Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Text Clustering

Text, Web and Social Media Analytics Lab

Prof. Dr. Diana Hristova

Exercise 5: Classification

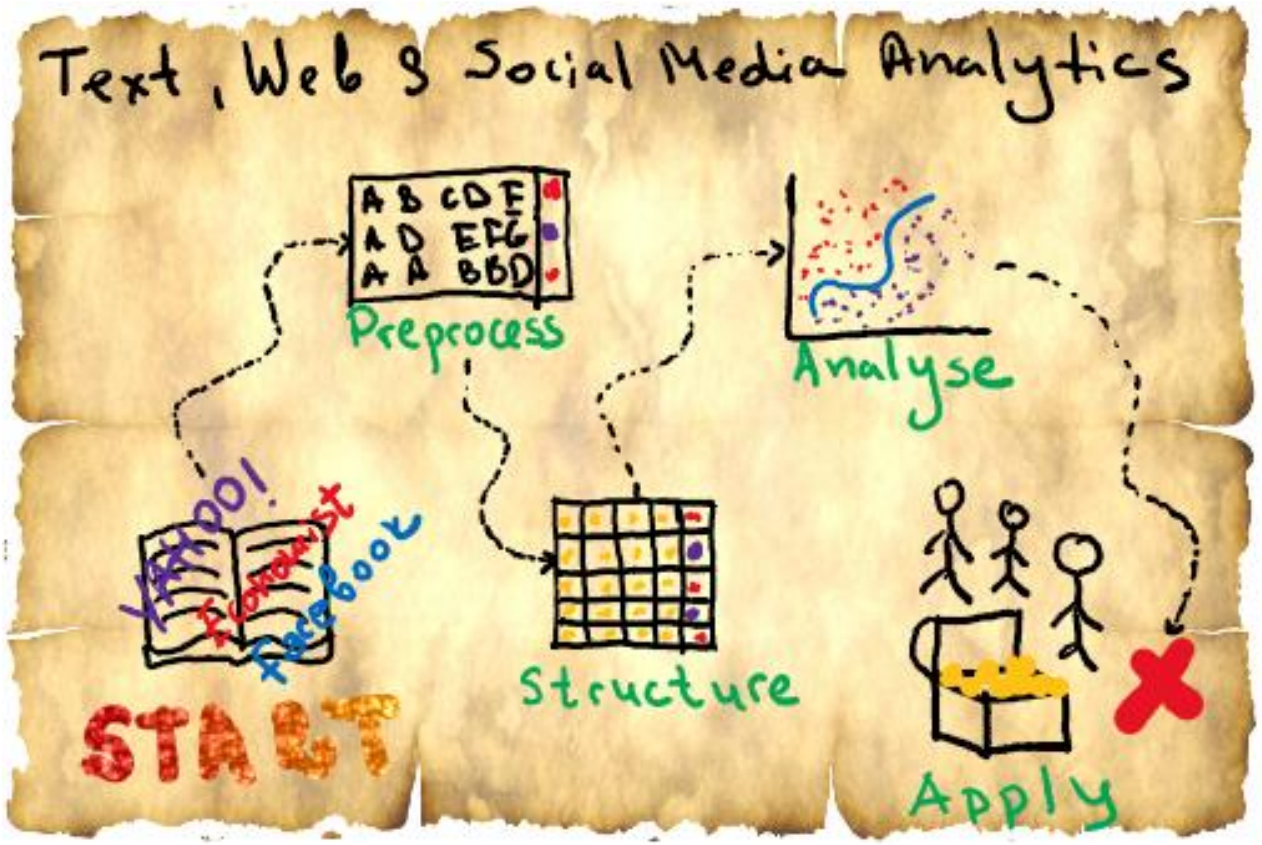
Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Can one group present please?

What did we learn last week?

