Exploratory Data Analysis: Clustering

Sandeep Soni

02/06/2024

Google News

Search for topics, locations & sources

Home | For you | Following | News Showcase | U.S. | World | Local | Business | Technology | Entertainment | Sports | Science | Health

Technology >

MacRumors

Apple Significantly Lowers Repair Fees for iPhone 15 Pro Models With Cracked Back...

4 hours ago

USA TODAY

iOS 17 release: Apple's new operating system has a lot to offer

13 hours ago

Newsmax

FDA to Look Into iPhone 12 Radiation Concerns

2 days ago

Entertainment >

THE WALL STREET JOURNAL.

Drew Barrymore, Other Talk Shows Halt Season Premieres Until Industry Strikes End

27 minutes ago

Variety

Billy Miller, 'The Young and the Restless' and 'General Hospital' Actor, Dies at 43

6 hours ago

Daily Mail

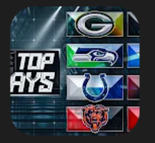
Teyana Taylor reveals split from Iman Shumpert after seven years of marriage and insists...

1 hour ago

Sports >

FOX SPORTS

NFL Week 2 top plays: Falcons, Bucs win, Giants-Cardinals, 49ers-Rams live



1 hour ago

newyorkjets.com

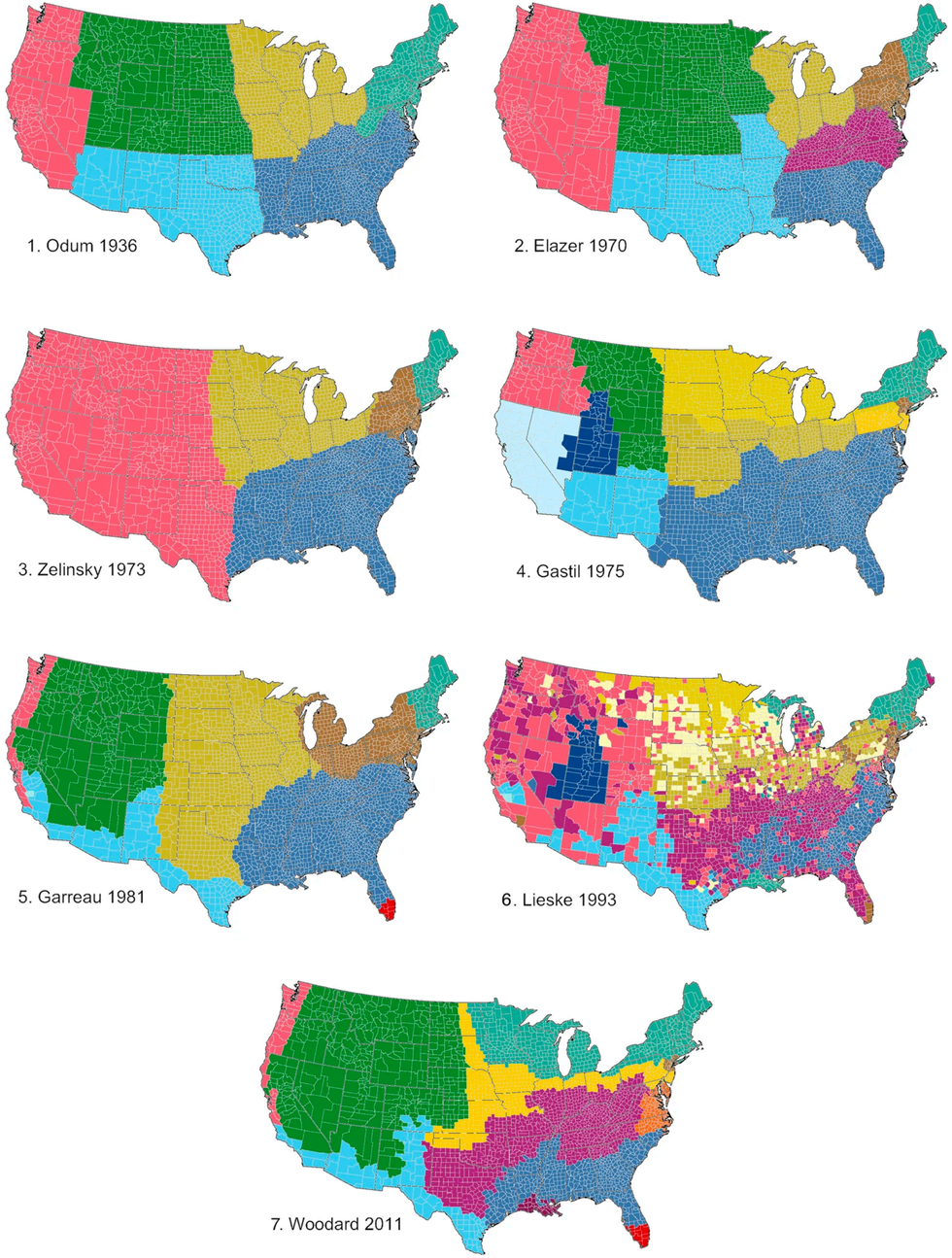
New York Jets vs. Dallas Cowboys Game Inactives - Week 2 2023

