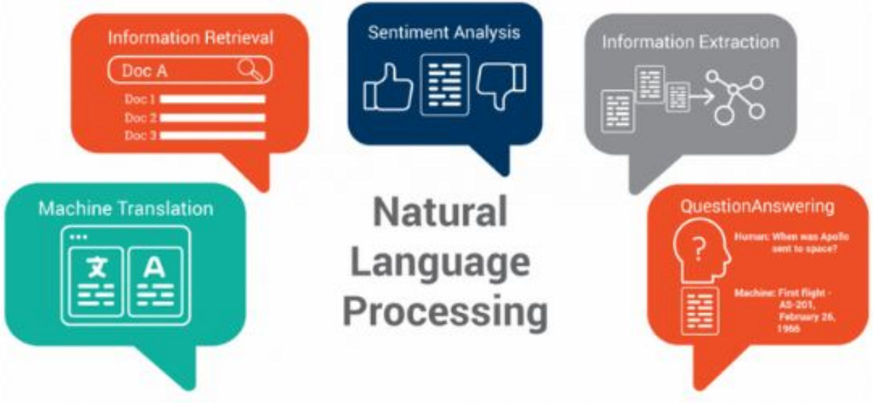
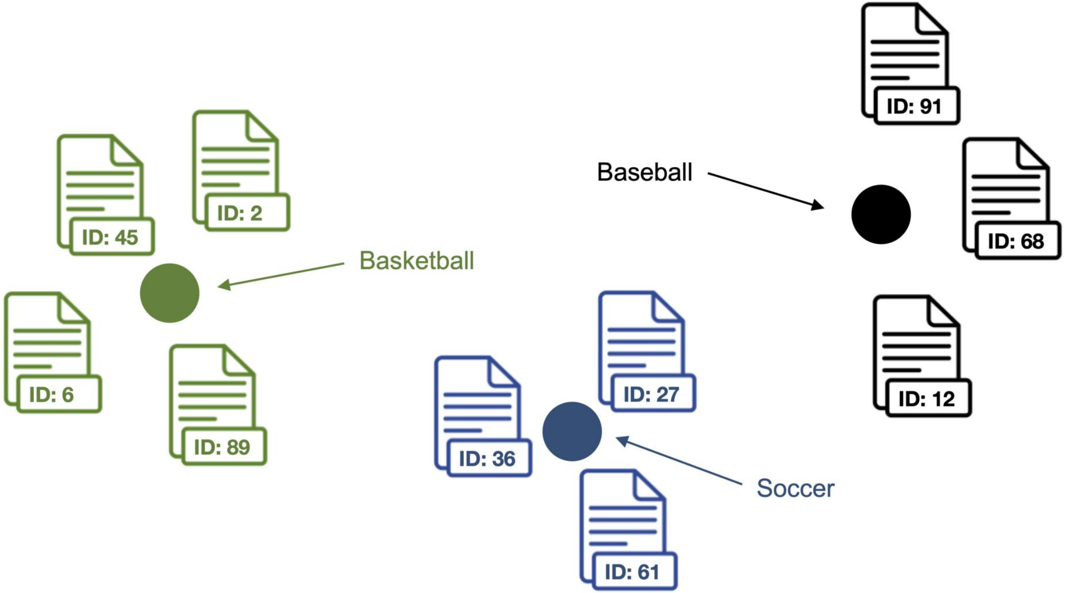


Solving Business Problems with NLP

Instructor: Juber Rahman

Topic Classification



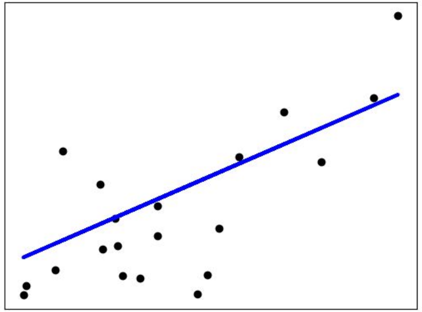
Reference: Towards Data Science blog

- ❑ Text Classification and relating with a topic
- ❑ Used in legal document classification, electronic health record classification, social media analytics, etc.
- ❑ Both supervised and unsupervised models can be used.
- ❑ Unsupervised approach is more convenient

Unsupervised vs Supervised Learning

Supervised learning: discover patterns in the data that relate feature attributes with a target (class/value).

