Machine Learning Techniques for Text

Module 3: Classifying Topics of Newsgroup Posts

Dr. Nikos Tsourakis

Course outline

- Module 0: Python Crash Course
- Module 1: Intro to Machine Learning
- Module 2: Detecting Spam Emails
- **Module 3: Classifying Topics of Newsgroup Posts**
- Module 4: Extracting Sentiments from Product Reviews
- Module 5: Recommending Music Titles
- Module 6: Teaching Machines to Translate
- Module 7: Summarizing Wikipedia Articles
- Module 8: Detecting Hateful and Offensive Language
- Module 9: Generating Text in Chatbots
- Module 10: Clustering Speech-to-Text Transcriptions

Overview

- The large volumes of unstructured text that large corporations and organizations need to sort daily necessitate automatizing tedious and time-consuming manual tasks
- We deal with how to tag a text document using a list of predefined topics. The aim is to assign each sample to one and only one label
  - We attack the problem by utilizing supervised and unsupervised ML techniques
  - We expand on the basic exploratory data analysis presented in the previous module and create richer visualizations with extra meaning and depth
  - The transformation of data from a high-dimensional space into a low-dimensional one
  - Then, we implement two classifiers and compare the different models
  - Finally, we introduce state-of-the-art word representation techniques with unique properties

Module objectives

After completing this module, you should be able to:

- Creating comprehensive plots
- Reducing the complexity of data either for visualization or classification
- Setting up a baseline model
- Training the classification models
- Fine-tuning the hyperparameters
- Understanding state-of-the-art word representation techniques

Machine Learning Techniques for Text

Section 1: Understanding topic classification

Topic classification

- Businesses deal with many other unstructured texts, such as news posts, support tickets, or customer reviews
- Failing to glean this data efficiently can lead to missed opportunities or, even worse, angry customers
- We focus on the problem of ***topic classification***, with the aim to assign a topic to some piece of text
- We focus on the problem of topic classification (***multiclass classification***), intending to assign a label (or topic) to a piece of text
- For this task, we use the ***20 newsgroups*** dataset available in the **scikit-learn** module, which comprises around 18,000 news posts on 20 topics

Machine Learning Techniques for Text

Section 2: Performing exploratory data analysis

Exploratory data analysis

- A primary concern during ***exploratory data analysis*** (EDA) is to verify that the dataset is appropriately formatted
- For instance, it is not uncommon to encounter missing or out-of-the range values
- Plotting the data or extracting various statistics can reveal this unpleasant situation
- We also might need to transform or exclude part of the data
- Having an imbalanced dataset where one class monopolizes the whole corpus is also a source of concern
  - The ML algorithm is overexposed and subsequently learns data of one class type well while having difficulty with samples from the less frequent classes

Dimensionality reduction

- Selecting the appropriate features for a given problem is not easy
  - We can end up with redundant or highly correlated features that unnecessarily tangle the ML algorithm
  - For example, consider the task of classifying planets based on two attributes, radius (***r***) and circumference (***2πr***)
  - We are using two highly correlated quantities, and there is no extra benefit to including both in the feature space
  - The solution is to either keep one of them or introduce a new feature that is a linear combination of radius and circumference
- This process is called **dimensionality reduction** and proves to be very helpful for speeding up the training of ML algorithms, filtering noise out of the data, performing feature extraction, and data visualization

Principal Component Analysis

- As part of the EDA, it can be helpful to visualize high-dimensional spaces in a way that our limited human brains can comprehend
- ***Principal component analysis*** (PCA) is dimensionality reduction technique that deals with unlabeled data, and for this reason, it is an unsupervised learning method
- The method creates a new coordinate system with a new set of orthogonal axes (principal components)
  - the first axis goes toward the highest variance in the data
  - the second one goes toward the second-highest variance

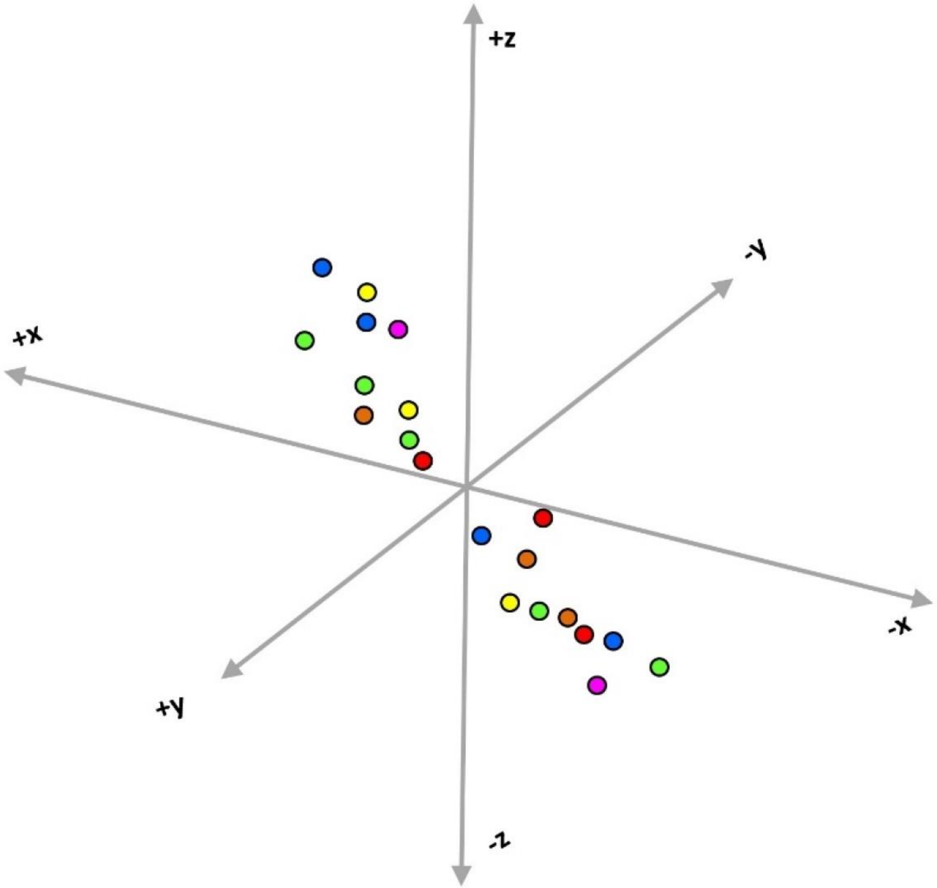
Principal Component Analysis

- As part of the EDA, it can be helpful to visualize high-dimensional spaces in a way that our limited human brains can comprehend
- ***Principal component analysis*** (PCA) is dimensionality reduction technique that deals with unlabeled data, and for this reason, it is an unsupervised learning method
- The method creates a new coordinate system with a new set of orthogonal axes (principal components)
  - the first axis goes toward the highest variance in the data
  - the second one goes toward the second-highest variance

tip *Variance is a statistical measure of dispersion that shows how far data points are spread out from their mean value*