4 hours ago

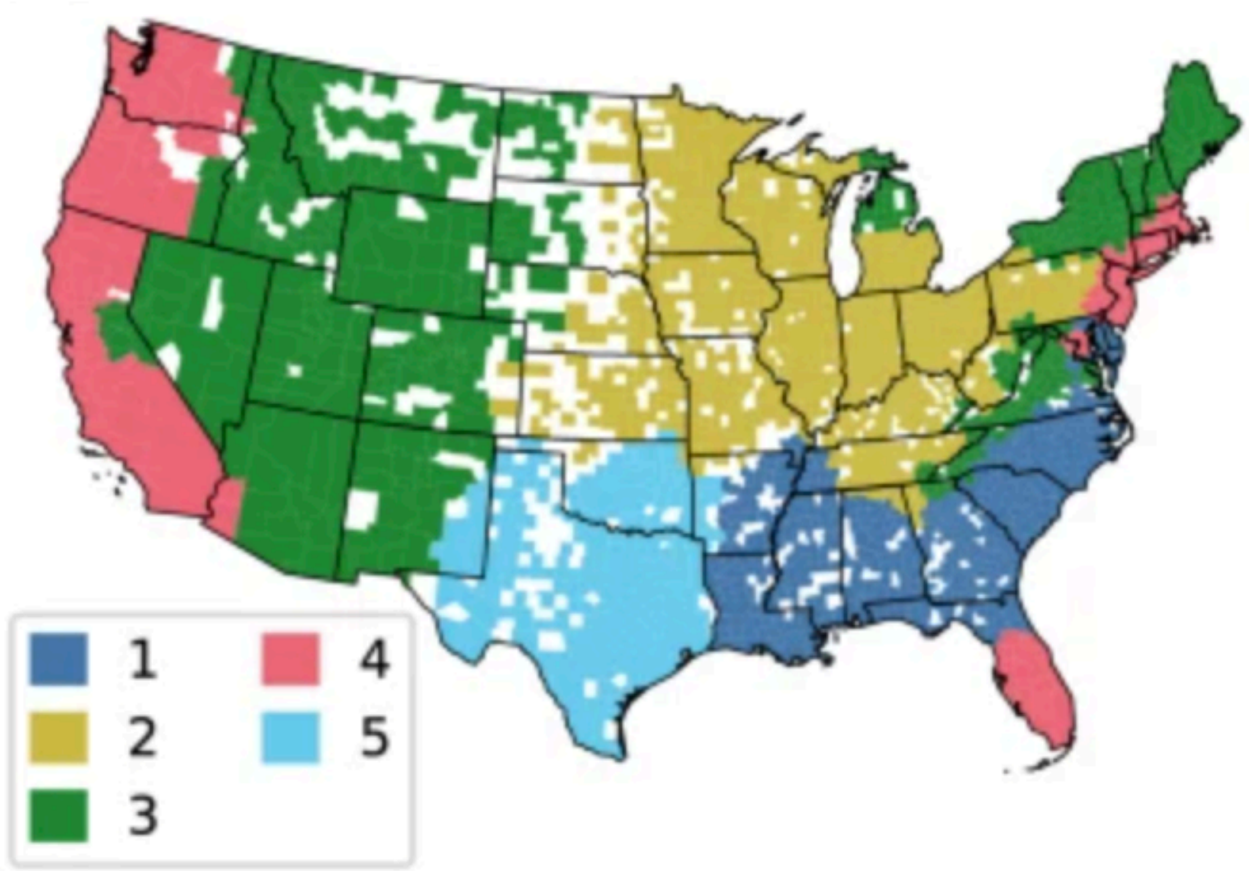
ESPN

49ers gamble with one second left, score on Brock Purdy sneak - ESPN

1 hour ago



American cultural regions from surveys



American cultural regions mapped from social media text

Story so far

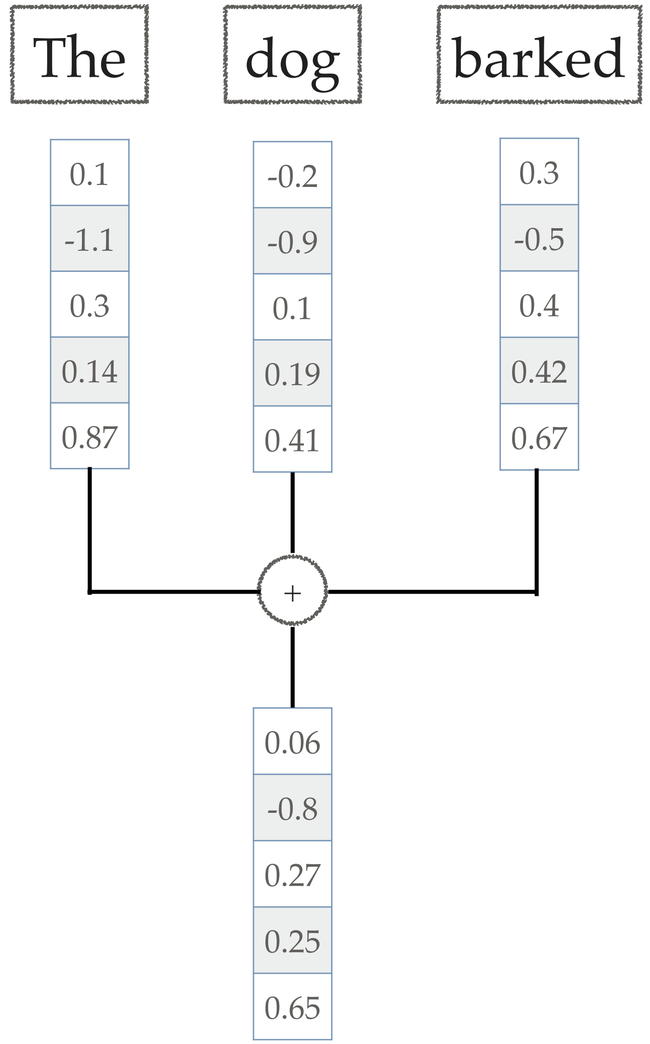
- The term document matrix is a representation to encode the statistics of a corpus
- In this matrix, words and documents can both be treated as vectors.

	d_1	d_2	d_3	…	d_m
w_1	2				1
w_2		3		1	
w_3	1				
…					
w_n			5	3	

Term-Document Matrix

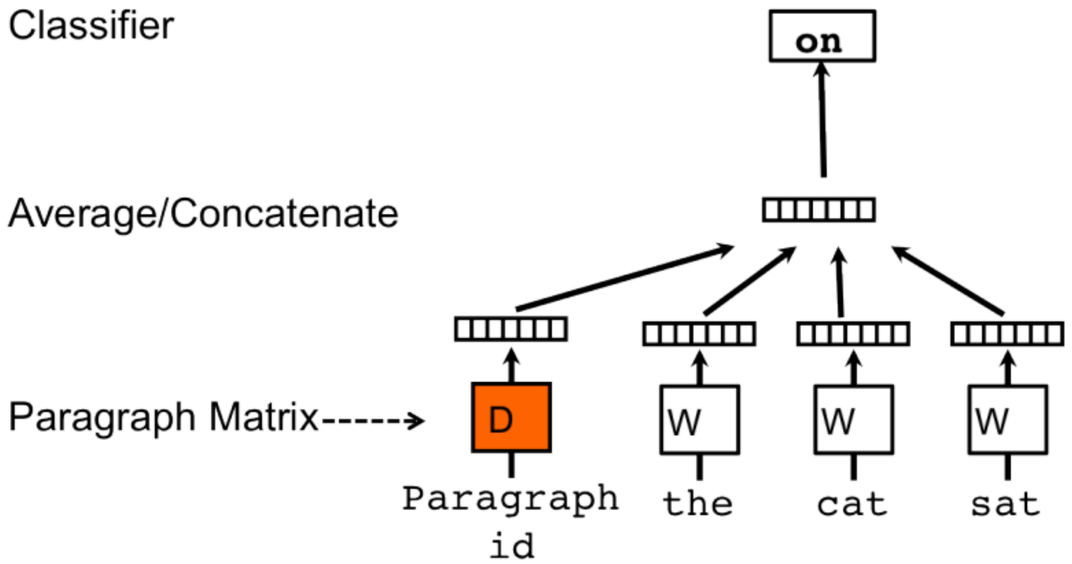
Sparse to dense representations

- We can learn dense representations of words using word2vec and then derive a document vector from the word vectors by simple averaging



Doc2Vec

- We can also directly learn a document vector by modifying the word2vec algorithm



Source: Le and Mikolov (2014) "Distributed Representations of Sentences and Documents"

Matrix Factorization

	d_1	d_2	d_3	…	d_m
w_1	2				1
w_2		3		1	
w_3	1				
…					
w_n			5	3	

Term-Document Matrix

$\approx$

w_1	0.1	-0.13	-1.2	0.15	1.23
w_2	0.2	-0.1	-0.2	0.45	1.4
w_3	…	…	…	…	…
…	…	…	…	…	…
w_n	0.15	-0.5	-0.3	0.23	-1.2

Word vectors

$\times$

	…	…	…	…	…
	…	…	…	…	…
	…	…	…	…	…
	…	…	…	…	…
	…	…	…	…	…

$\times$

	d_1	d_2	d_3	…	d_m
	0.1	-0.1	-0.2	0.11	1.21
	0.4	-0.2	-0.7	0.41	1.1
	…	…	…	…	…
	…	…	…	…	…
	0.35	-0.1	-0.2	0.32	-1.9

