INF5830 – 2015 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning & Lilja Øvrelid

Today

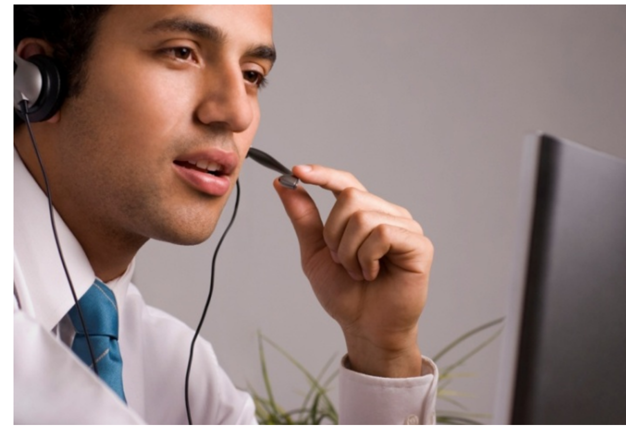
- Hour 1: Course overview
- Hour 2: ”Looking at data”:
  - Descriptive statistics

Name game

- Computational Linguistics
  - Traditional name, stresses interdisciplinarity
- Natural Language Processing
  - Computer science/AI/NLP
  - ”Natural language” a CS term
- Language Technology
  - Newer term
  - Stresses applicability
  - LT today is not SciFi (AI), but part of everyday app(lication)s
- The terms are more or less interchangeable

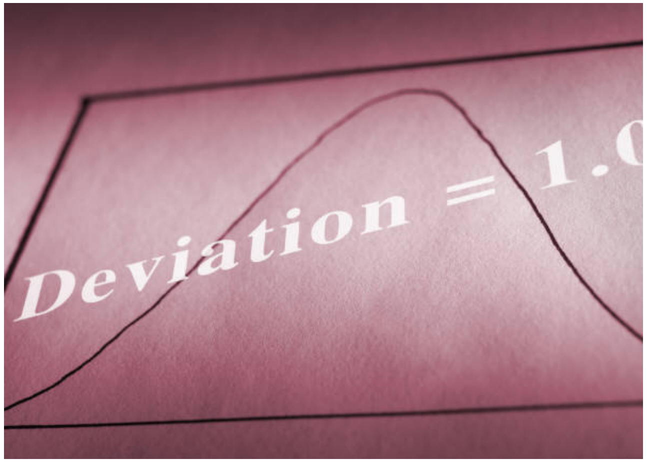
NLP applications - examples

- Translation (Google translate)
- Dialogue (Apple's Siri)
- Search
- Web analytics
- Intelligence
- Web recommendations (search, ads)
- Speech
- Mobile devices
- Language support



The place of INF5830

- Methods and modules across various applications
  - (INF5820 focus on applications)
- Main emphasis on statistical/empirical methods
  - (INF2820 non-statistical, symbolic, rule-based methods)
- Complement other courses
  - (INF1820, INF2820, INF4820, INF5830)



INF5830

- http://www.uio.no/studier/emner/matnat/ifi/INF5830/
- Recommended prior knowledge
  - INF4820 (may be studied the same semester)
- Advantage, but not assumed
  - INF2820
  - Some statistics
- Alternates with
  - INF5820 Language technological applications

Schedule

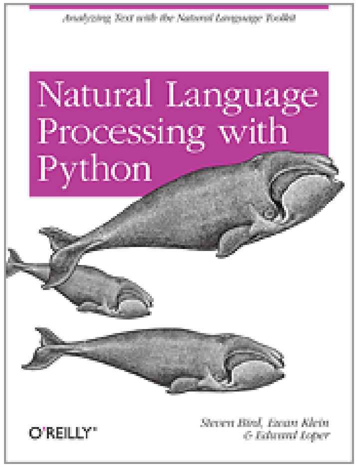
- Class
  - Monday14.15-16
  - Thursday 14.15-16 (not every week)
- Exam
  - Written
  - 8 Dec 2015, 0900
    - For they who fail: Exam spring 2016
      - Requires approved obligs this semester

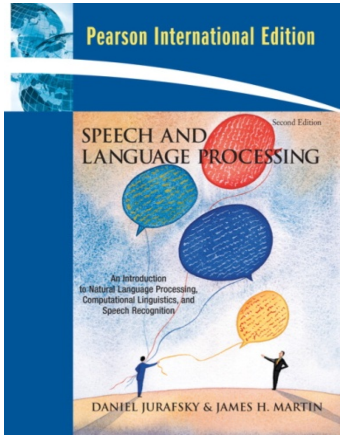
More on first part

Jan Tore

Syllabus

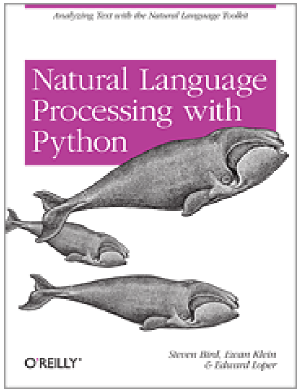
- Lectures: Presentations put on the web
- Parts of books:
  - S. Bird, E. Klein and E. Loper:
    - *Natural Language Processing with Python*
    - (Available online)
  - Jurafsky and Martin,
    - *Speech and Language Processing*
      - *2. ed*
      - *3. ed, chapters online*
- Statistics, a book may be useful, e.g.
  - Sarah Boslaugh:
    - *Statistics in a Nutshell*
  - (or find a free book on the web)
- Some articles/web-pages/distributed material





Computational "Work Bench"