Principal Component Analysis

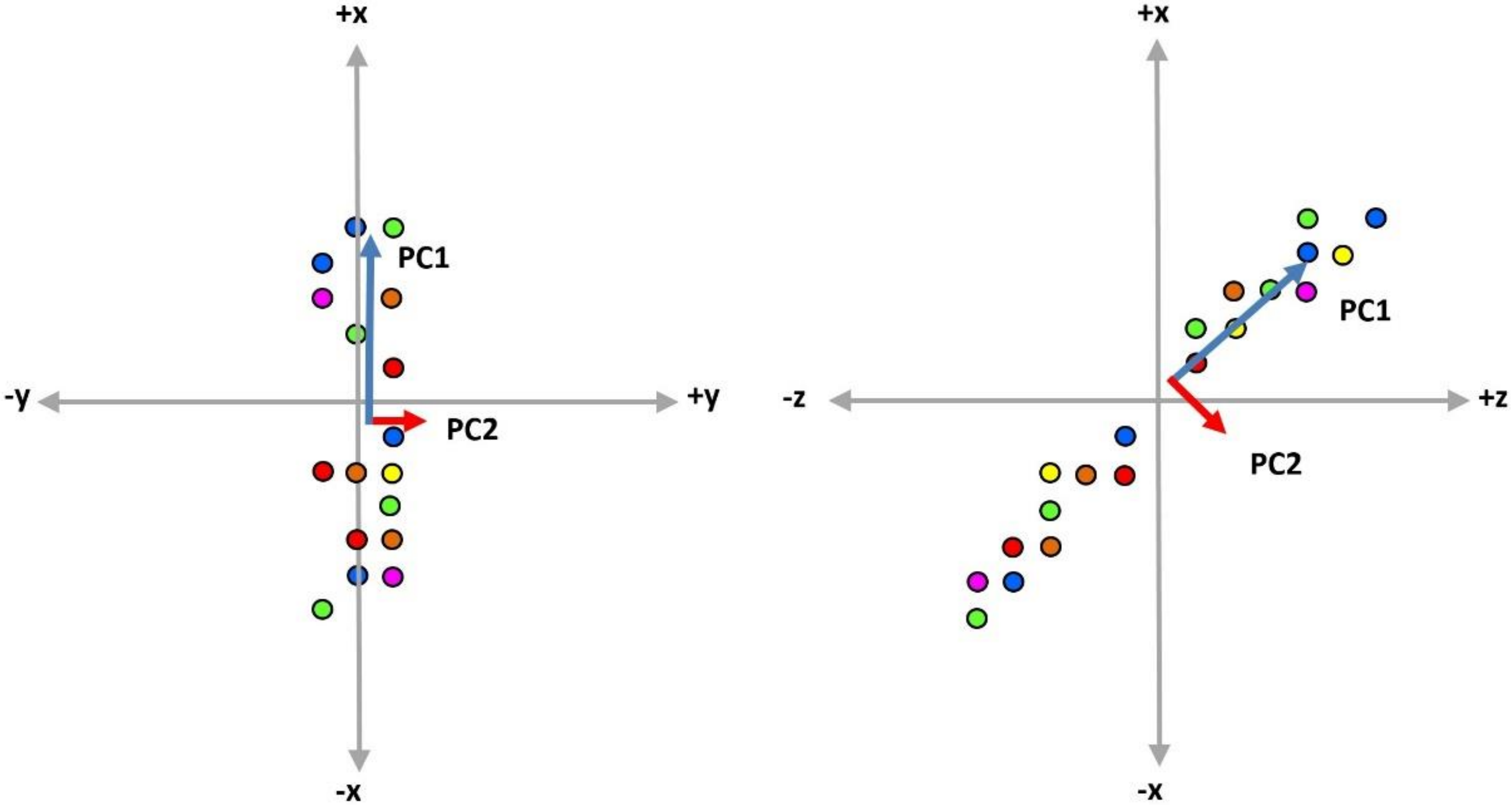
- Plot of 20 random points in a 3-D space



x	y	z
0.1	0.1	0.1
0.2	0	0.2
0.3	-0.1	0.3
0.4	-0.1	0.4
-0.1	0.1	-0.1
-0.2	0	-0.2
-0.3	0.1	-0.3
-0.4	0	-0.4
-0.5	0.1	-0.5
0.2	0	0.1
0.3	0.1	0.2
0.4	-0.1	0.5
0.5	0.1	0.4
0.5	0	0.6
0.3	-0.1	0.4
-0.2	-0.1	-0.1
-0.4	0.1	-0.3
-0.2	0.1	-0.3
-0.6	-0.1	-0.5
-0.5	0	-0.4

