Applied Natural Language Processing

Info 256

Lecture 8: Text classification (Sept. 20, 2023)

David Bamman, UC Berkeley

Project proposal — due 9/27

- Final project involving 1 to 3 students involving natural language processing -- involving natural language processing in support of an empirical research question.

- Proposal:

    - outline the work you're going to undertake
    - motivate its rationale as an interesting question worth asking
    - assess its potential to contribute new knowledge by situating it within related literature in the scientific community. (cite 5 relevant sources)
    - who is the team and what are each of your responsibilities (everyone gets the same grade)

Classification

A mapping h from input data x (drawn from instance space $\mathcal{X}$) to a label (or labels) y from some enumerable output space $\mathcal{Y}$

$\mathcal{X}$ = set of all documents

$\mathcal{Y}$ = {english, mandarin, greek, …}

x = a single document

y = ancient greek

Classification

$$h(x) = y$$

$$h(\text{μῆνιν ἄειδε θεὰ}) = \text{ancient grc}$$

Classification

Let h(x) be the "true" mapping. We never know it. How do we find the best $\hat{h}(x)$ to approximate it?

One option: rule based

if x has characters in
unicode point range 0370-03FF:
$\hat{h}(x)$ = greek

Classification

Supervised learning

Given training data in the form of <x, y> pairs, learn $\hat{h}(x)$

Text categorization problems

task	$\mathcal{X}$	$\mathcal{Y}$
language ID	text	{english, mandarin, greek, …}
spam classification	email	{spam, not spam}
authorship attribution	text	{jk rowling, james joyce, …}
genre classification	novel	{detective, romance, gothic, …}
sentiment analysis	text	{positive, negative, neutral, mixed}

Sentiment analysis

- Document-level SA: is the entire text positive or negative (or both/ neither) with respect to an implicit target?
- Movie reviews [Pang et al. 2002, Turney 2002]

Training data

positive

"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, Apocalypse Now

- "I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, North

negative

Sentiment analysis

- Is the text positive or negative (or both/neither) with respect to an explicit target within the text?

Feature: **picture**

Positive: 12

- Overall this is a good camera with a really good picture clarity.
- The pictures are absolutely amazing - the camera captures the minutest of details.
- After nearly 800 pictures I have found that this camera takes incredible pictures.

…

Negative: 2

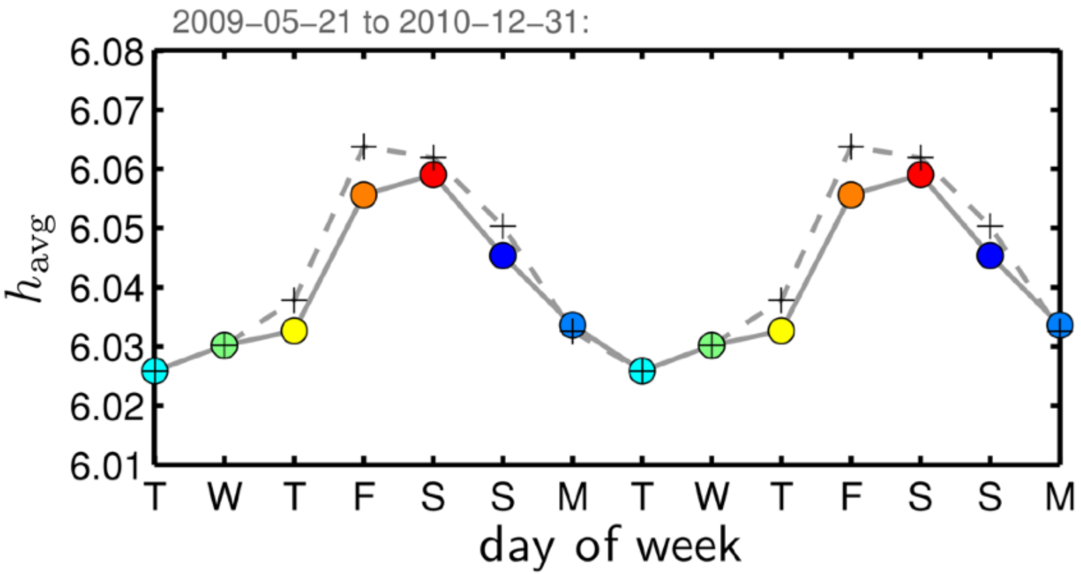
- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Hu and Liu (2004), "Mining and Summarizing Customer Reviews"

Sentiment as tone

- No longer the speaker's attitude with respect to some particular target, but rather the positive/negative tone that is evinced.