- Python, general programming language:
  - Well-suited for text
  - Readable, structured code
  - Packages, extensions
- NLTK:
  - Python toolbox for NLP
  - Ready programs for various NLP tasks
  - Emphasis on training
- Python packages: NumPy, SciPy (stats), plotting
  - Widely used for data science and machine learning
  - Working within the same Python universe
    - (in contrast to R or MatLab)



Content

- Statistical methods for NLP
  - Descriptive statistics
  - Probability theory
  - Stat. inference
  - Experiments, evaluation
  - Collocations
- Machine learning, classification applied to NLP:
  - Naive Bayes,
  - Decision trees
  - Maximum entropy
  - A general view

Statistics in NLP

Statistics and probability in NLP

1. "Choose the best":

- *bank* (Eng.) can translate to b.o. *bank* or *bredd* in No.
  - Which should we choose?
  - What if we know the context is "*river bank*"?
- *bank* can be Verb or Noun,
  - which tag should we choose?
  - What if the context is *they bank the money* ?

- We choose the most probable given the available information
- A sentence may be ambiguous:
  - What is the most probable parse of the sentence?

The benefits of statistics:

2. In constructing models from examples (ML):
   - What is the best model given these examples?

3. Evaluation:
   - Model1 is performing slightly better than model 2 (78.4 vs. 73.2), can we conclude that model 1 is better?
   - How large test corpus do we need?

Machine learning and classification

Example of classification tasks

- Word Sense diambiguation:
  - *bass* – fish, voice, instrument, …
- E-mail: spam or no spam
- Language class: Given text, which language?
- Genre
- Author attribution
- Search: Is this document relevant for the given search phrase?
- Textual entailment: does sentence A entail sentence B?
- Anaphora co-reference: Who is ”she”?

A class of methods (supervised, text-based):

- Propose features that may be relevant, e.g.
  - Words in context:
    - *music, sing, perform, soprano,…*
    - *fish, river, boat, eat, …*
  - Properties of these words, distance to target word etc.
- Training corpus with marked senses:
  - Count features in examples
- Construct the classifier from these counts
- Test the classifier on new material!
  - use a test corpus

Second part:
Dependency parsing and rôle labeling

Second part of the course (Lilja)

- theoretical background and practical experience with two NLP tasks
- “deeper processing”: syntactic and semantic analysis
  - data-driven dependency parsing, due Oct 23th
  - semantic role labeling (SRL), due Nov 6th

Why?

- Parsing provides “scaffolding” for semantic analysis
- Down-stream applications:
  - opinion mining
  - information extraction
  - syntax-informed statistical machine translation
  - sentence compression
  - etc…

Data-driven dependency parsing

- Increasing interest in dependency-based approaches to syntactic parsing in recent years:
  - new methods emerging
  - applied to a wide range of languages
  - CoNLL shared tasks (2006, 2007)

Project: training and evaluating parsers for several languages

Semantic role labeling

- Semantic argument classification
  - CoNLL08, 09 shared tasks: syntactic and semantic parsing of English (2008) and other languages (2009)
  - dependency representations for semantic role labeling

Project: system for argument classification with a focus on feature engineering (using syntactic analysis)

Syllabus: linguistics “classics” and research articles

Project will focus on:

- experimental methodology
- evaluation
- academic writing / reporting of results

Looking at data

Data

- Start by taking a look at your data
  - (But tuck away your test data first)
- General form:
  - A set of objects
  - To each object some associated features
- Later on:
  - Which features are interesting?
  - How do we extract them?

The feature – types

- Binary/Boolean:
  - Email: spam?
  - Person: 18 ys. or older?
  - Sequence of word: Grammatical English sentence?
- Categorical:
  - Person: Name
  - Word: Part of Speech (POS)
    - {Verb, Noun, Adj, …}
  - Noun: Gender
    - {Mask, Fem, Neut}

The feature – types

- Numeric
  - Discrete
    - Person: Years of age, Weight in kilos, Height in centimeters
    - Sentence: Number of words
    - Word: length
    - Text: number of occurrences of *great*, (42)
  - Continuous
    - Person: Height with decimals
    - Program execution: Time
    - Occurrences of a word in a text: Relative frequency (18.666…%)

Observations

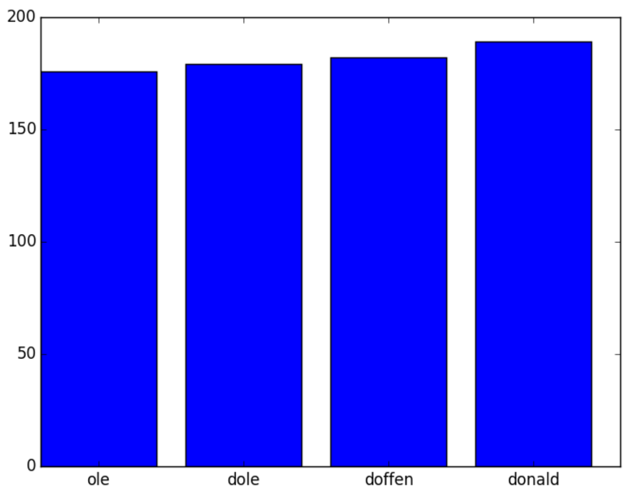
- The binary feature can be considered categoric and numeric {0,1}
- A discrete numeric feature has also all the properties of a categoric feature (when the value set is finite)
- We will see big differences between discrete and continuous features (variables) when we come to statistics.