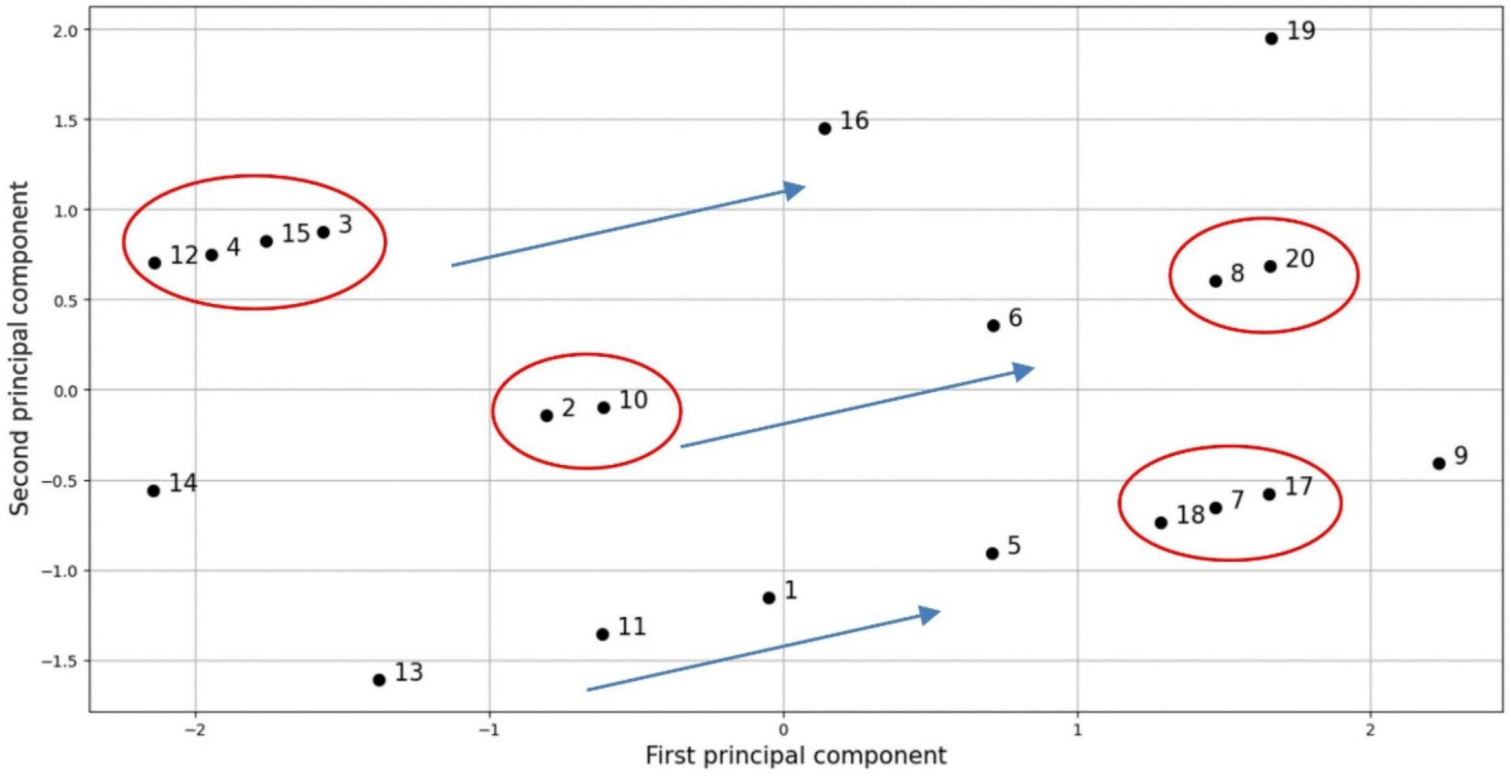
Principal Component Analysis

- Plot of 20 random points in a 3-D space



Principal Component Analysis

- A plot of the data points clusters in the new space



Let’s try!

Demos

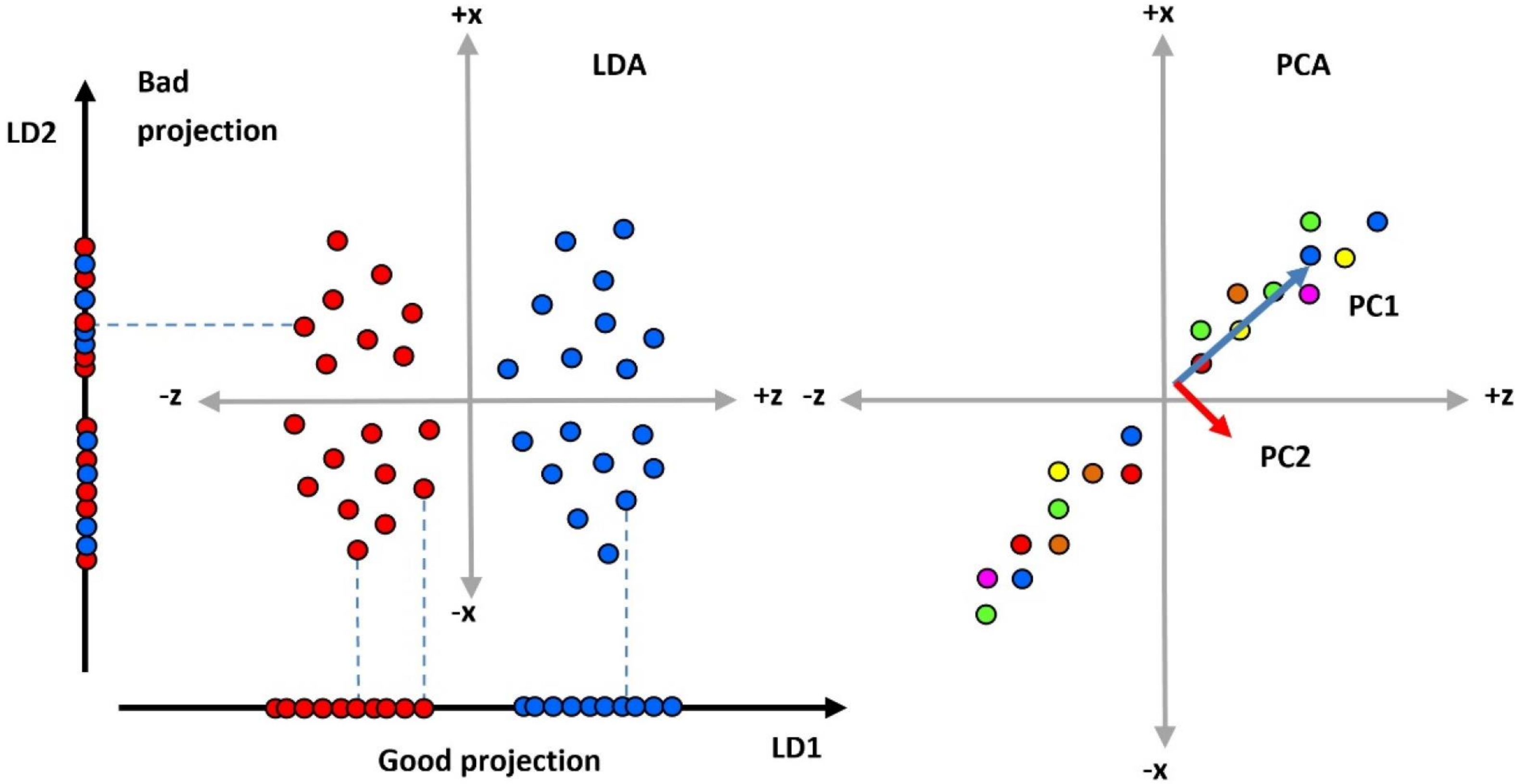
- https://setosa.io/ev/principal-component-analysis/
- https://projector.tensorflow.org/