Document vectors

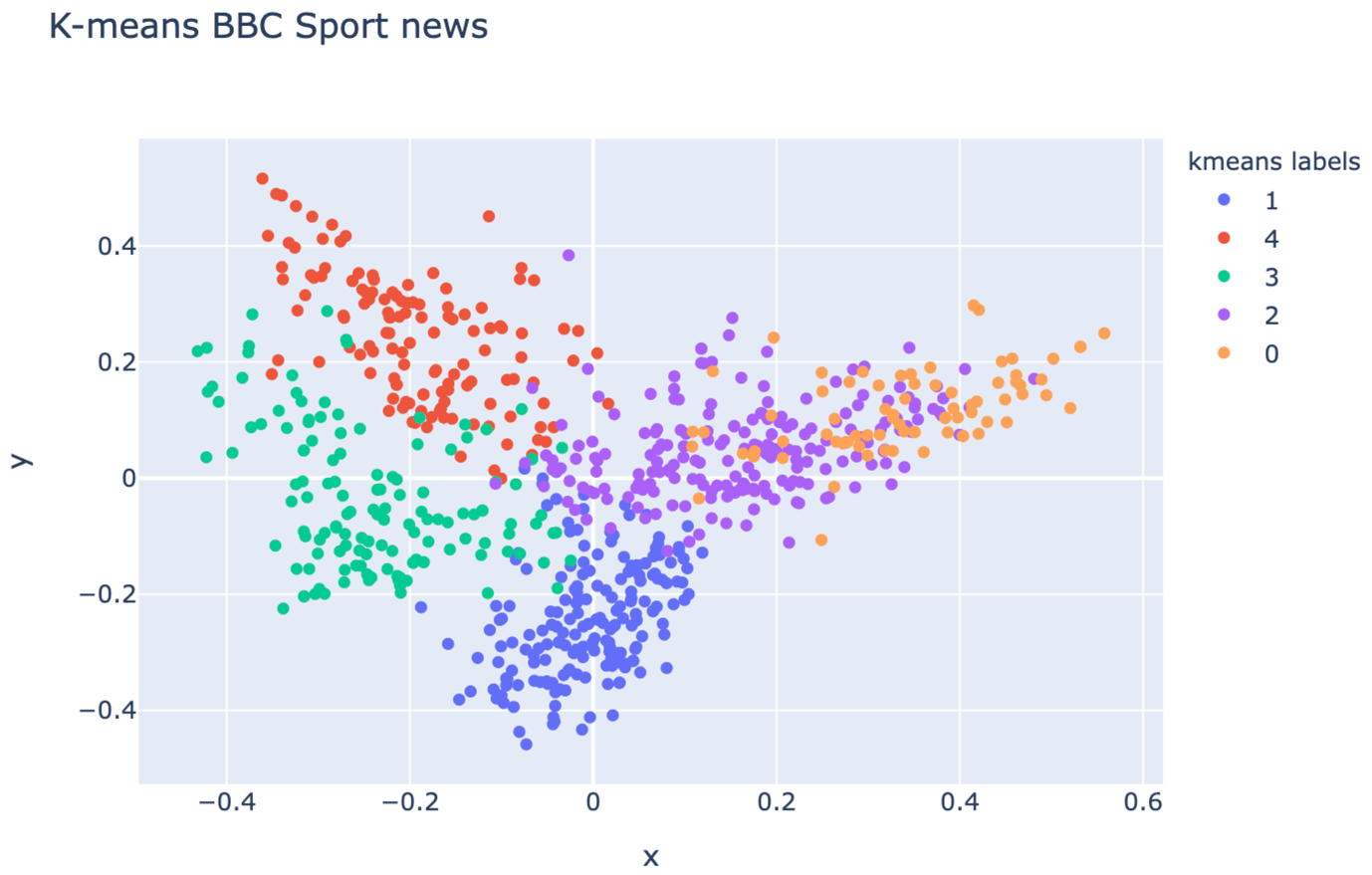
$\mathbb{R}^{n \times m}$ $\mathbb{R}^{n \times k}$ $\mathbb{R}^{k \times k}$ $\mathbb{R}^{k \times m}$

Question for the day

"How to group documents (e.g., speeches, lyrics, papers, tweets, etc) using their text?"