Graphical displays

and frequency distributions

Graphic displays – one feature

- To understand our data, it is useful to display them graphically in various ways
- With one parameter only, there is the Bar Chart ("søylediagram")
- Requires a numeric parameter



Height in centimeters

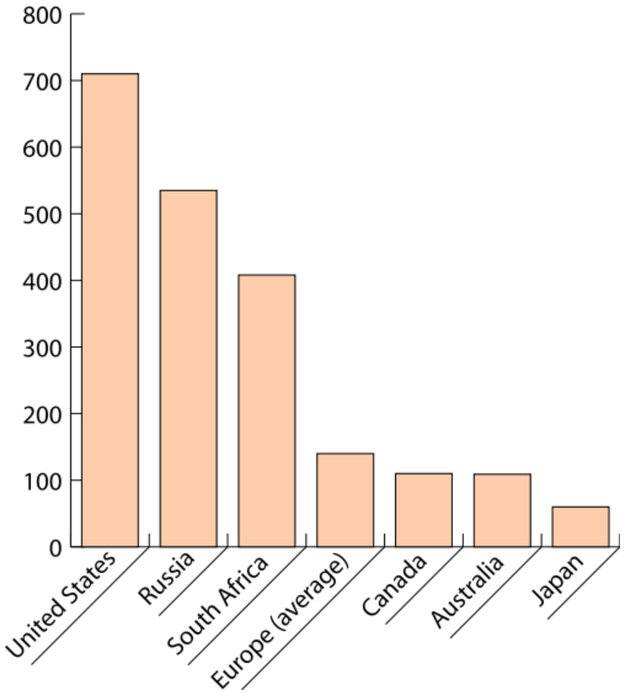


Diagram source: Wikipedia/Frequency

Frequencies

- Given:
  - A set of objects O
  - Which each has a feature f
  - Which takes values from a set V
- To each v in V, we can define two features
  - The absolute frequency of v in O:
    - the number of elements x in o such that x.f = v
      - (requires O finite)
  - The relative frequency of v in O:
    - The absolute frequency/the number of elements in O

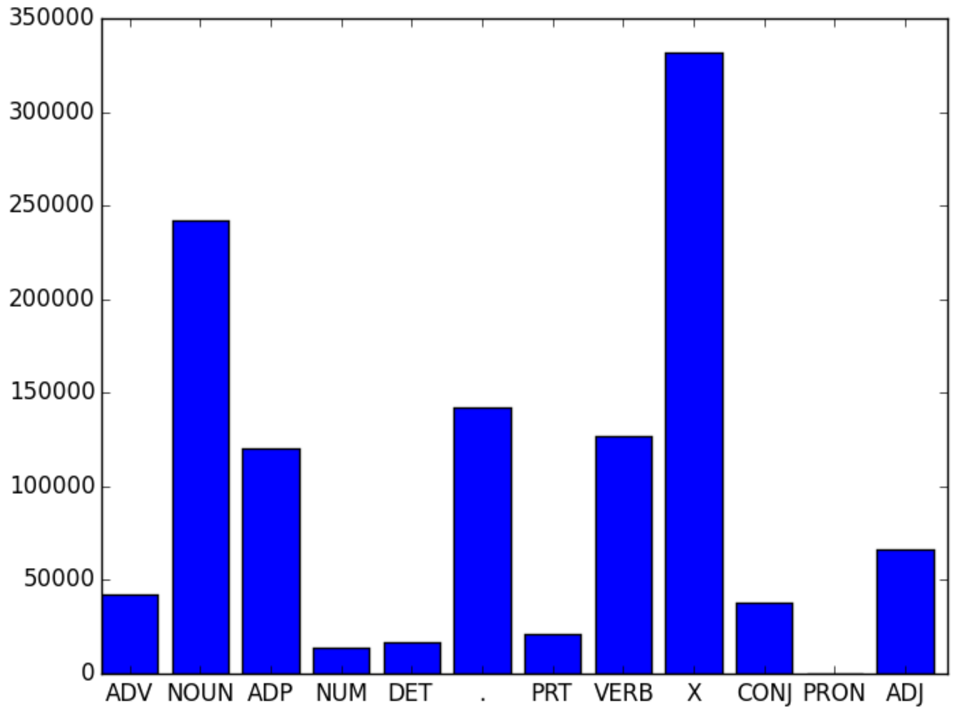
Example

- Brown corpus:
  - ca1.1 mill. words
- For each word occurrence:
  - feature: simplified tag
  - 12 different tags
- Frequency(absolute)
  - for each of the 12 values:
  - the number of occurrences in Brown
- Frequency (relative)
  - the relative number
    - Same graph pattern
    - Different scale

Frequency table

Cat	Freq
ADV	42 155
NOUN	242 056
ADP	120 557
NUM	13 510
DET	16 660
.	142 515
PRT	20 927
VERB	126 743
X	331 754
CONJ	37 718
PRON	252
ADJ	66 345

