
  parse trees (phrase structure, dependencies)

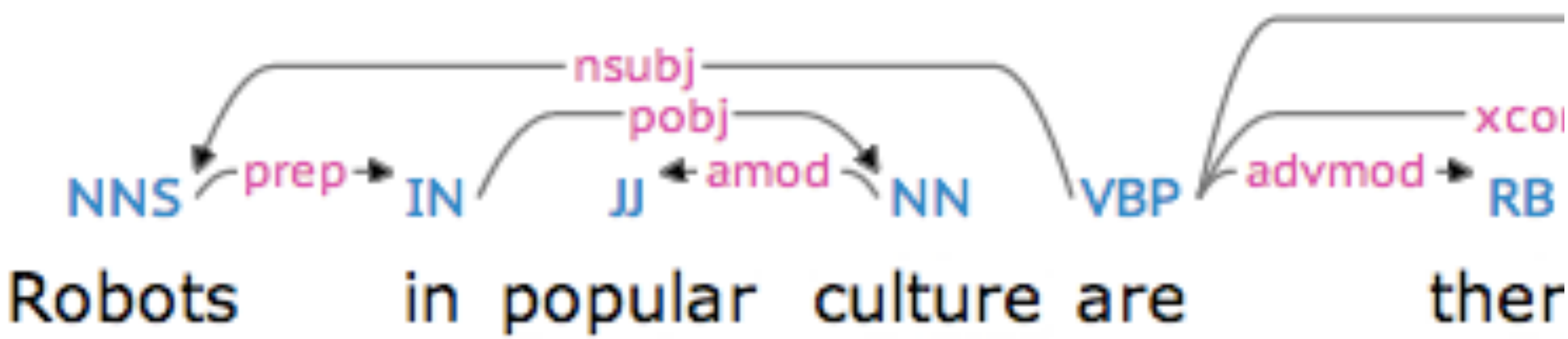
  coreference relationships among entity mentions (and pronouns)

  a huge range of semantic analyses

- Evaluation?

Dependency Parse





Frame-Semantic Parse

	Existence	Desirability	Evoking	People	Organization
Robots					
in	Entity		Stimulus		
popular		Desirability			
culture		Evaluee			
are	Existence				
there					
to					
remind			Evoking		
us			Phenomenon		
of					
the					
awesomeness					
of			Cognizer		
unbounded					
human				People	Descriptor
agency					Organization
.					

Run our Parsers!

