DIGIT.B4 – Big Data PoC

RTD – Health papers

D03.01.Data linguistic understanding

everis Spain S.L.U

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

5.2 Annual stability 13

6 Conclusions 14

Table of figures

Figure 1 - Example of text-mining treatment 6
Figure 2 - Top 10 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 65.000 repetitions 10
Figure 6 - Terms around ‘neuro’ and ‘micro’ topics 11
Figure 7 - Example of health specific terms 12
Figure 8 - Annual term frequency (2012-2014) 13
Figure 9 - Wordclouds for years 2012, 2013 and 2014 13

1 INTRODUCTION

1.1 Context of the project

The objective of the proof of concept that showcases the use of big data in the EC research domain, in cooperation with DG RTD, is to prove the usefulness and policy benefit that big data can bring.

This proof of concept shall also demonstrate the use of text mining techniques on large amounts of unstructured research papers as a means to identify areas of interest, as an additional input to consider prior to launching calls for grants.

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 a subset of 500.000 documents of the original sample, because our experience shows us that this is enough to carry out the data understanding task.

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 or two characters)
- 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: patient). This step should be applied in a latest phase.
- Verbs and nouns which do not have a specific meaning (for example: apply, use, compare, result, human…). This step should be applied in a latest phase.

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

Original abstract
rna degradation plays a fundamental role in maintaining cellular homeostasis whether it occurs as a surveillance mechanism eliminating aberrant mrnas or during rna processing to generate mature transcripts. 3'-5' exoribonucleases are essential mediators of rna decay pathways, and one such evolutionarily conserved enzyme is polynucleotide phosphorylase (pnpase). the human homologue of this fascinating enzymatic protein (hpnpaseold-35) was cloned a decade ago in the context of terminal differentiation and senescence through a novel overlapping pathway screening approach. since then, significant insights have been garnered about this exoribonuclease and its repertoire of expanding functions. the objective of this review is to provide an up-to-date perspective of the recent discoveries made relating to hpnpaseold-35 and the impact they continue to have on our comprehension of its expanding and diverse array of functions.

Text-mining treatment

Transformed abstract
rna degradation plays fundamental role maintaining cellular homeostasis whether occurs surveillance mechanism eliminating aberrant mrnas rna processing generate mature transcripts exoribonucleases essential mediators rna decay pathways one evolutionarily conserved enzyme polynucleotide phosphorylase pnpase human homologue fascinating enzymatic protein hpnpaseold cloned decade ago context terminal differentiation senescence novel overlapping pathway screening approach since significant insights garnered exoribonuclease repertoire expanding functions objective review provide date perspective recent discoveries made relating hpnpaseold impact continue comprehension expanding diverse array functions

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 documents in the HRCS categories.

There are 2.889 different terms in the sample of 500.000 documents with more than 2.500 occurrences - it makes more than 45 million terms. This scenario is more than enough to ensure convergence and quality of the models.

There are 22.408 different terms between 2.500 and 100 repetitions that will also be analysed and used as input for the statistical models.