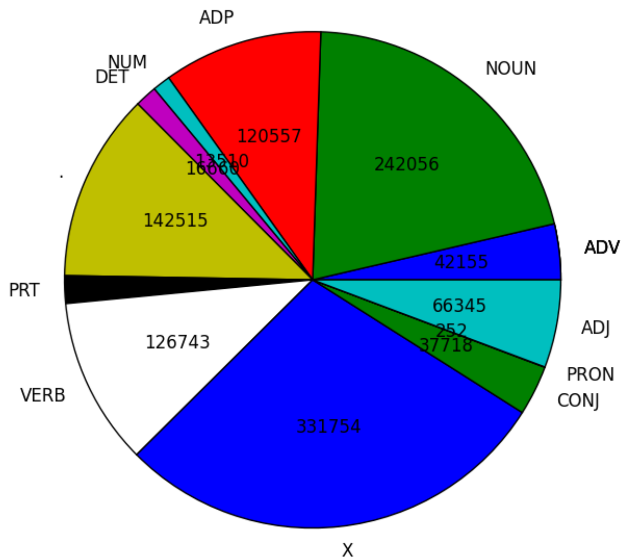
Example

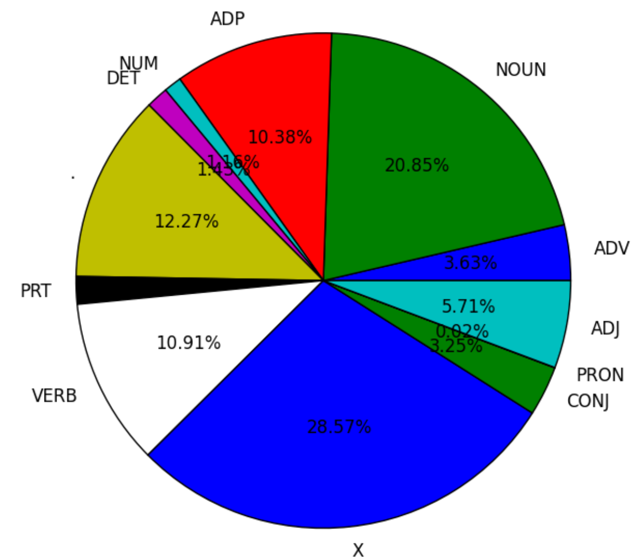


Bar chart

Cat	Freq
ADV	42 155
NOUN	242 056
ADP	120 557
NUM	13 510
DET	16 660
.	142 515
PRT	20 927
VERB	126 743
X	331 754
CONJ	37 718
PRON	252
ADJ	66 345

Pie chart





- A frequency distribution can also be displayed in a pie chart
  - – at least if the set of values isn't too big

Frequencies

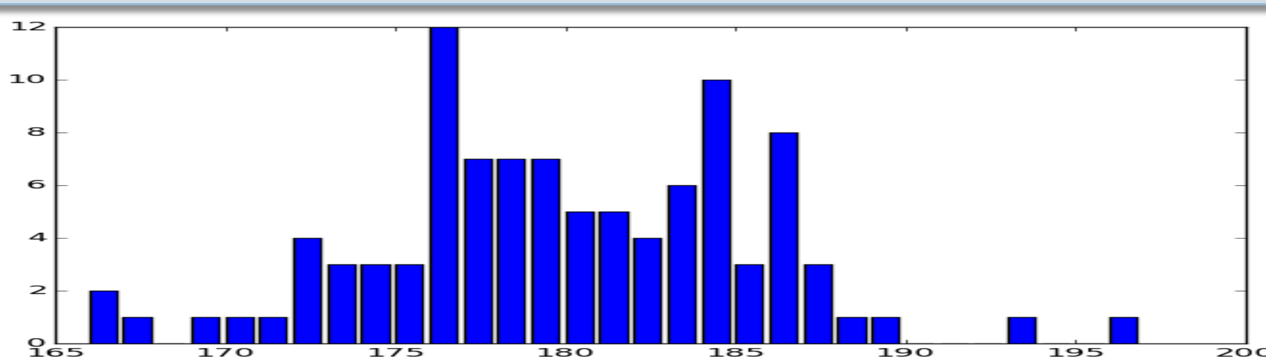
- Frequencies can be defined for all types of value sets V (binary, categoric, numeric) as long as there are only finitely many sets of observations or V is countable,
- But doesn't make much sense for continuous values or for numeric data with vary varied values.

Numerical data

Numeric values

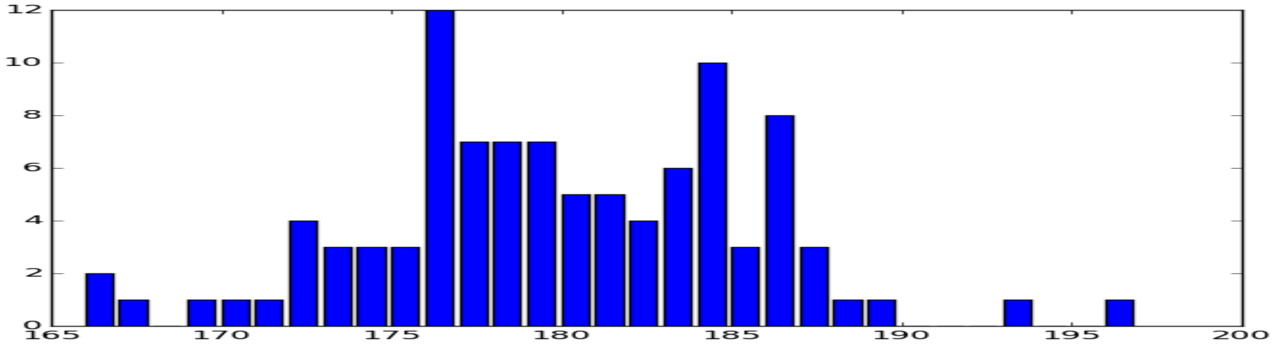
173 172 173 183 177 177 186 180 178 187 179 181 184 172 180 180 171 176 186 175 176 181 176 177 178 176 174 186 172 175 186 183 185 184 176 179 175 193 181 178 177 183 196 187 184 179 182 184 181 176 185 180 176 176 176 167 178 182 176 186 179 176 166 186 169 186 183 178 186 184 179 177 174 176 184 174 177 178 173 182 182 184 185 172 179 179 189 178 170 183 166 188 187 184 184 177 181 180 183 184

Ex 1



- When we have a set of objects with a numeric feature, we may ask more questions:
  - Max? 196
  - Min? 166
  - Middle, average?