Treasury map: Text Classification

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law



Course structure

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Date	Lecture	Exercise
12.04.2021	Introduction	Technical Installation
19.04.2021	Text Preprocessing	Projects kick-off
26.04.2021	Text Representation	Preprocessing Newsgroups
03.05.2021	Text Representation (2)	Text Representation Newsgroups
10.05.2021	Text Classification	Text Representation Newsgroups (2)
17.05.2021	Text Clustering/Capgemni	Newsgroups Topic Classification
31.05.2021	Text Mining in Social Media	Newsgroups Topic Clustering
07.06.2021	Mining Social Graphs	Sentiment Analysis and Time Series in Twitter
14.06.2021	Projects Status Update	Projects Status Update
21.06.2021	Web Analytics	Mining Social Graphs in Twitter
28.06.2021	Mock Exam	Web Analytics in E-commerce
05.07.2021	Final Presentation	Final Presentation
19.07.2021	Submit Code & Written report	
t.b.a.	Exam	

What will we learn today?

At the end of this lecture, you will:

1. Understand the main motivation behind unsupervised approaches for text analytics
2. Know the mechanics and the pros and cons of k-means clustering and LDA topic modelling

Introduction: Text Classification and Clustering

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Reminder

Supervised approaches

- **Idea:** the set of values of the target variable is known and the text data is annotated with them.
- **Example:**
  - "I require your financial assistance to transform you the heritage of the late Sir Johnson." ➔ spam
  - "We are meeting tomorrow at 5 o'clock." ➔ not spam
  - Aim: build a model that can classify each document in one of the categories.

➔ **Text classification**

Focus today

Unsupervised approaches

- **Idea:** the set of values of the target variable are not known, but need to be derived from the data.
- **Example:**
  - A: "The restaurant has terrible service."
  - B: "The meals were very satisfactory.
  - C: "The food was fantastic."
  - D: "The waitress took forever."

➔ A and D are similar as well as B and C

- Aim: build a model that can derive the unknown categories in each document..

➔ **Text clustering and topic modelling**

Unsupervised approaches: Main Idea

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

1. Unannotated corpus

Review
Wow... Loved this place.
Crust is not good.
Not tasty and the texture was just nasty.
Stopped by during the late May bank holiday of...
The selection on the menu was great and so wer...

2. Text preprocessing

prep
wow love place
crust good
tasti textur nasti
stop late bank holidai rick steve recommend love
select menu great price

