Chapter 2

Literature Review

2.1 Introduction

Our goal is to implement a system using machine learning for conspiracy checking from email data. To do so we have used some efficient algorithms and tools and study. It will be described in details here.

2.2 Conspiracy

According to the Oxford dictionary conspiracy is a secret plan by a group to do something unlawful and harmful. Ex: 'a conspiracy to destroy the government'. In another word the action of plotting or conspiring

Another definition of conspiracy is found in the Cambridge dictionary that a secret agreement made between two or more people or groups to do something bad or illegal that will harm someone else:

Conspiracy against sb It is my client's opinion that there has been a conspiracy against him.

conspiracy between sb (and sb) The group of optometrists denied there was any conspiracy between them and other industry associations to stop the sale of lenses to mail order houses.

conspiracy to do sth The four directors have denied conspiracy to defraud pensioners by misusing shares that belonged to pension funds.

A group of former housing counsellors has been indicted on fraud and conspiracy charges in one of the biggest real estate fraud cases ever seen in the state.

2.3 Conspiracy Theory

Conspiracy theories are omnipresent among members of modern and traditional societies (West & Sanders, 2003). A common definition of conspiracy theory is the conviction that a group of actors meets in secret agreement with the purpose of attaining some malevolent goal (e.g., Bale, 2007). Contrary to the view that belief in such theories is pathological (Hofstadter, 1966), large portions of the human population believe conspiracy theories. In 2004, 49% of New York City residents believed the U.S. government to be complicit in the 9/11 terrorist attacks (Sunstein & Vermeule, 2009). In addition, in a nationally representative sample of the U.S. population, 37% answered "agree" to the following statement: "the Food and Drug Administration is deliberately preventing the public from getting natural cures for cancer and other diseases because of pressure from drug companies." Another 31% answered "neither agree nor disagree," and only 32% disagreed with this statement (Oliver & Wood, 2014). Belief in conspiracy theories is thus a widespread societal phenomenon and has increasingly drawn the research attention of social scientists (for overviews, see Brotherton, 2015; Douglas, Sutton & Cichocka, 2017; van Prooijen, 2018).

Although the definition provided above is rather general, here we explicate the specific underlying features of conspiracy theories. We argue that a conspiracy theory contains at least five critical ingredients.

First, conspiracy theories make an assumption of how people, objects, or events are causally interconnected. Put differently, a conspiracy theory always involves a hypothesized pattern (see Shermer, 2011; Whitson & Galinsky, 2008).

Second, conspiracy theories stipulate that the plans of alleged conspirators are deliberate. Conspiracy theories thus ascribe intentionality to the actions of conspirators, implying agency (Douglas, Sutton, Callan, Dawtry, & Harvey, 2016; Imhoff & Bruder, 2014).

Third, a conspiracy theory always involves a coalition, or group, of actors working in conjunction. An act of one individual, a lone wolf, does not fit the definition of a conspiracy theory (van Prooijen & van Lange, 2014).

Fourth, conspiracy theories always contain an element of threat such that the alleged goals of the conspirators are harmful or deceptive (Hofstadter, 1966)

Fifth, and finally, a conspiracy theory always carries an element of secrecy and is therefore often difficult to invalidate. Conspiracy theories that turn out true—such as Watergate or the Iran-Contra scandal—are no longer conspiracy "theories." Hence, in judging the validity of conspiracy theories, there is always room for error.

People hold many beliefs that share some of the key elements of conspiracy theories, such as supernatural beliefs. Indeed, conspiracy theories and supernatural beliefs are positively correlated (Darwin, Neave, &Holmes, 2011; Swami et al., 2011). What distinguishes conspiracy theories from supernatural beliefs is that they necessarily involve a coalition element of deceptive or potentially dangerous other human beings acting in unison (Bale, 2007). For conspiracy theories to occur, however, these nonhuman stimuli need, at the very least, to be connected to the real or suspected presence of a coordinated group of deliberate actors. Unlike other forms of beliefs, a hostile coalition is a prerequisite of any conspiracy theory (van Prooijen & van Lange, 2014).

