NLP analysis of Incident and Problem Descriptions

Attila Soti (soti.attila@freemail.hu)
Szechenyi István University Győr

Zoltan Dobos (zoltan.dobos@t-online.hu)

Abstract

In this article, incident, problem ticket RCA (root case analysis) are analyzed with Natural Language Processing (NLP). Considering that incident and problem description are mainly described in an unstructured-way, there is a need to apply text mining techniques to gain insights or detect focusing area. This case is more complex if incident or problem tickets related to large organization having wide-variety of services. This article shows how NLP can help to find patterns, area to focus by applying semantic analysis.

Keywords: Statistical Natural Language Processing, Incident and Problem descriptions, RCA

Incident and problem management

Incident and problem management is very important activity in IT services of large companies. The goal of the incident management process is to restore a normal service operation as quickly as possible and to minimize the impact on business operations, thus ensuring that the best possible levels of service quality and availability are maintained. When multiple occurrences of related incidents are observed, a problem record should be created. The management of a problem differs from the process of managing an incident and is typically performed by different staff and controlled by the problem management process. Root cause analysis (RCA) is part of problem resolution. - see figure 1 (Philip L Yuson 2017).

It is widely estimated that 80 percent of all business-relevant information resides in unstructured and semi-structured text data. In other words, without the text analyses to discover the wealth of data represented in that 80 percent, all the business information and behaviour data goes to waste. If incident and problem ticket descriptions could be analysed it would be very useful.

They are usually written by non-professionals and many times they are not formulated in adequate English, but they contain a lot of information. Using this information in a thorough and systematic way is increasingly necessary to understand customer behaviour and attitudes. For example, if the cause of most incidents were known, measures could be taken to decrease the occurrences. There are no standard rules for writing text so that a computer can understand it. The language and meaning for every piece of text vary depending on the purpose. The only way to accurately

include unstructured data in a data-mining project is to understand the language and the context within which the text was created. (Mark Johnson 2016)

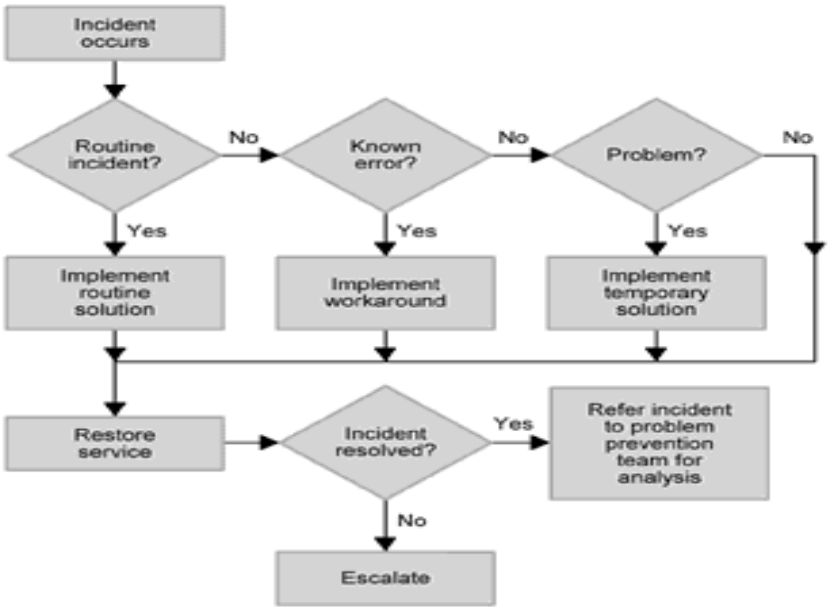

Figure 1 – Process of Incident and Problem management

NLP definition (Natural Language Processing)

Understanding human language is based on linguistics, commonly referred to as Natural Language Processing (NLP). NLP is a way for computers to analyse, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, and topic segmentation. NLP algorithms are typically based on machine learning algorithms. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analysing a set of examples. A system that incorporates NLP can intelligently extract terms, including compound phrases, and permit classification of terms into related groups. (Martinez L, Ruan D, Herrera F -2010). Linguistic systems are knowledge-sensitive: the more information contained in the linguistic resources (dictionaries), the higher the quality of results. Modification of the dictionary content, such as synonym definitions, can simplify the resulting information and focus attention on the most relevant concepts. A Statistical NLP approach seeks to solve problems by automatically learning lexical and structural preferences from corpora and if the existing dictionary is extended (rules, types, synonyms) to support this process we might be able to analyse our non-formal descriptions. (See Figures below 1, 2) Text context is crucial. (Matt Kiser 2016)