3. Text representation

```
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

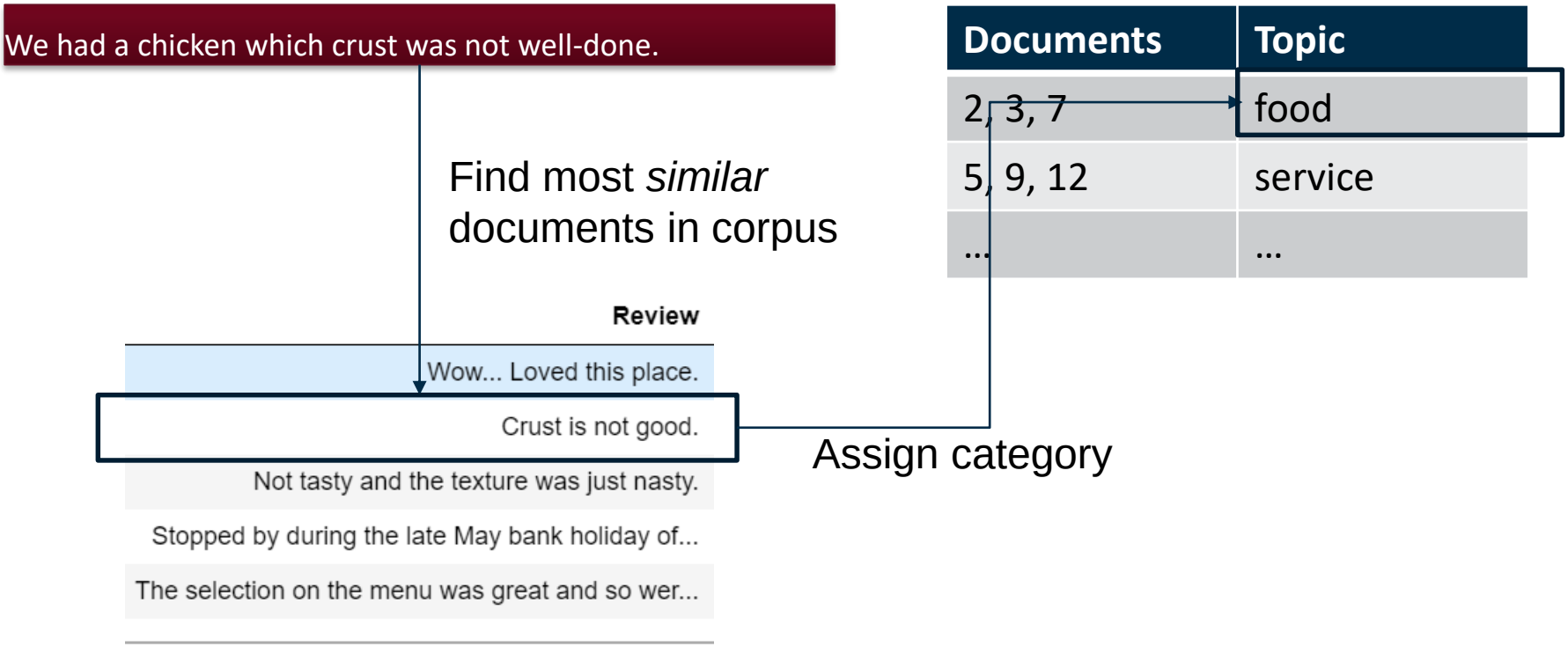
Model

4. Annotated corpus

Documents	Topic
2, 3, 7	food
5, 9, 12	service
…	…

Unsupervised approaches: Main Idea

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

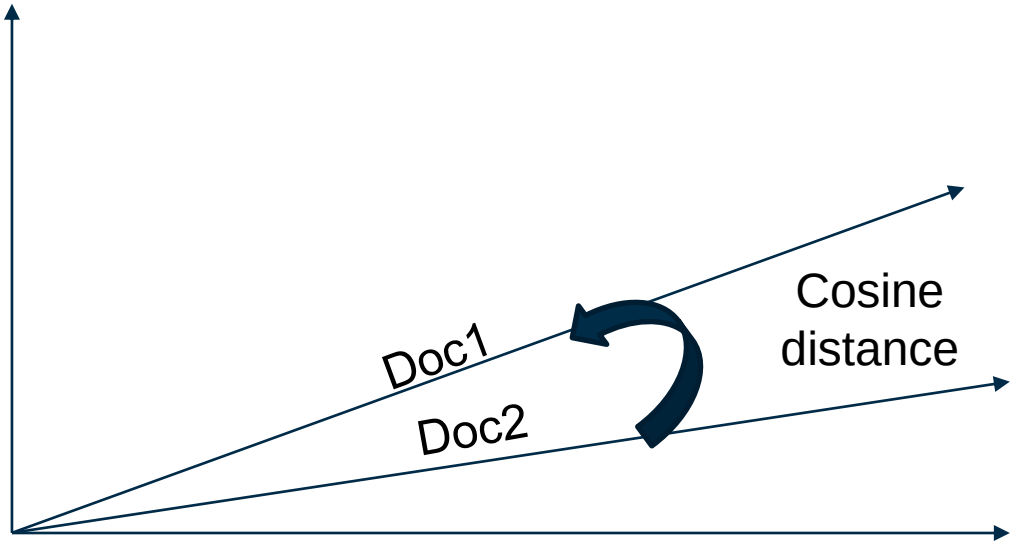


? What does *similar* mean?

Embeddings Word2Vec (2)

Reminder

- The resulting embeddings are the same for all usage of the word.
- They allow calculating the similarity between two words and even mathematical operations.
- **Cosine similarity (based on the cosine distance)**



Embeddings Word2Vec Cosine similarity

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Reminder

- The resulting embeddings are the same for all usage of the word.
- They allow calculating the similarity between two words and even mathematical operations.
- **Cosine similarity (based on the cosine distance)**
- $A = (a_1, a_2)$ and $B = (b_1, b_2)$

$$CosSim(A, B) = \frac{a_1 b_1 + a_2 b_2}{\sqrt{{a_1}^2 + {a_2}^2}\sqrt{{b_1}^2 + {b_2}^2}}$$

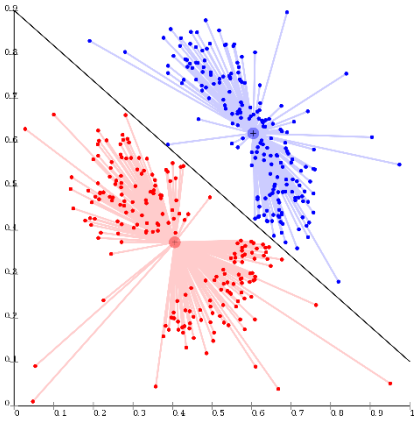
Note

There are also other types of distances which we don‘t consider here.

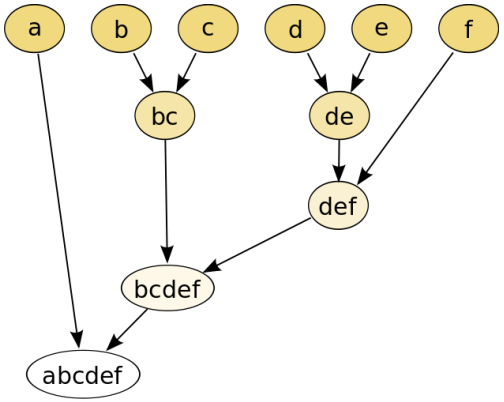
Text Clustering: Methods

There are two types of clustering methods:

- **Partitive clustering:** partition documents into non-overlapping groups of similar documents. Focus here: k-means
