Applied Natural Language Processing

Info 256

Lecture 17: Text clustering (Oct. 23, 2023)

David Bamman, UC Berkeley

Topic Models

- **Input**: set of documents, number of clusters to learn.
- **Output**:
  - topics
  - topic ratio in each document
  - topic distribution for each word in doc

{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play
{god, call, give}	**{company, market, business}**	**{math, number, function}**
god	company	math
call	market	number
give	business	function
man	year	code
time	product	set
{city, large, area}	**{math, energy, light}**	**{law, state, case}**
city	math	law
large	energy	state
area	light	case
station	field	court
include	star	legal

Anaphora Resolution	resolution anaphora pronoun discourse antecedent pronouns coreference reference definite algorithm
Automata	string state set finite context rule algorithm strings language symbol
Biomedical	medical protein gene biomedical wkh abstracts medline patient clinical biological
Call Routing	call caller routing calls destination vietnamese routed router destinations gorin
Categorial Grammar	proof formula graph logic calculus axioms axiom theorem proofs lambek
Centering*	centering cb discourse cf utterance center utterances theory coherence entities local
Classical MT	japanese method case sentence analysis english dictionary figure japan word
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phonetic phonology pronunciation vowels phonemes
Comp. Semantics*	semantic logical semantics john sentence interpretation scope logic form set
Dialogue Systems	user dialogue system speech information task spoken human utterance language
Discourse Relations	discourse text structure relations rhetorical relation units coherence texts rst
Discourse Segment.	segment segmentation segments chain chains boundaries boundary seg cohesion lexical
Events/Temporal	event temporal time events tense state aspect reference relations relation
French Function	de le des les en une est du par pour
Generation	generation text system language information knowledge natural figure domain input

Hall et al. 2008, “Studying the History of Ideas Using Topic Models”

Document clustering

iPad Pro: hands-on with Apple's new all-screen tablet
The Verge - 1 hour ago
Apple announced a new, completely redesigned **iPad** Pro moments ago in New York, and I just got to spend a few minutes trying it out.

The New **iPad** Pro Ditches the Home Button For a Giant Display
WIRED - 1 hour ago

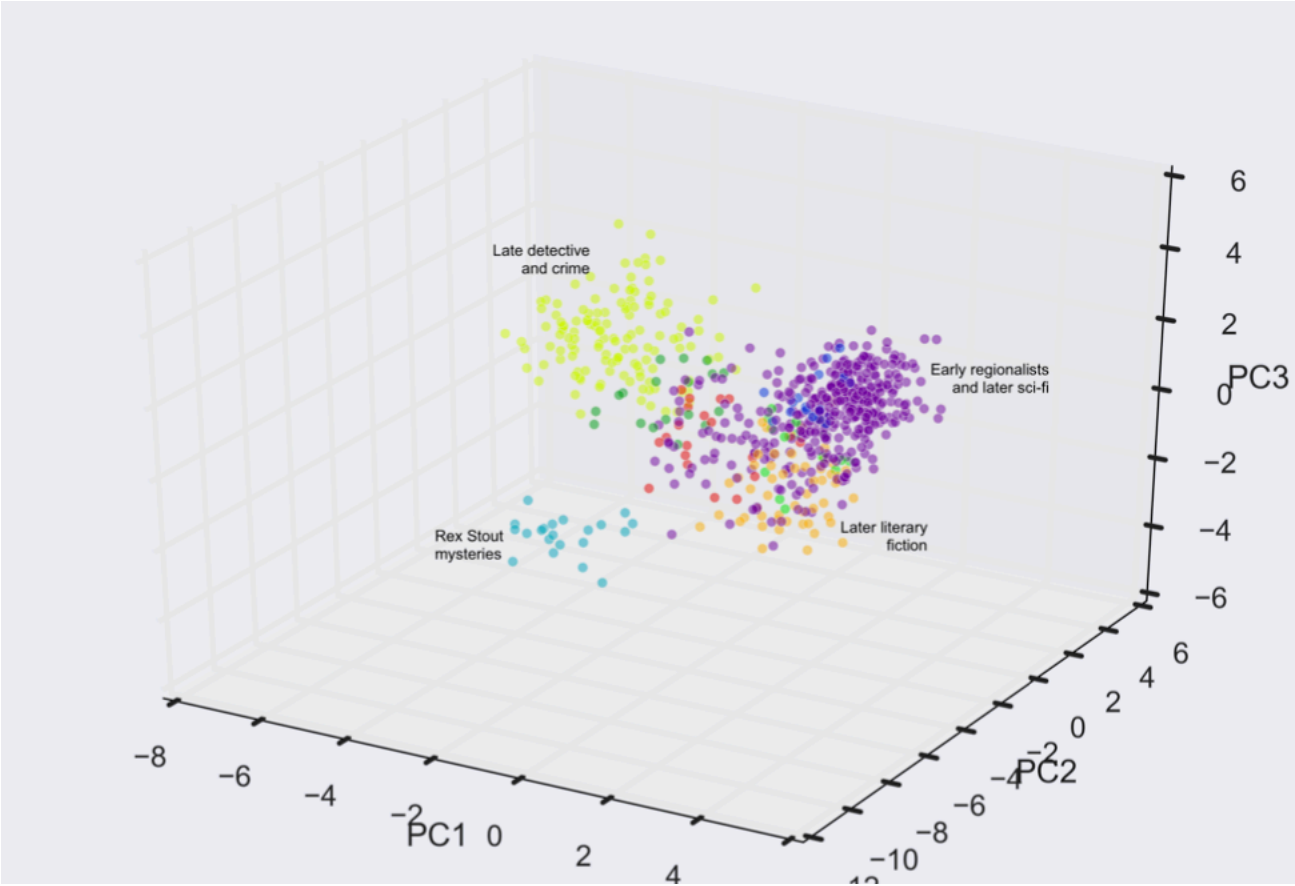
Apple redesigns the **iPad** Pro, breathes new life in the MacBook Air
Opinion - Washington Post - 42 minutes ago

Apple announces new **iPad** Pros and MacBook Air with Retina Display
In-Depth - USA TODAY - 1 hour ago

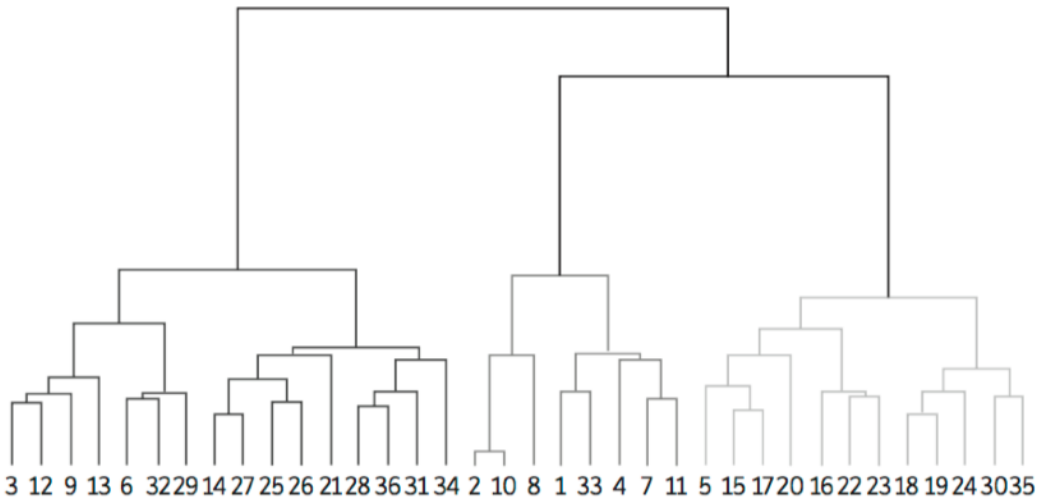
Apple's October 2018 Event: Looking for a New MacBook and **iPad** ...
Live Updating - Wall Street Journal - 52 minutes ago