http://demo.ark.cs.cmu.edu/parse

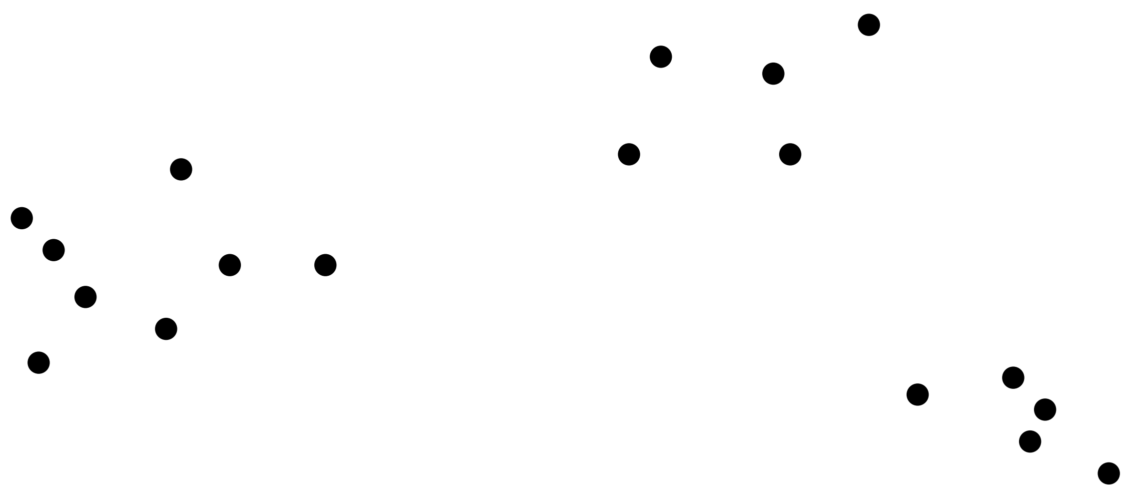
Outline

✓ Automatically categorizing documents

✓ Decoding sequences of words

3. Clustering documents and/or words

Clustering Real Data

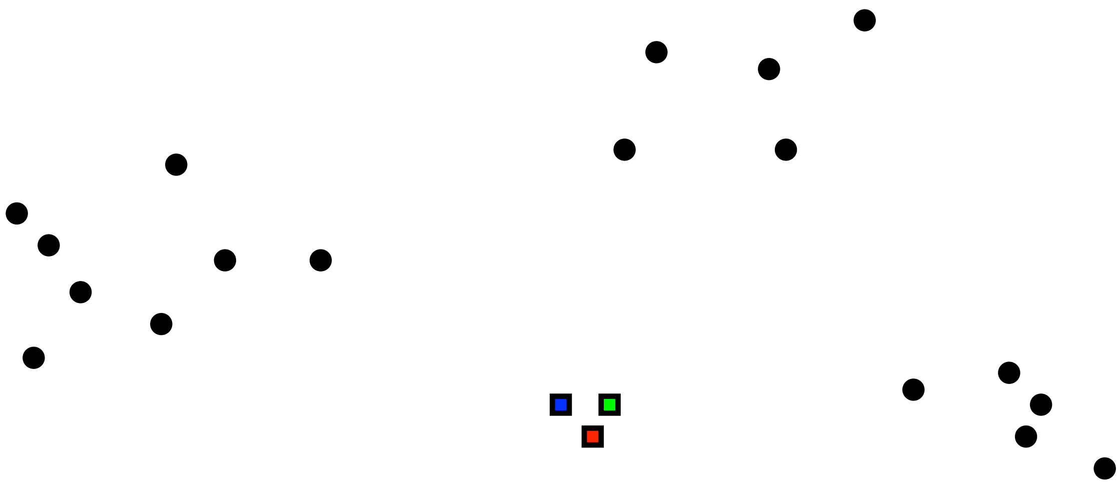


K-Means

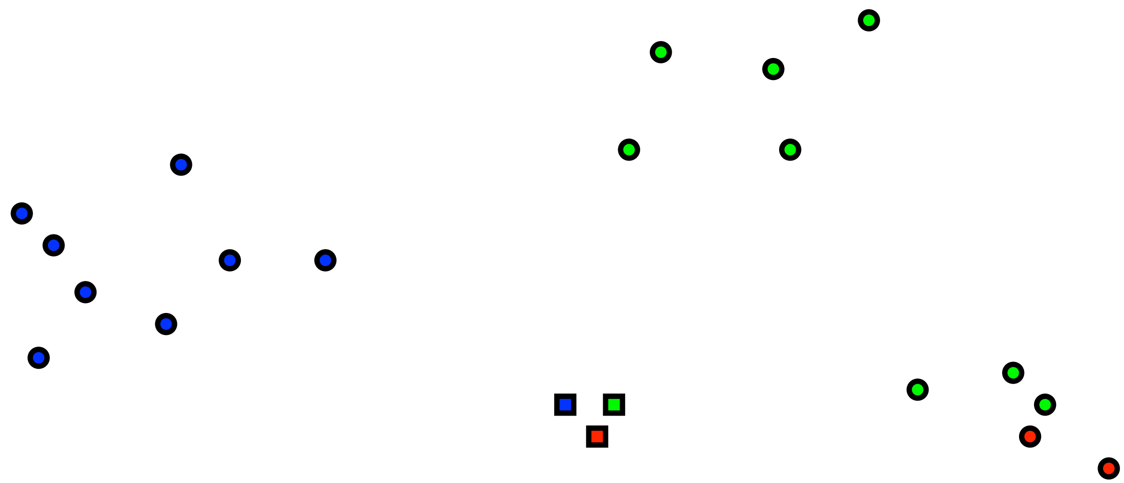
Given: points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, K (number of clusters)