- **Hierarchical clustering:** iteratively generate a set of nested groups consisting of similar documents.



https://de.wikipedia.org/wiki/Datei:KMeans-density-data.svg



https://de.wikipedia.org/wiki/Datei:Hierarchical_clustering_simple_diagram.svg

Text Clustering: k-means

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Aim: separate documents in k groups such that:

- Each group is represented by its mean
- The documents in each group have maximum similarity
- The documents in different groups have maximum difference

Algorithm:

1. Randomly choose initial cluster means (centroids) c_k (e.g. k=2)
2. For each document d determine its cluster as $k = \text{arg}max_k CosSim(d, c_k)$[1]
3. Calculate the new cluster mean as the average among all documents in the cluster
4. Go-back to 2.
5. Stop at convergence.

[1]Other similarity measure are possible.

Text Clustering: k-means example

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

A: “The food was fantastic.”

B: “The meals were very satisfactory.

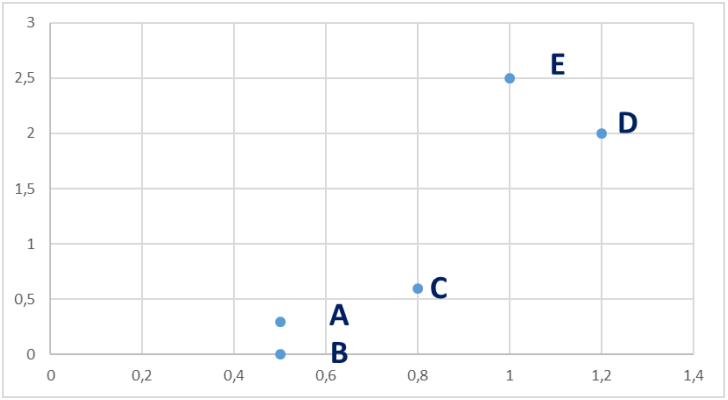
C: “The restaurant has a nice location.”

D: “The waitress took forever.”

E: “The restaurant has terrible service.”

Preprocessing
Embeddings

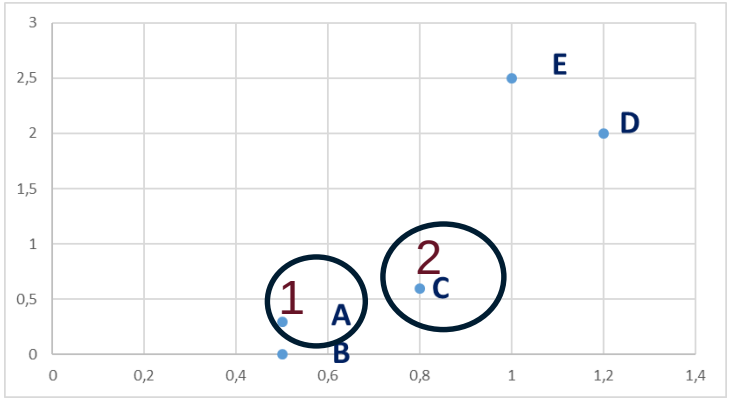
Doc	v1	v2
A	0.5	0.3
B	0.5	0
C	0.8	0.6
D	1.2	2
E	1	2.5



Text Clustering: k-means example (2)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