One can find many lay theories that fit the key ingredients of a conspiracy theory (patterns, agency, coalitions, threats, secrecy). They usually involve powerful groups such as societal leaders, governmental institutions (e.g., secret services), influential branches of industry (e.g., oil companies, the pharmaceutical industry), or stigmatized minority groups (e.g., Muslims, Jews). Besides the context of citizens' perception of geopolitical events, conspiracy theories emerge frequently in the microlevel setting of organizations, as employees often suspect their managers of conspiring toward evil goals such as pursuing their self-interest at the expense of employees and the organization (van Prooijen & de Vries, 2016)

Furthermore, although the term conspiracy theory may sometimes be used to invalidate legitimate accusations of corruption

2.4 Psychology of conspiracy theories

Conspiracy theories explain events as the result of secret, deliberate actions and cover ups at the hands of malicious and powerful groups (Goertzel, 1994; McCauley & Jacques, 1979; Sunstein & Vermeule, 2009).

Psychologists have also begun to consider what some of the potential consequences of conspiracy theories might be. In particular, whilst conspiracy theories may allow individuals to question social hierarchies and require elites be more transparent (e.g., Clarke, 2002; Fenster, 1999; Swami & Coles, 2010), recent empirical findings suggest that they may have important negative societal consequences. It is therefore becoming clear that conspiracy theories cannot be dismissed as trivial notions that affect the lives of only a small handful of individuals and marginalized communities

2.5 Organizational conspiracy theories

We define organizational conspiracy theories as notions that powerful groups (e.g., managers) within the workplace are acting in secret to achieve some kind of malevolent objective. For example, managers may deliberately conspire to hire a preferred candidate for a job, or work together to have an employee fired. We differentiate organizational conspiracy theories from various associated concepts. Specifically, rumour and gossip more often implicate individuals than groups and do not necessarily allege conspiracies between individuals (Allport & Postman, 1947; DiFonzo & Bordia, 2007). General mistrust, whilst correlated with conspiracy belief about societal events (e.g., Goertzel, 1994) refers to broader negative feelings about individuals or groups rather than specific allegations of dishonesty and wrongdoing by groups. Like general conspiracy beliefs, organizational conspiracy beliefs may also thrive under conditions of powerlessness (Whitson & Galinsky, 2008) and uncertainty (Van Prooijen & Jostmann, 2013). Specifically, in situations where workers lack control (e.g., have little responsibility, or little control over their duties) or under conditions of uncertainty (e.g., new management, concern about the motives of managers), organizational conspiracy theories may prosper.

What effects are such organizational conspiracy theories likely to have on the workplace? Our investigation focuses on one of the most important outcomes for organizations – employee turnover. Turnover represents a significant challenge to organizations and can be very costly, resulting in financial losses associated with training employees who opt to leave, associated recruitment and other administrative costs (Cascio, 2006; Shaw, Gupta, & Delery, 2005; Weisberg & Kirschenbaum, 1991) as well as the potentially disastrous loss of valuable individuals. Turnover can also affect organizational performance outcomes, including customer service, profits, and revenues (Hancock, Allen, Bosco, McDaniel, & Pierce, 2013). Specifically, workers may be more likely to consider leaving their organization to the extent that they view it as a negative place where groups act secretly and maliciously in the pursuit of their own selfish interests. Indeed, some recent research shows support for this idea, demonstrating that belief in workplace-related conspiracies – as a result of despotic or laissez-faire leadership – is associated with turnover intentions (Van Prooijen & de Vries, 2016). As yet, however, much is unknown about the relationship between conspiracy theorizing in organizations and workers' intentions to leave their workplace.

2.5.1 Organizational identification

Organizational identification refers to individuals' self-definition as members of a particular organization (Ashforth & Mael, 1989; Mael & Ashforth, 1995). Organizational identificationhasbeenfoundtouniquelypredictorganizationaloutcomes(e.g.,Abrams& RandsleydeMoura,2001;Haslam,Postmes,&Ellemers,2003;Ouwerkerk,Ellemers,&De Gilder, 1999) and attitudes and behaviors at work (Lee, Park, & Koo, 2015; Riketta, 2005). For example, it has been associated with workers' well-being (Van Dick, 2004; Wegge, Van Dick, Fisher, Wecking, & Moltzen, 2006), performance (Lee et al., 2015; Van Dick, 2004), and, most relevant to the current investigation, turnover intentions (Mael & Ashforth, 1995; Van Dick, 2004).