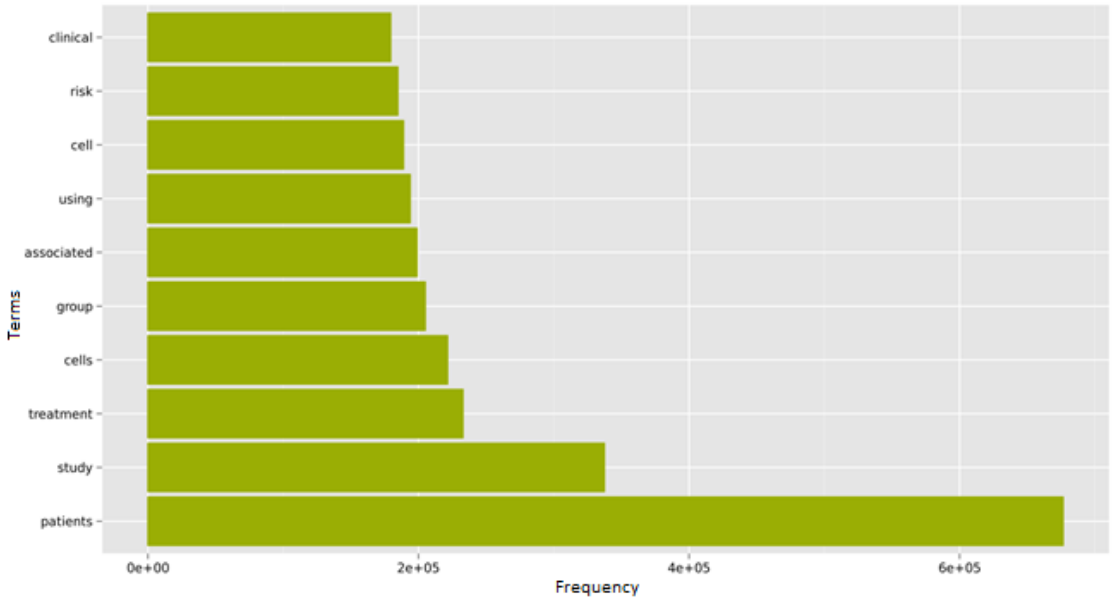

Figure 2 - Top 10 frequency terms

Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
patients	676.573	increased	127.807	performed	87.311	included	70.708
study	337.481	levels	127.399	total	86.277	surgery	69.787
treatment	232.904	time	125.539	low	85.802	primary	69.113
cells	221.551	use	120.331	respectively	85.206	survival	68.990
group	205.027	patient	117.827	rate	84.495	system	68.814
associated	198.957	one	116.517	tumor	83.112	outcomes	68.179
using	193.901	higher	113.704	months	81.770	review	67.222
cell	188.956	however	112.801	including	81.362	diagnosis	67.065
risk	184.771	human	111.083	three	81.257	findings	66.863
clinical	179.477	control	105.649	specific	81.234	evidence	66.643
cancer	179.098	showed	105.496	blood	81.030	population	65.581
disease	170.070	among	103.074	role	80.756	without	65.482
may	168.395	factors	102.096	first	80.558	research	64.761
data	162.795	effects	101.891	level	80.425	number	64.738
used	153.770	related	101.730	type	76.660	changes	64.306
analysis	153.657	cases	100.785	reported	76.530	factor	64.292
compared	149.898	found	99.542	potential	76.397	test	63.417
results	149.496	care	98.312	year	76.376	within	62.113
years	147.845	groups	97.655	lower	74.970	treated	61.759
significantly	144.588	activity	96.867	response	74.631	early	61.638
studies	143.334	protein	95.642	identified	74.521	differences	60.672
high	143.156	well	94.742	development	74.386	association	60.066
also	140.504	different	92.719	new	74.092	increase	59.673
significant	139.810	women	91.972	function	74.018	case	59.486
can	138.720	children	90.824	observed	73.771	quality	59.125
expression	138.321	therapy	90.423	induced	73.619	follow	58.851
two	137.030	non	89.728	present	73.085	mortality	58.627
age	135.609	mean	89.503	positive	71.544	participants	58.498
health	134.378	effect	87.807	gene	71.098	pain	57.859
based	129.968	model	87.341	important	71.008	infection	57.689

Figure 3 - Top 120 frequency terms

4 QUALITY

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

Less than 1% of wrong terms have been found. In the context of the project (huge amount of documents and terms) it is a small number of mistakes and does not represent a 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
nephrotic	nephritic
origine	origin
absortion	absorption
achived	achieved
acitvates	activates
acros	across
additionaly	additionally
adjusment	adjustment
anatom	anatomy
aortography	orthography
appllied	applied
assoicated	associated
fuorescence	fluorescence
identfiy	identify
identfying	identifying
identic	identical
identication	identification
laboratoires	laboratories
neurosci	neurosis

Figure 4 - Example of errors and corrections

5 BUSINESS STANDPOINT

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.

If there are no specific terms that can separate the documents in categories, the algorithms will give very generic solutions that will not suit with the defined categories.

The best way to evaluate whether the information analysed is good enough from a business standpoint is to find terms related to health 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 cancer, cells, blood or cardiac in the top 100 most frequent terms as showed in this wordcloud.

