

DIGIT.B4 – Big Data PoC

DIGIT 01 – Social media topics

D03.01.Data linguistic understanding

Table of contents

1 Introduction	4
1.1 Context of the project	4
1.2 Objective	4
2 Text-mining treatment.....	5
2.1 Basic transformations.....	5
2.2 Stop words.....	5
2.3 Meaningless terms.....	5
3 Completeness.....	7
4 Quality.....	9
5 Business standpoint.....	10
5.1 Topics	11
6 Conclusions.....	12

Table of figures

Figure 1 - Example of text-mining treatment	6
Figure 2 - Top 15 frequency terms	7
Figure 3 - Top 120 frequency terms	8
Figure 4 - Example of errors and corrections	9
Figure 5 - Wordcloud for terms with more than 10 in a tf-idf range	10
Figure 6 – Preliminary 4-topics classification	11

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 analyses and processes carried out during the data understanding under the CRISP-DM methodology. In this phase of the applied methodology, the data is explored and analysed in order to validate the quality of the information and ensure the viability of the project.

This phase is structured in:

- Text treatment: transform the text to an input for analysis and models.
- Completeness: the volume of documents and different words must be sufficient to allow a reliable analysis.
- Quality: the words used for the analysis must be grammatically correct.
- Business standpoint: the most frequent terms are analysed to validate quality from a business perspective.

2 TEXT-MINING TREATMENT

The data understanding phase will be carried out with 1.861 tweets.

Before starting with the analysis the text must be cleaned in order to:

- Reduce the number of terms
- Focus the analysis on the main words that give sense to the text
- Group terms to obtain more relevant and specific terms
- Optimize the input for clustering and classification algorithms

2.1 Basic transformations

The first transformations applied to the text are:

- Convert text to lowercase
- Remove punctuation symbols (!"#\$%&'()*+,-./:;<=>?@[\\]^_`{|}~)
- Remove numbers
- Remove extra white spaces

2.2 Stop words

Stop words are meaningless terms that do not give extra information so they are removed.

There is no single universal list of stop words. However, this list is usually made of prepositions, pronouns, articles, adverbs, conjunctions and some verbs.

The list of stop words used in the project (stop words package in R library TM) is: a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, cannot, can't, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, her, here, here's, hers, herself, he's, him, himself, his, how, how's, I, I'd, if, I'll, I'm, in, into, is, isn't, it, its, it's, itself, I've, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours, ourselves, out, over, own, same, shan't, she, she'd, she'll, she's, should, shouldn't, so, some, such, than, that, that's, the, their, theirs, them, themselves, then, there, there's, these, they, they'd, they'll, they're, they've, this, those, though, to, too, under, until, up, very, was, wasn't, we, we'd, we'll, were, we're, weren't, we've, what, what's, when, when's, where, where's, which, while, who, whom, who's, why, why's, with, won't, would, wouldn't, you, you'd, you'll, your, you're, yours, yourself, yourselves, you've

2.3 Meaningless terms

A selection of terms without meaning is used to create an open list where words are added as new analysis is made – this list will be increased along the project. All these terms will be also removed.

This list is made of:

- Short terms (one character)
- Prepositions, pronouns, articles, adverbs and conjunctions not included in the stop word list. Does not apply in this phase.
- Terms that appear in most of the documents and cannot be used to separate them in different categories (for example: #DIGITconf). This step should be applied in a latest phase.
- Verbs and nouns which do not have a specific meaning (for example: apply, use, compare,...). This step should be applied in a latest phase.

Below an example showing all the transformations applied to the texts.

Original tweet
"We have to tap into the power of the millennials - we need reverse mentoring in @EU_Commission" Director-General @stephen_quest #DIGITconf



**Text-mining
treatment**

Transformed tweet
"tap power millennials need reverse mentor eu commission director general stephen quest digitconf"

Figure 1 - Example of text-mining treatment

3 COMPLETENESS

One of the most important requirements for the statistical models is to be stable and consistent. To achieve this objective, the analysis and inputs for the algorithm must be performed with a sufficient number of terms.

Clustering and classification algorithms use significant text to create consistent rules and groups in order to categorize documents. If the available terms are not enough, the algorithms will not be able to create small and specific groups to categorize the tweets into topics.

There are 797 different terms in the sample of 1.861 documents with more than 5 occurrences - it makes more than 2.854 terms. This scenario is more than enough to ensure convergence and quality of the models.

There are 603 different terms with a significant distribution that will also be analysed and used as input for the statistical models.