We argue that organizational conspiracy theories will decrease organizational identification. If an organization is riddled with perceptions of conspiracy, such as beliefs that managers are deliberately trying to harm employees, this is likely to weaken the importance of the organization to the individual and reduce the positive self-esteem they derive from it. Research has shown that

self-esteem and identification are positively influenced by receiving procedural justice from a group (Tyler, Degoey and Smith, 1996)

2.5.2 Organizational commitment

Similarly, organizational commitment can be understood as the psychological link people have to the organization. Research suggests that commitment is also a significant predictor of turnover intentions. Further, Van Prooijen and de Vries (2016) found that organizational commitment was associated with belief in workplace-related conspiracies and that it mediated the association between organizational conspiracy belief and turnover intentions. It remains to be seen whether organizational conspiracy belief exerts a causal influence on organizational commitment. Specifically, Jolley and Douglas (2014a) showed that conspiracy theories concerning the government weakened political engagement by generating feelings of powerlessness and uncertainty.

Like political engagement, organizational commitment entails willingness to act on behalf of the interests of the collective.

2.5.3 Job satisfaction

Job satisfaction is the evaluation that employees make of their job and includes their attitudes to specific aspects of the job (VanDicket al., 2004).Research has found that, like organizational identification and organizational commitment, job satisfaction is associated with turnover intentions, and that more satisfied workers are less likely to want to leave their jobs (Eby et al., 1999; Hom & Kinicki, 2001; Randsley de Moura et al., 2009; Schwepker, 2001; Van Dick et al., 2004).

Like organizational identification and organizational commitment, we argue that organizational conspiracy theories will decrease job satisfaction. A perceived climate of conspiracy or specific

beliefs that groups are conspiring against employees are likely to lead to disappointment and dissatisfaction in the job, or at least with specific aspects of the job.

2.5.4 Implications

More generally, our findings suggest that managers and employees may need to be mindful of the effects that conspiracy theories could have on the workplace. The current research suggests that, like broader societal conspiracy theories (Jolley & Douglas, 2014a,b), organizational conspiracy theories may have clear and detrimental consequences for employees and the organization as a whole. Management especially should be mindful of workplace conspiracy theories that may not only damage their reputation, but force them to lose valuable employees, or even keep disengaged employees on their team.

It is plausible that interventions that would focus on improving specific aspects of the organizational climate could strengthen employees' commitment to the organization, their satisfaction with their job, and, consequently, decrease intentions to leave the organization.

2.6 Machine Learning

Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision. Computer Vision and Machine Learning are two core branches of Computer Science that can function, and power very sophisticated systems that rely on CV and ML algorithms exclusively but when we combine the

two, we can achieve even more. Machine learning tasks are typically classified into two categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are –

2.6.1 Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. The training data consists of a set of examples which is called training examples. In order to solve a given problem using supervised learning there are some steps:

- Deciding what kind of data will be in the training set.

- Gathering a set of input object and related output object in training set.

- Determining input feature representation of the learned function.

- Determining the structure of the learned function and corresponding learning algorithm.

- Running the learning algorithm in previously defined training set.

- Evaluating the accuracy.

2.6.2 Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end. For example cluster analysis is an unsupervised learning method. This explores data analysis and find hidden patterns or grouping in data.

We did out project using supervised learning.

2.7 Text Classifier –The basic building blocks

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email.

Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

There are two types of learners in classification as lazy learners and eager learners.

1. Lazy learners

Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.

Ex. k-nearest neighbor, Case-based reasoning

2. Eager learners

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.

Ex. Decision Tree, Naive Bayes, Artificial Neural Networks [15].

2.8 Sentiment Analysis

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.

Sentiment analysis is a text mining challenge which deals with determining the opinion expressed by the author of a text. This task has become pressing in recent years, as the amount of opinionated texts online, such as blogs, editorials, and reviews, has skyrocketed. Being able to quickly and accurately measure people's opinions from the material the write and post online would allow governments and companies to better tailor and adapt their policies and products. Most research on sentiment analysis has been carried out on subjective texts such as blogs and product reviews. Authors of such text typically express their opinion freely, using clearly positive or negative language. The situation is different when dealing with news articles: in order to maintain an appearance of objectivity, journalists will often refrain from using clearly opinionated vocabulary [13]. The writing style is also different than in opinion pieces, and uses more complex sentences and specialized vocabulary.

Sentiment analysis is the automated process of understanding an opinion about a given subject from written or spoken language. In a world where we generate 2.5 quintillion bytes of data every day, sentiment analysis has become a key tool for making sense of that data. This has allowed companies to get key insights and automate all kind of processes.

Sentiment Analysis also known as *Opinion Mining* is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

1. *Polarity*: if the speaker express a *positive* or *negative* opinion,

2. *Subject*: the thing that is being talked about,

3. *Opinion holder*: the person, or entity that expresses the opinion.