Our Apple **iPad** Pro Event Liveblog Is Right Here
Blog - Gizmodo - 20 hours ago

Google News



Wilkens, “Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction” (2016)



A Midsummer Night's Dream (3)
Twelfth Night (12)
Much Ado About Nothing (9)
Two Gentlemen (13)
Measure for Measure (6)
Othello (32)
Julius Caesar (29)

The Winter's Tale (14)
Cymbeline (27)
Antony and Cleopatra (25)
Coriolanus (26)
Henry VIII (21)
Hamlet (28)
Troilus and Cressida (36)
Macbeth (31)
Timon of Athens (34)

All's Well That Ends Well (2)
Taming of the Shrew (10)
Merry Wives of Windsor (8)
A Midsummer Night's Dream (1)
Romeo and Juliet (33)
Comedy of Errors (4)
Merchant of Venice (7)
The Tempest (11)

Love's Labours' Lost (5)
1 Henry IV (15)
2 Henry IV (17)
Henry V (20)
1 Henry VI (16)
King John (22)
Richard II (23)

2 Henry VI (18)
2 Henry VI (19)
Richard III (24)
King Lear (30)
Titus Andronicus (35)

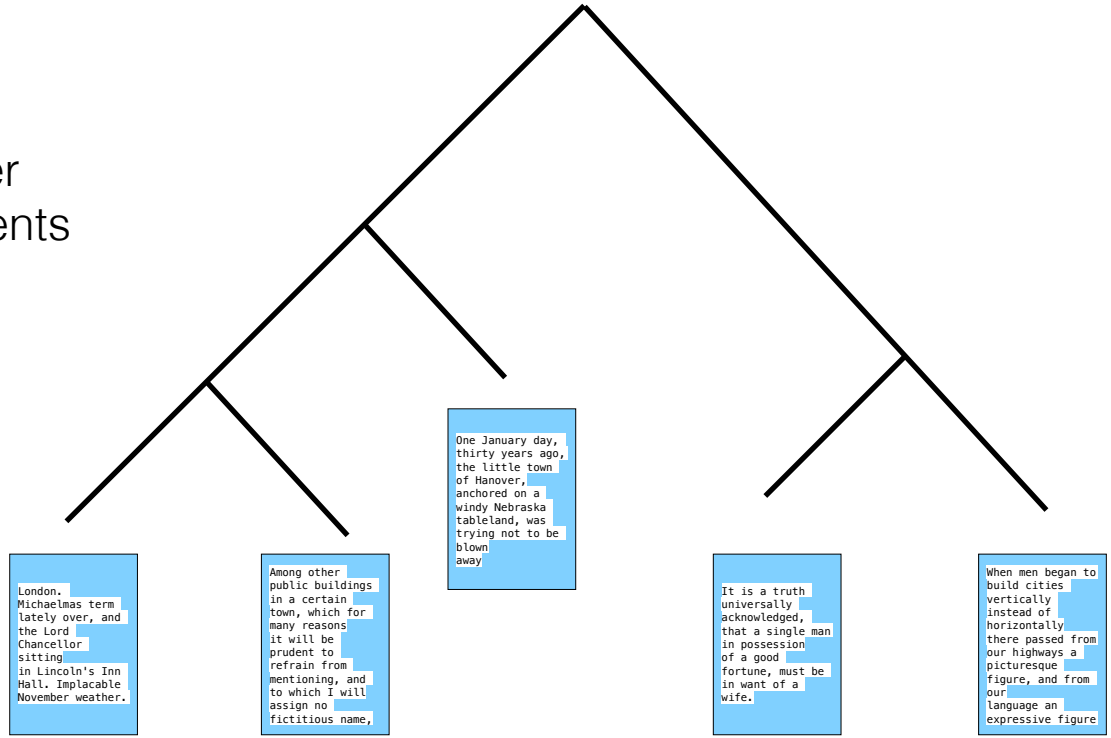
Allison et al. 2009

Clustering

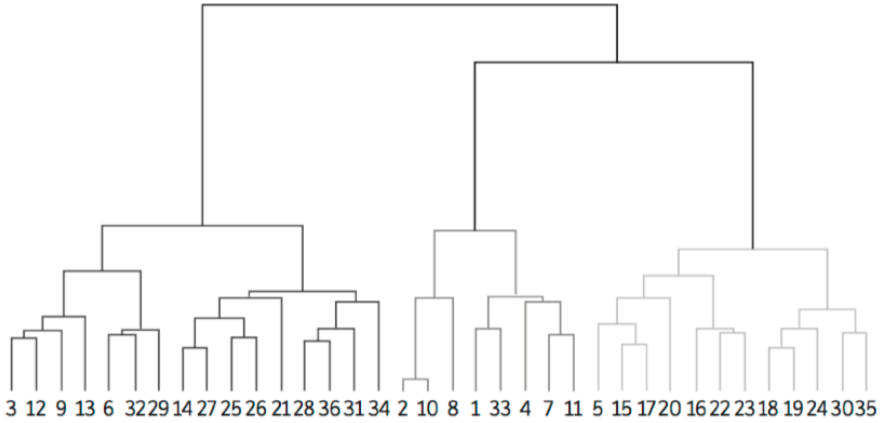
- Clustering is designed to learn structure in the data:
    - Hierarchical structure between data points
    - Natural partitions between data points

Hierarchical Clustering

- *Hierarchical* order among the elements being clustered



Hierarchical clustering



A Midsummer Night's Dream (3)
Twelfth Night (12)
Much Ado About Nothing (9)
Two Gentlemen (13)
Measure for Measure (6)
Othello (32)
Julius Caesar (29)

The Winter's Tale (14)
Cymbeline (27)
Antony and Cleopatra (25)
Coriolanus (26)
Henry VIII (21)
Hamlet (28)
Troilus and Cressida (36)
Macbeth (31)
Timon of Athens (34)

All's Well That Ends Well (2)
Taming of the Shrew (10)
Merry Wives of Windsor (8)
A Midsummer Night's Dream (1)
Romeo and Juliet (33)
Comedy of Errors (4)
Merchant of Venice (7)
The Tempest (11)

Love's Labours' Lost (5)
1 Henry IV (15)
2 Henry IV (17)
Henry V (20)
1 Henry VI (16)
King John (22)
Richard II (23)

2 Henry VI (18)
2 Henry VI (19)
Richard III (24)
King Lear (30)
Titus Andronicus (35)

Allison et al. 2009

Bottom-up clustering

Algorithm 1 Hierarchical agglomerative clustering

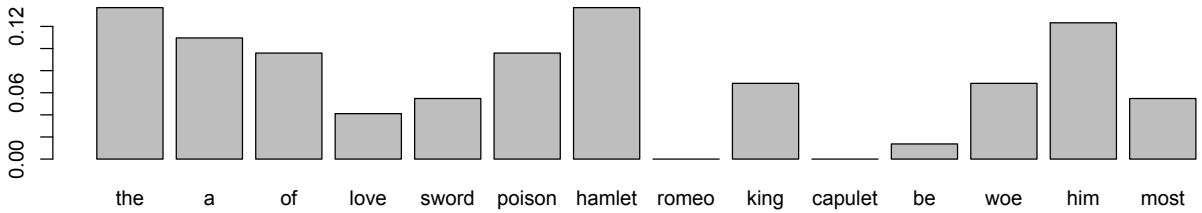
1: Data: N training data points $x \in \mathbb{R}^F$
2: Let X denote a set of objects x
3: Given some linkage function $d(X, X') \to \mathbb{R}$
4: Initialize clusters $\mathcal{C} = \{C_1, \ldots, C_N\}$ to singleton data points
5: **while** data points not in one cluster **do**
6: Identify X, Y as clusters with smallest linkage function among clusters in $\mathcal{C}$
7: Create new cluster $Z = X \cup Y$
8: remove X, Y from $\mathcal{C}$
9: add Z to $\mathcal{C}$
10: **end while**

Similarity

$$\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$$

- What are you comparing?
- How do you quantify the similarity/difference of those things?