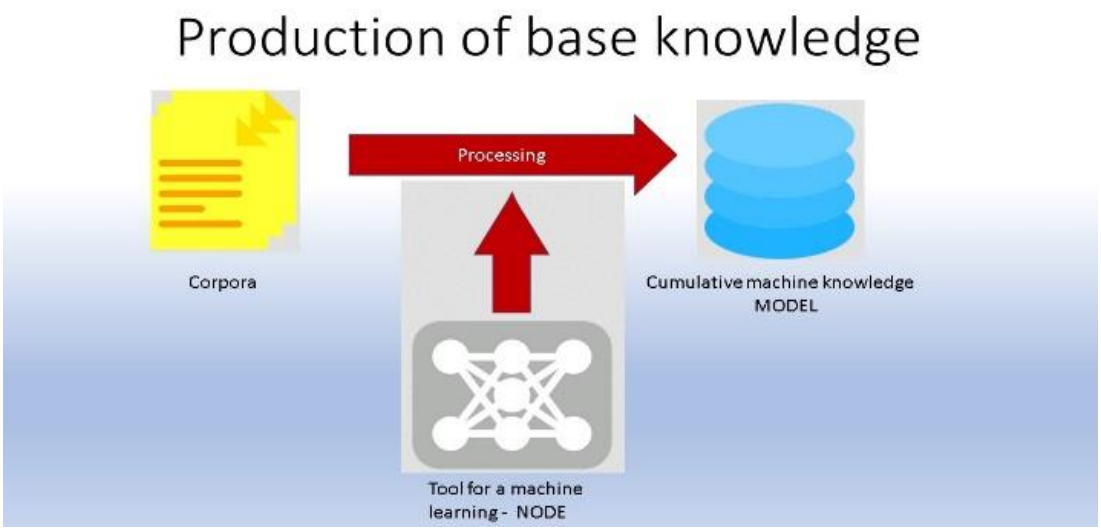

Figure 2 – Statistical NLP approach phase I

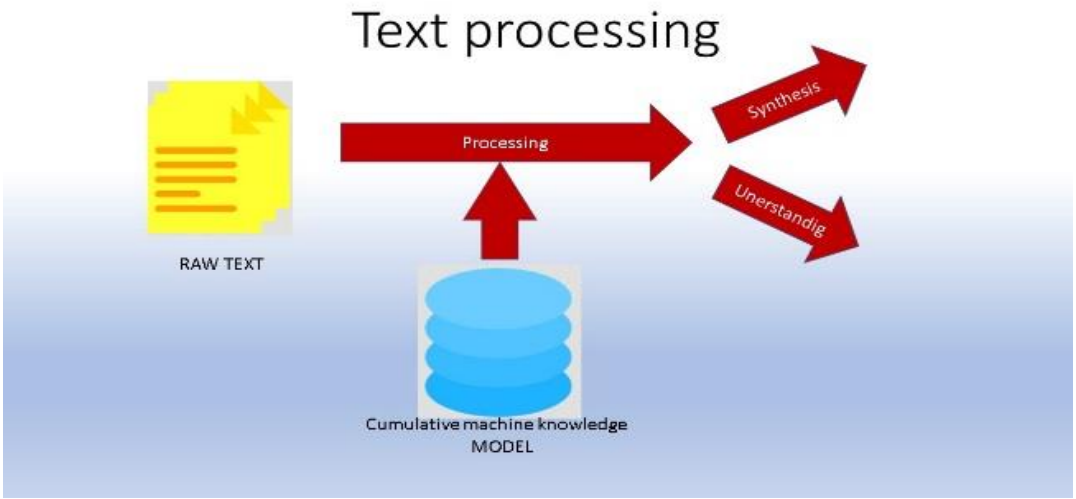

Figure 3 – Statistical NLP approach phase II

Text mining must consider the universal fact that languages contain ambiguities. The same words can be different parts of speech (nouns, pronouns, verbs, adjectives, adverbs, etc.), and therefore play different roles in meaning. (Matt Kiser 2016) The same word, even when used as the same part of speech, can have different meanings depending on how it is used and the context within which it appears. Linguistic analysis involves the study of the elements, structure, and meaning of language: (Fraser, N. M. and Hudson, R. A. 1992).