Linear Discriminant Analysis

- ***Linear discriminant analysis*** (LDA) is also a dimensionality reduction technique
- While PCA aims to identify the combination of principal components that maximize the variance in a dataset, LDA maximizes the separability between different classes by projecting the points onto a lower-dimensional space
- It aims to find the linear projection of the data in this subspace that optimizes some measure of class separation
- In contrast to the PCA algorithm, LDA is a supervised method

Linear Discriminant Analysis

- The aim of both is to find the right components PCA: highest variance, LDA: highest separability



Occam's razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Occam’s razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Simpler can be better!

Problem: *Find the next number in the sequence: 1, 3, 5, 7, ?*

Occam’s razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Simpler can be better!

Problem: *Find the next number in the sequence: 1, 3, 5, 7, ?*
Answer: *9*

Occam's razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Simpler can be better!

Problem: *Find the next number in the sequence: 1, 3, 5, 7, ?*

Answer: *9*

Wrong! The correct answer is 217341. As according to our model:

$$f(x) = 9055.5 * x^4 - 90555 * x^3 + 316942.5 * x^2 - 452773 * x + 217331$$

$$f(1) = 1, f(2) = 3, f(3) = 5, f(4) = 7 \; and \; f(5) = 217341$$

Occam's razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Simpler can be better!

Problem: *Find the next number in the sequence: 1, 3, 5, 7, ?*
Answer: *9*

Wrong! The correct answer is 217341. As according to our model:

$$f(x) = 9055.5 * x^4 - 90555 * x^3 + 316942.5 * x^2 - 452773 * x + 217331$$

$$f(1) = 1, f(2) = 3, f(3) = 5, f(4) = 7 \text{ and } f(5) = 217341$$

In this case a much simpler model will work better:

$$f(x) = f(x-1) + 2, \text{where } f(1) = 1 \text{ and } x \text{ is positive natural number.}$$

Occam’s razor

- A much smaller representation with less features can provide the same performance as a model with many features
- How can we choose between several possible and more complex alternatives for solving a particular problem?

Simpler can be better!

Problem: *Find the next number in the sequence: 1, 3, 5, 7, ?*
Answer: *9*

Wrong! The correct answer is 217341. As according to our model:

$$f(x) = 9055.5 * x^4 - 90555 * x^3 + 316942.5 * x^2 - 452773 * x + 217331$$

$$f(1) = 1, f(2) = 3, f(3) = 5, f(4) = 7 \text{ and } f(5) = 217341$$

In this case a much simpler model will work better:

$$f(x) = f(x-1) + 2, \text{where } f(1) = 1 \text{ and } x \text{ is positive natural number.}$$

tip

Precedence should be given to simplicity; the simpler explanation of the problem must be preferred

Let’s practice!

Tasks

- Exploratory data analysis
- Dimensionality reduction

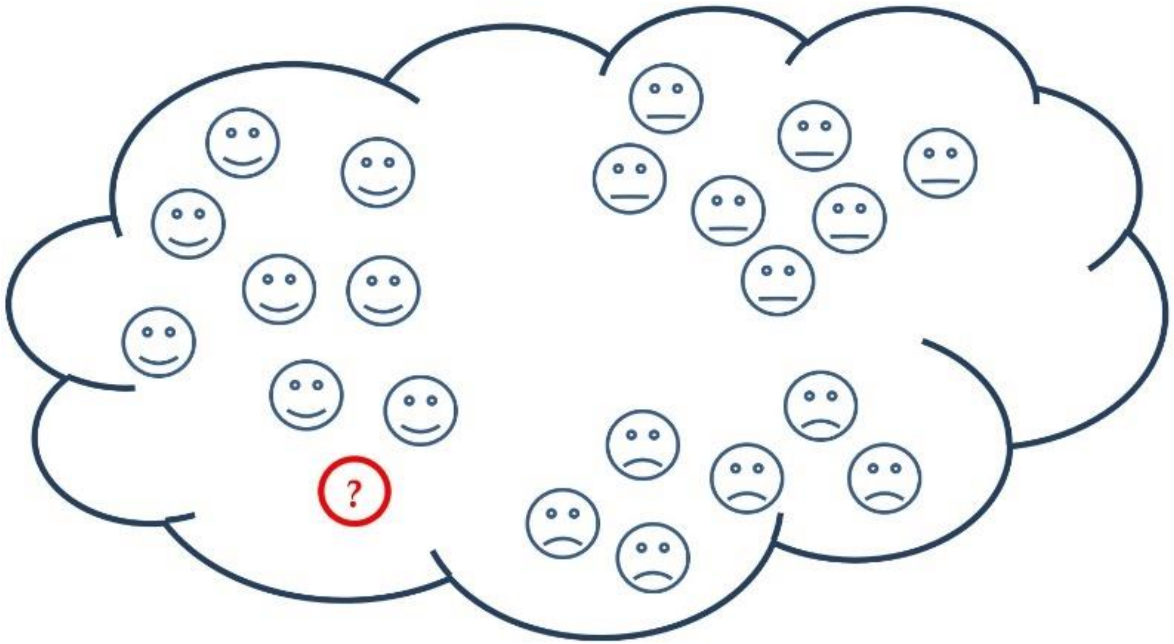
https://colab.research.google.com/github/PacktPublishing/Machine-Learning-Techniques-for-Text/blob/main/chapter-03/topic-classification.ipynb

Machine Learning Techniques for Text

Section 3: Performing classification

K-Nearest Neighbors

- Consider the cloud that contains three types of smiley faces – happy, sad, and neutral
- There is also a hidden face depicted by a question mark. If you had to guess what its actual type was, what would that be?