Unigram probability



TF-IDF

- Term frequency ($tf_{t,d}$) = the number of times term t occurs in document d; several variants (e.g., passing through log function).

- Inverse document frequency = inverse fraction of number of documents containing (D_t) among total number of documents N

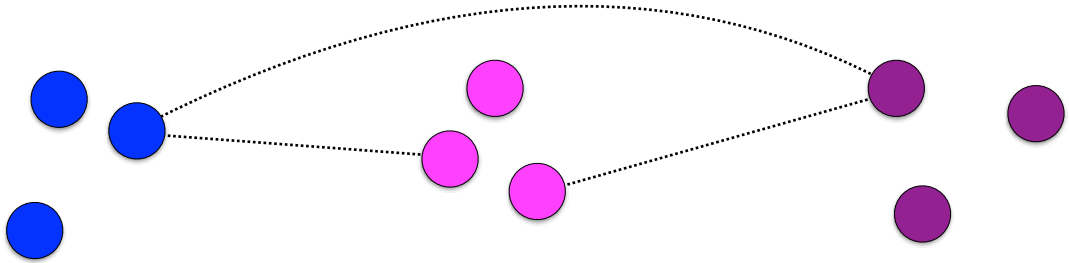
$$tfidf(t,d) = tf_{t,d} \times \log \frac{N}{D_t}$$

Hierarchical clustering

We know how to compare data points with distance metrics.

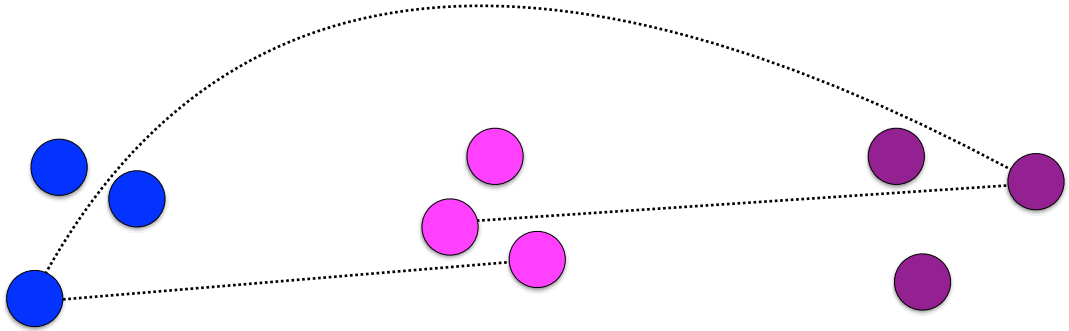
How do we compare sets of data points?

Single linkage



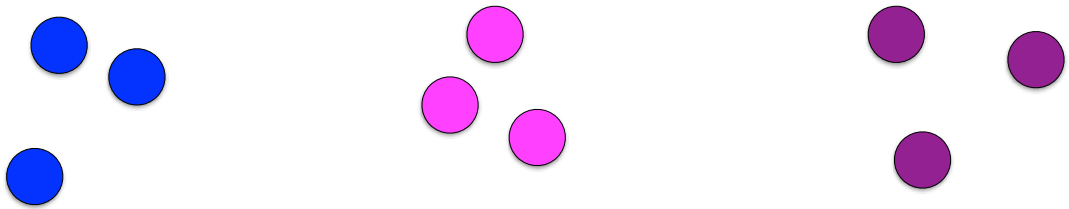
$$\min_{x \in A,\, y \in B} \mathrm{Dis}(x, y)$$