Standard methodology

Text mining is the process of extracting knowledge and information from natural

language texts. Text mining proceeds in two stages.

Stage 1: Key concepts/terms are extracted from the text that represents the essence of information the text contains.
Stage 2: These concepts/terms are grouped into categories that represent the higher-level ideas contained in the text.

Semantic Analysis

As an RCA is focusing on identifying root causes -, while incident description is simple describing the circumstance or symptom of an issue - to analyze the problem description can provide insight about problematic area where the enterprise needs to focus. Also, an RCA focuses on the problem, which is typically not a one-time incident, but repetitive issues, which means being able to identify problematic area on enterprise level can bring larger benefit. There are several ways how NLP can help to gather insights out of any text description.

Segmentation of text description (by creating clusters) can help to identify problem area, where we can detect which parts of the service have the most impact on operation.

Another way is to analyze the problem description is based on semantic analysis (stage-1.).

The study of the meaning of words, phrases, sentences, and texts. The best examples are synonyms and homonyms. For example, to ring (call or phone) versus to ring (some noise in my ear) versus ring (on my finger) or ring (sport ground for boxing). Semantic analysis is the most difficult task for text mining, and involves in part the use of dictionaries, the sauries, glossaries, lexicons, typologies, and so forth.

Each part of speech explains not what the word is, but how the word is used. Traditional English grammar classifies words based on eight parts of speech: verb, noun, pronoun, adjective, adverb, preposition, conjunction, and interjection. However, part-of-speech (PoS) tagging in Text Analytics uses the following tags:

I. N: **Noun**

A word used to name a person, place, thing, quality, or action and can function as the subject or object of a verb.

II. V: **Verb**

A word that expresses existence, action, or occurrence such as be, sell, and happen.

III. A: **Adjective**

A word used to modify a noun by limiting, qualifying, or specifying it.

IV. B: **Adverb**

A word that modifies a verb, an adjective, or another adverb.

V. O: **Coordination**

A conjunction such as "and" and "or".

VI. D: **Determiner**

A noun modifier including articles, demonstratives, possessive adjectives, and words such as any, both, or whose.

VII. G: **Gerund**

A noun derived from a verb. In English ending in the suffix "ing", as smoking is harmful

VIII. P: **Participle**

A verb used as an adjective, most often ending in "ing" (present) or "ed" (past), as in returning home.

IX. C: **Preposition**

A word placed before a noun that indicates the relation of that noun to a verb, an adjective, or another noun. For example, the word "of" is a preposition. Most prepositions are tagged as S or Stop words.

X. X: **Auxiliary**

A verb such as is, have, can, could, or will that usually accompany a main verb in a clause.

XI. S: **Stop word**

A very large category of words used to exclude from extraction. It contains all pronouns, particles, and prepositions (except "of")

Semantic analysis has several advantages, it can detect connection among different tags, there are several combinations, which have powerful meaning like verb-noun connections.

In case of noun, we can increase the significance if we identify not a single term, but a sequence. A typical noun sequence is a bi-gram, but that can be any number of nouns (n-grams). Advantage of the n-grams, that comparing with daily-speech environment the occurrence of n-grams is relatively low, while in technical environments these occurrences are high (typically identify technical terms or expressions) and have higher business meaning value, what we can use for RCA. One example, how bi-grams helps to identify areas, what we can see here that several RCA refers to hardware problems or issues related to filesystems or problems describing cases related to firmware.