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is

constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media.

With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service [14].

2.9 Dataset

The term data set refers to a file that contains one or more records. The record is the basic unit of information used by a program running on z/OS.

Any named group of records is called a data set. Data sets can hold information such as medical records or insurance records, to be used by a program running on the system. Data sets are also used to store information needed by applications or the operating system itself, such as source programs, macro libraries, or system variables or parameters. For data sets that contain readable text, you can print them or display them on a console (many data sets contain load modules or other binary data that is not really printable). Data sets can be catalogued, which permits the data set to be referred to by name without specifying where it is stored.

In simplest terms, a record is a fixed number of bytes containing data. Often, a record collects related information that is treated as a unit, such as one item in a database or personnel data about one member of a department. The term field refers to a specific portion of a record used for a particular category of data, such as an employee's name or department.

The records in a data set can be organized in various ways, depending on how we plan to access the information. If you write an application program that processes things like personnel data, for example, your program can define a record format for each person's data [16].

2.10 Features

Features are the variables found in the given problem set that can strongly/sufficiently help us build an accurate predictive model.

Eg : To predict the sale price of a house, size of the house is a feature.

1. Features are a column of data given as the input. They are also called as attributes or might sometimes be referred as dimensions.

2. A particular problem data set can have several features tagging to them. It is important to select the features that are more relevant to our problem so that the accuracy of the model improves. It also reduces the complexity of the model as we avoid the least significant / unnecessary feature data. The simpler model is simpler to understand and explain.

3. This Process is called feature engineering / selection and is one of the crucial step of pre-processing. *Different algorithms* can be used to implement it.

4. The Features can be of different types.

5. Simple Supervised selection where they are simple values like numbers and characters.

6. Eg: Size of the house (number).

7. In unsupervised learning, the model is itself trained to recognize the features and work on it.

8. Eg: In character recognition, features may include histograms counting the number of black pixels along horizontal and vertical directions, number of internal holes, stroke detection and many others.

Eg. : Loan Granting Problem

Let us build a model that tells us if to give loan to a particular customer or not.

Now its data will have many features/attributes attached to it:

Loan id, Cust. Name, Cust. id, Cust.Addr, Employed (or) not, Age, Marital Status, Has already avail loan, Annual Income, no.of open accounts, tax liens, credit score, current balance and so on.

Using the feature selection it can be observed that for a particular Customer,

Employed (or) not, age, current credit score, annual income, already availed loan can significantly explain / contribute to the model accuracy better than the others.

Thus they become the features for building our model for this particular problem.

In Artificial Intelligence, features are observable or derived properties of instances in a model's domain. For instance, for an intelligent agent which learns to classify types of fish, features could be size, colour, shape, scale patterns, etc. of fish. The model then learns the correlation between these features and the class of fish, in time becoming able to determine which class of fish a given instance is only by looking at these features.

At last we can say that features are those properties of a problem based on which you would like to predict results.

2.11 Data Preprocessing

Data Scientists across the word have endeavored to give meaning to Data preprocessing. However, simply put, data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

How is this done? Just like medical professionals getting a patient prepped for surgery so is data preprocessing, it prepares raw data for further processing. Below are the steps to be taken in data preprocessing

1. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

2. Data integration: using multiple databases, data cubes, or files.

3. Data transformation: normalization and aggregation.

4. Data reduction: reducing the volume but producing the same or similar analytical results.

5. Data discretization: part of data reduction, replacing numerical attributes with nominal ones. [17]

2.12 Python

Dating from 1991, the Python programming language was considered a gap-filler, a way to write scripts that "automate the boring stuff" (as one popular book on learning Python put it) or to rapidly prototype applications that will be implemented in other languages. However, over the past few years, Python has emerged as a first-class citizen in modern software development, infrastructure management, and data analysis. It is no longer a back-room utility language, but a major force in web application creation and systems management, and a key driver of the explosion in big data analytics and machine intelligence [18].

2.12.1 Advantages

Python is easy to learn and use: The number of features in the language itself is modest, requiring relatively little investment of time or effort to produce your first programs. The Python syntax is designed to be readable and straightforward. This simplicity makes Python an ideal teaching language, and it lets newcomers pick it up quickly. As a result, developers spend more time thinking about the problem they're trying to solve and less time thinking about language complexities or deciphering code left by others [18].