Complete linkage

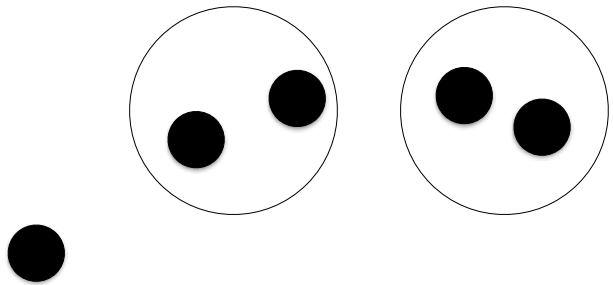


$$\max_{x \in A,\, y \in B} \mathrm{Dis}(x, y)$$

Average linkage

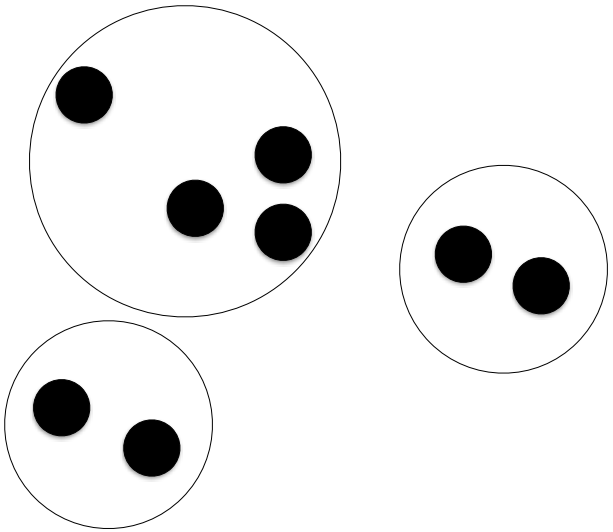


$$\frac{\sum_{x \in A,\, y \in B} \mathrm{Dis}(x, y)}{|A| \times |B|}$$



Single linkage may link bigger clusters together before outliers

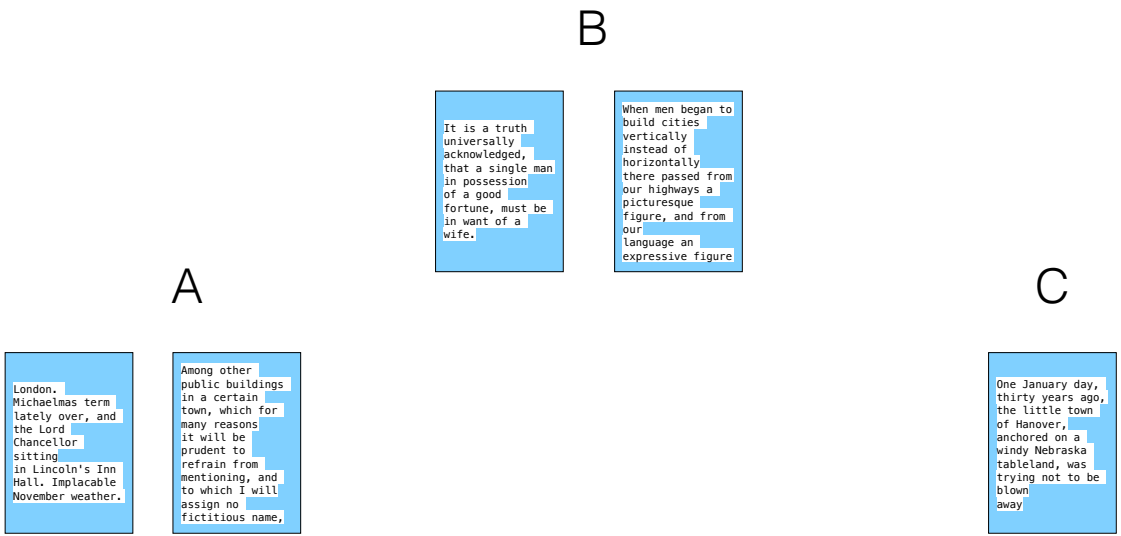
Complete linkage



Complete linkage may *not* link close clusters together because of outliers

Flat Clustering

- Partitions the data into a set of *K* clusters

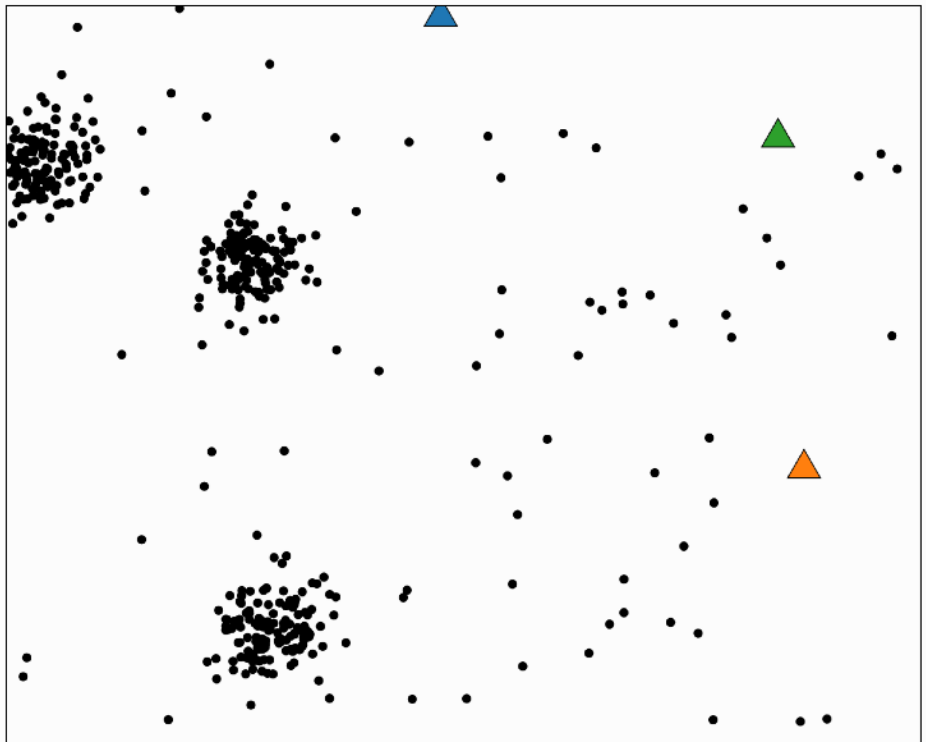


K-means

Algorithm 1 K-means

Data: training data $x \in \mathbb{R}^F$
Given some distance function $d(x, x') \to \mathbb{R}$
Select k initial centers $\{\mu_1, \ldots, \mu_k\}$
while not converged **do**
    for $i = 1$ to N **do**
        Assign x_i to $\arg\min_c d(x_i, \mu_c)$
    end for
    for $i = 1$ to K **do**
        $\mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i$
    end for
end while