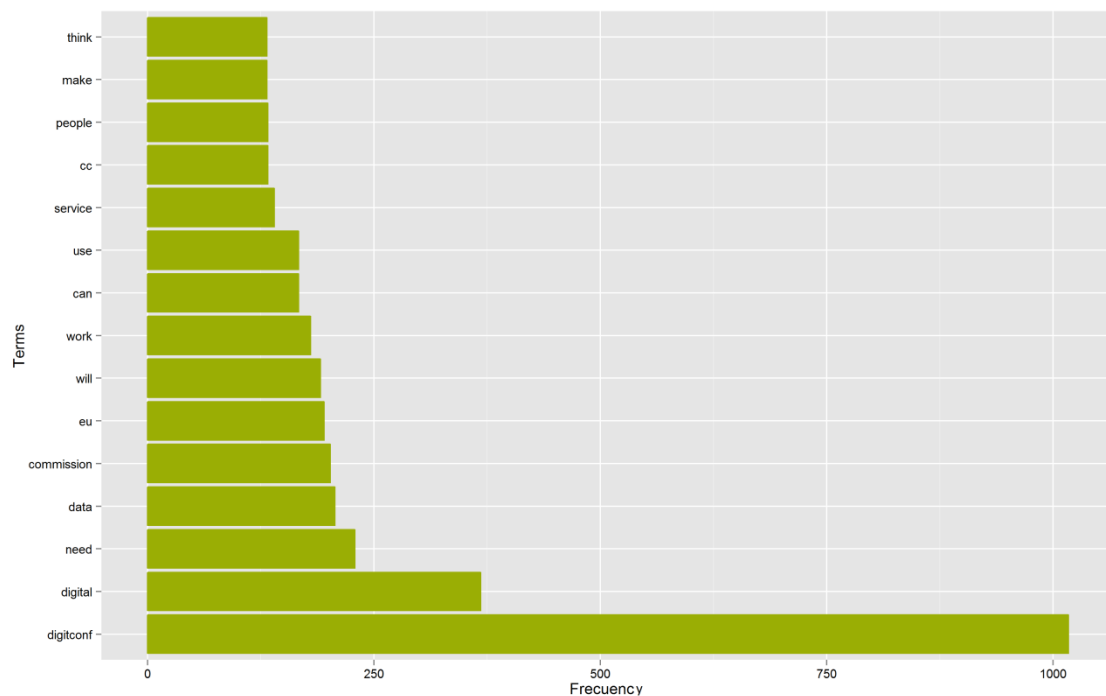


Figure 2 - Top 15 frequency terms

Term	Frequency	Term	Frequency	Term	Frequency
digitconf	1017	thank	99	interest	69
digital	368	transformation	98	year	69
need	229	also	97	year	69
data	207	innovation	96	big	67
commission	202	user	96	europa	66
eu	195	ec	95	much	63
will	191	now	94	project	63
work	180	stephen	93	share	63
work	180	yammer	90	digit	61
can	167	yammer	90	find	60
use	167	http	88	mean	60
service	140	don	87	organisation	60
cc	133	us	87	question	59
people	133	quest	85	want	59
make	132	goettingereu	84	day	58
think	132	new	84	real	58
well	124	process	83	email	57
just	123	great	82	learn	57
mtbracken	118	like	81	staff	57
good	117	information	78	really	56
change	114	know	78	social	56
time	111	cloud	77	citizen	55
one	110	government	76	come	55
way	106	talk	76	fabiozib	55
open	105	may	75	breton	54
get	104	look	74	policy	54
see	103	take	73	start	54
go	101	tool	71	twitter	54
public	100	idea	70	ask	52
say	99	thing	70	com	52

Figure 3 - Top 120 frequency terms

4 QUALITY

The terms used as input for analysis and algorithms must be orthographically correct.

More than 40% of 'wrong' terms have been considered by the corrector. In the context of the project (tweets with lots of symbols) most of them are not real errors (emojis, natural language, etc.) and does not represent a big problem.

Some of the errors will be corrected in the next phases to use as much information as possible. This will allow us to have the maximum number of terms in a document to classify it in the best way possible.

Below an example of errors founded on the abstracts and corrections made:

Original term	Edited Term
@stephen_quest	stephen quest
alexandr	alexander
conmplexity	complexity
buy/sell	buy sell
ltittle	little

Figure 4 - Example of errors and corrections

It is important to have a large amount of different terms related to the different categories in order to use all the power of the algorithms.

The best way to evaluate whether the information analysed is good enough from a business standpoint is to find terms related to relevant topics. In the analysis made plenty of these terms have been identified, not only in the most frequent terms but also in terms related to specific concepts.

We can find terms such as innovation, data, transformation or the speaker's names in the top most important terms (frequents and not uniformly distributed) as showed in this wordcloud.

