CS5056 DATA ANALYTICS TEXT ANALYTICS

HÉCTOR G. CEBALLOS, FRANCISCO CANTÚ

CEBALLOS@TEC.MX, FCANTU@TEC.MX

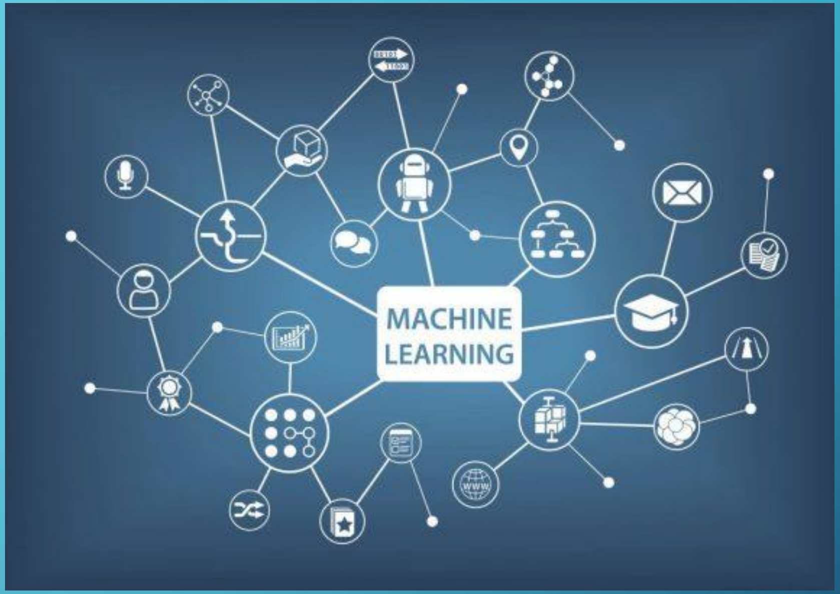


Image: https://www.mba-madrid.com/empresas/impacto-del-machine-learning-ambito-empresarial/

AGENDA

- Text Analytics
- Text Analytics exercise

THEORY

Image: https://www.muycomputerpro.com/2018/02/07/glosario-terminos-basicos-machine-learning

TEXT ANALYTICS

- It involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

https://en.wikipedia.org/wiki/Text_mining

TEXT ANALYTICS TASKS

- Text categorization
- Text clustering
- Concept extraction
- Production of granular taxonomies
- Sentiment analysis
- Document summarization
- Learning of entity relation

https://en.wikipedia.org/wiki/Text_mining

TEXT ANALYTICS TASKS

- Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics.
- The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

https://en.wikipedia.org/wiki/Text_mining

TEXT ANALYTICS APPLICATIONS

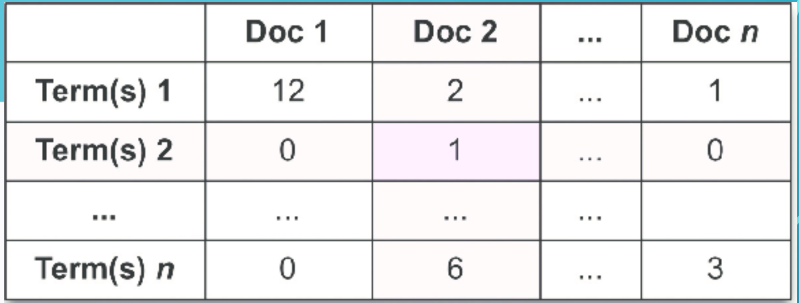
- Security applications
- Biomedical applications
- Software applications
- Online media applications
- Business and marketing applications
- Sentiment analysis
- Scientific literature mining and academic applications
- Digital humanities and computational sociology