1. Arbitrarily select $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.
2. Assign each $\mathbf{x}_i$ to the nearest $\boldsymbol{\mu}_j$.
3. Select each $\boldsymbol{\mu}_j$ to be the mean of all $\mathbf{x}_i$ assigned to it.
4. If all $\boldsymbol{\mu}_j$ have converged stop; else go to 2.

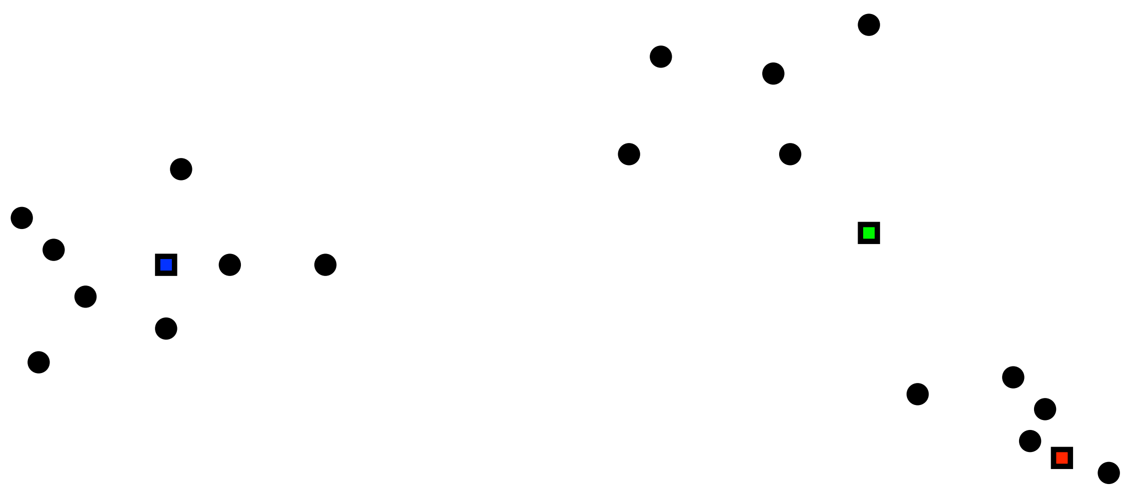
K-Means, Visualized



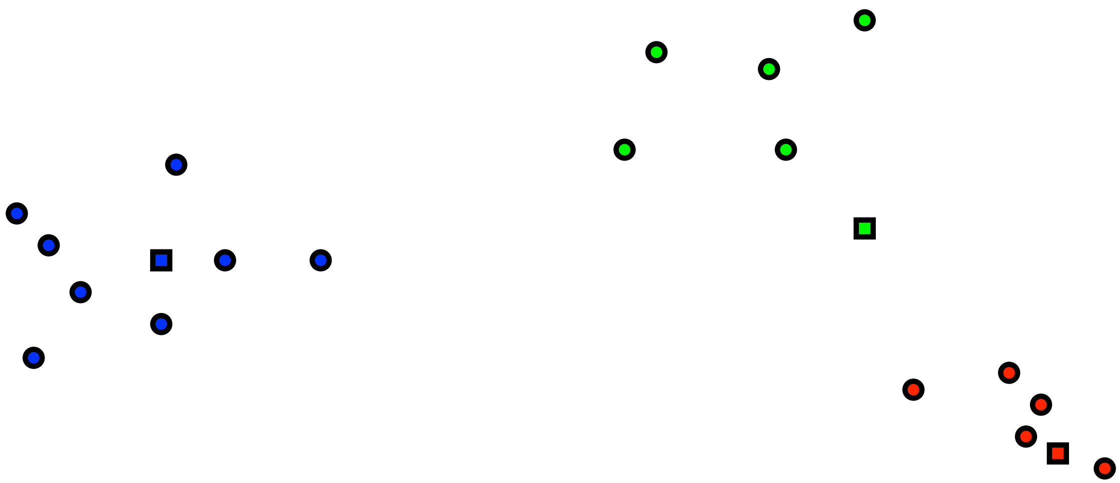
K-Means, Visualized



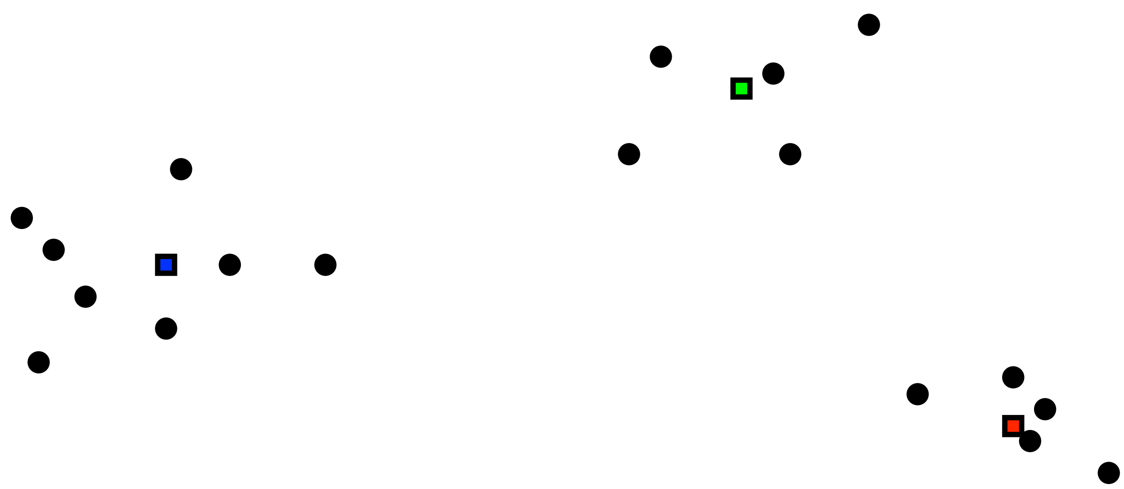
K-Means, Visualized



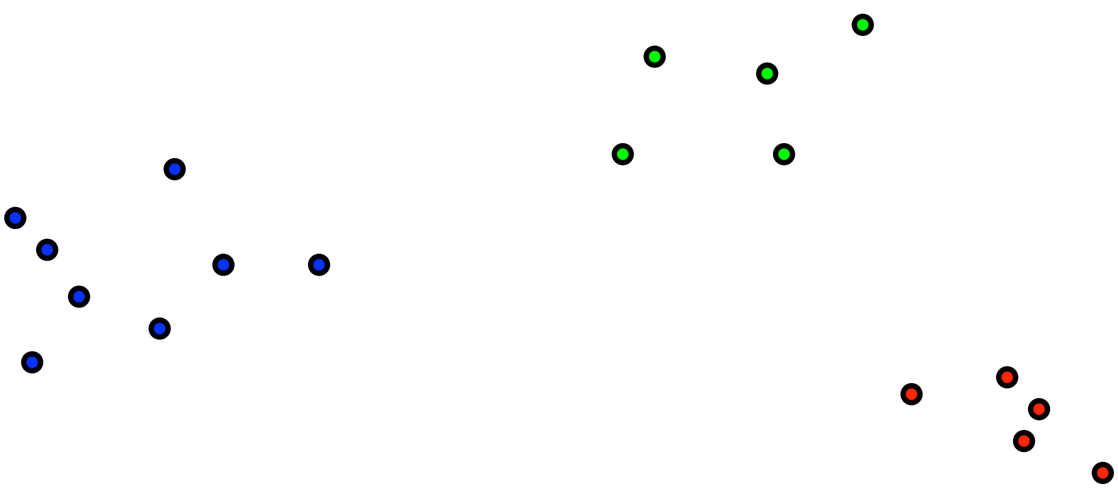
K-Means, Visualized



K-Means, Visualized



K-Means, Visualized



K-Means for Text?

- Documents
  - Use the same ***f*** we might use for classification.
- Words
  - Use “context” vectors ...

Where’s the *beef*?

```
1   fertility.    Organ meats such as beef and chicken liver, tongue  and hear
2   controlling scours.  _HOW TO FEED: BEEF AND DAIRY CALVES_     - 0.2 gram Dy