Visualizing K-Means Clustering

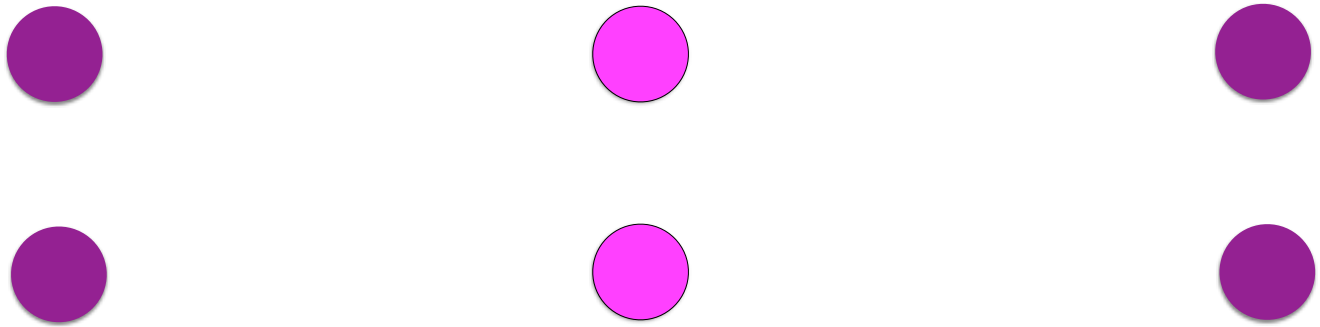


https://github.com/karanveerm/kmeans

Problems

K-means

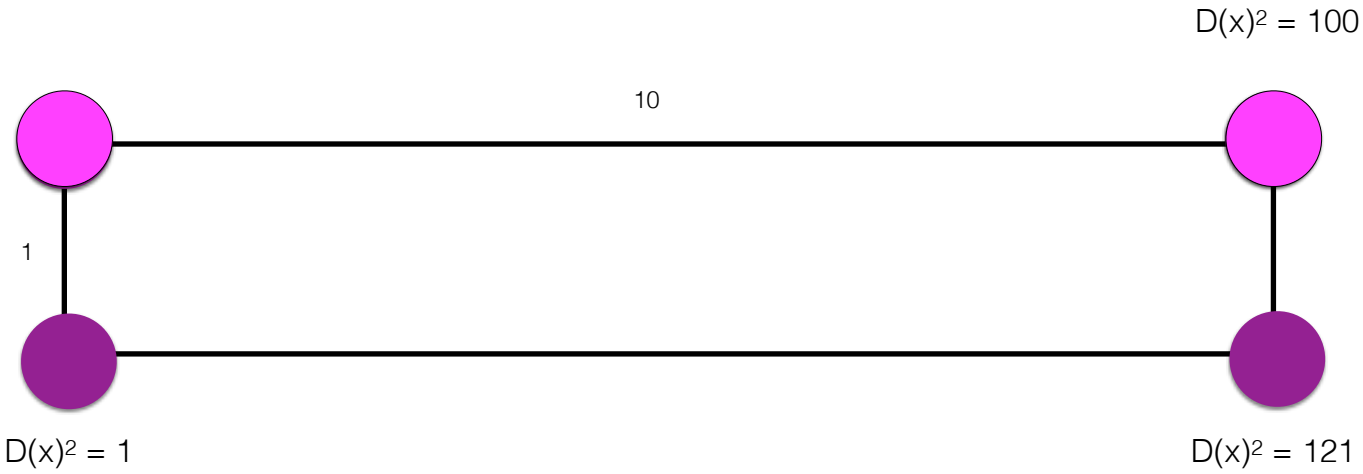
initial cluster centers



K-means++

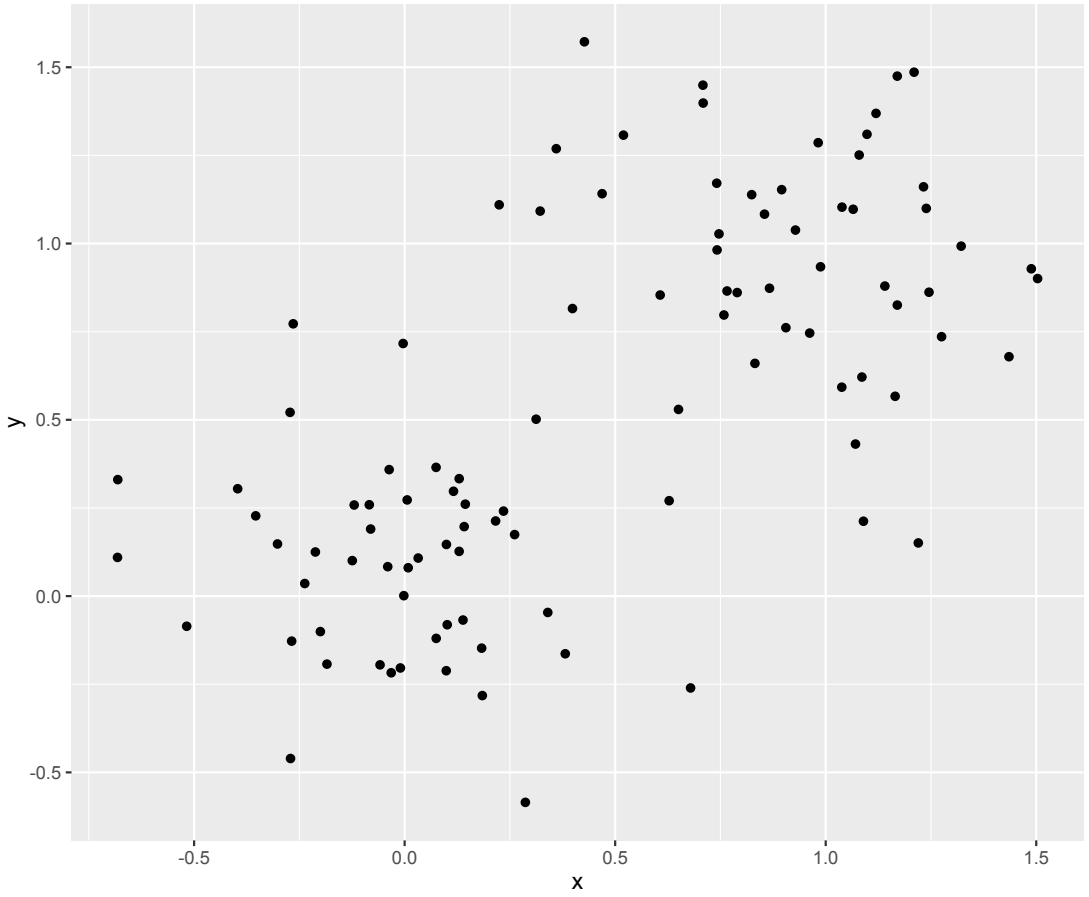
- Improved initialization method for K-means:
  - Choose data point at random as first center
  - For all other data points x, calculate the distance D(x) between x and the nearest cluster center
  - Choose new data point x as next center, with probability proportional to $D(x)^2$
  - Repeat until *K* centers are selected