1. Randomly choose initial cluster means (centroids) c_k (e.g. k=2)



2. For each document d determine its cluster as $k = \arg max_k CosSim(d, c_k)$

Doc	Cos to A	Cos to C	Cluster
B	0.86	0.80	
D	0.88	0.93	
E	0.80	0.85	

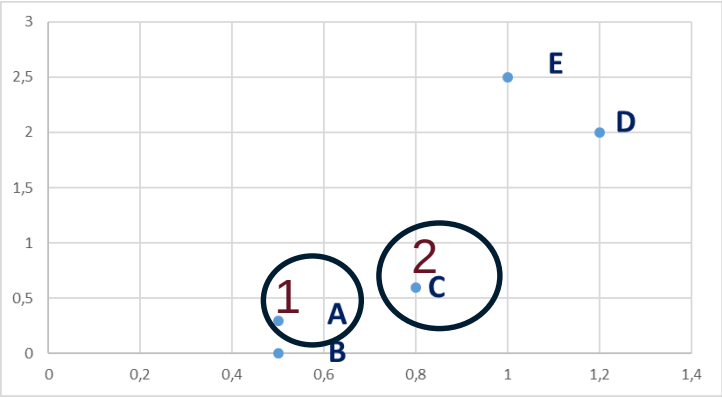
Doc	v1	v2
A	0.5	0.3
B	0.5	0
C	0.8	0.6
D	1.2	2
E	1	2.5

? What are the cluster values?

Text Clustering: k-means example (2)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

1. Randomly choose initial cluster means (centroids) c_k (e.g. k=2)



2. For each document d determine its cluster as $k = \arg max_k CosSim(d, c_k)$

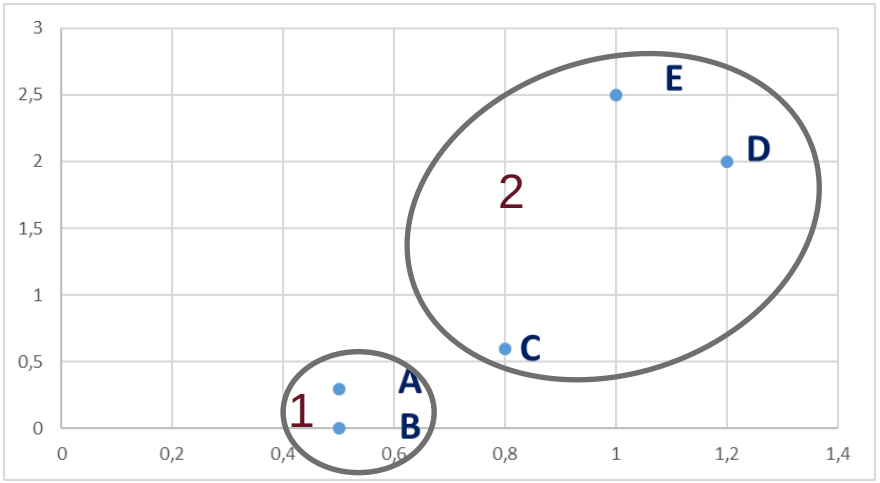
Doc	Cos to A	Cos to C	Cluster
B	0.86	0.80	1
D	0.88	0.93	2
E	0.80	0.85	2

Doc	v1	v2
A	0.5	0.3
B	0.5	0
C	0.8	0.6
D	1.2	2
E	1	2.5

Text Clustering: k-means example (3)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

3. Calculate the new cluster mean as the average among all documents in the cluster

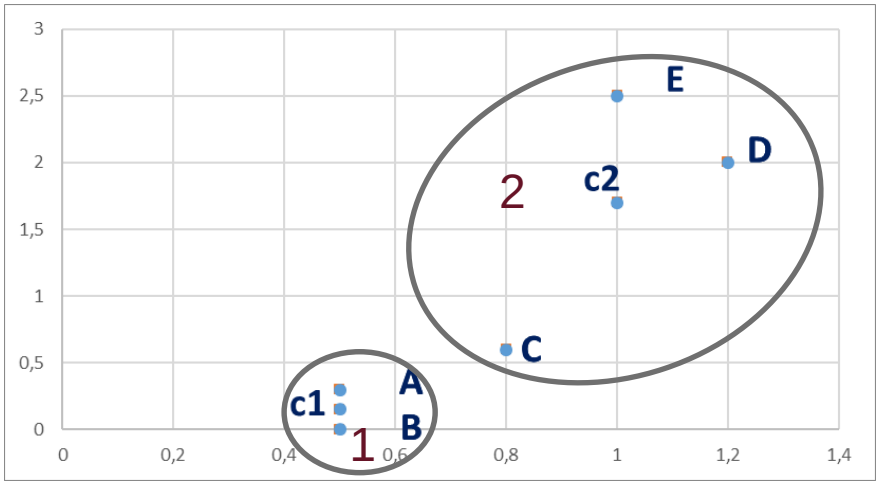


Doc	v1	v2	Cluster	Mean cluster	
A	0.5	0.3	1		
B	0.5	0	1	0.5	0.2
C	0.8	0.6	2		
D	1.2	2	2		
E	1	2.5	2	1.0	1.7

Text Clustering: k-means example (4)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

3. Calculate the new cluster mean as the average among all documents in the cluster



2. For each document d determine its cluster as $k = \arg max_k CosSim(d, c_k)$

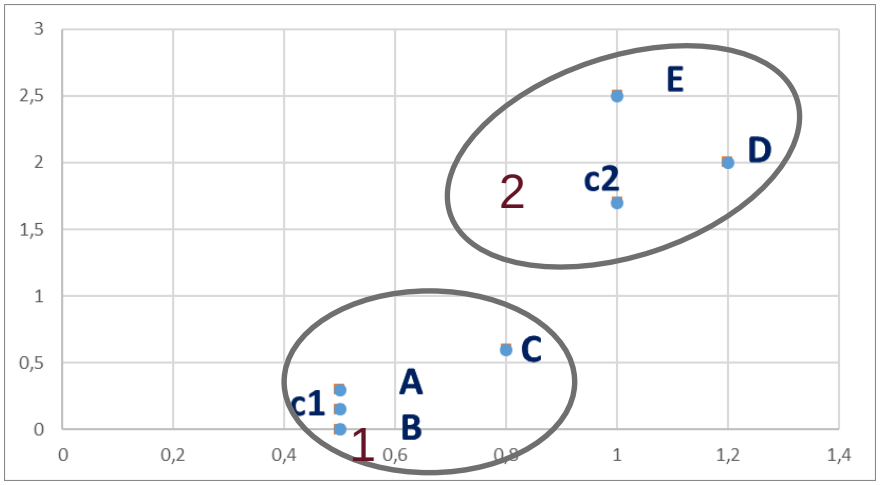