Sentiment as tone



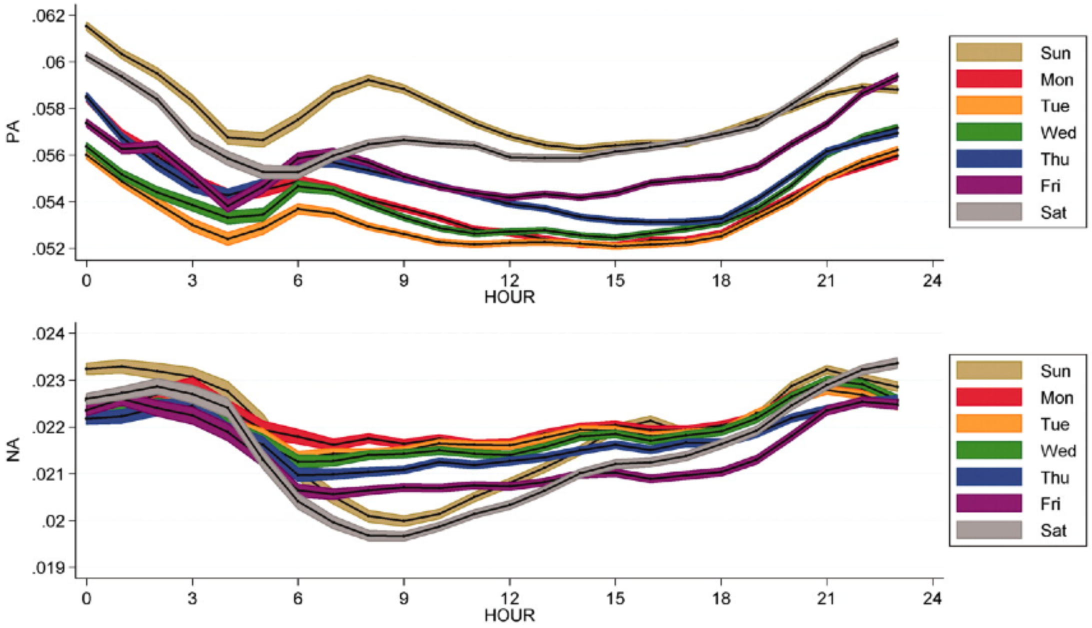
Dodds et al. (2011), "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter" (PLoS One)

Sentiment Dictionaries

- General Inquirer (1966)
- MPQA subjectivity lexicon (Wilson et al. 2005) http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- LIWC (Linguistic Inquiry and Word Count, Pennebaker 2015)
- AFINN (Nielsen 2011)
- NRC Word-Emotion Association Lexicon (EmoLex), Mohammad and Turney 2013

pos	neg
unlimited	lag
prudent	contortions
superb	fright
closeness	lonely
impeccably	tenuously
fast-paced	plebeian
treat	mortification
destined	outrage
blessing	allegations
steadfastly	disoriented

Sentiment as tone



Golder and Macy (2011), "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures," *Science*. Positive affect (PA) and negative affect (NA) measured with LIWC.

Why is SA hard?

- Sentiment is a measure of a speaker's private state, which is unobservable.
- Sometimes words are a good indicator of sentiment (love, amazing, hate, terrible); many times it requires deep world + contextual knowledge

"*Valentine's Day* is being marketed as a Date Movie. I think it's more of a First-Date Movie. If your date likes it, do not date that person again. And if you like it, there may not be a second date."

Roger Ebert, Valentine's Day

Classification

Supervised learning

Given training data in the form of <x, y> pairs, learn $\hat{h}(x)$

x	y
loved it!	positive
terrible movie	negative
not too shabby	positive

$\hat{h}(x)$

- The classification function that we want to learn has two different components:
  - the formal structure of the learning method (what's the relationship between the input and output?) → Naive Bayes, logistic regression, convolutional neural network, etc.
  - the representation of the data


RNNs
Decision trees
Transformers
LLMs
Random forests
Probabilistic graphical models
Support vector machines
Logistic regression
CNNs
Perceptron

Representation for SA

- Only positive/negative words in MPQA
- Only words in isolation (bag of words)
- Conjunctions of words (sequential, skip ngrams, other non-linear combinations)
- Higher-order linguistic structure (e.g., syntax)

"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, Apocalypse Now

"I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, North

Bag of words

Representation of text only as the counts of words that it contains

	Apocalypse now	North
the	1	1
of	0	0
hate	0	9
genius	1	0
bravest	1	0
stupid	0	1
like	0	1
…		

Bag of words

For short documents, a binary representation can often suffice: only notes the *existence* of word in the document and not its count.