Mean, median, mode

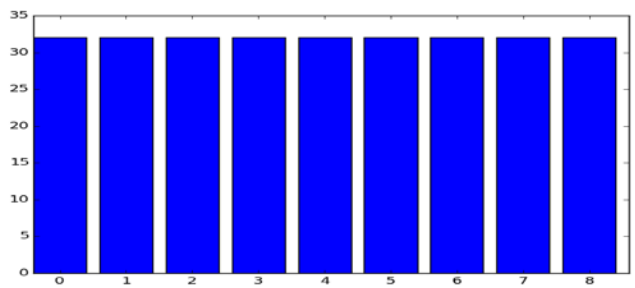


- 3 ways to define "middle", "average"
  - **Median**: equally many above and below, in the example: 179
    - Formally, if the objects are ordered $x_1, x_2, \ldots, x_n$, then the median is $x_{(n/2)}$ if n is even and $(x_{(n-1)/2} + x_{(n+1)/2})/2$ if it is odd.
  - **Mean**: ex: 179.54
    - $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - **Mode**, the most frequent one, ex: 176

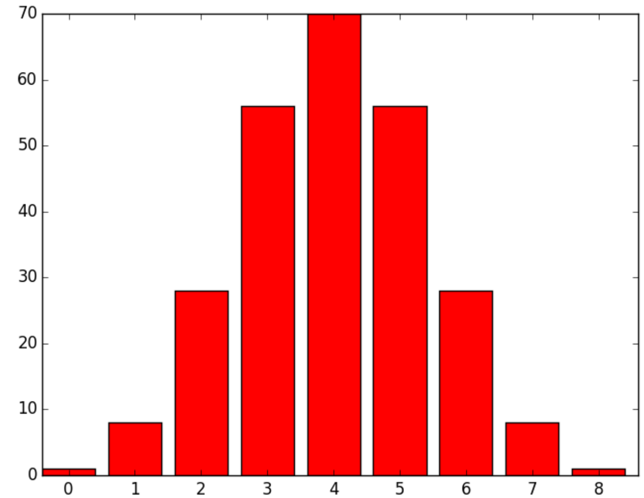
Observe:
Mean and median may be different, e.g.
- Sentence length
- Income

Dispersion

- Median or mean does not say everything
- Nor does max, mean or range (=max-min)
- Example:
  - Two sets
  - The same median=mean=4, min:0, max:8



Ex 2: Uniform



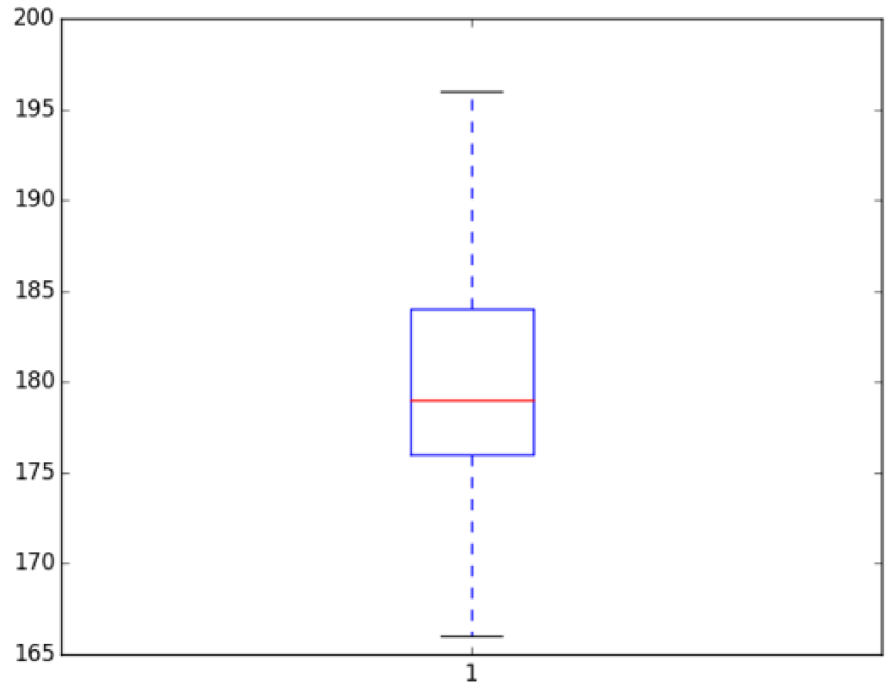
Ex 3: Binomial

Median, quartile, percentile

- The n-percentile p:
  - n percent of the objects are below p
  - $(100–n)$ percent are above p
  - (where $0<n<100$)
- Median is the 50-percentile
- Quartiles are the 25-, 50-, 75-percentiles
  - Split the objects into 4 equally big bins
  - Example 1: 176, 179, 184
  - Example 2: 2, 4, 6; Example 3: 3, 4, 5

Boxplot

- Example 1:
  - Max 196
  - Quartiles:
  - 176, 179, 184
  - Min 166
- Also good for continuous data
- (The exact definition varies, “outlayers”)