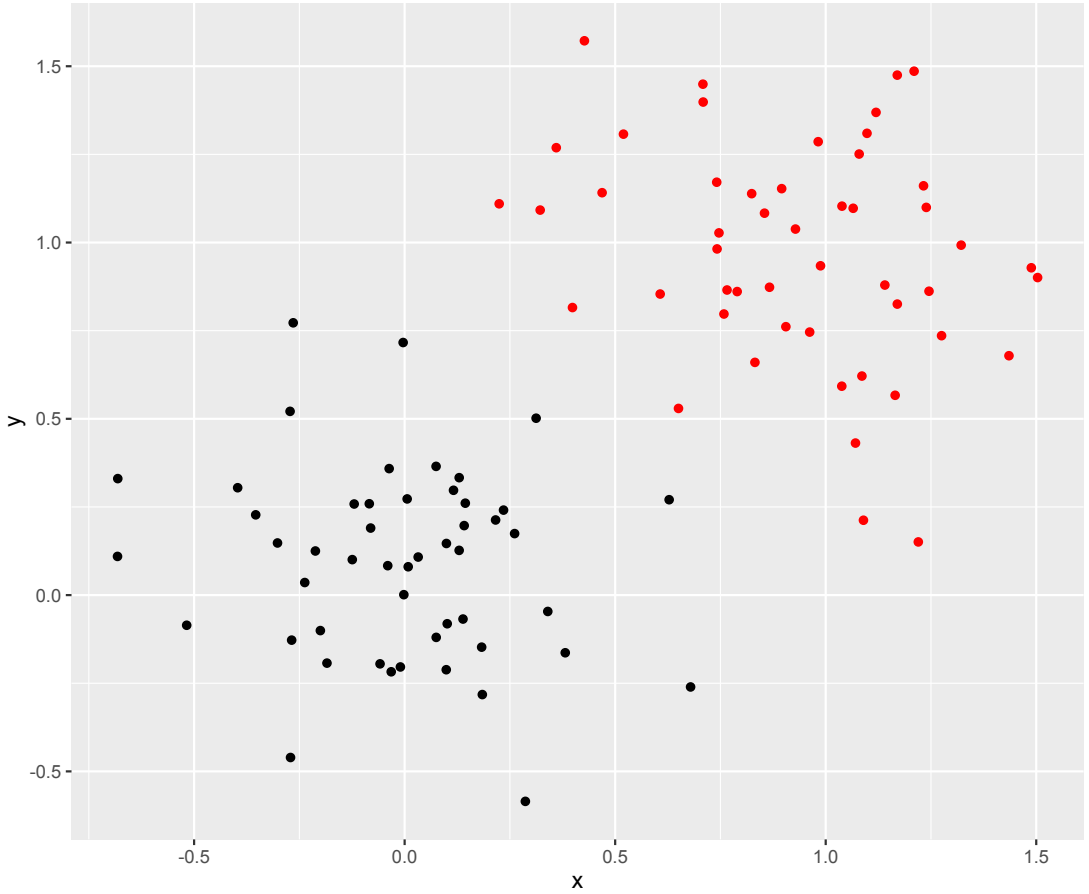
K-means++

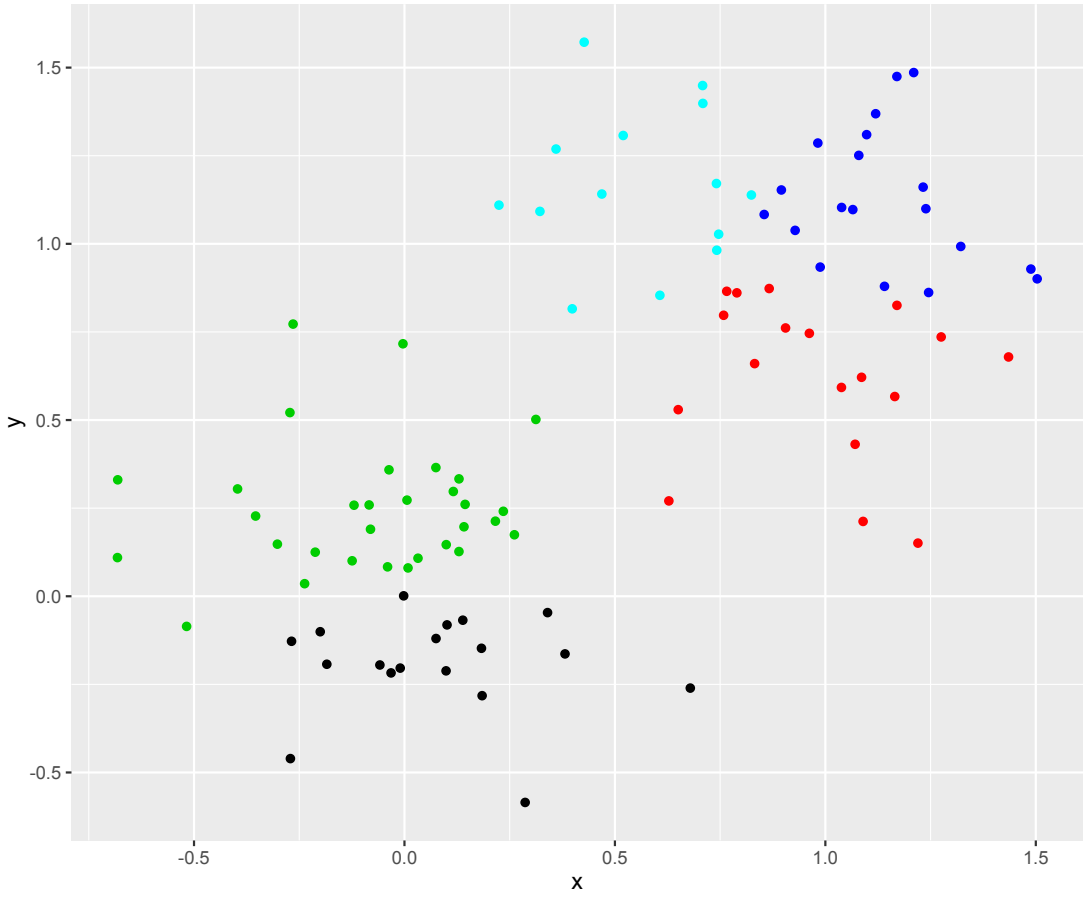


Choosing K

- how do we choose K?

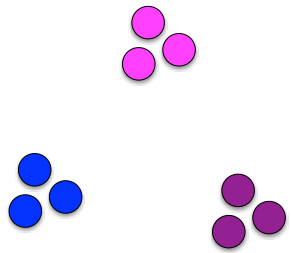

1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x


1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x

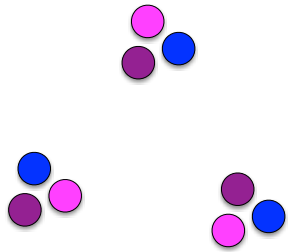

1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x

The “elbow”

Core idea: clusters should minimize the within-cluster variance



good



bad

The “elbow”

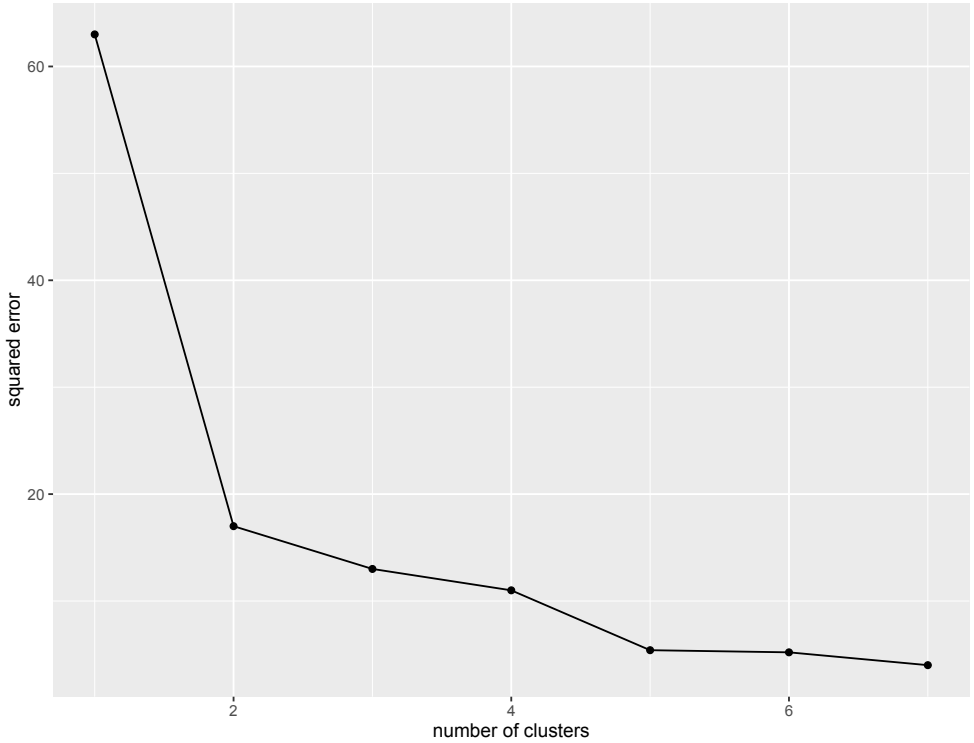
Core idea: clusters should minimize the within-cluster variance

within-cluster
sum of squares

$$\sum_{i=1}^{F} (x_i - \mu_i)^2$$

for each cluster

The “elbow”



Representation

$$x \in \mathbb{R}^F$$

[x is a data point characterized by F real numbers, one for each feature]

- This is a huge decision that impacts what you can learn

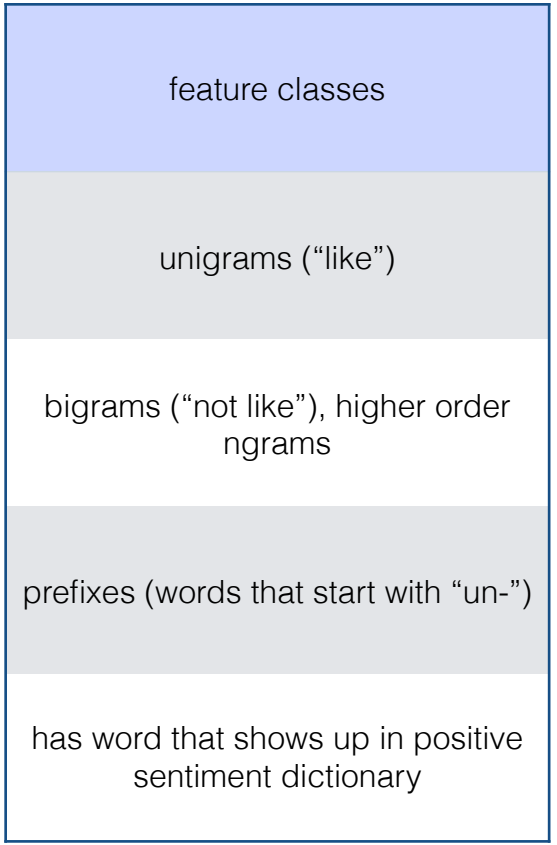
Bag of words

Representation of text only as the counts/relative frequencies/binary indicators of words that it contains

	doc1	doc2
the	0.04	0.04
of	0	0
hate	0	0.03
genius	0.005	0
bravest	0.002	0
stupid	0	0.001
like	0	0.01
…		