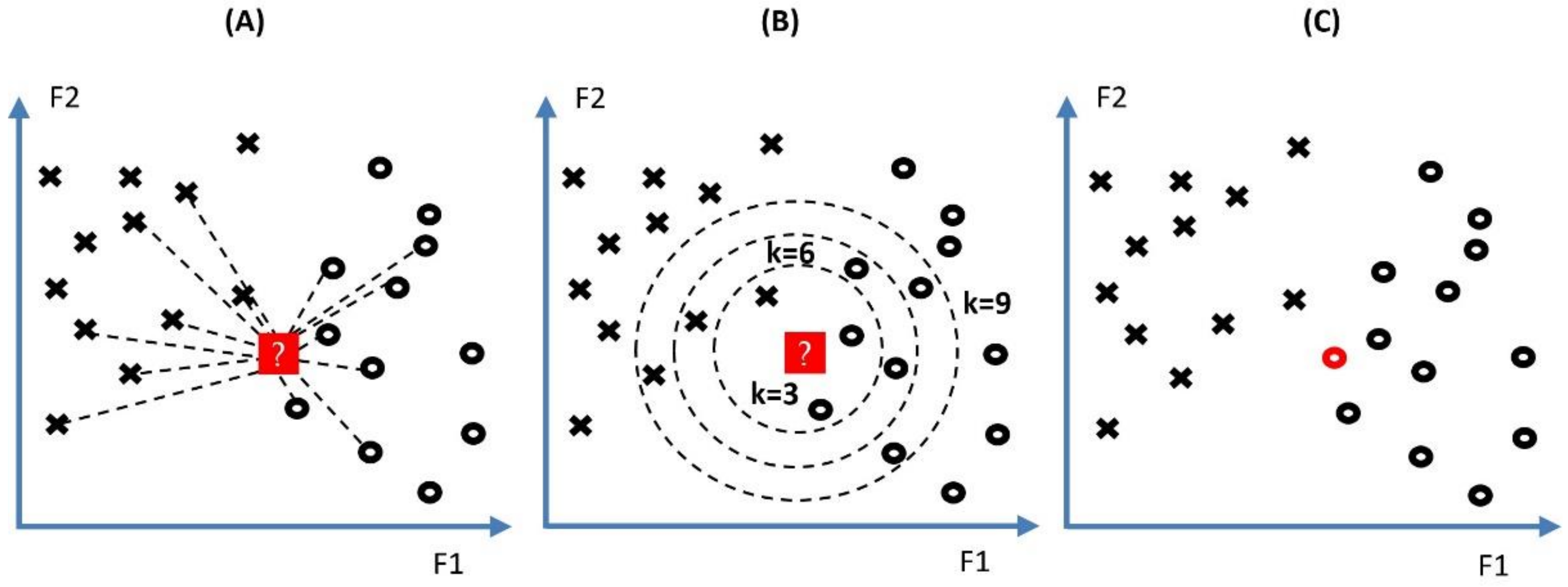


K-Nearest Neighbors

- ***K-Nearest Neighbors*** (KNN) is a non-parametric and lazy learning method that stores the position of all data samples and classifies new cases based on some similarity measure
- Lazy learning means that the algorithm takes almost zero time to learn in this case
- The training samples are stored and used to classify new observations based on a majority vote
- ***K*** is the only hyperparameter of KNN and specifies the number of closest neighbors to be considered
  - when $K = 1$, the nearest neighbor class is assigned to the new sample
  - when $K = 3$, the three closest neighbors are examined

K-Nearest Neighbors

- We choose different values for ***K*** and examine the data points in each neighborhood



Cross-validation

- What should the value of ***K*** be?
- Fine-tuning it using ***cross-validation***
- Three basic steps
  - Partitioning the data into several subsets (folds)
  - Holding out one of the subsets each time and training the model with the rest
  - Evaluating the model with the holdout test
- 5-fold cross-validation:

