

DIGIT.B4 – Big Data PoC

DIGIT 01 – Social media topics

D03.02.Dictionaryes

everis Spain S.L.U

Table of contents

1 Introduction	3
1.1 Context of the project	4
1.2 Objective	4
2 Dictionaries.....	5
2.1 Unification of terms	5
2.2 Meaningless terms.....	6
2.3 Relevant terms.....	7
3 Examples of corpus	9
4 Conclusions.....	10
5 ANNEX	11
5.1 Dictionaries	11

Table of figures

Figure 1 - Transformation into 'electronic'	6
Figure 2 – Terms discarded	7
Figure 3 - Relevant terms	8
Figure 4 - Example of corpus transformed	9
Figure 5 - Ten of the most frequency terms	10

1 INTRODUCTION

1.1 Context of the project

This proof of concept shall demonstrate the use of text mining techniques on large amounts of social media posts as a means to identify areas of interest for the 2016 ICT conference.

1.2 Objective

The purpose of this document is to reflect the processes carried out during the data preparation under the CRISP-DM methodology. In this phase of the applied methodology, the data is cleaned and transformed in order to optimize the input for the segmentation algorithms.

2 DICTIONARIES

Creating a dictionary of terms is one of the most important activities in text-mining techniques. It is even more necessary when we have tweets and posts which use an informal (even incorrect) language. The purpose of the dictionary is to unify terms that belong to the same context in order to guide the segmentation and to get the hidden topics on posts.

There are three key points in this process:

- Unification of terms
- Remove meaningless terms
- Identification of relevant terms

More than 2.000 terms have been analysed to create a final dictionary with approximately 500 terms to make the segmentation.

The dictionary is created from the cleaned corpus created as an output of the 'data linguistic understanding' (without stopwords, symbols, numbers...)

2.1 Unification of terms

All terms around a topic are transformed with the unification of terms in the same context. These tasks also give more relevance to a global term that includes all those terms.

To guide the algorithm the global terms defined try to include the specifics terms in:

- Terms identified with the prefix e- are linked to 'electronic'
- Hashtags with two words together have been separated in two. For example: expressions like #word1Word2 are being changed into word1 word2
- Digital terms have also been separated in two words. Expressions like digitalword2 are being changed into digital word2.

These actions give the intelligence that a machine cannot have itself to perform the segmentation process.

The figure below shows an example of terms transformed into 'electronic stuff':

Terms	Translation
e-	electronic
edelivery	electronic delivery
eforms	electronic forms
egovernance	electronic governance
egov	electronic government
egovernment	electronic government
egoverment	electronic government
eid	electronic id
einvoicing	electronic invoicing
esig	electronic signature
esignature	electronic signature
electronicsignature	electronic signature
eskills	electronic skills

Figure 1 - Transformation into 'electronic'

2.2 Meaningless terms

All the terms without a specific meaning and which will not help to categorize the documents will be removed from the corpus in order to reduce the number of terms. It also will facilitate the convergence of the algorithm.

The figure below shows an example of terms discarded:

Term	Frequency
be	2217
1	1033
have	470
do	416
need	232
will	223
use	173
good	157
how	156
other	110
very	86
yammer	85
new	85
great	82
time	81
talk	78
look	74
take	73
year	69
interest	69
why	60

Figure 2 – Terms discarded

2.3 Relevant terms

Relevant terms referred to important topics or DIGIT's strategy are not transformed as they have a meaning themselves and also have enough frequency as an input to the algorithm.

The figure below shows an example of main terms without transformation:

Term	Frequency
digital	374
work	173
data	169
government	144
people	133
change	115
public	110
way	107
cloud	90
open	84
tool	81
process	79
know	78
project	67
share	63
email	60
electronic	53
next	52
create	51
internal	50
learn	50

Figure 3 - Relevant terms

3 EXAMPLES OF CORPUS

After applying all the rules mentioned before related to the creation of the dictionary, it is easier to identify the topic of the posts. These actions also allow the classification algorithm to find a better solution for the segmentation.

The figure below shows a few examples of corpus before and after the application of the dictionary.