https://en.wikipedia.org/wiki/Text_mining

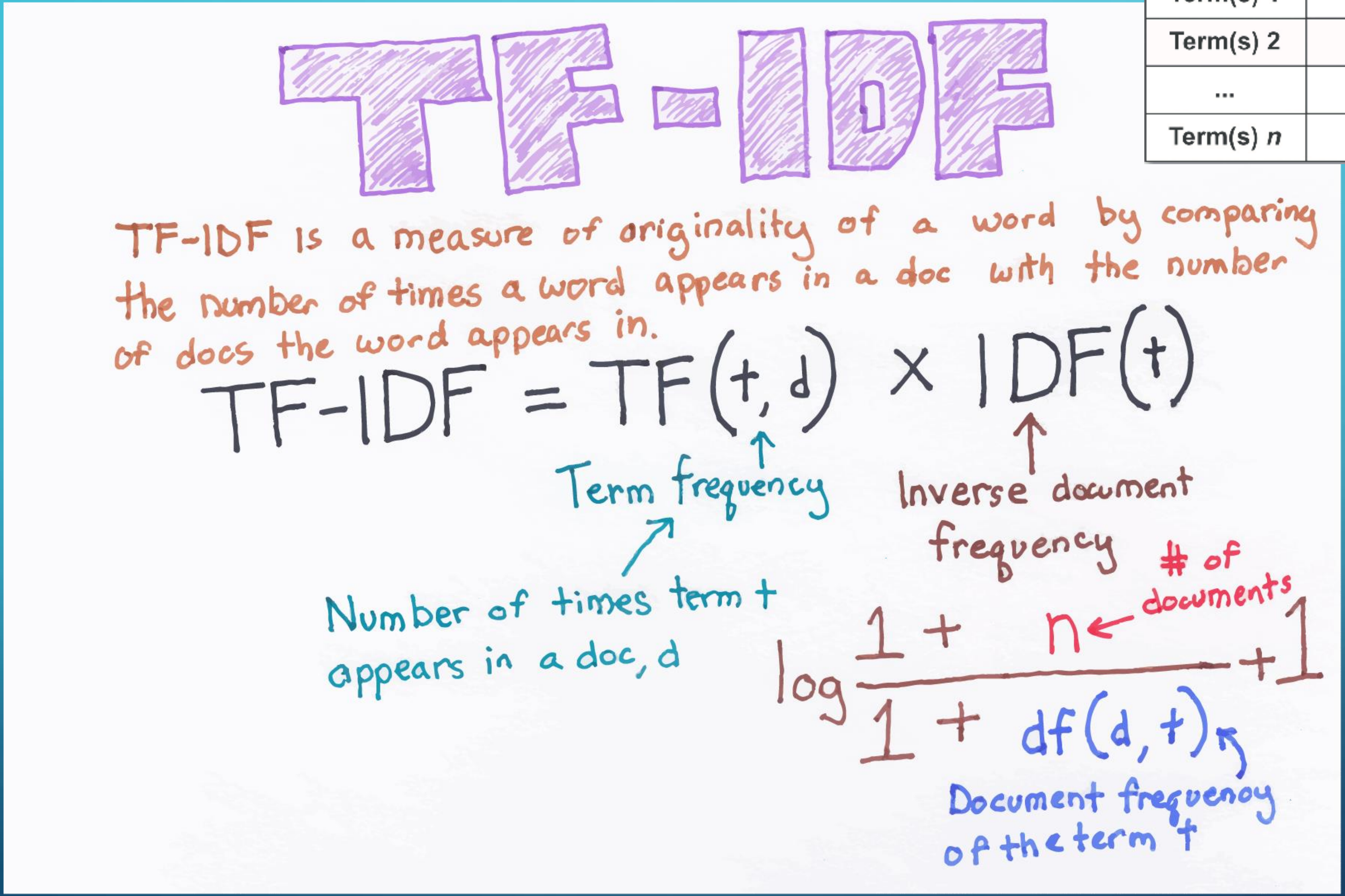
FEATURE ENGINEERING

- Count Vectors as features
- TF-IDF Vectors as features
  - Word level
  - N-Gram level
  - Character level

- Word Embeddings as features
- Text / NLP based features
- Topic Models as features



	Doc 1	Doc 2	...	Doc *n*
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	...	...	...	
Term(s) *n*	0	6	...	3



TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = TF(t, d) \times IDF(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

$$\log \frac{1 + n}{1 + df(d, t)} + 1$$

# of documents

Document frequency of the term t

https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

TF-IDF VECTORS AS FEATURES

- TF-IDF Vectors can be generated at different levels of input tokens (words, characters, n-grams)
  - Word Level TF-IDF : Matrix representing tf-idf scores of every term in different documents
  - N-gram Level TF-IDF : N-grams are the combination of N terms together. This Matrix representing tf-idf scores of N-grams
  - Character Level TF-IDF : Matrix representing tf-idf scores of character level n-grams in the corpus

TEXT / NLP-BASED FEATURES

- A number of extra text based features can also be created which sometimes are helpful for improving text classification models. Some examples are:
    - Word Count of the documents – total number of words in the documents
    - Character Count of the documents – total number of characters in the documents
    - Average Word Density of the documents – average length of the words used in the documents
    - Puncutation Count in the Complete Essay – total number of punctuation marks in the documents

TOPIC MODELS AS FEATURES

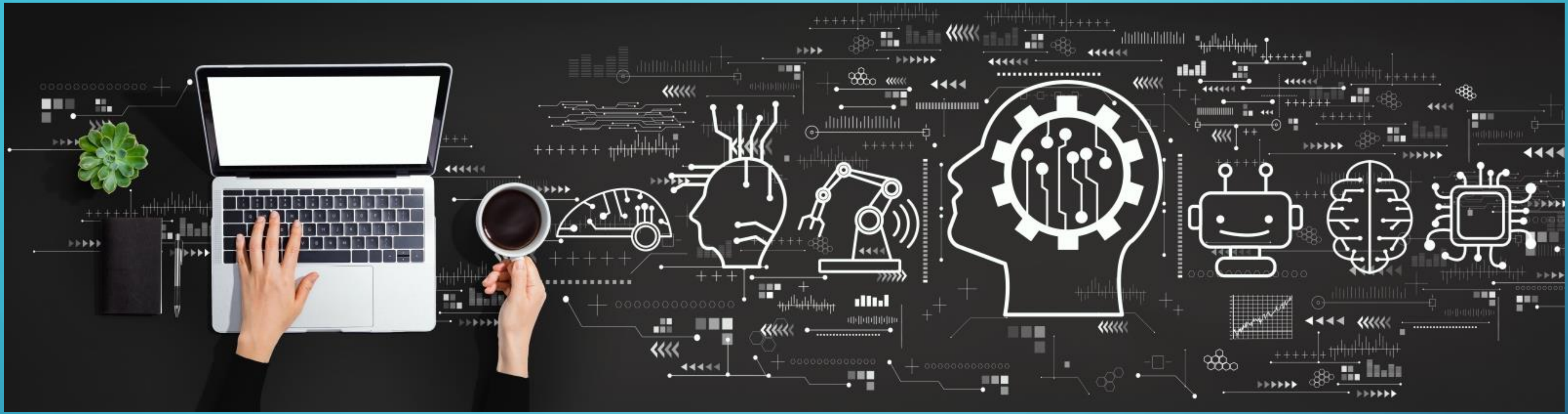
- Topic Modelling is a technique to identify the groups of words (called a topic) from a collection of documents that contains best information in the collection
- Latent Dirichlet Allocation (LDA) is a technique commonly used for generating Topic Modelling Features.

TOPIC MODELS AS FEATURES

- LDA is an iterative model which starts from a fixed number of topics
- Each topic is represented as a distribution over words, and each document is then represented as a distribution over topics
- Although the tokens themselves are meaningless, the probability distributions over words provided by the topics provide a sense of the different ideas contained in the documents

MODEL BUILDING

- Supervised learning
  - Classify text into categories
  - Estimate the influence of certain feature in a given score.
- Unsupervised learning
  - Identify topics / Keywords
  - Cluster documents



PRACTICE

TEXT MINING EXERCISE

Credits: https://expertsystem.com/machine-learning-definition/

EXERCISE

- Load a dataset with 3.6M Amazon reviews.
- Extract features using: count vectors, TF-IDF (Word, NGram, Characters)
- Learn classification models: Logistic Regression, SVM, Random Forest.
- Compare the accuracy of the models and the features.