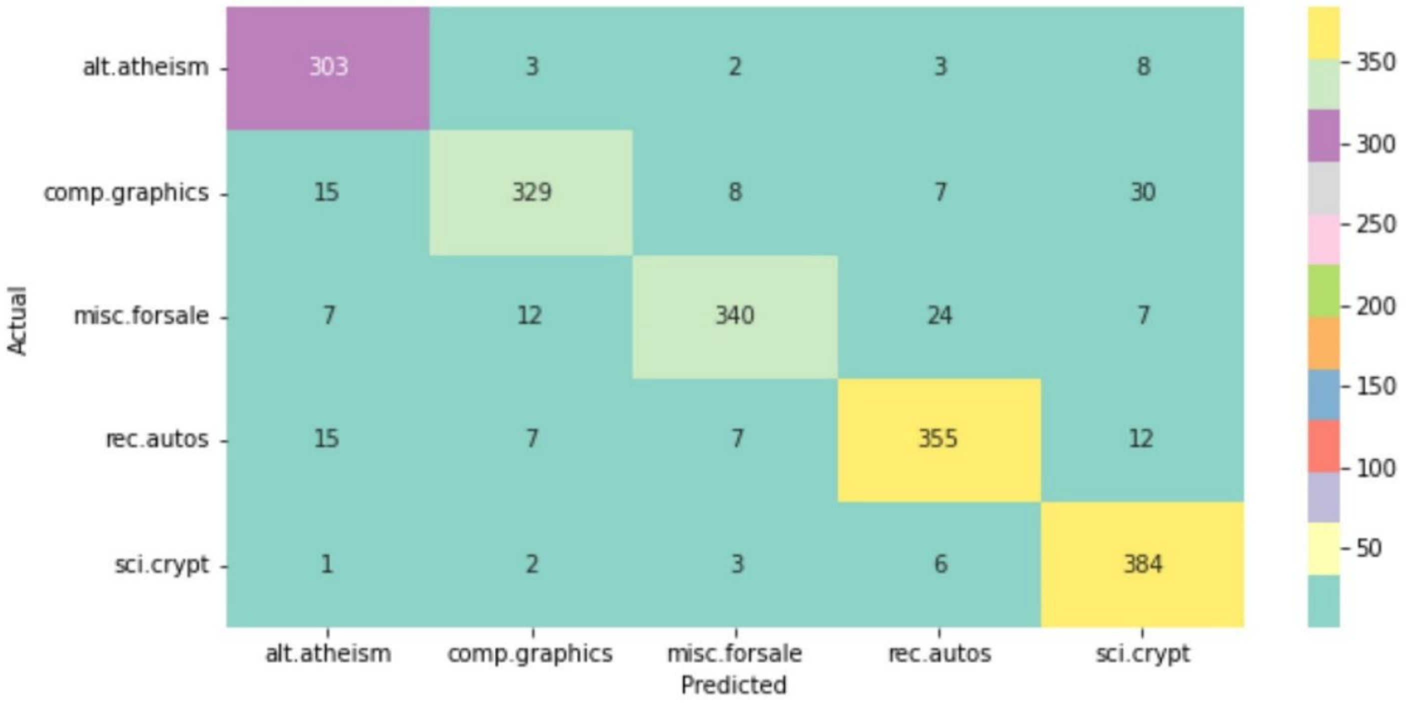
Train Fold
Test Fold

Training data

Iteration 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Score
Iteration 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Score
Iteration 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Score
Iteration 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Score
Iteration 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Score

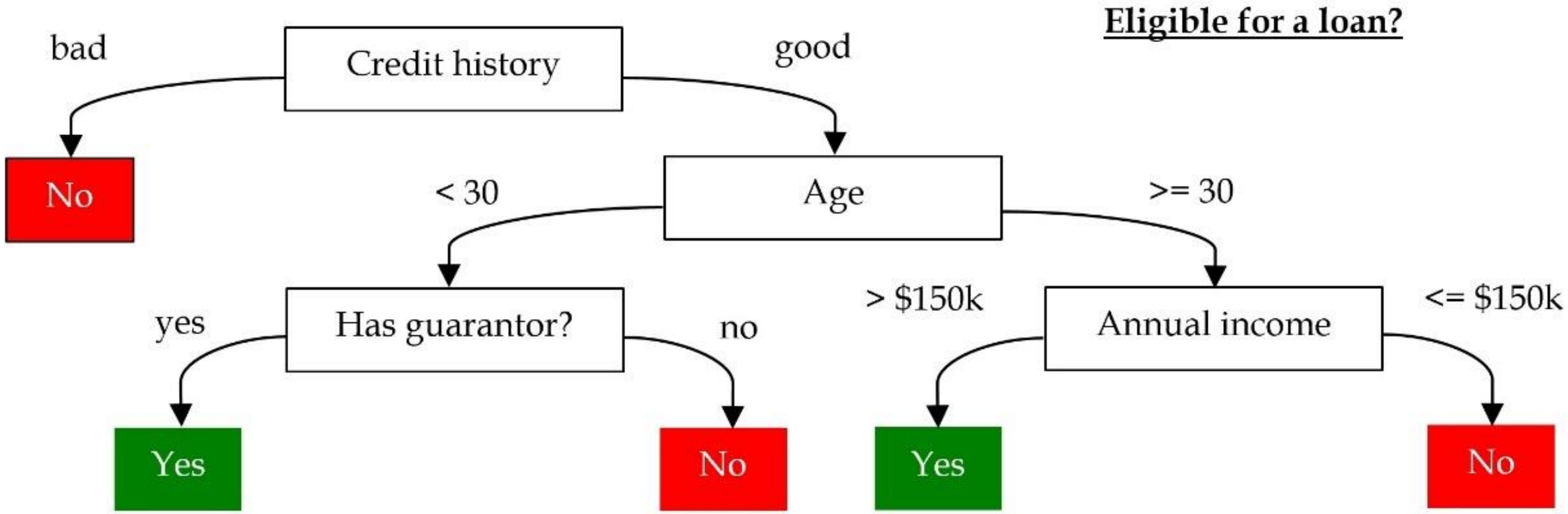
Confusion matrix

- The ***confusion matrix*** provides a better analysis of the strengths and weaknesses of the model
- Each row (or column) represents the instances in the actual class, while each column (or row) represents the instances in the predicted one



Decision trees

- ***Decision trees*** are one of the most popular supervised ML algorithms because their models are intuitive and easy to explain
- The data is represented in a tree hierarchy where:
  - each internal (non-leaf) node is labeled with an input feature
  - the arcs in the internal nodes signify possible values for a specific feature
  - each leaf represents a class



Random forest

- In ***ensemble learning***, multiple classifiers are generated and combined to solve a particular problem

or

- The ***random forest*** method exploits the benefits of ensemble learning by constructing a multitude of decision trees on randomly selected data samples
- Each decision tree produces its own prediction and the method is responsible for choosing the best result by voting

Singular Value Decomposition

- PCA and LDA help to visualize high-dimensional data
- Techniques of this kind can also be applied during classification to reduce the feature space of the problem
  - Too many features can degrade the performance of ML algorithms while increasing computation and memory requirements
- A suitable method for dimensionality reduction is the ***Singular Value Decomposition*** (SVD)
  - expresses the feature space in a new components system
  - works well with sparse matrices frequently encountered in text classification

Singular Value Decomposition

- PCA and LDA help to visualize high-dimensional data
- Techniques of this kind can also be applied during classification to reduce the feature space of the problem
  - Too many features can degrade the performance of ML algorithms while increasing computation and memory requirements
- A suitable method for dimensionality reduction is the ***Singular Value Decomposition*** (SVD)
  - expresses the feature space in a new components system
  - works well with sparse matrices frequently encountered in text classification

Let’s practice!