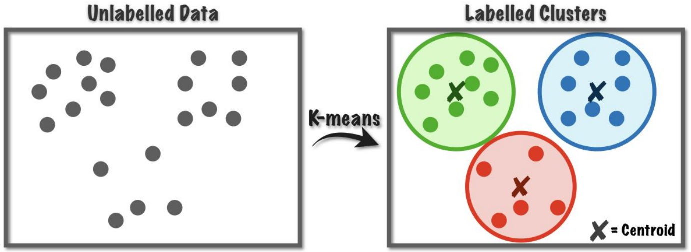
These patterns are then utilized to predict the values of the target attribute in future data instances.



Example- Linear Regression

Unsupervised learning : The data have no target attribute.

We want to explore the data to find some intrinsic structures in them.



Example- KMeans Clustering

Popular NLP methods for Topic Classification

❑ Latent Dirichlet Allocation (LDA)

❑ Non-negative Matrix Factorization (NMF)

❑ Latent Semantic Indexing (LSI)

Packages:

- Scikit-learn
- Gensim

Pre-processing:

- Cleaning, stopword removal, vectorizing
- Same as supervised learning

N.B. Most of the topic classification methods are generative and not discriminative - same document may belong to different topics

Discriminative vs Generative Models

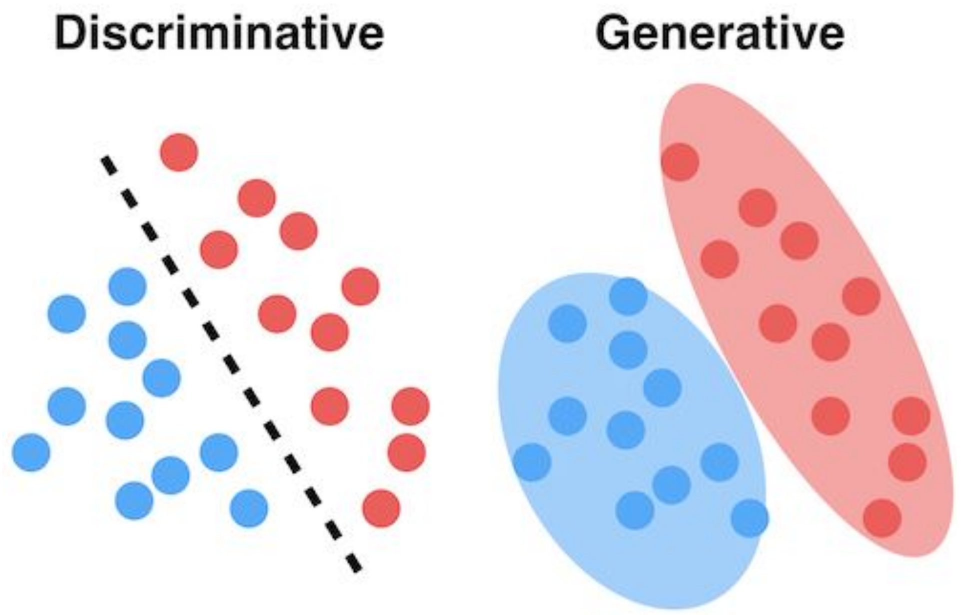
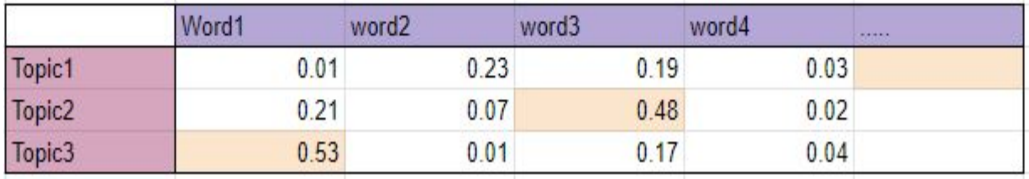


Image Source:
https://dataisutopia.com/blog/discremenet-generative-models

- Discriminative models (SVM, RF, LR) also called *conditional models*, tend to learn the boundary between classes/labels in a dataset.

- Generative models (HMM, NB, LDA, etc.) are models where the focus is the distribution of individual classes in a dataset and the learning algorithms tend to model the underlying patterns/distribution of the data points.