Python is broadly adopted and supported: Python is both popular and widely used, as the high rankings in surveys like the Tiobe Index and the large number of GitHub projects using Python attest. Python runs on every major operating system and platform, and most minor ones

too. Many major libraries and API-powered services have Python bindings or wrappers, letting Python interface freely with those services or directly use those libraries. Python may not be the *fastest* language, but what it lacks in speed, it makes up for in versatility [18].

2.12.2 Uses [18]

Python is used in many ways:

1) General application programming with Python

2) Data science and machine learning with Python

3) Web services and RESTful APIs in Python

4) Metaprogramming and code generation in Python

2.12.3 Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, we have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users [19]

2.12.4 Numpy [20]

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

　　1) A powerful N-dimensional array object

2) Sophisticated (broadcasting) functions

3) Tools for integrating C/C++ and Fortran code

4) Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases

2.12.5 Scikit-learn [21]

Scikit-learn is a library function that is used in python for machine learning.

1) Simple and efficient tools for data mining and data analysis

2) Accessible to everybody and reusable in various contexts

3) Built on NumPy, SciPy and matplotlib

4) Open source, commercially usable – BSD license

2.12.5.1 History [22]

Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.

Later Matthieu Brucher joined the project and started to use it as apart of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010.

The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation.

2.12.6 Pandas

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. *Pandas* is a NumFOCUS sponsored project. This will help ensure the success of development of pandas as a world-class open-source project, and makes it possible to donate to the project [23].

2.12.6.1 Usage of Pandas

Python has long been great for data munging and preparation, but less so for data analysis and modeling. Pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R. Combined with the excellent IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate. *Pandas* does not implement significant modeling functionality outside of linear and panel regression; for this, look to statsmodels and scikit-learn. More work is still needed to make Python a first class statistical modelling environment, but we are well on our way toward that goal [23].

2.12.7 PyMySQL

This package contains a pure-Python MySQL client library, based on PEP 249.

Most public APIs are compatible with mysqlclient and MySQLdb.

NOTE: PyMySQL doesn't support low level APIs _mysql provides like data_seek, store_result, and use_result. You should use high level APIs defined in PEP 249. But some APIs like auto commit and ping are supported because PEP 249 doesn't cover their use case.[24]

Requirements

- Python – one of the following:

  - CPython : 2.7 and >= 3.4

  - PyPy : Latest version

- MySQL Server – one of the following:

  - MySQL >= 5.5

  - MariaDB >= 5.5

2.12.8 wordCloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

For generating word cloud in Python, modules needed are – matplotlib, pandas and wordcloud. To install these packages, run the following commands:

Pip install matplotlib

Pip install pandas

Pip install wordcloud

2.12.9 BeautifulSoup

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

2.12.10 CountVectorizer and LogisticRegression

CountVectorizer builds a count matrix where rows are occurrences counts of different words taking into account the high-dimensional sparsity.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

2.12.11 TF-IDF

TF-IDF, which stands for term frequency—inverse document frequency, is a scoring measure widely used in information retrieval (IR) or summarization. TF-IDF is intended to reflect how relevant a term is in a given document.

The intuition behind it is that if a word occurs *multiple times in a document*, we should boost its relevance as it should be more meaningful than other words that appear fewer times (TF). At the same time, if a word occurs many times in a document but also *along many other documents*, maybe it is because this word is just a frequent word; not because it was relevant or meaningful (IDF).

We can come up with a more or less subjective definition driven by our intuition: a word's relevance is proportional to the amount of information that it gives about its context (a sentence, a document or a full dataset). That is, the most relevant words are those that would help us, as humans, to better understand a whole document without reading it all.

As pointed out, relevant words are not necessarily the most frequent words since stopwords like "the", "of" or "a" tend to occur very often in many documents.

There is another caveat: if we want to summarize a document compared to a whole dataset about a specific topic (let's say, movie reviews), there will be words (other than stopwords, like *character* or *plot*), that could occur many times in the document as well as in many other documents. These words are not useful to summarize a document because they convey little discriminating power; they say very little about what the document contains compared to the other documents [25].

2.13 Private Mail Server

A private server is a physical computer that you own and operate, and has all the operating systems, software and programs in place to provide essential services, including email. In a textbook definition *A private server is a machine or virtual machine that is privately administrated. As servers need an adequate Internet connection, power and can be noisy, they are often located in a colocation center.*

A private email server would be the email system that's offered by the private server.

In other words, with a private email server one can have his own email system, from computers to programs. One can run it, use it, manage it and limit (allow and prevent) access to it.