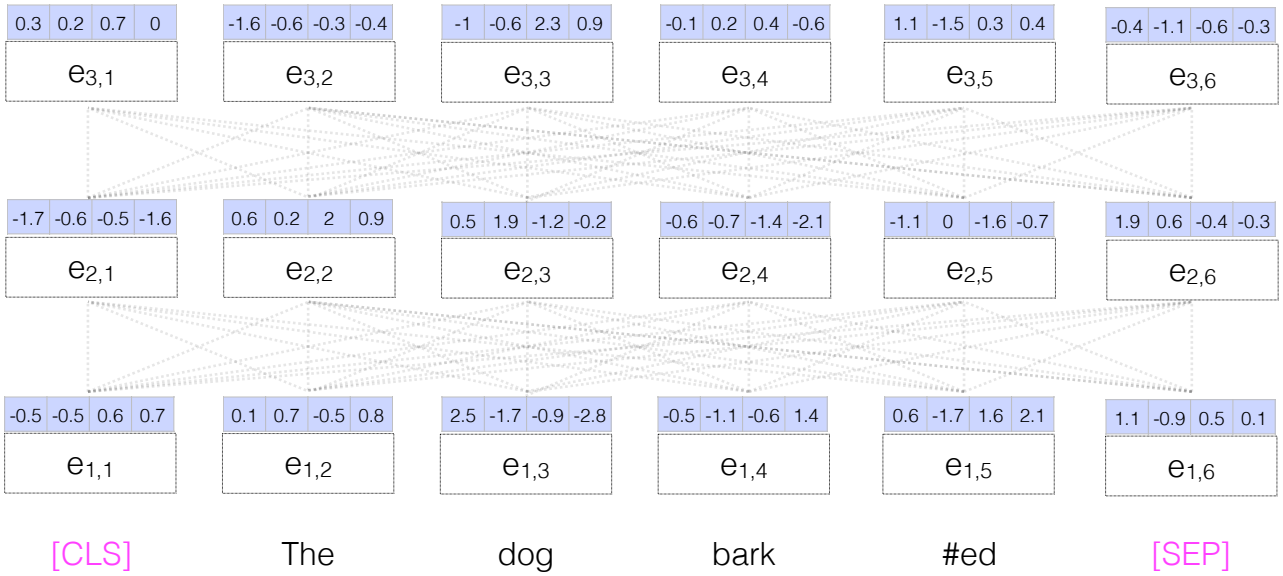
Features

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

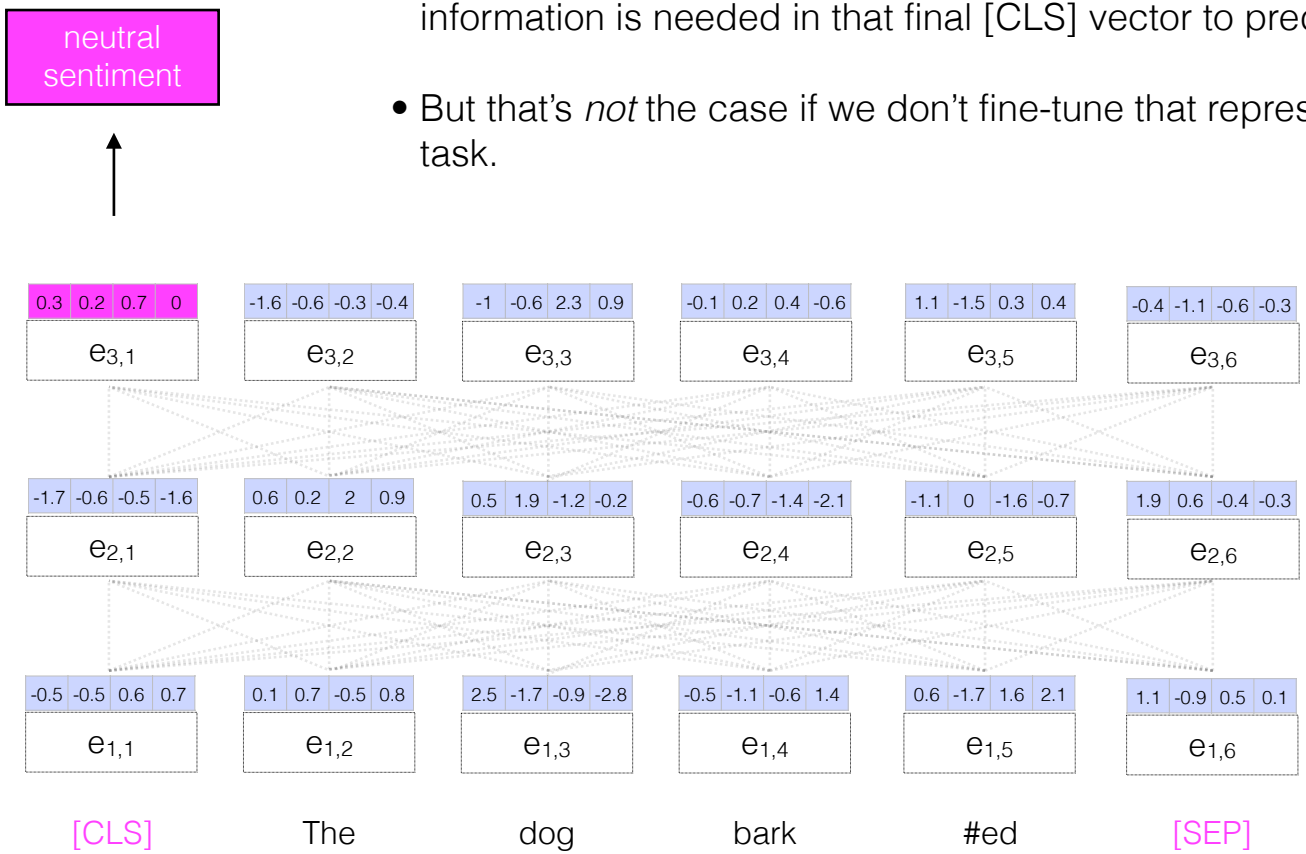


- BERT also encodes each sentence by appending a special token to the beginning ([CLS]) and end ([SEP]) of each sequence.

- This helps provides a single token that can be optimized to represent the entire sequence (e.g., for document classification)

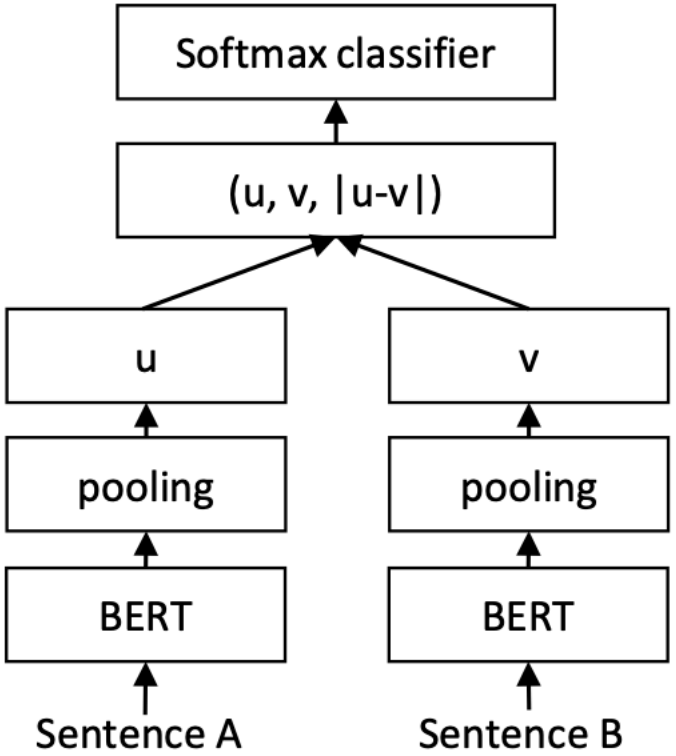


- When used for a supervised problem (e.g. sentient analysis), ALL of the parameters of BERT are optimized to encode whatever information is needed in that final [CLS] vector to predict the label.
- But that's *not* the case if we don't fine-tune that representation for a task.