3   ing process discolors the treated  beef and liquid accumulates in prepackag
4   say. He  did say she could get her beef and vegetables in cans  this summer
5    and feed efficiency of fattening  beef animals.  _HOW TO FEED:_     At the
6    steaks, chops, chicken and prime  beef as well as Tom's favorite dish, stu
7   ross  from him was surmounted by a beef barrel with ends  knocked out. In t
8    counter of boards laid across two beef barrels.  There was, of course, no
9   Because Holstein  cattle weren't a beef breed, they were rarely seen  on a
10   2-5 grams of phenothiazine daily; beef  calves- .5 to 1.5 grams daily depe
11  ties of  this drug.  _HOW TO FEED: BEEF CATTLE (FINISHING RATION)_     - To
12   dairy cows and lesser amounts to  beef cattle and poultry. About 90 percen
13   raises enough  poultry, pigs, and beef cattle for most of their needs.  Lo
14  on of liver abscesses  in feed-lot beef cattle. Prevention of bacterial pne
15  pal feed bunk  types for dairy and beef cattle: (1) Fence-line bunks-  catt
16  es feed efficiency.  _HOW TO FEED: BEEF CATTLE_     - 10 milligrams of diet
17   the rations you are feeding  your beef, dairy cattle, and sheep are adequa
18  itive business more profitable for beef, dairy,  and sheep men.     The tar
19  o bear. She was ready  to kill the beef, dress it out, and with vegetables
20  . She had raised a calf,  grown it beef-fat. She had, with her own work-wea
21  with feeding  low-moisture corn in beef-feeding programs. Several  firms ar
22  he shelf life (at 35  F)  of fresh beef from 5 days to 5 or 6 weeks. Howeve
23   canned pork products.  Tests with beef have been largely unsuccessful beca
24   for eggs, pigs to eat garbage,  a beef herd and wastes of all kinds. Separ
25    their money's worth. A good many beef-hungry settlers  were accepting the
```