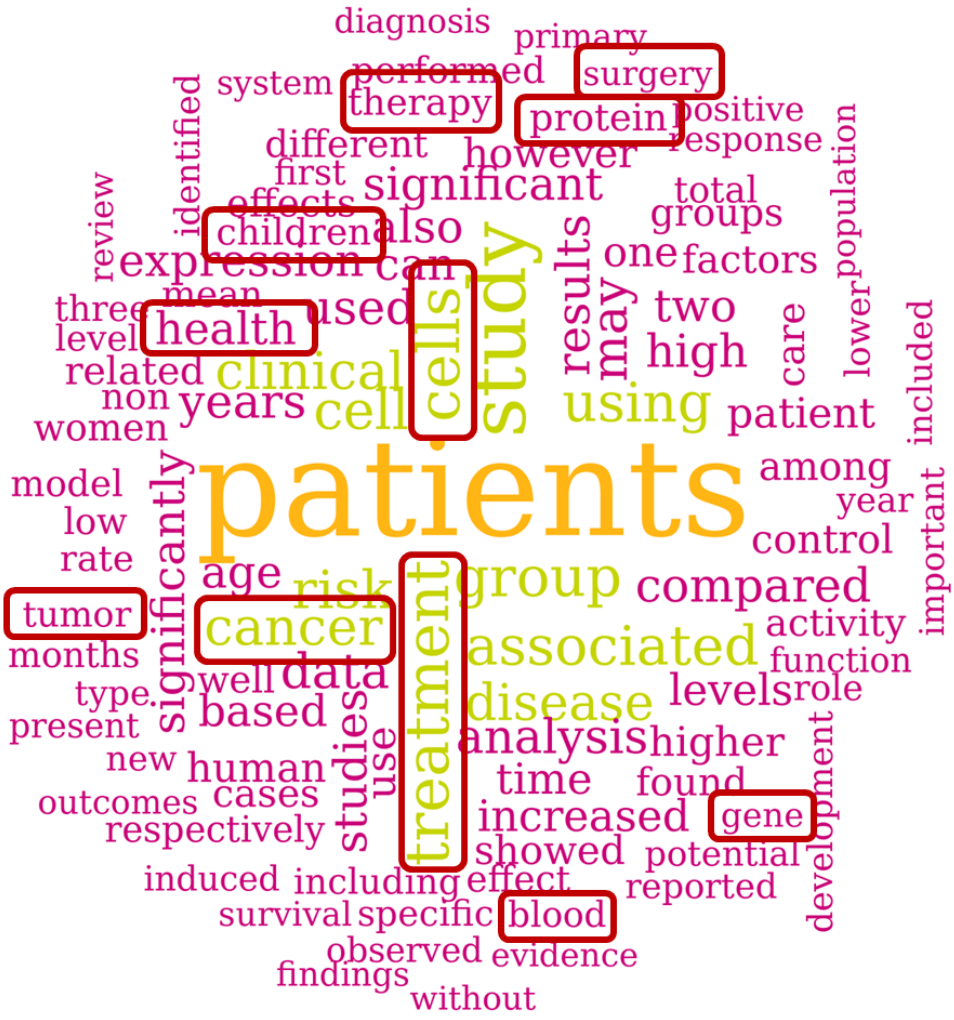

Figure 5 - Wordcloud for terms with more than 65.000 repetitions

5.1 Topics

The variety of terms is one of the keys to take advantege 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 diferent topics that will allow to classify the documents in groups and make a consistent segmentation.

For example, if we analyse terms around ‘neuro’ we find enough differents terms (with suficient size) that allow the algorithm to – probabily – classify most documents containing any of these words into 'neurological' category

neuro	191
neurobiological	188
neuroblastoma	330
neurocognitive	311
neurodegeneration	301
neurodegenerative	848
neurodevelopmental	285
neuroendocrine	402
neurogenic	186
neuroimaging	571
neuroinflammation	167
neurologic	585
neurological	2.057
neurology	178
neuromuscular	448
neuron	429
neuronal	1.877
neurons	2.311
neuropathic	351
neuropathy	710
neurophysiological	167
neuroprotection	152
neuroprotective	434
neuropsychiatric	355
neuropsychological	579
neuroscience	201
neurosurgery	144
neurosurgical	193
neurotoxicity	189
neurotransmission	142
neurotransmitter	211
neurotransmitters	127
neurotrophic	260
neurovascular	146

Figure 6 - Terms around ‘neuro’ topics

Additionally, a large number of very specific terms that will make better and easier the results of the algorithm have been found. Below there is an example of some health specific terms found.

Term
immunohistochemical
encephalomyelitis
immunocytochemistry
neovascularization
neurodegenerative
clinicopathological
hypercholesterolemia
anesthesiologists
gastrointestinal
fluorodeoxyglucose
electroencephalography
echocardiographic
musculoskeletal
atherosclerotic
ultrasonography
pcr (polymerase chain reaction)
mri (magnetic resonance imaging)
mrna (messenger Ribonucleic acid)
mmp (matrix metalloproteinases)
snps (single nucleotide polymorphism)

Figure 7 - Example of health specific terms

5.2 Annual stability

Once finalized, the generated model must properly classify new documents. Therefore, the results of the models must be consistent when a new document is classified.

To ensure the stability of the segmentation and classification model, the variability over time of the most common terms (top 100 in 2014) has been analyzed.

The figure below shows that the term distribution does not change along years. This means that we can use a random sample to train the models, and the classification will work with any document added later (regardless of the date of publication).