	Apocalypse now	North
the	1	1
of	0	0
hate	0	~~0~~ 1
genius	1	0
bravest	1	0
stupid	0	1
like	0	1
…		

Refresher

$$\sum_{i=1}^{F} x_i \beta_i = x_1\beta_1 + x_2\beta_2 + \ldots + x_F\beta_F$$

$$\prod_{i=1}^{F} x_i = x_i \times x_2 \times \ldots \times x_F$$

$$\exp(x) = e^x \approx 2.7^x \qquad \exp(x+y) = \exp(x)\exp(y)$$

$$\log(x) = y \rightarrow e^y = x \qquad \log(xy) = \log(x) + \log(y)$$

Binary logistic regression

$$P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{F} x_i \beta_i\right)}$$

output space $\mathcal{Y} = \{0, 1\}$

x = feature vector

Feature	Value
the	0
and	0
bravest	0
love	0
loved	0
genius	0
not	0
fruit	1
BIAS	1

β = coefficients

Feature	β
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
not	-3.0
fruit	-0.8
BIAS	-0.1

Multiclass logistic regression

$$P(Y = y \mid X = x; \beta) = \frac{\exp(x^\top \beta_y)}{\sum_{y' \in \mathcal{Y}} \exp(x^\top \beta_{y'})}$$

output space $\mathcal{Y} = \{1, \ldots, K\}$

x = feature vector

Feature	Value
the	0
and	0
bravest	0
love	0
loved	0
genius	0
not	0
fruit	1
BIAS	1

β = coefficients (one set for each output class)

Feature	β_1	β_2	β_3	β_4	β_5
the	1.33	-0.80	-0.54	0.87	0
and	1.21	-1.73	-1.57	-0.13	0
bravest	0.96	-0.05	0.24	0.81	0
love	1.49	0.53	1.01	0.64	0
loved	-0.52	-0.02	2.21	-2.53	0
genius	0.98	0.77	1.53	-0.95	0
not	-0.96	2.14	-0.71	0.43	0
fruit	0.59	-0.76	0.93	0.03	0
BIAS	-1.92	-0.70	0.94	-0.63	0

Binary logistic regression

	BIAS	love	loved
β	-0.1	3.1	1.2

	BIAS	love	loved	$a=\sum x_i\beta_i$	exp(-a)	1/(1+exp(-a))
x^1	1	1	0	3	0.05	95.2%
x^2	1	1	1	4.2	0.015	98.5%
x^3	1	0	0	-0.1	1.11	47.4%

Features

- As a discriminative classifier, logistic regression doesn't assume features are independent like Naive Bayes does.
- Its power partly comes in the ability to create richly expressive features without the burden of independence.
- We can represent text through features that are not just the identities of individual words, but any feature that is scoped over the entirety of the input.

features
contains like
has word that shows up in positive sentiment dictionary
review begins with "I like"
at least 5 mentions of positive affectual verbs (like, love, etc.)

Features

- Features are where you can encode your own domain understanding of the problem.

feature classes
unigrams ("like")
bigrams ("not like"), higher order ngrams
prefixes (words that start with "un-")
has word that shows up in positive sentiment dictionary

Features

Task	Features
Sentiment classification	Words, presence in sentiment dictionaries, etc.
Keyword extraction	
Fake news detection	
Authorship attribution	

Features

Feature	Value
the	0
and	0
bravest	0
love	0
loved	0
genius	0
not	1
fruit	0
BIAS	1

Feature	Value
like	1
not like	1
did not like	1
in_pos_dict_MPQA	1
in_neg_dict_MPQA	0
in_pos_dict_LIWC	1
in_neg_dict_LIWC	0
author=ebert	1
author=siskel	0

β = coefficients

How do we get good values for β?