Input: documents/words

Output: partitions

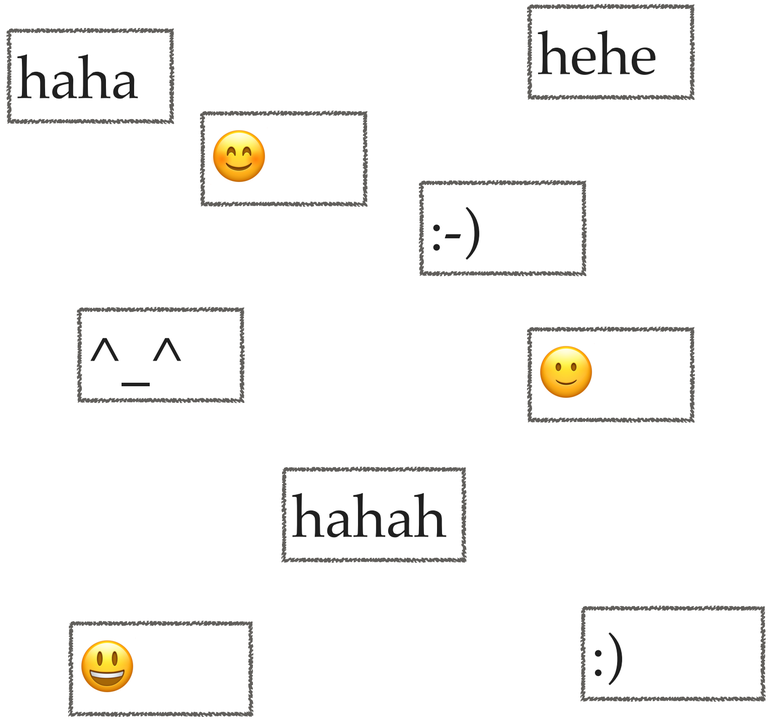


https://github.com/jbesomi/texthero

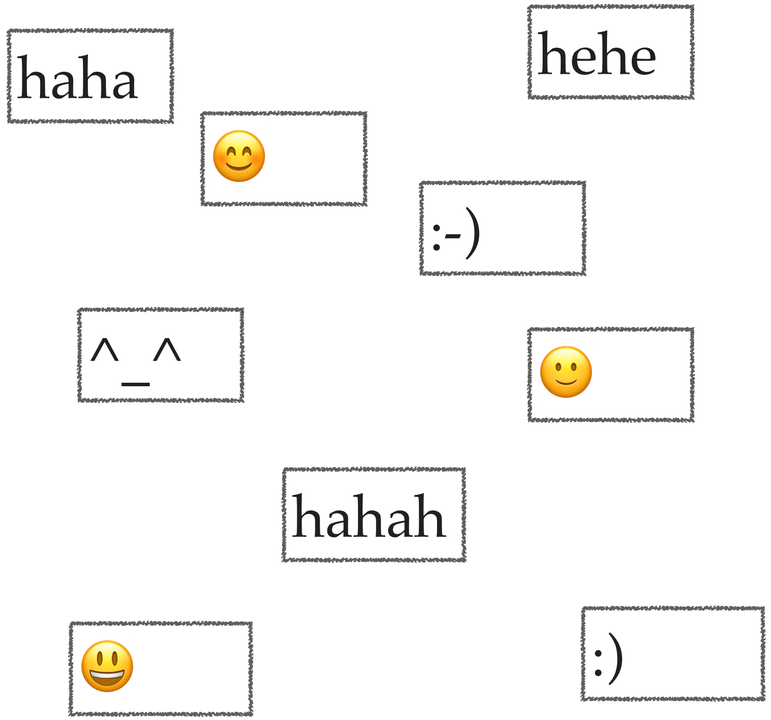
Agenda

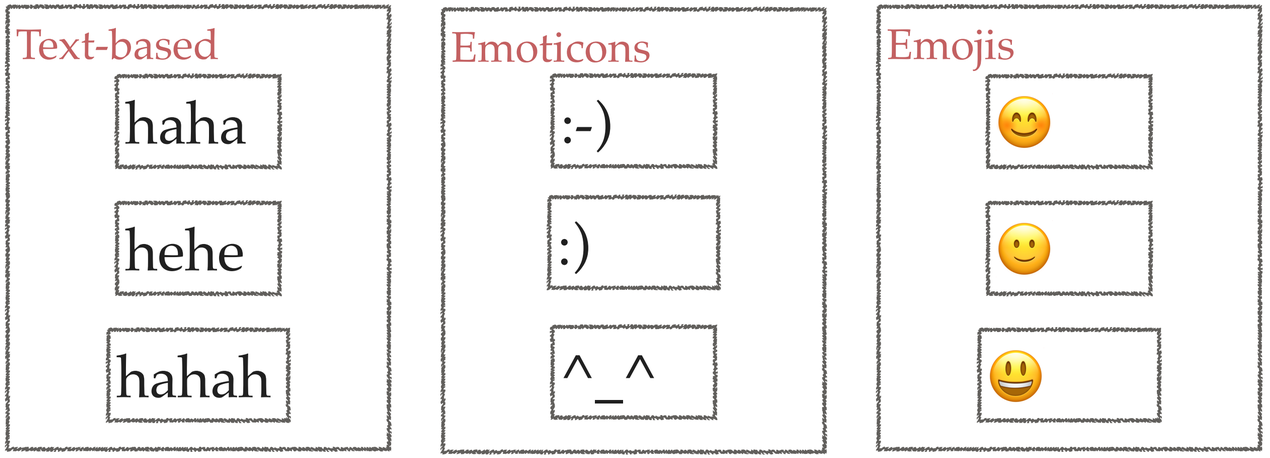
- What are the techniques of grouping documents?
- What are the challenges in grouping documents?
- How to evaluate grouping?