SentenceBERT

- Given training data in the form of related sentence pairs (from SNLI/MNLI), optimize the representation to predict the natural language inference relation between sentences (contradiction/entailment/neutral).



Reimers and Gurevych (2019) "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks"

SentenceBERT

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

- Good performance on STS (semantic textual similarity — how similar are two sentences?) is a by-product of this training.

Representation

- Books (e.g., to learn genres)
- News articles (e.g., to learn articles about the same event)

Evaluation

- Much more complex than supervised learning since there's often no notion of "truth"

Internal criteria

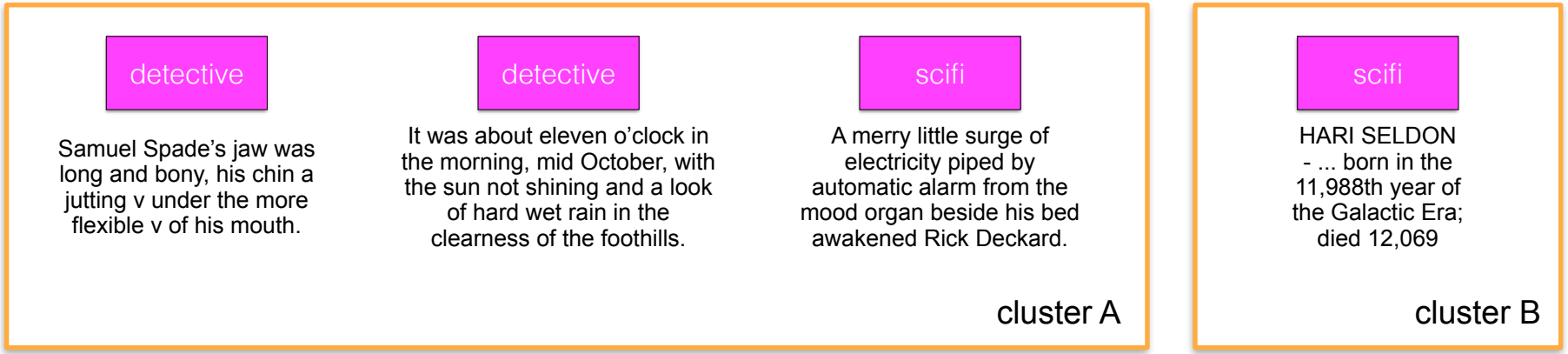
- Elements within clusters should be more similar to each other
- Elements in different clusters should be less similar to each other

External criteria

- How closely does your clustering reproduce another (“gold standard”) clustering?

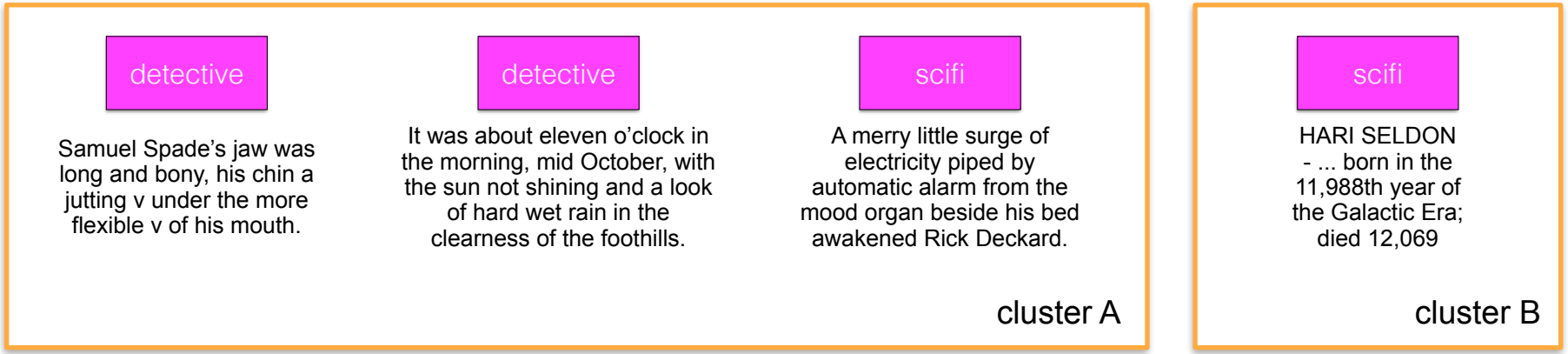
External criteria

- Let's presume that we've run clustering over our data (to generate clusters A and B) and that we also have a separate human-labeled clustering (here, assigning each data point to detective or sci-fi clusters.



External criteria

- How much does the cluster for a data point tell you about its likely label (and vice versa)?



External criteria

- How much does the cluster for a data point tell you about its likely label (and vice versa)?

	Samuel Spade's jaw was long and bony, his chin a jutting v under the more flexible v of his mouth.	It was about eleven o'clock in the morning, mid October, with the sun not shining and a look of hard wet rain in the clearness of the foothills.	A merry little surge of electricity piped by automatic alarm from the mood organ beside his bed awakened Rick Deckard.	HARI SELDON - ... born in the 11,988th year of the Galactic Era; died 12,069
Gold label	detective	detective	science fiction	science fiction
Cluster	A	A	A	B

Mutual information

- Mutual information provides a measure of how independent two variables (X and Y) are.

	Samuel Spade's jaw was long and bony, his chin a jutting v under the more flexible v of his mouth.	It was about eleven o'clock in the morning, mid October, with the sun not shining and a look of hard wet rain in the clearness of the foothills.	A merry little surge of electricity piped by automatic alarm from the mood organ beside his bed awakened Rick Deckard.	HARI SELDON - ... born in the 11,988th year of the Galactic Era; died 12,069
Gold label	detective	detective	science fiction	science fiction
Cluster	A	A	A	B

Mutual information

	Samuel Spade's jaw was long and bony, his chin a jutting v under the more flexible v of his mouth.	It was about eleven o'clock in the morning, mid October, with the sun not shining and a look of hard wet rain in the clearness of the foothills.	A merry little surge of electricity piped by automatic alarm from the mood organ beside his bed awakened Rick Deckard.	HARI SELDON - ... born in the 11,988th year of the Galactic Era; died 12,069
G	detective	detective	science fiction	science fiction
C	A	A	A	B

$$\text{MI(G, C)} = \sum_{i=1}^{|G|} \sum_{j=1}^{|C|} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$$

G = {detective, scifi}
C = {A, B}

$$P(\text{detective}) = \frac{\text{count(detective)}}{N}$$

$$P(\text{A}) = \frac{\text{count(A)}}{N}$$

$$P(\text{detective, A}) = \frac{\text{count(detective, A)}}{N}$$

Normalized mutual information

- MI is bounded by 0 below (compete independence), but can range to ∞.
- Normalized mutual information bounds that value between [0,1] by normalizing by the average entropy of the two groups.

$$\text{NMI}(G, C) = \frac{2 \times \text{MI(G,C)}}{H(G) + H(C)}$$

$$H(G) = -\sum_{i=1}^{|G|} P(i) \log P(i)$$

Activity

`https://bit.ly/anlp-clustering`

- Use SentenceBERT + K-means to cluster movie summaries/book titles; upload this at the end of class.