Latent Dirichlet Allocation (LDA)



	Word1	word2	word3	word4	.....
Topic1	0.01	0.23	0.19	0.03	
Topic2	0.21	0.07	0.48	0.02	
Topic3	0.53	0.01	0.17	0.04	

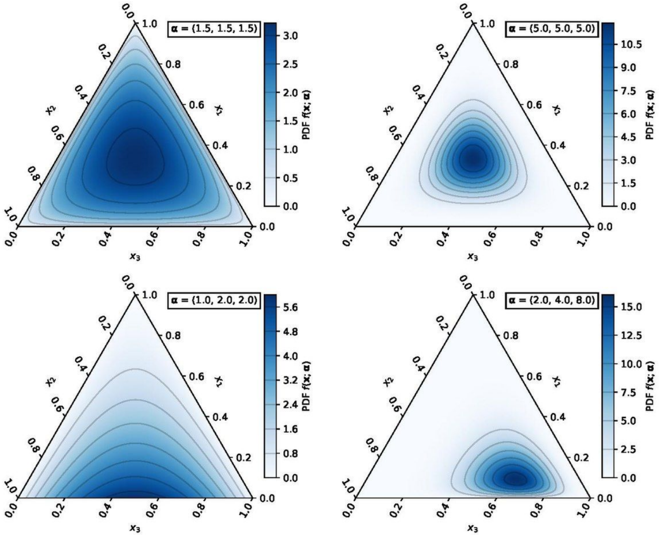
- In LDA, documents are represented as a mixture of topics and a topic is a bunch of words.
- Topic = multinomial distribution of words
- Each row in the table represents a different topic
- each column a different word in the corpus.
- Each cell contains the probability that the word(column) belongs to the topic(row).
-

LDA explained

‘Latent’ indicates that the model discovers the ‘yet-to-be-found’ or hidden topics from the documents. **‘**

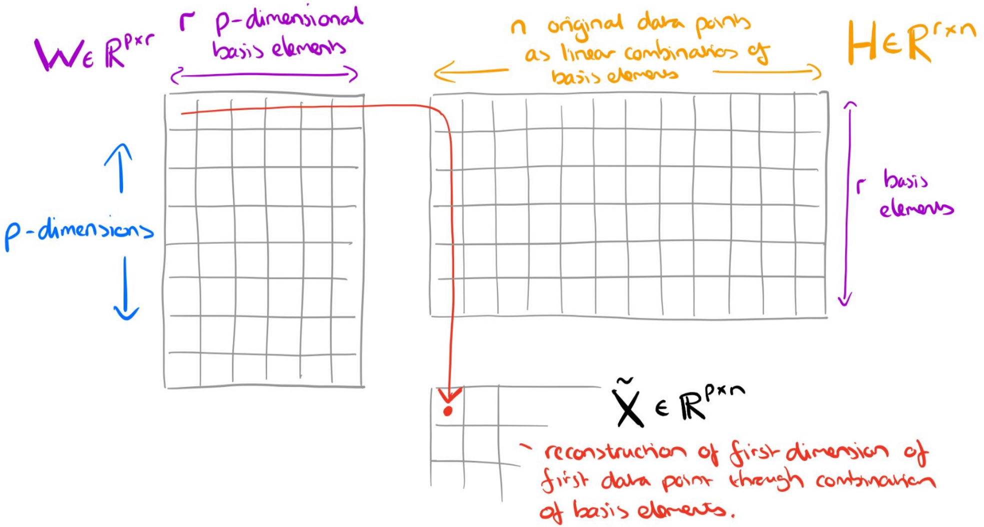
Dirichlet’ indicates LDA’s assumption that the distribution of topics in a document and the distribution of words in topics are both Dirichlet distributions. ‘

Allocation’ indicates the distribution of topics in the document.



Non-negative Matrix Factorization

- Nonnegative matrix factorization (NMF) has become a widely used tool for the analysis of high dimensional data as it automatically extracts sparse and meaningful features from a set of nonnegative data vectors.
- NMF approximates a matrix X with a low-rank matrix approximation such that


W ∈ R^{p×r}
r p-dimensional basis elements
p-dimensions
n original data points as linear combinations of basis elements
H∈R^{r×n}
r basis elements
X̃ ∈ ℝ^{p×n}
reconstruction of first dimension of first data point through combination of basis elements.

X ~WH

$$\underbrace{X(:,j)}_{j\text{th document}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{k\text{th topic}} \underbrace{H(k,j)}_{\substack{\text{importance of } k\text{th topic} \\ \text{in } j\text{th document}}} , \quad \text{with } W \geq 0 \text{ and } H \geq 0.$$

Supplemental Materials:

1. https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture20-topic-models.pdf
2. https://personal.utdallas.edu/~nrr150130/cs6347/2017sp/lects/Lecture_18_LDA.pdf
3. https://towardsdatascience.com/dirichlet-distribution-a82ab942a879
4. https://blog.acolyer.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/
5.