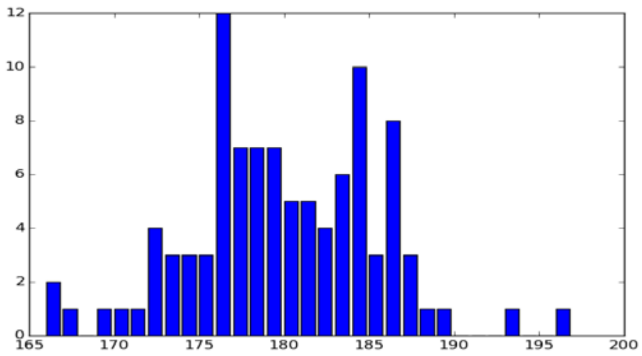


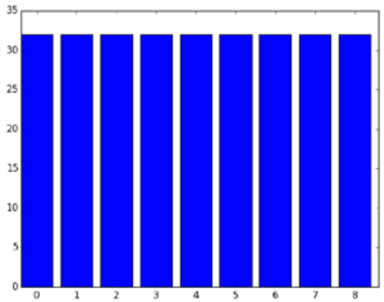
Variance

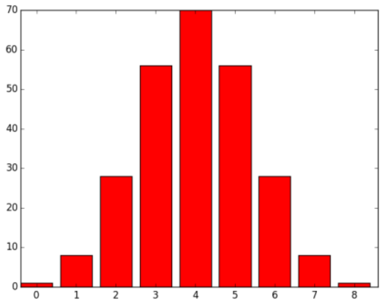
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Idea:
  - Measure how far each point is from the mean
  - Take the average
  - Square – otherwise the average would be 0
- Standard deviation: square root of the variance
  - "Correct dimension and magnitude"

Beware:
For some purposes we will later on divide by (n-1) instead of n.
We return to that!

The examples

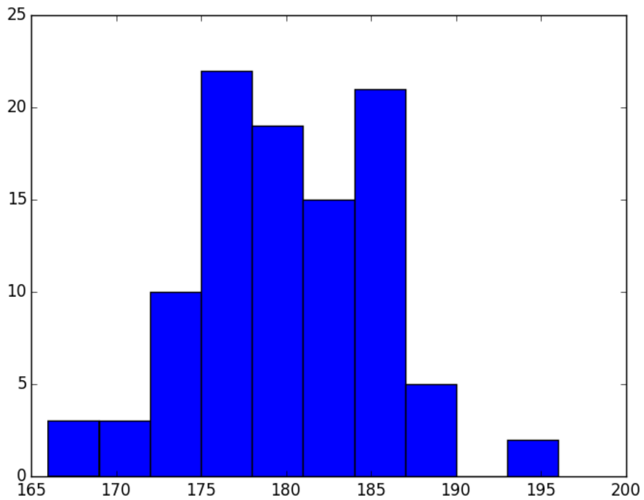




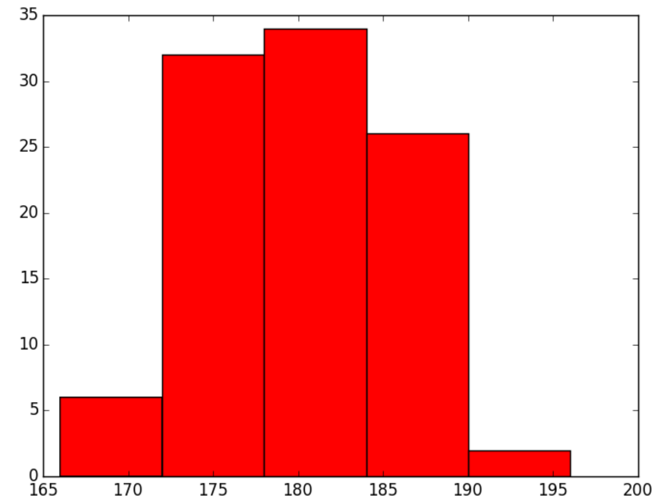


EX	Min	25%	Median	75%	Max	Mean	Vari.	s.d
1	166	176	179	184	196	179.54	30.33	5.5
2	0	2	4	6	8	4	6.67	2.58
3	0	3	4	5	8	4	2.0	1.414

Histogram (≠ bar chart)



Ex 1: 10 bins

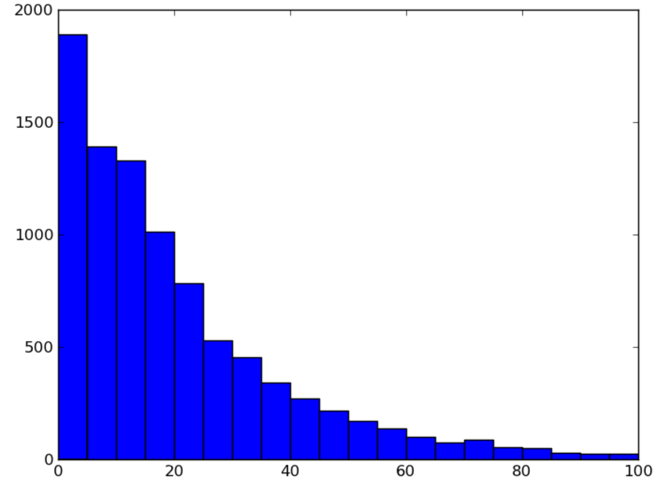


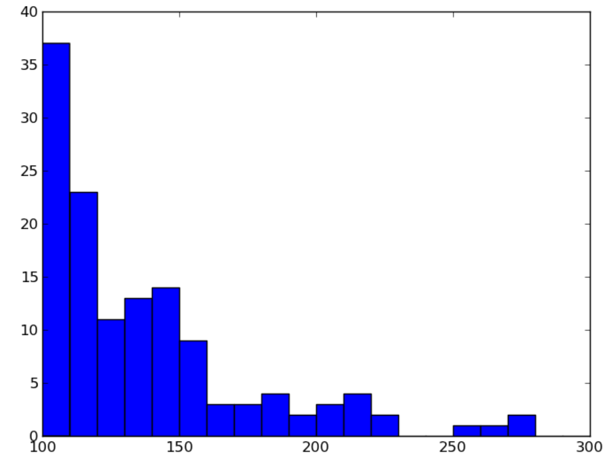
Ex 1: 5 bins

- Shows how many items which takes a value between an m and an n
- Also good for continuous values
  - in contrast to frequency distributions and bar charts