chicken

1 y the irradiated and refrigerated chicken. Acceptance of radiopasteurization
2 torehouse". Glendora dropped a chicken and a flurry of feathers, and went
3 will specialize in steaks, chops, chicken and prime beef as well as Tom's fa
4 ard as the one concerned with the chicken and the egg. Which came first? Is
5 he millions of buffalo and prairie chicken and the endless seas of grass that
6 "! "Come on, there's some cold chicken and we'll see what else". They wen
7 ves to extend the storage life of chicken at a low cost of about 0.5 cent per
8 CHICKEN CADILLAC# Use one 6-ounce chicken breast for each guest. Salt and pe
9 ion juice, to about half cover the chicken breasts. Bake slowly at least one-
10 d, in butter. Sprinkle over top of chicken breasts. Serve each breast on a th
11 around, they had a hard time". #CHICKEN CADILLAC# Use one 6-ounce chicken
12 successful, and the shelf life of chicken can be extended to a month or more
13 ay from making a cake, building a chicken coop, or producing a book, to found
14 , they decided, but a deck full of chicken coops and pigpens was hardly suita
15 im. "Johnny insisted on cooking a chicken dinner in my honor- he's always bee
16 nutes. Kid Ory, the trombonist chicken farmer, is also one of the solid a
17 y Johnson reaching around the wire chicken fencing, which half covered the tr
18 yes glittering behind dull silver chicken fencing. "That was Tee-wah I was t
19 wine in the pot roast or that the chicken had been marinated in brandy, and
20 yed this same game and called it "Chicken". He could not go through the f
21 f the Mexicans hiding in a little chicken house had passed through his head,
22 I'll never forget him cleaning the chicken in the tub". A story, no doubt
23 . Organ meats such as beef and chicken liver, tongue and heart are planne
24 p. "Miss Sarah, I can't cut up no chicken. Miss Maude say she won't". Aga
25 pot. "What is it"? he asked. "Chicken", Mose said, and theatrically licke
26 im"? Adam shook his head. "Chicken", Mose said. She was a child too m

Hypothetical Counts based on Syntactic Dependencies

	Modified-by-ferocious(adj)	Subject-of-devour(v)	Object-of-pet(v)	Modified-by-African(adj)	Modified-by-big(adj)
Lion	15	5	0	6	15
Dog	7	3	8	0	12
Cat	1	1	6	1	9
Elephant	0	0	0	10	15
…					

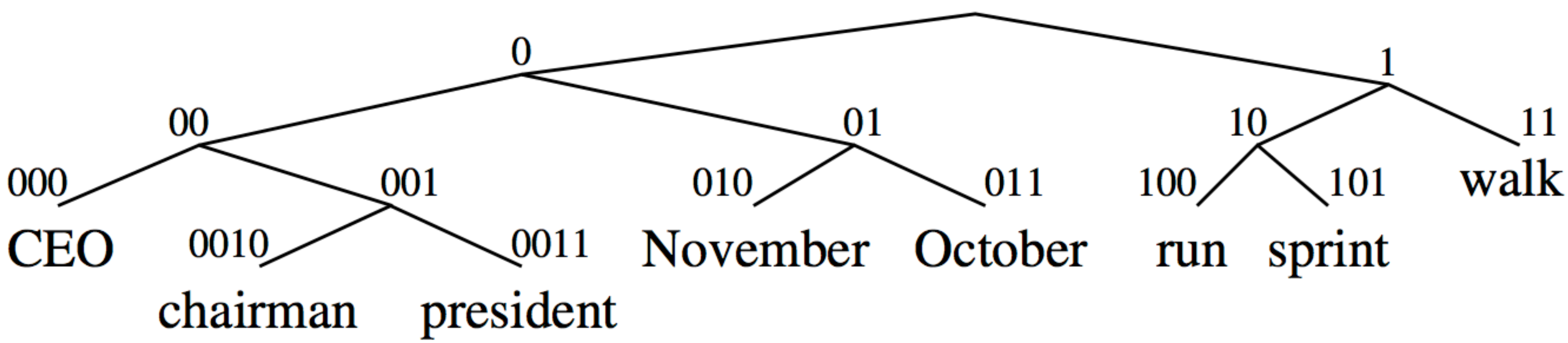
Brown Clustering

Given: corpus of length N, K

1. Assign each word to its cluster (V clusters)
2. Repeat V – K times:
   - Find the single merge (c_j, c_k) that results in a new clustering with the highest *Quality* score
   - Prepend c_j's bitstring with 0 and c_k's with 1 (and the same for all their descendents)

Mini-Example

Bitstrings that share a prefix are in the same cluster, at some level of granularity.