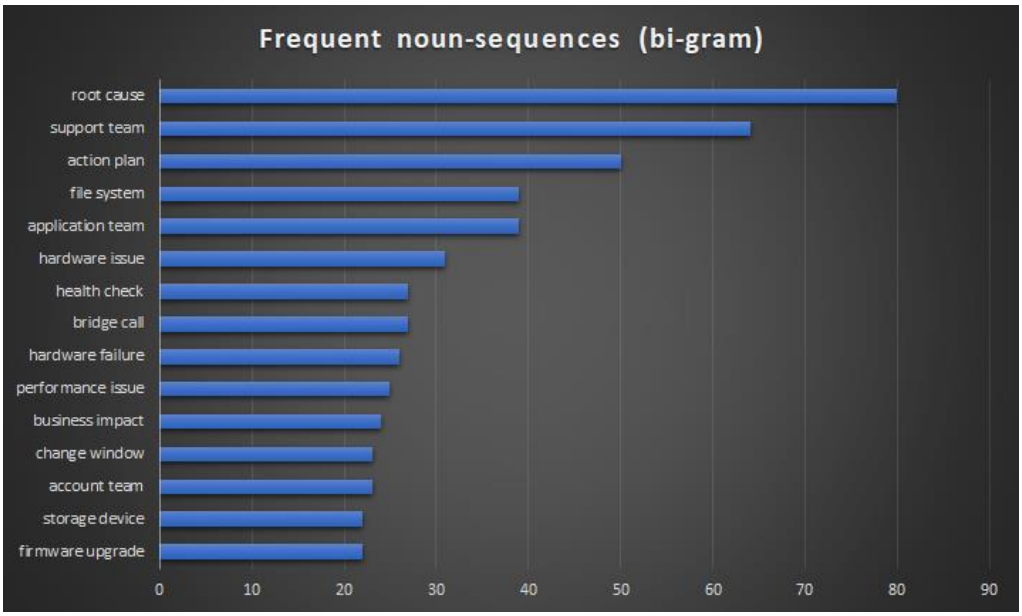

Figure 4 – Frequent noun-sequence

As problems are not occurring at the same time, enriching our RCA data with timeline, can easily help to identify trends for each of frequent grammar tags.

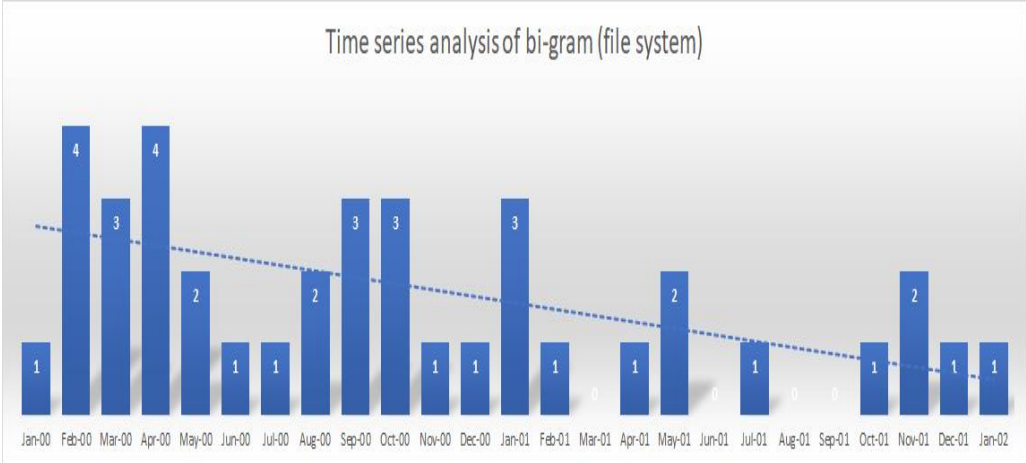

Figure 5 – Time series analysis of bi-gram (file system)

Same timeseries can be analyzed from deviations points of view, which helps to identify those points in the time, which behaves differently in time that what we would predict from previous data points.