Example I

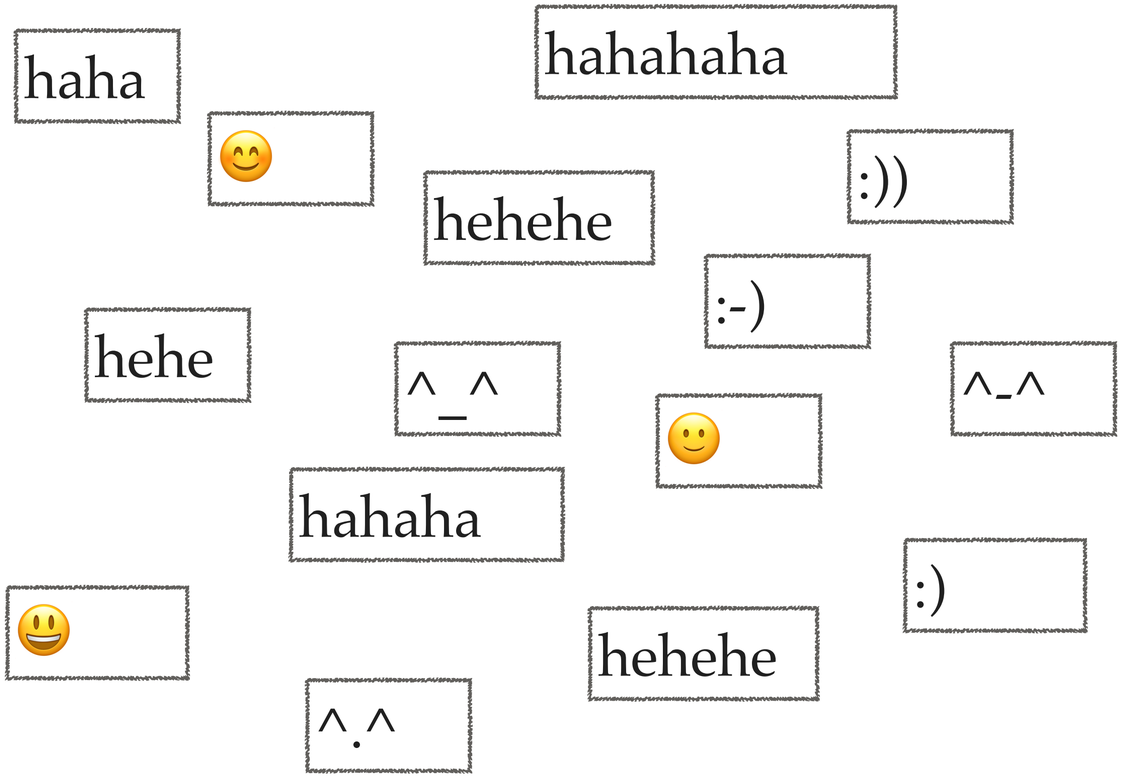


Example I

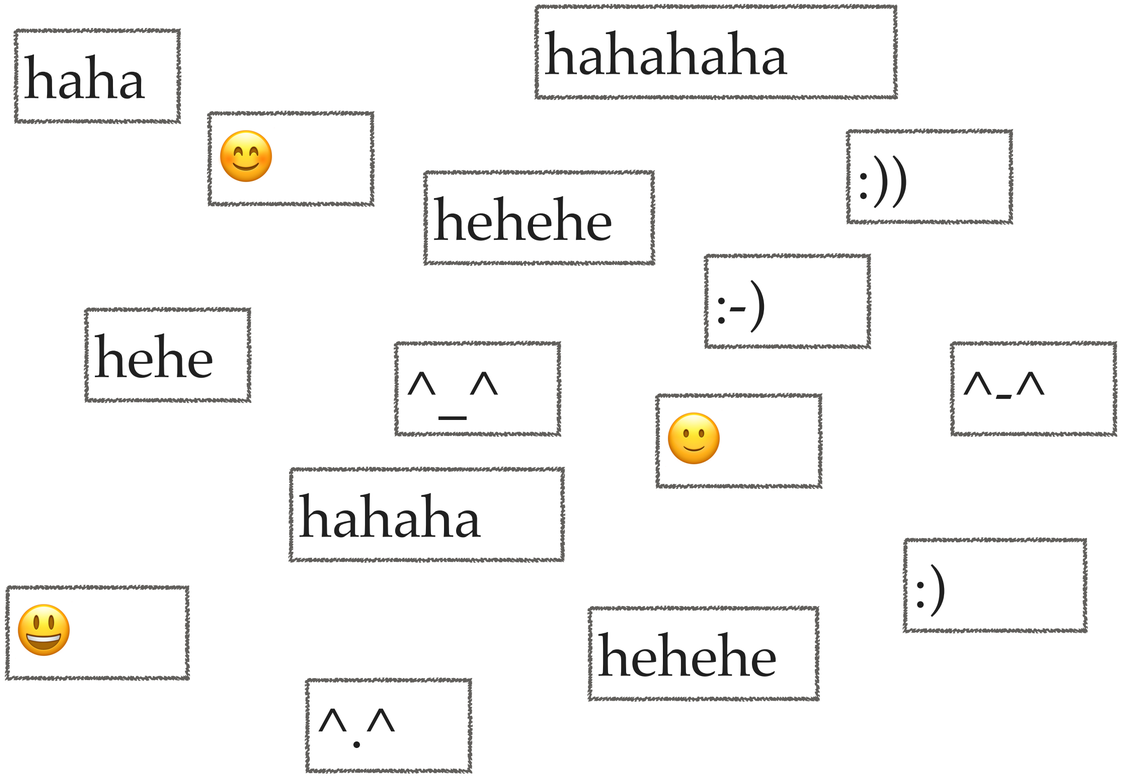

haha
hehe
😊
:-)
^_^
🙂
hahah
😀
:)


Text-based
haha
hehe
hahah
Emoticons
:-)
:)
^_^
Emojis
😊
🙂
😀

Example II



Example II



Text-based

ha-variation
haha
hahaha
hahahaha

he-variation
hehe
hehehe
hehehehe

Emoticons

Sideface emoticons
:)
:))
:-)

Frontface emoticons
^_^
^.^
^_^

Emojis

Clustering techniques

- Many ways but the most popular is kmeans clustering and its variations

KMeans

KMeans

- Let d be a document; $d \in \mathbb{R}^m$
- D is a set of documents; $D = [d^{(1)}; d^{(2)}; \cdots; d^{(|D|)}] \in \mathbb{R}^{|D| \times m}$

KMeans

- Let d be a document; $d \in \mathbb{R}^m$
- D is a set of documents; $D = [d^{(1)}; d^{(2)}; \cdots; d^{(|D|)}] \in \mathbb{R}^{|D| \times m}$
- Objective: Find a partition set $\Pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$ such that:
  - $\pi_i \cap \pi_j = \phi$ and $\pi_i \cup \pi_j = D$
  - dist($d_i^{(q)}$,$d_j^{(r)}$) should be small if $i = j$
  - dist($d_i^{(q)}$,$d_j^{(r)}$) should be big if $i \neq j$

KMeans

KMeans

- Choose k

KMeans

- Choose k
- Initialize k vectors of m dimensions as cluster centers

KMeans

- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:

KMeans

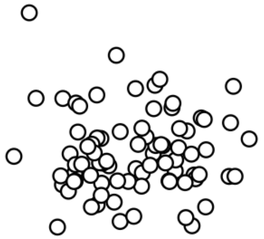
- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
  - For every document, assign closest cluster center

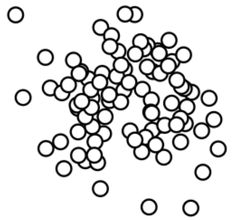
KMeans

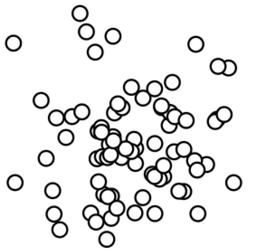
- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
  - For every document, assign closest cluster center
  - Recalculate cluster centers based on recent assignments