Clusters from Brown et al. (1992)

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
June March July April January December October November September August
people guys folks fellows CEOs chaps doubters commies unfortunates blokes
down backwards ashore sideways southward northward overboard aloft downwards adrift
water gas coal liquid acid sand carbon steam shale iron
great big vast sudden mere sheer gigantic lifelong scant colossal
man woman boy girl lawyer doctor guy farmer teacher citizen
American Indian European Japanese German African Catholic Israeli Italian Arab
pressure temperature permeability density porosity stress velocity viscosity gravity tension
mother wife father son husband brother daughter sister boss uncle
machine device controller processor CPU printer spindle subsystem compiler plotter
John George James Bob Robert Paul William Jim David Mike
anyone someone anybody somebody
feet miles pounds degrees inches barrels tons acres meters bytes
director chief professor commissioner commander treasurer founder superintendent dean custodian
liberal conservative parliamentary royal progressive Tory provisional separatist federalist PQ
had hadn't hath would've could've should've must've might've
asking telling wondering instructing informing kidding reminding bothering thanking deposing
that tha theat
head body hands eyes voice arm seat eye hair mouth

Clusters from Owoputi et al. (2013) (56M Tweets)

acronyms for laughter	lmao lmfao lmaoo lmaooo hahahahaha lool ctfu rofl loool lmfaoo lmfaooo lmaoooo lmbo lololol
onomatopoeic laugher	haha hahaha hehe hahahaha hahah aha hehehe ahaha hah hahahah kk hahaa ahah
affirmative	yes yep yup nope yess yesss yessss ofcourse yeap likewise yepp yesh yw yuup yus
negative	yeah yea nah naw yeahh nooo yeh noo noooo yeaa ikr nvm yeahhh nahh nooooo
metacomment	smh jk #fail #random #fact smfh #smh #winning #realtalk smdh #dead #justsaying

Clusters from Owoputi et al. (2013) (56M Tweets)

second person pronoun	u yu yuh yhu uu yuu yew y0u yuhh youh yhuu iget yoy yooh yuo yue juu dya youz yyou
prepositions	w fo fa fr fro ov fer fir whit abou aft serie fore fah fuh w/her w/that fron isn agains
“contractions”	tryna gon finna bouta trynna boutta gne fina gonn tryina fenna qone trynaa qon
going to	gonna gunna gona gna guna gnna ganna qonna gonnna gana qunna gonne goona
so+	soo sooo soooo sooooo soooooo sooooooo soooooooo sooooooooo soooooooooo

Clusters from Owoputi et al. (2013) (56M Tweets)

mischevious	;) :p :-) xd ;-) ;d (; :3 ;p =p :-p =)) ;] xdd #gno xddd >:) ;-p >:d 8-) ;-d
happy	:) (: =) :)) :] :') =] ^_^ :))) ^.^ [: ;)) ((: ^__^ (= ^-^ :))))
sad	:(:/ -_- -.- :-(:'(d: :| :s -__- =(=/ >.< -___- :-/ </3 :\ -____- ;(/: :((>_< =[:[#fml
love	<3 xoxo <33 xo <333 #love s2 <URL-twitition.com> #neversaynever <3333
F-word + ing	fucking fuckin freaking bloody freakin friggin effin effing fuckn fucken frickin fukin f'n fckn flippin fkn motherfucking fckin f*cking fricken fukn fuccin fcking fukkin

Browse our Twitter Clusters!

http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html

Additional Notes

- Soft clustering allows items to have *mixed membership* in different clusters.
  - Typically accomplished with probabilistic models
  - Latent Dirichlet allocation is a popular model built on Bayesian inference
- Evaluation?
- One view of clusters: feature creation!

Summary

supervised
classification

*(5 steps: data, features,
prediction function,
learning, evaluation)*

*local factoring +
dynamic programming*

structured
prediction

*alternating
or greedy
optimization*

unsupervised
clustering

Classes

- 11-411/11-611: Natural Language Processing
- 11-711: Algorithms for NLP
- 11-761: Language and Statistics
- 11-713: Advanced NLP Seminar