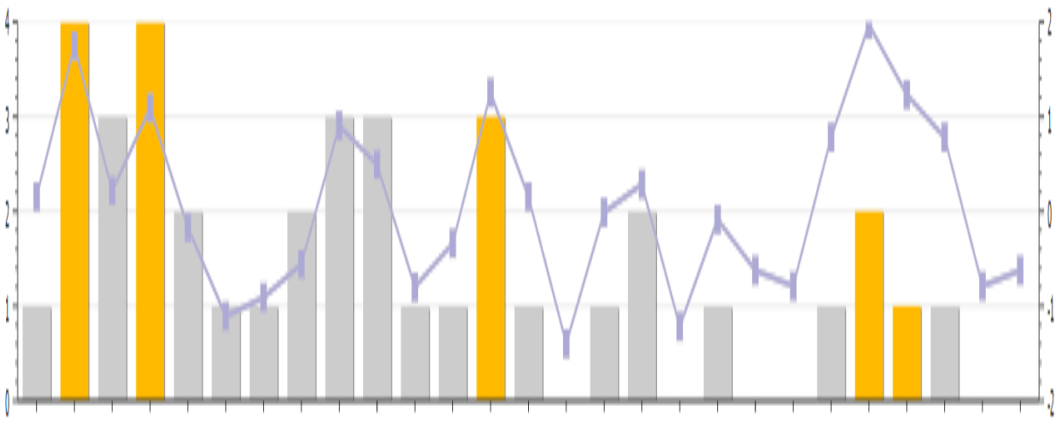

Figure 6 – Deviation detection

Semantic analysis can help to detect meaningful connection among different grammatical part of the RCA description.

Let's see two examples for connection by detecting correlation between noun sequences (bi-gram) and modified nouns, we gain more insight by seeing that a larger part of filesystem issues are connecting to Unix team, so we can identify that this problem occurs mainly around Unix environment.

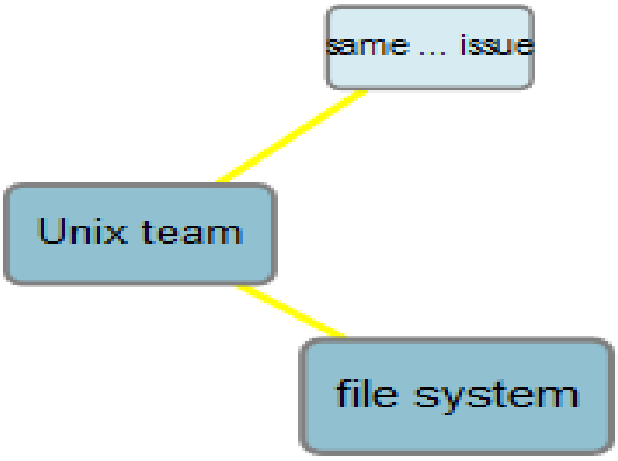

Figure 7 - Connection network sample I

Another example, shows that there is correlation between hardware failure and firmware upgrade, which can help us to identify common issues with firmware level:

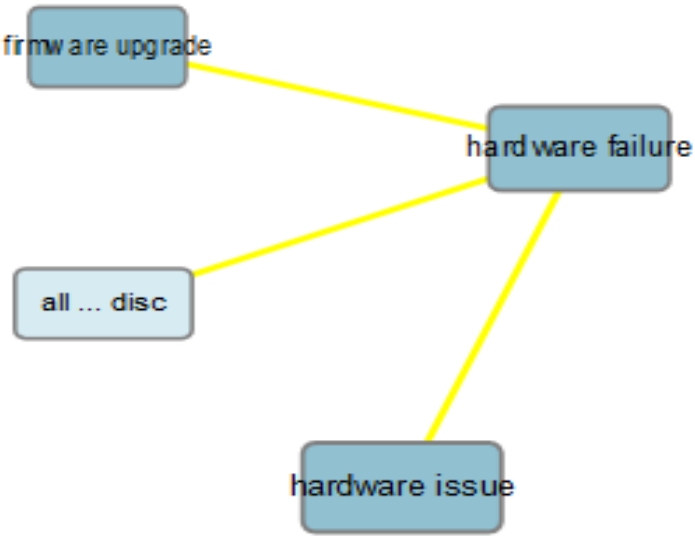

Figure 8 - Connection network sample II

Results

NLP is an efficient way to discover patterns on a large set of problem and root cause analysis records, simply identifying grammatical structures supports to detect trends and deviations related to the provided service or the applied service components. There are grammatical structures, which are more valuable for this purpose, basically nouns and verbs are the key language components from this point of view.