KMeans

We may have a gaussian data distribution



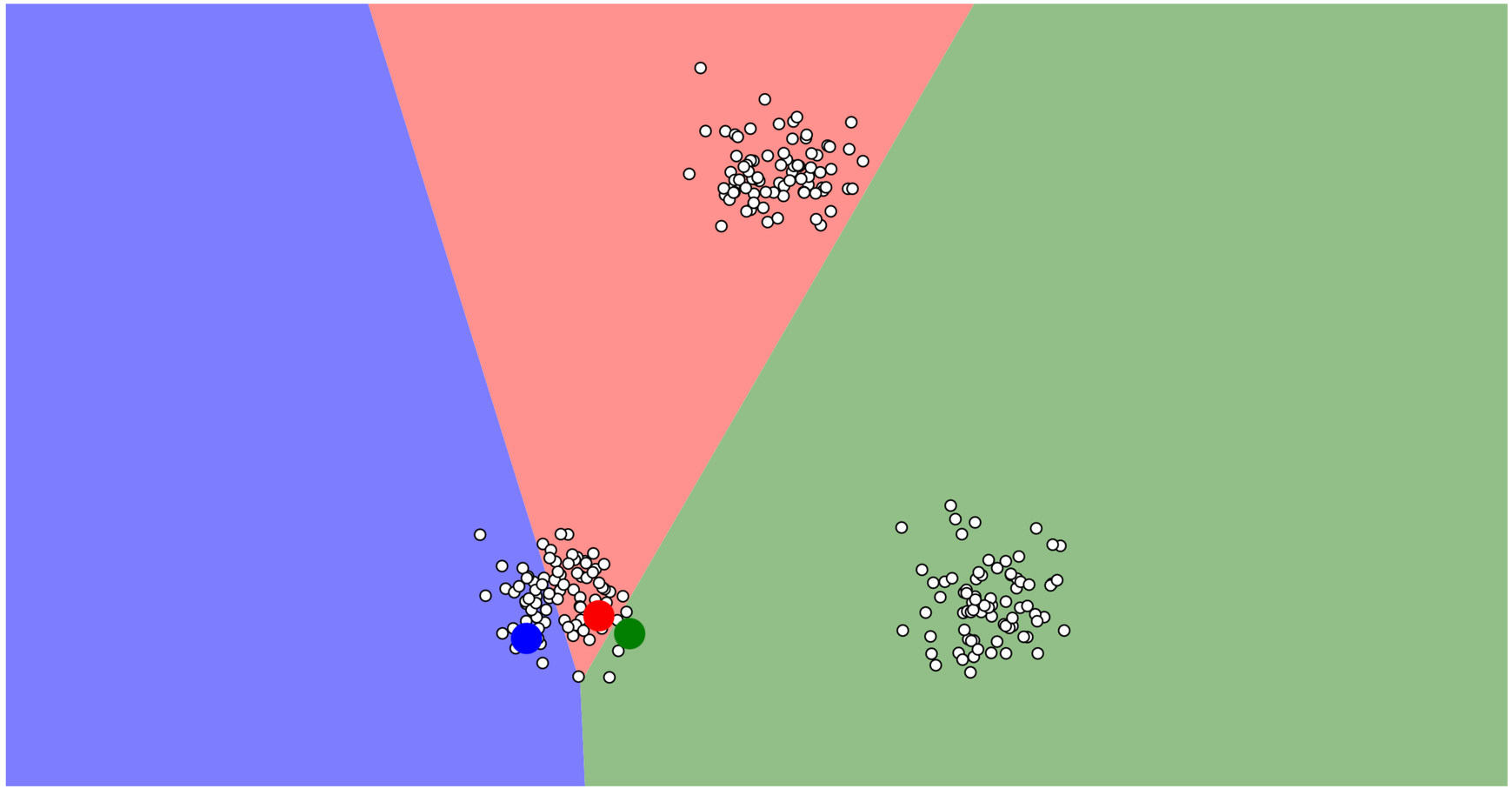




Source: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

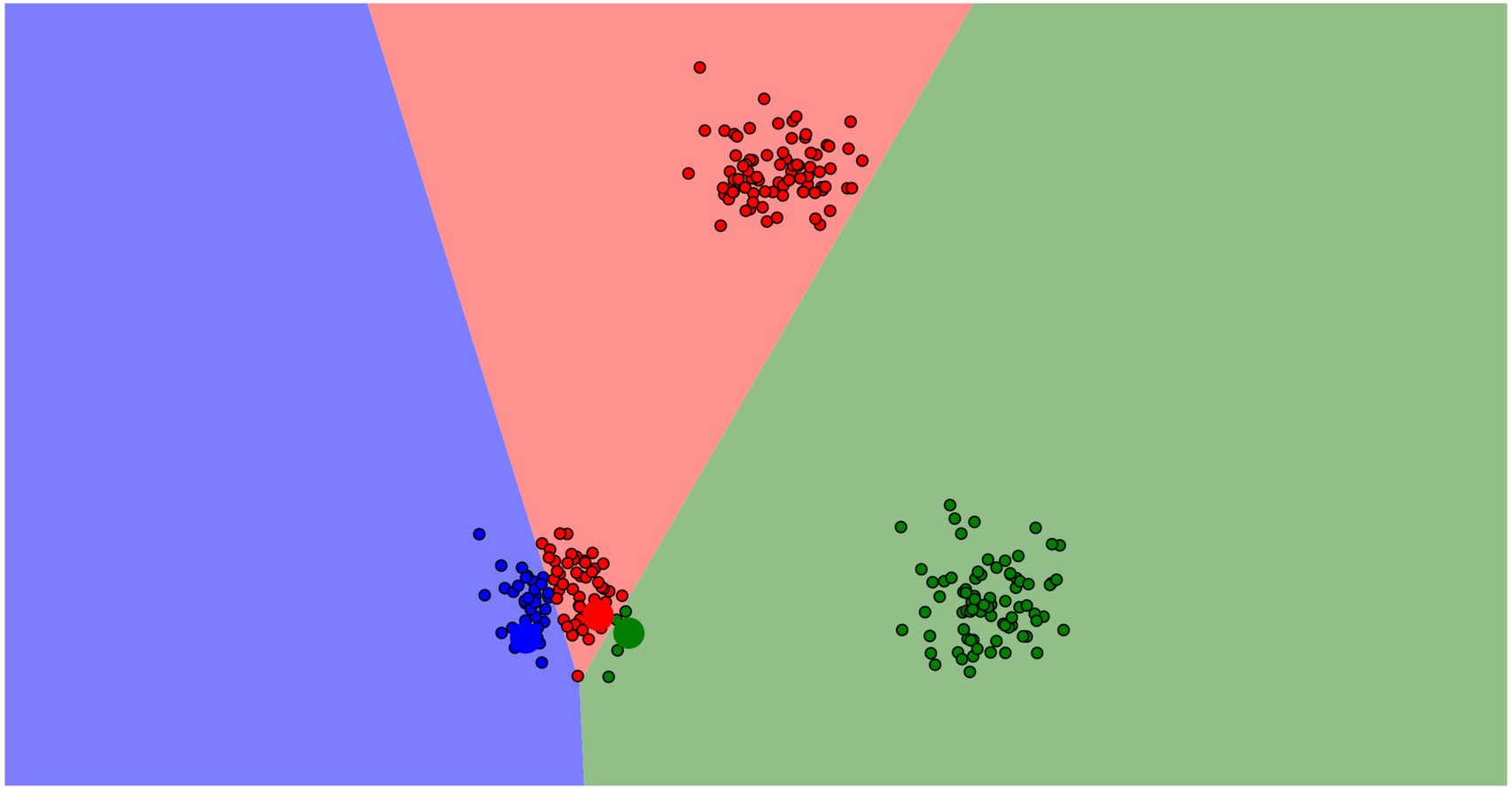
KMeans

Initialize the cluster centers



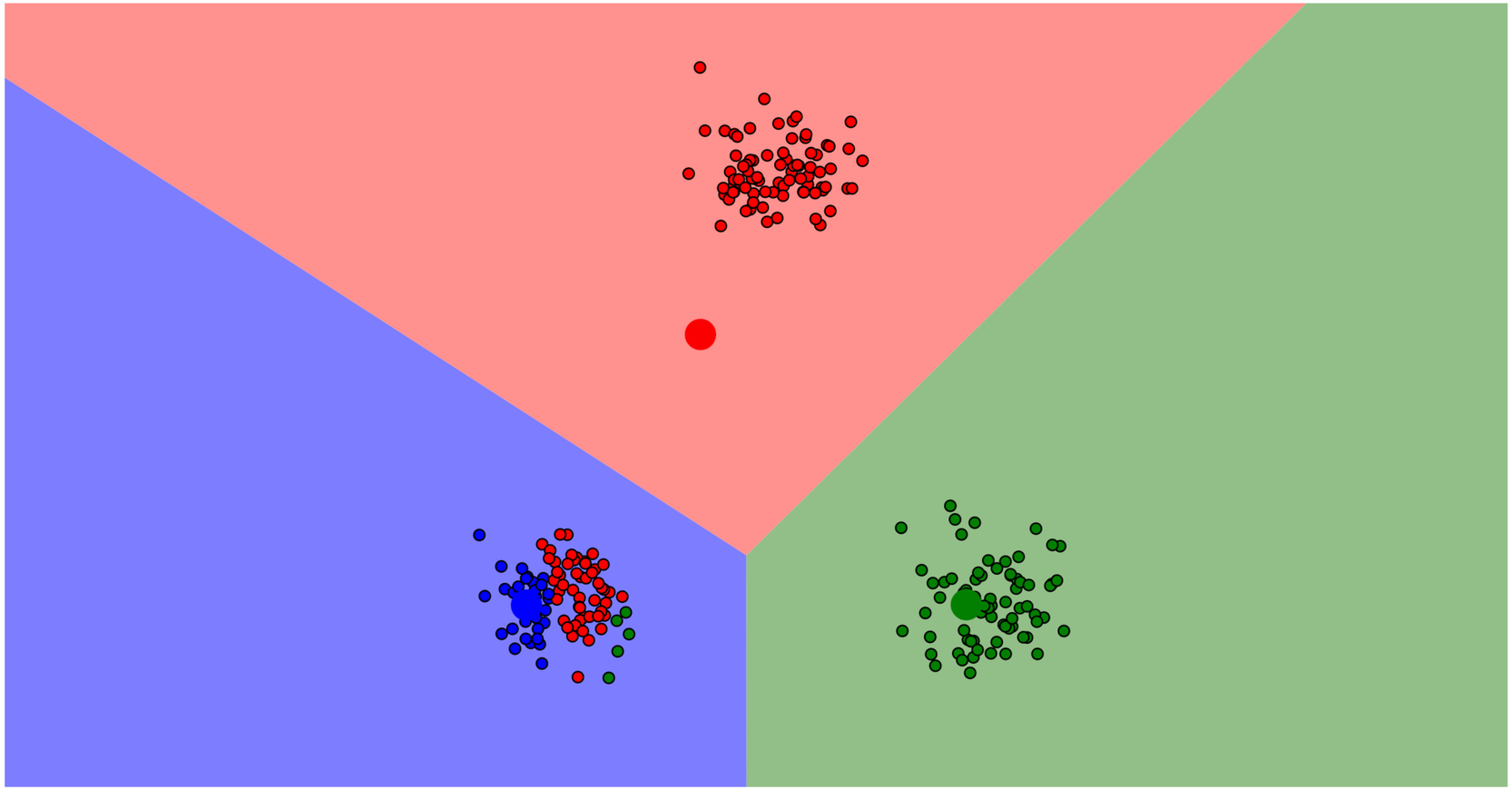
KMeans

Partitions formed based on cluster centers



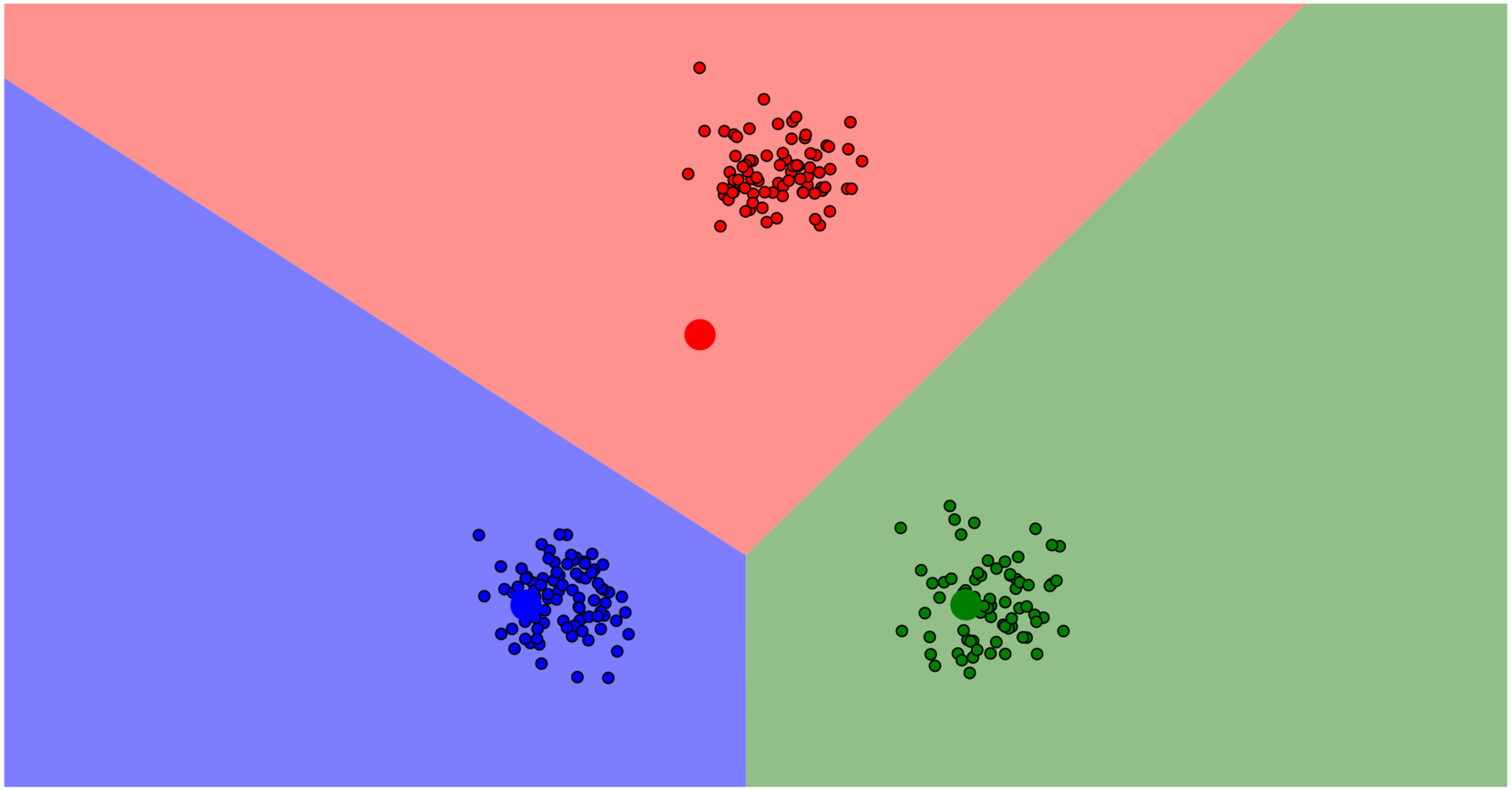
KMeans

Update the cluster centers



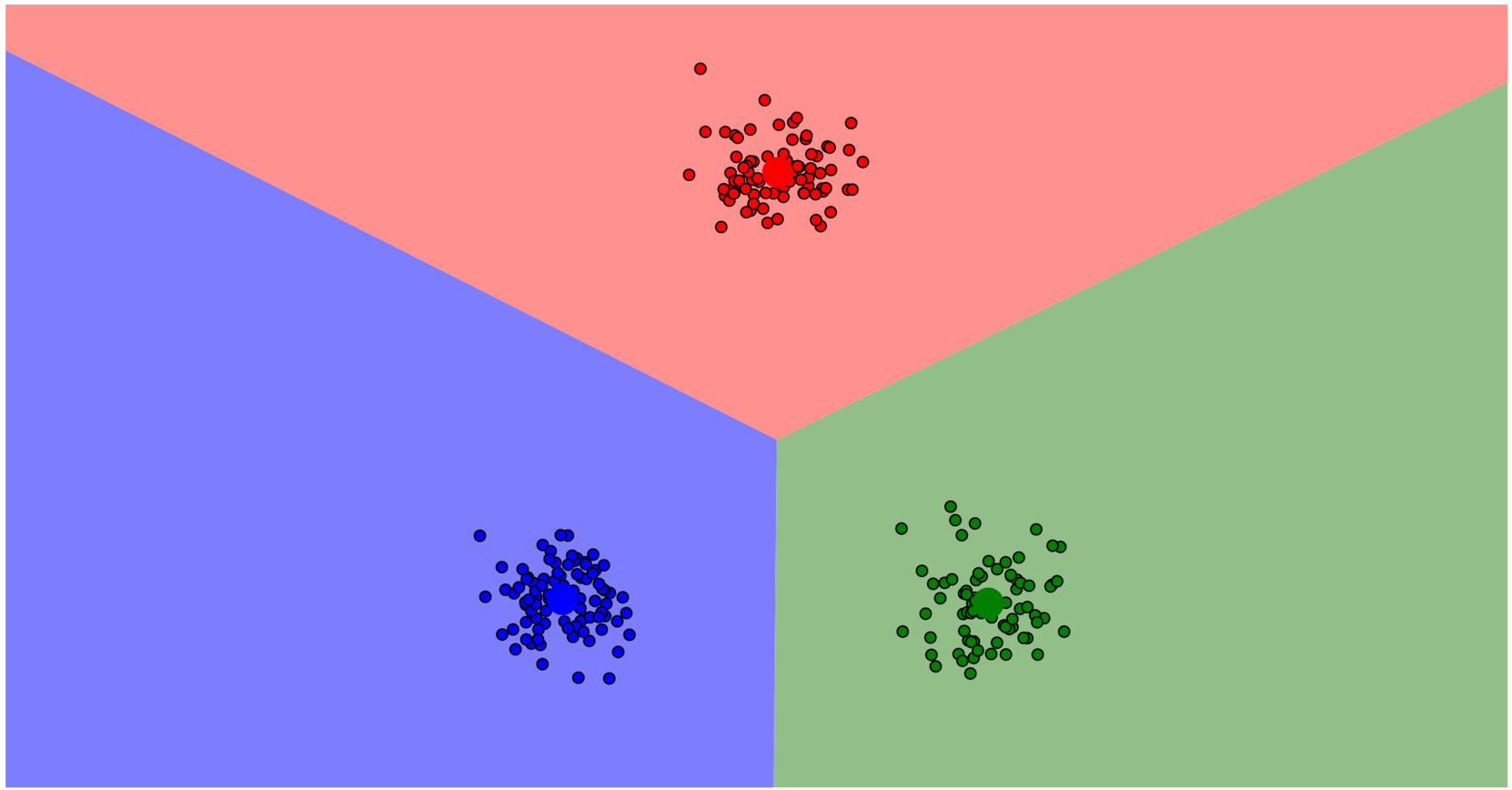
KMeans

Reassign the data points



KMeans

Repeat
until
convergence

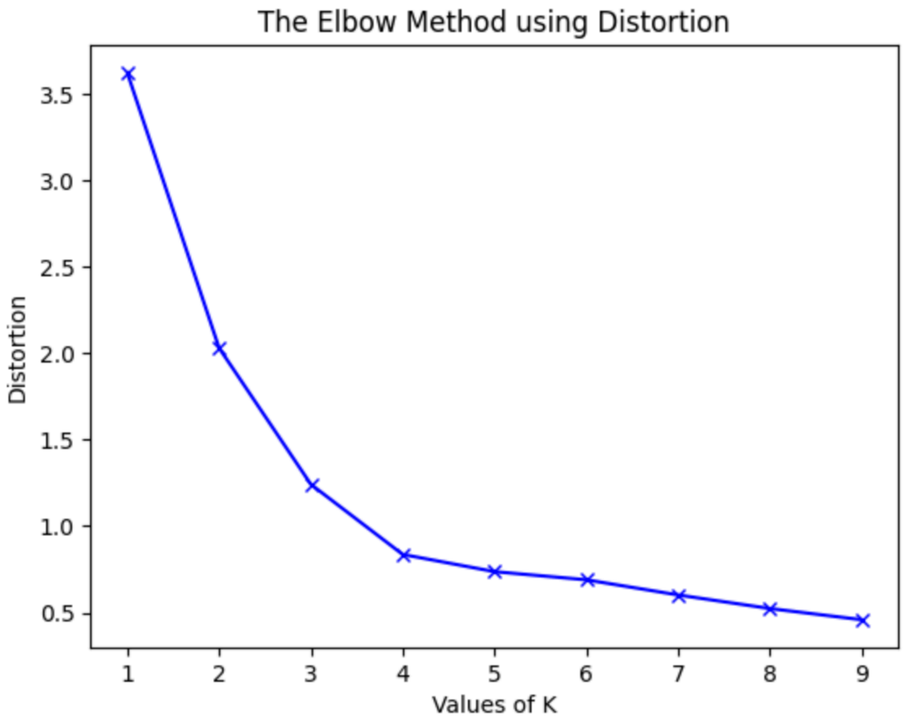


KMeans

- Choose k
- Initialize k vectors of m dimensions as cluster centers
- Repeat until convergence:
  - For every document, assign closest cluster center
  - Recalculate cluster centers based on recent assignments

Choosing K