Feature	β
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
not	-3.0
fruit	-0.8
BIAS	-0.1

Conditional likelihood

$$\prod_{i}^{N} P(y_i \mid x_i, \beta)$$

For all training data, we want the probability of the true label y for each data point x to be high

	BIAS	love	loved	$a=\sum x_i\beta_i$	exp(-a)	1/(1+exp(-a))	true y
x^1	1	1	0	3	0.05	95.2%	1
x^2	1	1	1	4.2	0.015	98.5%	1
x^3	1	0	0	-0.1	1.11	47.5%	0

Conditional likelihood

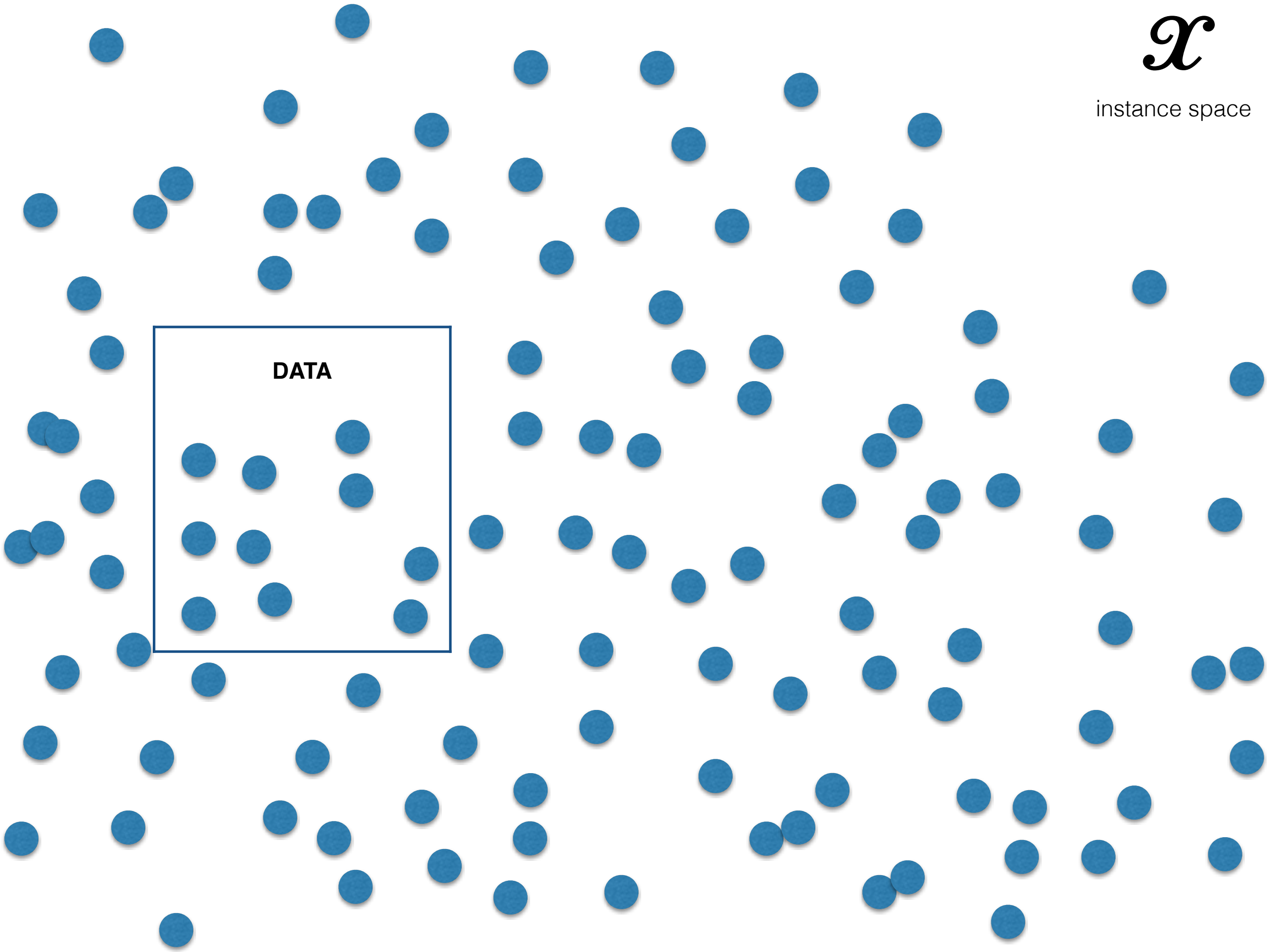
$$\prod_{i}^{N} P(y_i \mid x_i, \beta)$$