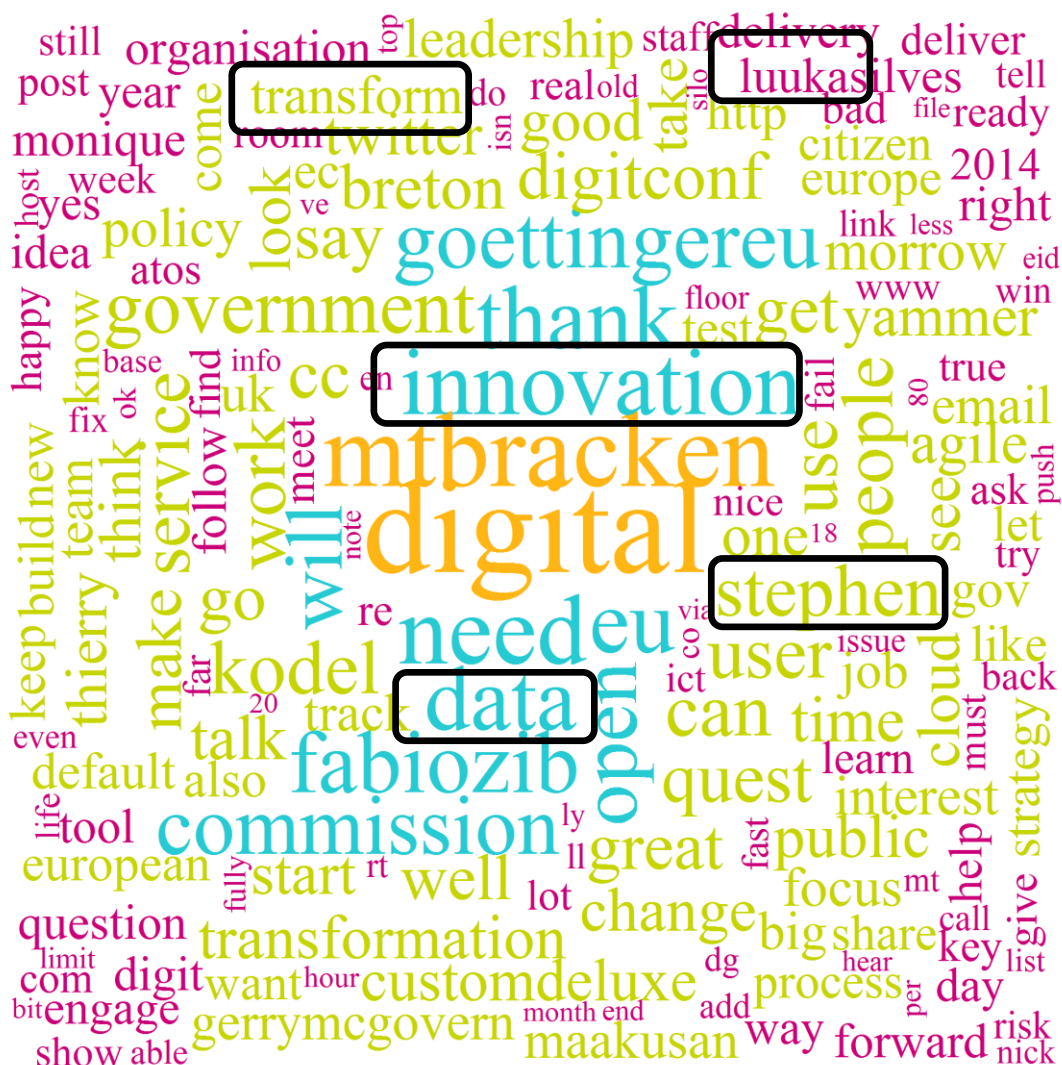


Figure 5 - Wordcloud for terms with more than 10 in a tf-idf range

5.1 Topics

The variety of terms is one of the keys to take advantage of all the capacity of algorithms. The greater the variability and specificity of the terms is, the easier will be to segment the documents into categories.

We have found lots of different topics that will allow to classify the documents in groups and make a consistent segmentation.

For example, if we make a preliminary analysis around only 4 topics, our algorithm classifies tweets into: **data**, **CC**, **EU** and **digital**.

	Topics			
	data	cc	eu	digital
N. Docs	156	153	271	1.266

Figure 6 – Preliminary 4-topics classification

As we observe, there are tweets that are unclassified due to the corrections that could eliminate complete tweets.

6 CONCLUSIONS

The documents extracted for the project are good enough to allow the development of clustering and topic models.

- Completeness: 1.861 tweets with more than 2.000 terms will give statistical support to the models, and it will make possible the topics extraction.
- Quality: although more than 40% of the terms had been marked as incorrect, most of them were symbols or expressions that could be treated to be useful.
- Business point of view: the most frequent terms are related to the conference's speakers, data science, computer science or DIGIT itself, so we can assume topics people prefer are going to be easily distinguishable.