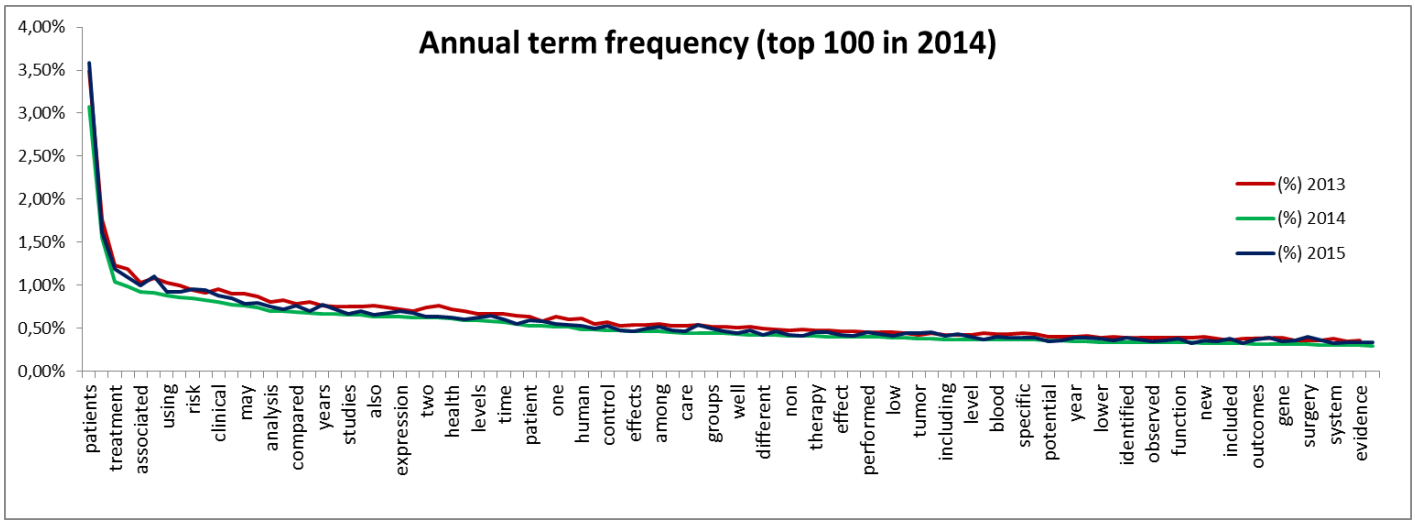

Figure 8 - Annual term frequency (2012-2014)

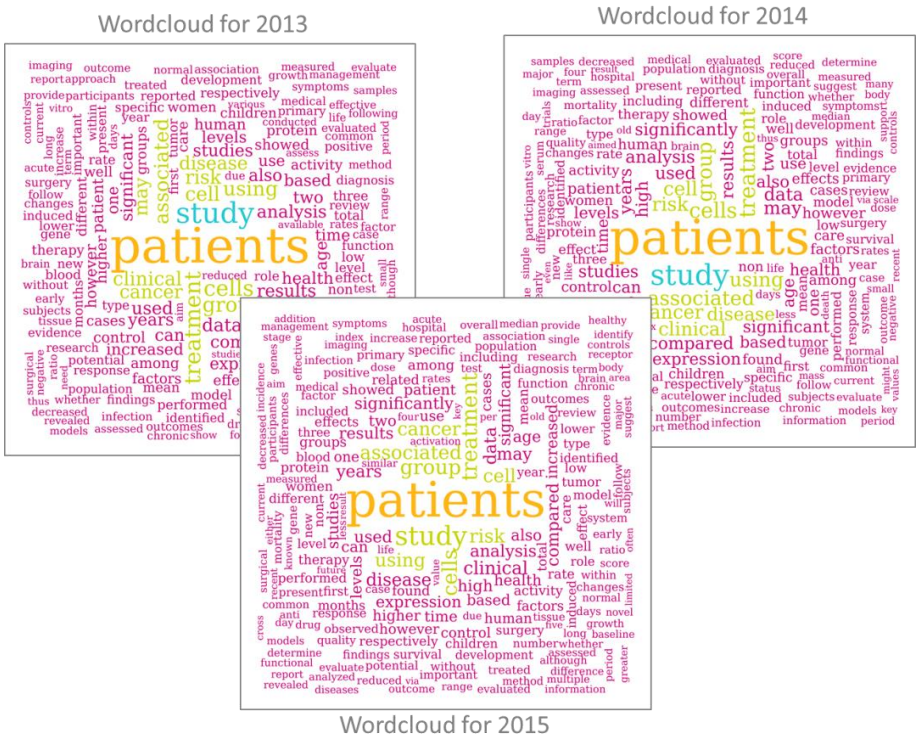

Figure 9 - Wordclouds for years 2012, 2013 and 2014

6 CONCLUSIONS

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

- Completeness: more than one million of documents and more than 20.000 terms will give statistical support to the analysis and models; clustering and classification models will use smaller samples though.

- Quality: more than 99% of the terms are correct (only a really small number of terms will not be used in the models).

- Business point of view: the most frequent terms are related to health and there are lot of different topics with enough amounts of terms. Additionally, the frequency of terms does not change annually, thus making the segmentation and the classification rules stable in time.