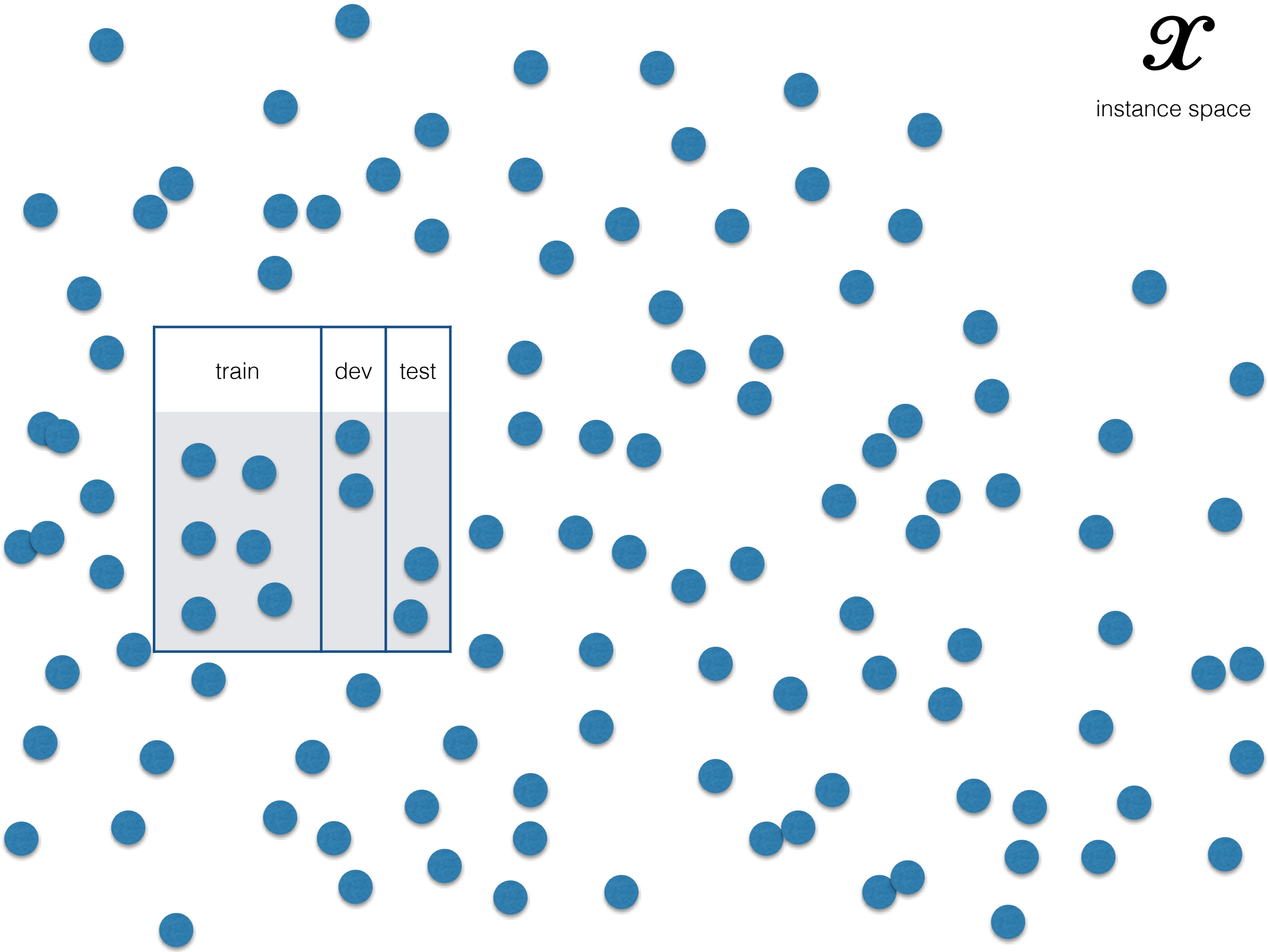
For all training data, we want the probability of the true label y for each data point x to be high

Pick the values of parameters β to maximize the conditional probability of the training data <x, y> using gradient ascent.

Evaluation

- For all supervised problems, it's important to understand how well your model is performing
- What we try to estimate is how well you will perform in the future, on new data also drawn from $\mathcal{X}$
- Trouble arises when the training data <x, y> you have does not characterize the full instance space.
    - n is small
    - sampling bias in the selection of <x, y>
    - x is dependent on time
    - y is dependent on time (concept drift)


$\mathcal{X}$
instance space
DATA



$\mathcal{X}$
instance space
train
dev
test

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Multiclass confusion matrix

True (y) \ Predicted (ŷ)	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Accuracy

$$\frac{1}{N}\sum_{i=1}^{N} I[\hat{y}_i = y_i]$$

$$I[x] \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Predicted (ŷ)

True (y)	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Precision

Precision(POS) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = POS)}{\sum_{i=1}^{N} I(\hat{y}_i = POS)}$$

Precision: proportion of predicted class that are actually that class.