Example: sentence length

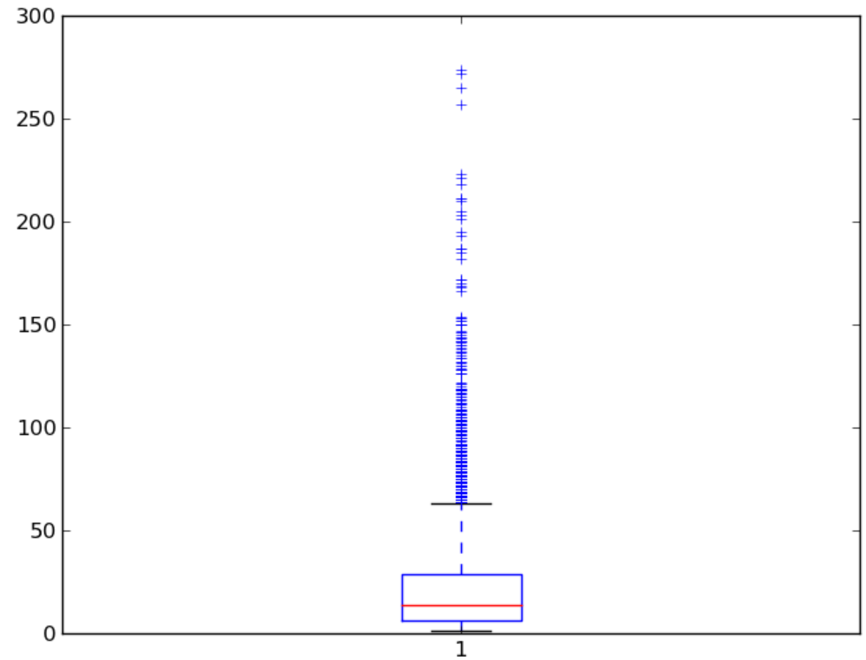
- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86





Example: sentence length

- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86



More than one feature

Example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Observations, O, all occurrences of the five modals in Brown
- For each observations, two parameters
  - f1, which modal, V1 = {can, could, may, might, must, will}
  - f2, genre, V2={news, religion, hobbies, sci-fi, romance, humor}

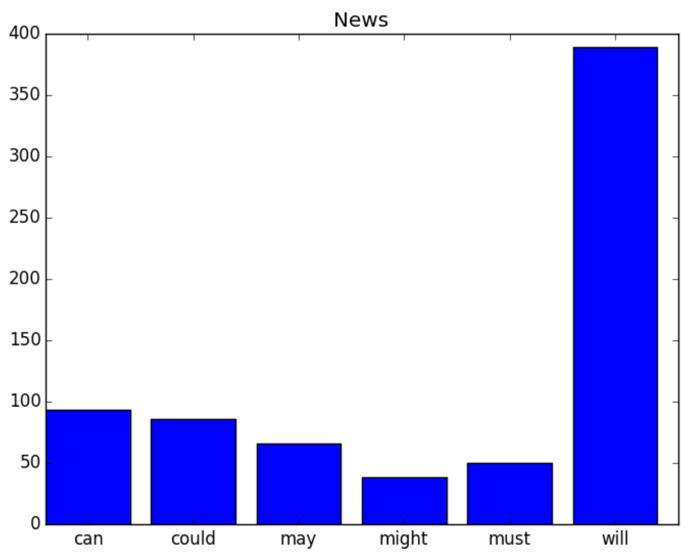
Example NLTK, sec. 2.1

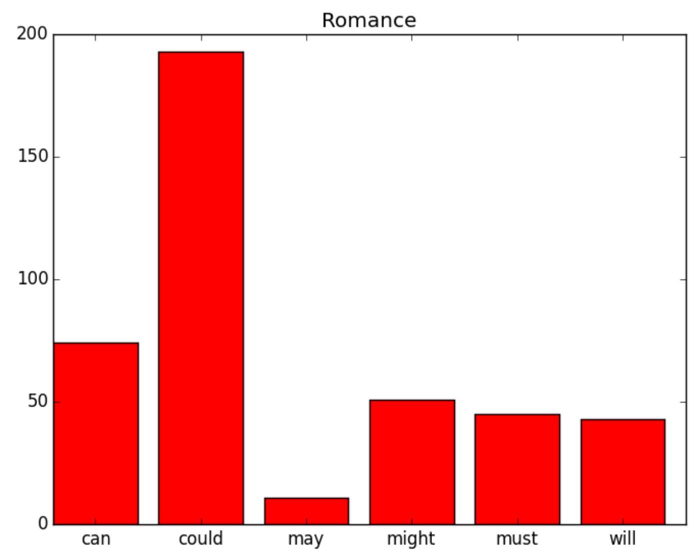
	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Each row and each column is a frequency distribution
- We can make a chart for each row and inspect the differences