Corpus	Corpus after dictionary
we need to build bridges within the @EU_Commission & a collaborative environment connecting DGs sharing tasks & focus on delivery #DIGITconf	build collaborative environment connect share focus delivery
It's all about leadership at the #DIGITconf	leadership
the best move of the UK gov was the move from egov to DIGITAL gov. Think DIGITAL. DIGITAL is the new default #DIGITconf	government electronic government digital government digital digital
Use DATA to understand vulnerabilities + your critical assets @colemasec #cybersecurity #DIGITconf	data understand critical cyber security
Good to hear @GOettingerEU confirms Internet of Things, Big Data and Cloud crucial to successfull EU digital transformation #DIGITconf #IoE	internet thing bigdata cloud digital transformation
It has also been my observation. She regularly publishes articles with pics of her. She seems to have very good communication skills. Which is a must for us.	communication skill
Err, Government? RT @StefDzhumalieva: Any suggestions for alternative word for e-Government? @luukasilves #DIGITconf	government suggestion alternative electronic government
@stephen_quest one of the highlights of #DIGITconf : "it's not about fixing IT, it's about transforming government"	fix information_technology transform government
Open data seems to be the mantra of the day till now. #DIGITconf	open_data

Figure 4 - Example of corpus transformed

4 CONCLUSIONS

A dictionary with more than 500 relevant and specific terms have been created out of 2000 individuals terms

The following bar graph shows ten of the most used terms in posts:

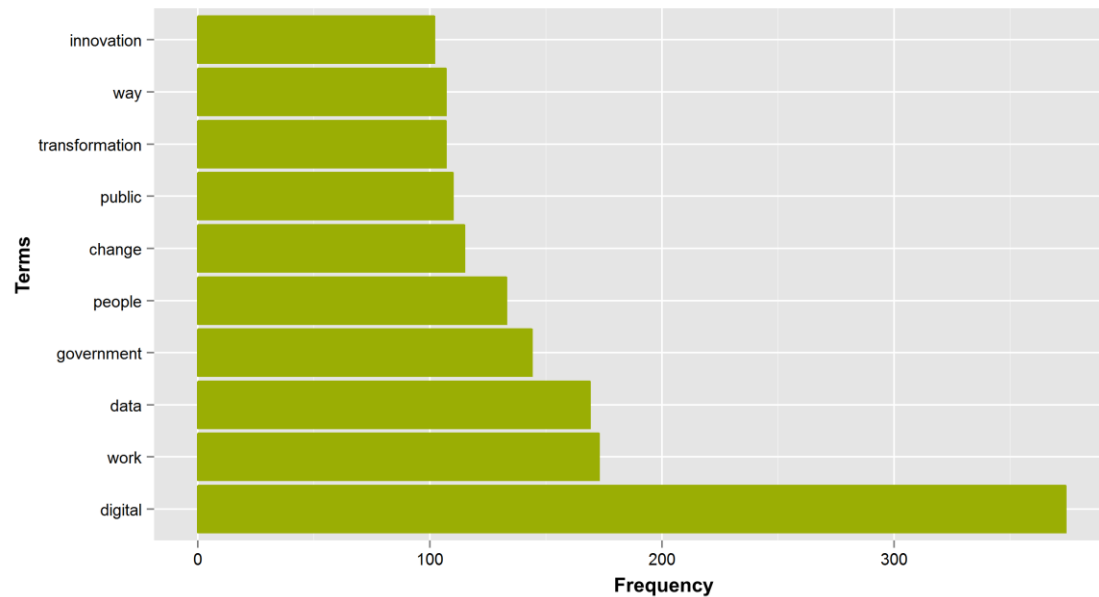


Figure 5 - Ten of the most frequency terms

The data preparation process execution allows to obtain the input for the optimized and guided segmentation in order to identify the suggested topics for the next ICT conference, which will be the ones that are mentioned the most by the audience.

5 ANNEX

5.1 Dictionaries

The following excel includes three sheets:

- Sheet 1: Dictionary with the keywords and its global term in the right column, called translation. It is the file with all the transformation used in '2.1 Unification of terms'.
- Sheet 2: All terms which have been removed from the original posts. It is the file used in '2.2 meaningless terms'.
- Sheet 3: A collection of terms, which are contained in the posts, and its frequency of occurrence.



D03.02.Dictionaries_
v.1.0.xlsx