- Sometimes we know the value k (e.g., political parties)
- In general, use your judgment or some heuristics to find k
- Bayesian non-parametric methods find the optimal k



Initialization

- Initialization may affect convergence time and optimal clustering
- There are some smart initializations but this is a non-trivial problem [see Caleb et. al]

Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications 40.1 (2013): 200-210.

Repeat

- Each cluster assignment and centroid recalculation step is repeated several times
- Since the algorithm is sensitive to initialization, the entire procedure may be repeated to get stability.

Measuring closeness

- Typically euclidean distances are used to measure closeness
- Non-euclidean metrics or kernel transformations of vectors useful when data is not linearly separable

Text preprocessing

Text preprocessing

- Preprocessing steps such as stop words removal, lemmatization, normalization, etc are crucial

Text preprocessing

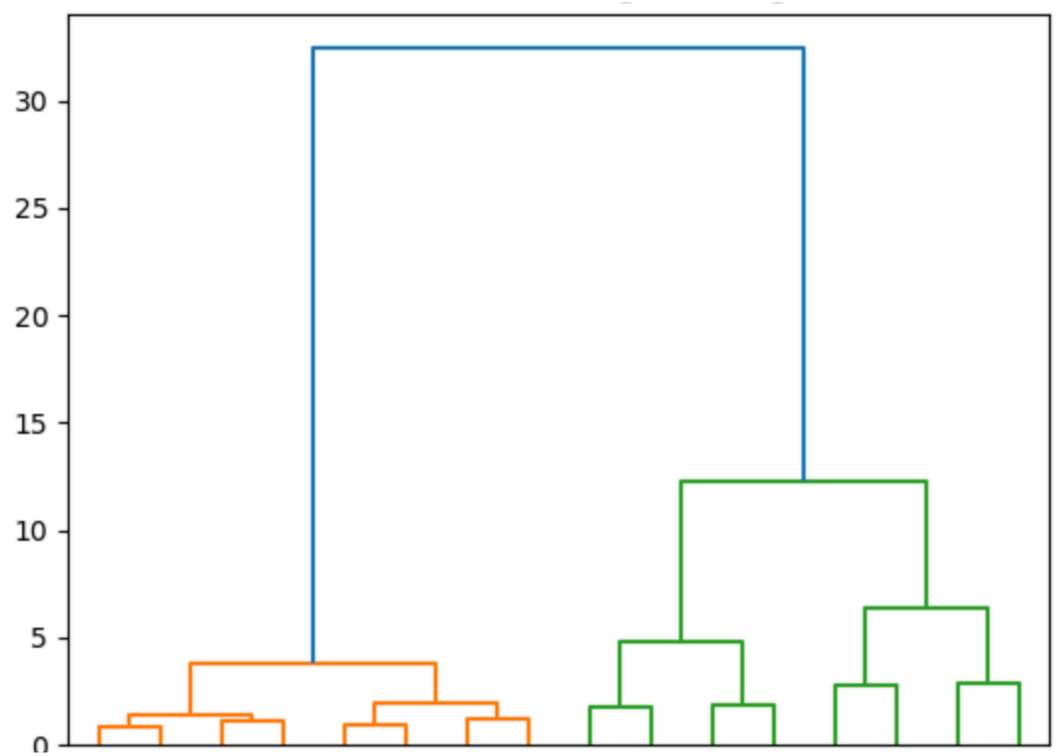
- Preprocessing steps such as stop words removal, lemmatization, normalization, etc are crucial
- In general, clustering works best with dense features than sparse, so some dimensionality reduction technique can also be useful

Shortcomings

- Knowing K is not easy (e.g., how many genres of books?)
- Outliers can have a strong effect on cluster formation
- Kmeans does not extract a hierarchical structure from the data

Other methods

- Agglomerative clustering can extract hierarchical clustering
- Affinity propagation algorithm on a graph



In class

- Clustering demo