Example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

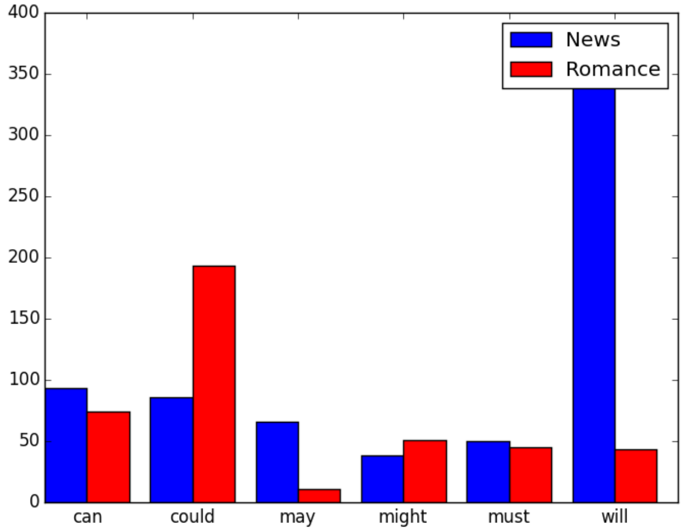




Example NLTK, sec. 2.1

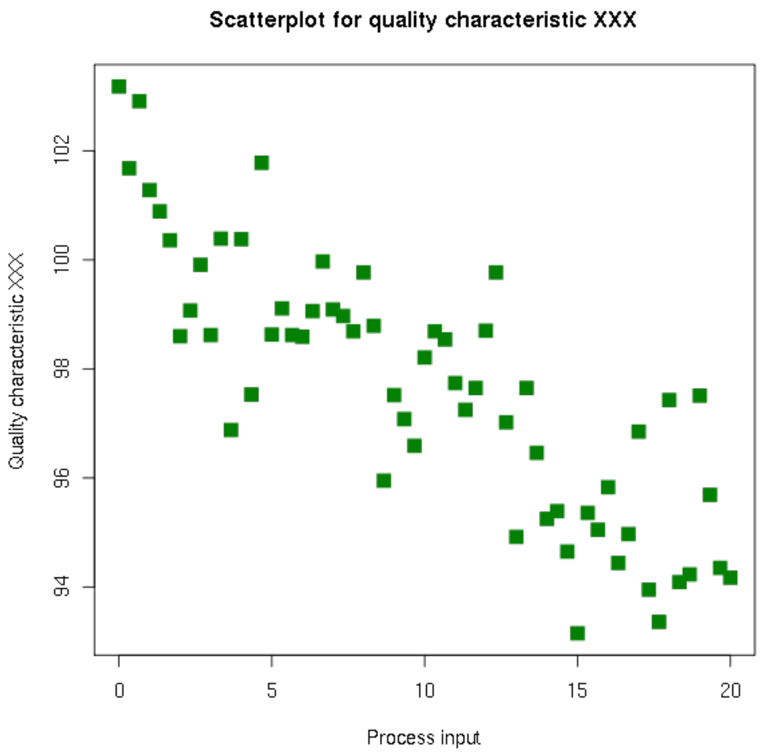
	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Or one may combine several frequency distributions into one chart in some way
- Which way depends on the data



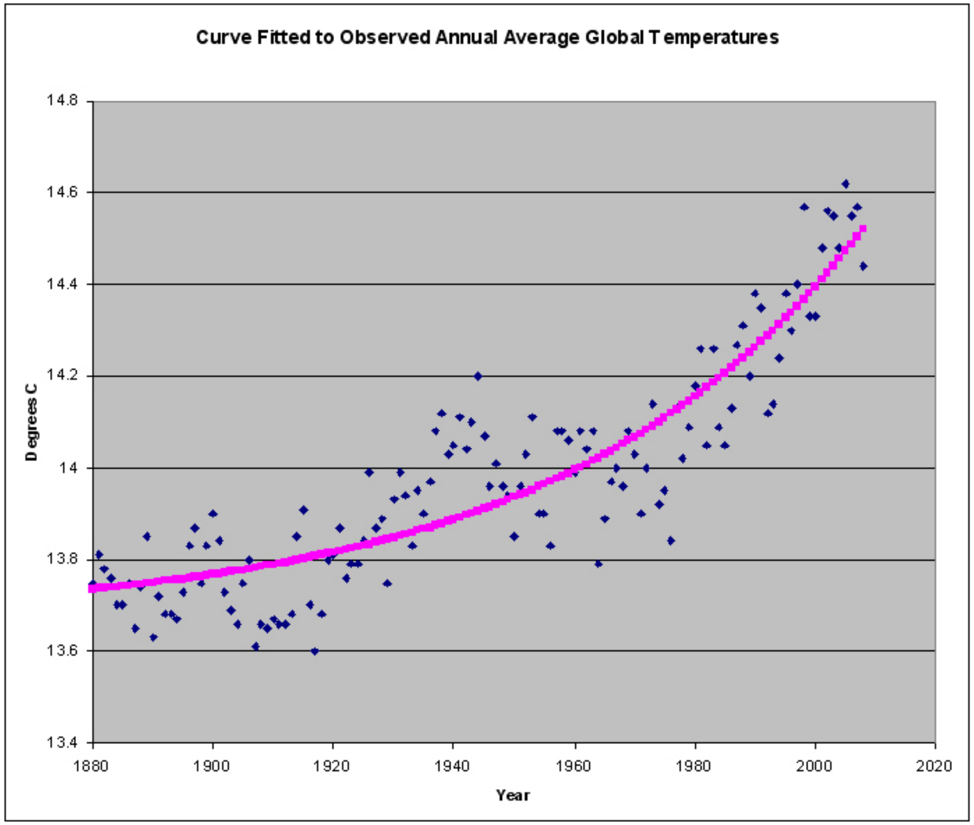
Scatterplots

- With two numerical features, (x, y), the data may be displayed in a scatterplot



Machine learning(ML)

- Two types of ML reflected in scatterplots:
- Is there a law-like connection between f1 and f2 such that we can predict f2 from f1 for unseen events?



ML 2

- The goal is to predict a third categorical feature f3, from f1 and f2:
- Is there a straight line (or some other curve) that does this for us?