Beyond individual words we can identify structures such as noun sequences, modified nouns or nouns with predicates, which helps us identify more relevant text structures describing our problem space. Example, not real data see figure below

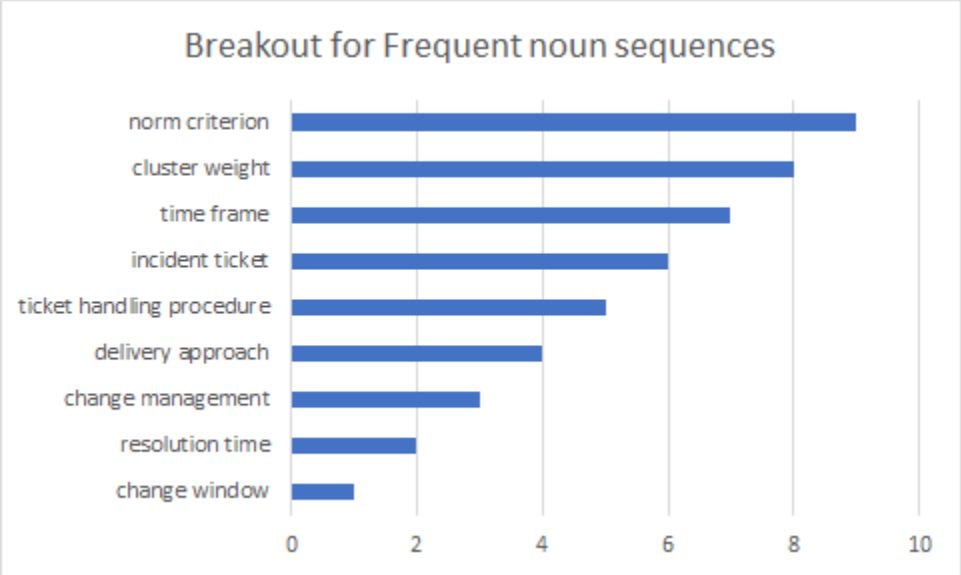

Figure 9 -

Frequency of those terms and text structures can reflect the severity of the problematic area.

Conclusion

A problem and RCA record individually describes the cause of a problem, but contains more information in a complex enterprise environment. As RCA is highly unstructured data we need to apply techniques to extend information about it. By applying NLP procedures, we can identify trends and deviation focusing on business relevant expressions and terms.

References

[1] Briscoe, T. and Carroll, J. (1993). Generalized Probabilistic LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics,19(1), 25–59.
[2] Chen, S. F., Seymore, K., and Rosenfeld, R. (1998). Topic adaptation for language modelling using unnormalized exponential models. In IEEE ICASSP-98, pp. 681–684. IEEE.
[3] Christopher D. Manning and Hinrich Schütze, 1999, Foundations of Statistical Natural Lang-uage Processing,MITPress, Cambridge, MA. http://www.aclweb.org/archive/misc/what.html
[4] D. Jurafsky, J. H. Martin, Speech and Language Processing, 2000
http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf
http://progmat.hu/tananyagok/alkalmazott_mesterseges_intelligencia/book.html#d5e1408 05.02.2017
[5] Fraser, N. M. and Hudson, R. A. (1992). Inheritance in word grammar. Computational Linguistics,18(2), 133–158
[6] Matt Kiser:: Introduction to Natural Language Processing (NLP) 2016
http://blog.algorithmia.com/introduction-natural-language-processing-nlp/
[7] Mark Johnson: Introduction to Computational Linguistics and Natural Language Processing
http://web.science.mq.edu.au/~mjohnson/papers/Johnson14MLSS-talk.pdf 05.02.2017
[8] Marko Bohanec's Data. 1997. Available at http://archive.ics.uci.edu/ml/machine-learning- databases/car/car. Data (access date: 06/21/2016).

[9] Philip L Yuson: Incident and Problem: What is the Difference (2010) http://www.conceptsolutionsbc.com/it-service-management-mainmenu-60/30-it-service-management/182-incident-and-problems-what-is-the-difference 2017-05.