Doc	Cos to c1	Cos to c2	Cluster
A	0.97	0.88	1
B	0.96	0.51	1
C	0.94	0.92	1
D	0.74	1.00	2
E	0.62	0.99	2

Text Clustering: k-means example (5)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

3. Calculate the new cluster mean as the average among all documents in the cluster

2. For each document d determine its cluster as $k = \arg max_k CosSim(d, c_k)$



Doc	Cos to c1	Cos to c2	Cluster
A	0.97	0.88	1
B	0.96	0.51	1
C	0.94	0.92	1
D	0.74	1.00	2
E	0.62	0.99	2

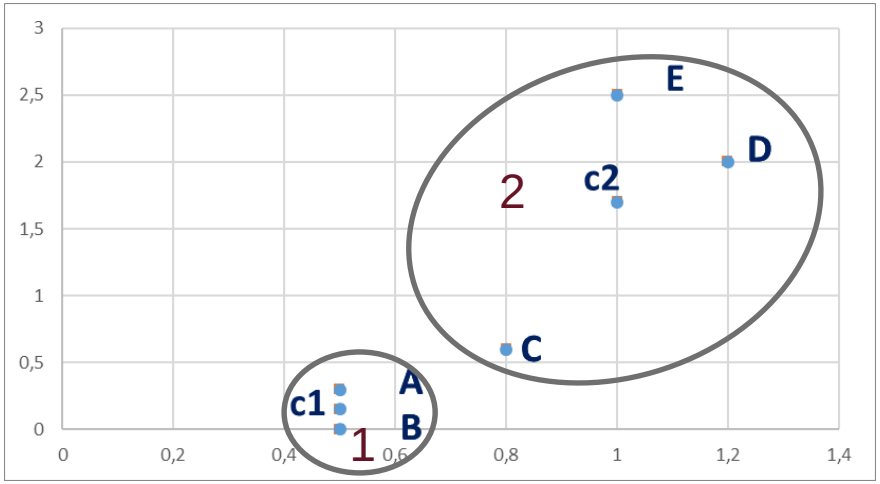
? What is a disadvantage of k-means clustering?

Text Clustering: k-means example (6)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

3. Calculate the new cluster mean as the average among all documents in the cluster

2. For each document d determine its cluster as $k = \arg max_k CosSim(d, c_k)$



Doc	Cos to c1	Cos to c2	Cluster
A	0.97	0.88	1
B	0.96	0.51	1
C	0.94	0.92	1
D	0.74	1.00	2
E	0.62	0.99	2

- **Advantage:** simple and quick
- **Disadvantage:** strong dependence on the initial assignment and the number of clusters k

➔ Hierarchical clustering

Topic modelling with LDA

- You can use Latent Dirichlet Allocation (LDA) for topic modelling
- **Idea:** Each document is represented by a mixture of topics and each topics is defined as a mixture of words.
- Example:
  - A: “The restaurant has terrible service and food.” ➔ topics: 30% service, 30% eating, 40% negative