Tasks

- Exploratory data analysis
- Dimensionality reduction
- Classification

https://colab.research.google.com/github/PacktPublishing/Machine-Learning-Techniques-for-Text/blob/main/chapter-03/topic-classification.ipynb

Machine Learning Techniques for Text

Section 4: Extracting word embedding representation

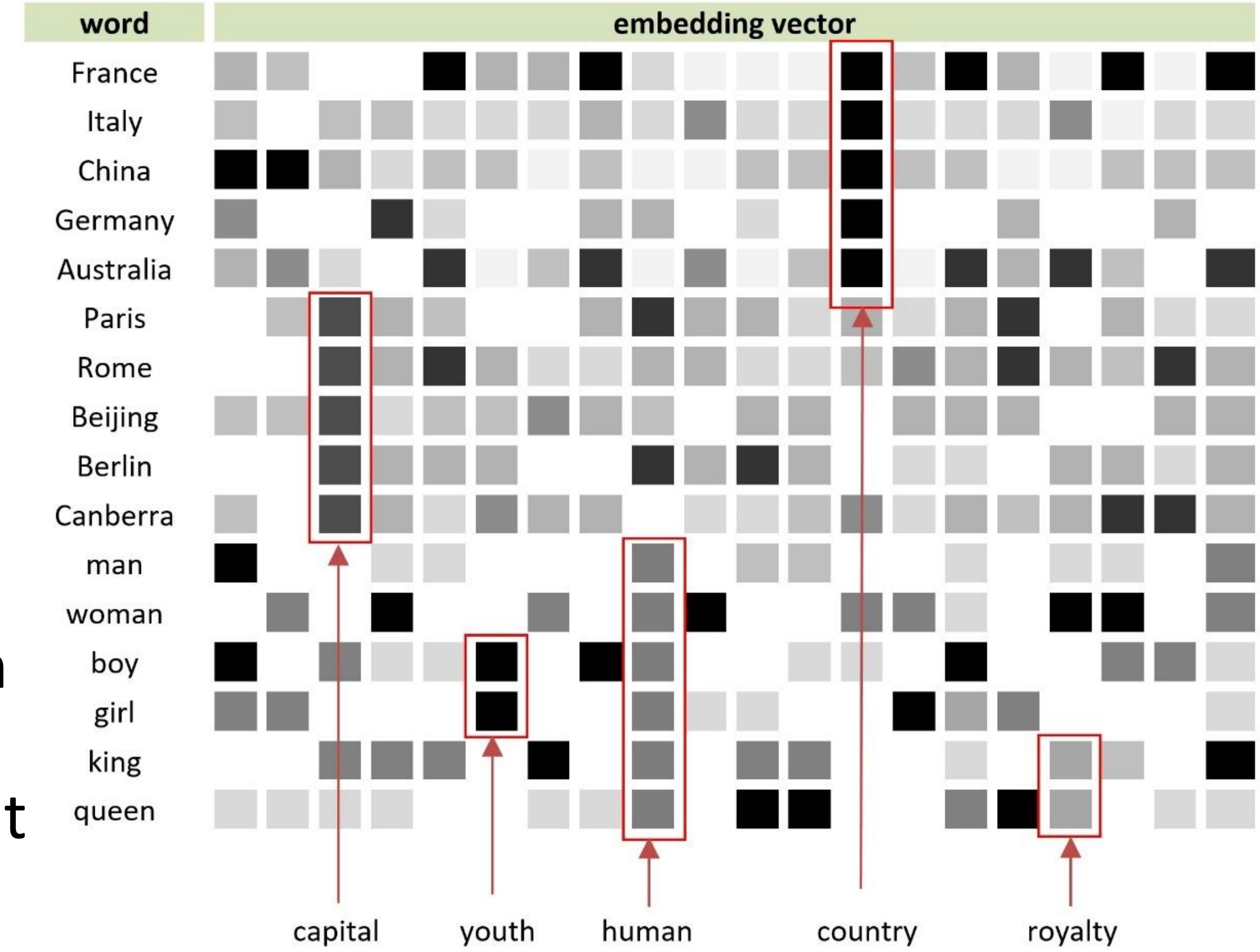
Match profiles

- You are assigned to create the matching algorithm for a new dating service
- This algorithm must identify people with similar characteristics (Big Five) and propose candidate profiles

*The **Big Five** personality traits is a taxonomy for human personality and psyche*