		Predicted (ŷ)	
True (y)	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Recall

Recall(POS) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = POS)}{\sum_{i=1}^{N} I(y_i = POS)}$$

Recall: proportion of true class that are predicted to be that class.

True (y) \ Predicted (ŷ)	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

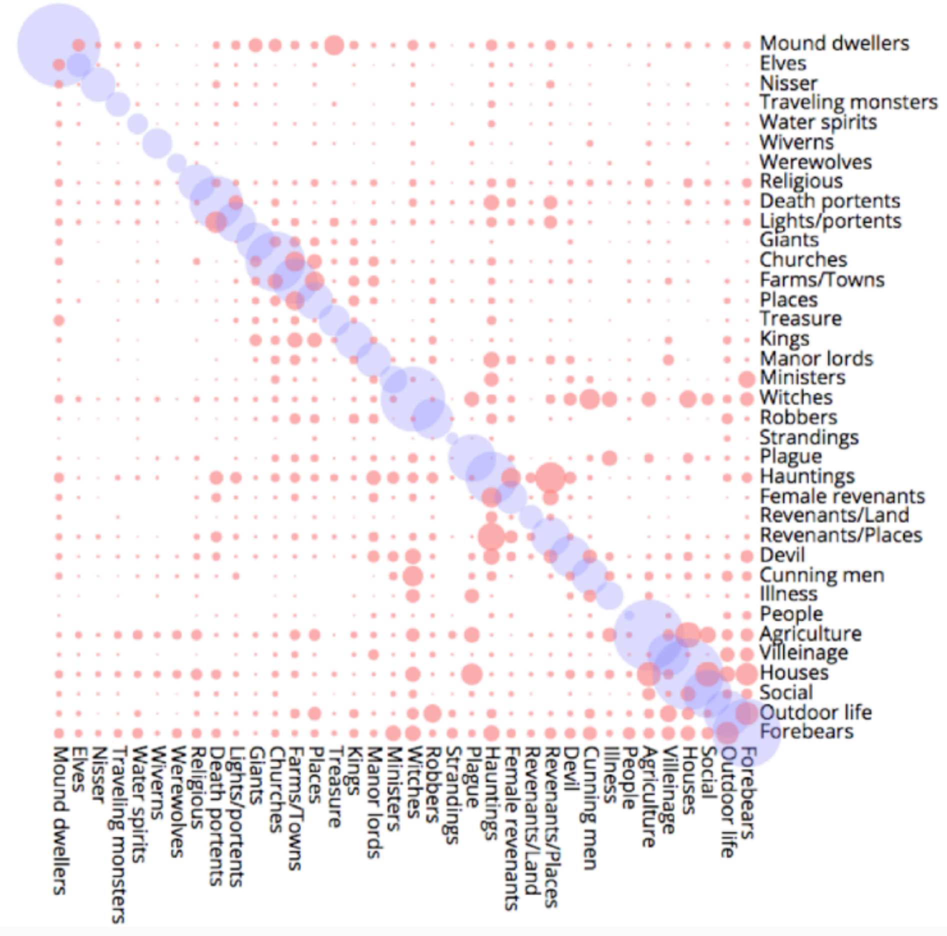
Majority class baseline

- Pick the label that occurs the most frequently in the training data. (Don't count the test data!)
- Predict that label for every data point in the test data.

Challenging classification

Classification

- Peter M. Broadwell, David Mimno and Timothy R. Tangherlini (2017): Using classification to explore the boundaries between categories in Danish folk tales.



Haiku

Whitecaps on the bay:
A broken signboard banging
In the April wind.

— Richard Wright

Long and So (2016), “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning,” Critical Inquiry

Three spirits came to me
And drew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

— Ezra Pound

Activity

`6.classification/Classification.ipynb`

- Design features for predicting the genre of a movie

Parameters vs. Hyperparameters

Parameters whose values are *learned*

Feature	β
the	0.01
and	0.03
bravest	1.4
love	3.1
loved	1.2
genius	0.5
BIAS	-0.1

Hyperparameters whose values are *chosen*

Hyperparameter	value
minimum word frequency	5
max vocab size	10000
lowercase	TRUE
regularization strength	1.0