  - B: “The meals were very satisfactory.” ➔ topics: 60% eating, 40% positive
  - C: “The food was fantastic.” ➔ topics: 50% eating, 50% positive
  - D: “The waitress took forever.” ➔ topics: 60% service, 40% negative

Topic modelling with LDA (2)

- Example (contd.):
  - Topic Eating: 60% food, 20% meal, 20% dish
  - Topic Service: 50% service, 20% waitress, 20% waiter
  - Topic Positive: 50% fantastic, 30% satisfactory, 20% good
  - Topic Negative: 60% terrible, 20% forever, 20% alwful

- **Advantage:** mostly interpretable
- **Disadvantage:** can take long for big datasets, number of topics has to be determined

Summary and Outlook

Summary:

- Unsupervised text analytics approaches can be used to derive an annotated corpus necessary for many applications
- k-means clustering is based on partitioning documents in groups where each group is presented by its centroid document
- LDA derives topic distribution for each document where each topic is defined as a distribution over the dictionary
- **Outlook:** You can apply both supervised and unsupervised approaches to analyse social media data

Questions?

Exercise 5

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

In a minute, six break-out rooms will be created. Choose the room that corresponds to your group in Moodle e.g. Room 1= Group 1. In your project group discuss and document the solution for Exercise 4 (in Moodle).