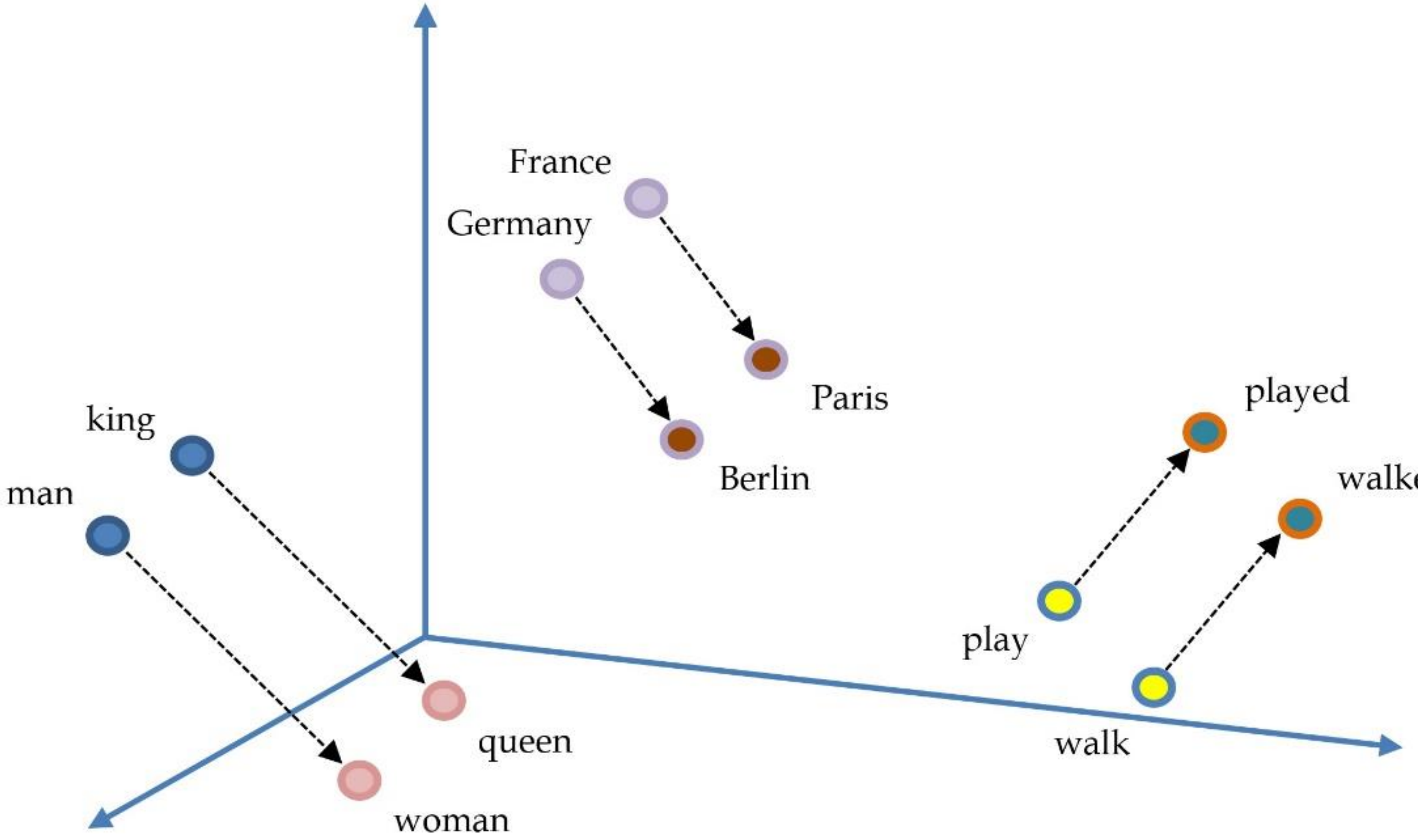
Word embedding

- Just as the five traits represent each person as a unique point in a five-dimensional space, ***word embedding*** represent words in a multidimensional space, typically in the order of hundreds
- Following the same approach as before, we show the embedding vector of different English words



Word embedding

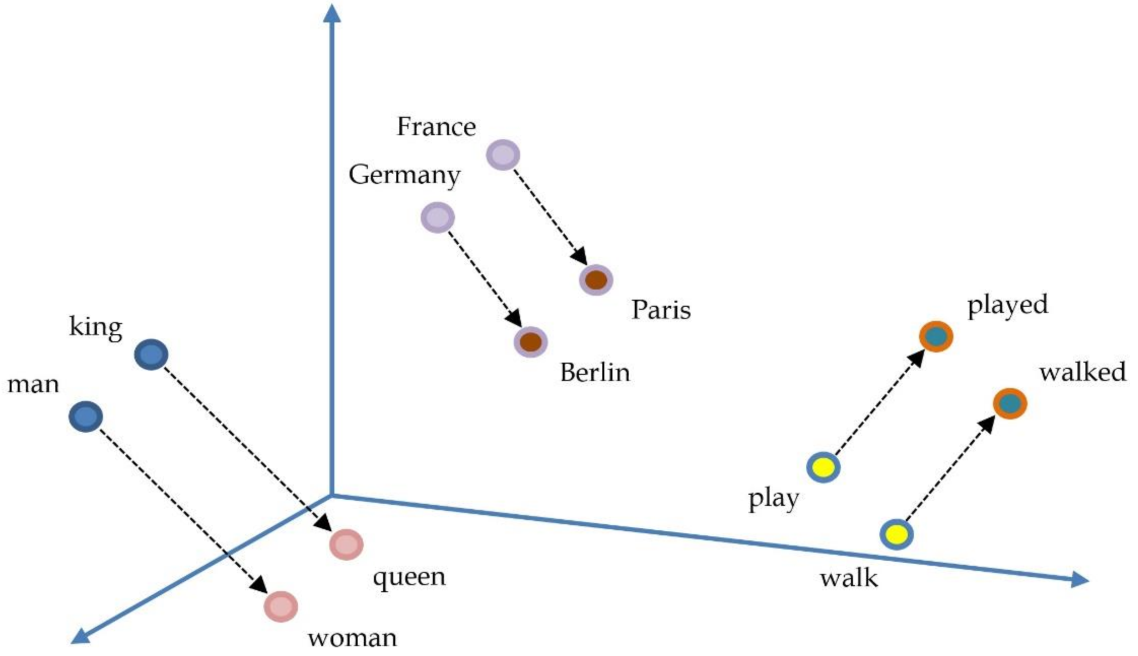
Embed the points of a set of English words into a three-dimensional space



Vector arithmetic

- We can build word analogies using statements "*a is to b as c is to d*". For example:
  - "Paris is to France as Berlin is to Germany"
  - "King is to man as queen is to woman"
  - *etc.*

- Essentially, we subtract embedding vectors in all these equations, a process called ***vector arithmetic***. For example:
  - `man – psychiatrist = woman - psychologist`



Let’s practice!

Tasks

- Exploratory data analysis
- Dimensionality reduction
- Classification
- Word embedding

https://colab.research.google.com/github/PacktPublishing/Machine-Learning-Techniques-for-Text/blob/main/chapter-03/topic-classification.ipynb

Key takeaways

Visualizations

- N-gram frequencies
- Pie charts
- Scatter plots
- Heatmaps

Dimensionality reduction

- Principal Component Analysis
- Linear Discriminant Analysis
- Singular Value Decomposition

ML algorithms & models

- ZeroR
- K-Nearest Neighbor
- Random Forest
- Decision Trees

Text representations

- Word2Vec

ML concepts

- Unsupervised learning
- Cross-Validation

Tools

- fastText

Machine Learning Techniques for Text

Questions?