DIGIT.B4 – Big Data PoC

RTD – Health papers

D03.02.Dictionaries

everis Spain S.L.U

Table of contents

Table of figures

1 INTRODUCTION

1.1 Context of the project

The objective of this proof of concept is to prove how big data techniques can be applied in the research domain and to demonstrate the policy benefits big data can bring.

Specifically, this proof of concept demonstrates the use of text mining techniques on large amounts of unstructured research papers as a means to identify trending topics in the health research field. This analysis can be used as an additional input prior to launching calls for grants.

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 methodology, the data is cleaned and transformed in order to optimize the input for the segmentation algorithms.

2 DICTIONARY

The creation of a term dictionary is one of the most important activities in text-mining thecniques and it is even more necessary when we want to make a segmentation of the documents and when we know the final categories that we want to discover. The purpose of the dictionary is to unify terms that belong to the same context in order to guide the segmentation and to obtain the categories we are looking for.

There are three key points in this process:

- Unification of terms
- Removal of meaningless terms
- Identification of relevant terms

More than 5.000 terms have been analysed to create the final dictionary with approximately 3.000 terms needed 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 include all those terms.

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

- Diseases and disorders linked to an HRCS category : identified with the suffix '_disease'. For example: 'eye_disease' and 'cardiovascular_disease'
- Medicaments linked to an HRCS category: identified with the suffix '_med'. For example: 'inflammatory_med' and 'cardiovascular_med'
- Treatments linked to an HRCS category: identified with the suffix '_treat'. For example: 'neurological_treat' and 'oral_gastro_treat'
- Global and important concepts related to health. For example:
    - 'drug' including 'cocaine', 'heroine' and 'cannabis'
    - 'virus' including 'adenovirus', 'herv' and 'hbov'
    - 'pancreas' including 'pancreatic', 'insulin' and 'glucagon'

For example, all the terms around 'eye' are transformed into the term 'eye' and 'eye_disease' so we can join all the documents with these words. The algorithm has no intelligence and does not know that the terms 'retinal', 'pupils', 'maculopathy' or 'strabismus' belongs to the same category; with the dictionary we can unify 'retinal' and 'pupils' to the global term 'eye' and 'maculopathy' and 'strabismus' to the global term 'eye_disease'. These actions give the intelligence that a machine cannot have itself to perform the segmentation process.

The figures below show an example of terms transformed into ‘eye’ and ‘eye_disease’:

Term	Translation
lycium	eye
ophthalmology	eye
optotypes	eye
pentacam	eye
trabeculectomy	eye
blink	eye
choroidal	eye
conjunctiva	eye
corneas	eye
fovea	eye
intraocular	eye
iols	eye
lentis	eye
lutein	eye
macula	eye
melanopsin	eye
miosis	eye
mydriasis	eye
ophthalmic	eye
optic	eye
optica	eye
optical	eye
parafoveal	eye
photoreceptor	eye
pupils	eye
retinal	eye
visual	eye

Figure 1 - Transformation into 'eye'

Term	Translation
amblyopia	eye_disease
anisometropia	eye_disease
antiglaucoma	eye_disease
blind	eye_disease
drusen	eye_disease
hyperopia	eye_disease
hyperopic	eye_disease
keratoconus	eye_disease
keratomileusis	eye_disease
keratopathy	eye_disease
macular_degeneration	eye_disease
maculopathy	eye_disease
neuromyelitis	eye_disease
nrem	eye_disease
nystagmus	eye_disease
poag	eye_disease
presbyopia	eye_disease
pterygium	eye_disease
retinopathy	eye_disease
retinoschisis	eye_disease
saccade	eye_disease
strabismus	eye_disease
trachoma	eye_disease

Figure 2 - Transformation into 'eye_disease'

2.2 Meaningless terms

All the terms without an specific meaning will not help to categorize the documents and are removed from the corpus in order to (e.g.: when a term appears in every document):

- reduce the number of terms of the corpus
- facilitate the convergence of the algorithm.

The figure below shows an example of terms discarded:

Term	Frequency
patient	31.867
study	20.590
use	19.027
group	12.265
high	11.168
treatment	9.840
increase	9.587
year	8.986
result	8.409
level	8.357
associate	8.116
risk	8.081
show	7.888
compare	7.461
effect	7.397
control	7.390
age	7.325
clinical	7.258
include	7.041
factor	6.696
may	6.660
case	6.339
data	6.279
analysis	6.144
rate	5.990
expression	5.935
report	5.839
significantly	5.806
model	5.721
also	5.706
health	5.598

Figure 3 - Terms discarded

2.3 Relevant terms

Relevant terms referred to important parts/organs of the body or important diseases with enough frequency are not transformed. The algorithm itself will recognize and link this relevant terms with the global terms defined in '2.1 Unification of terms' that appear in similar documents.

The figure below shows an example of the main relevant terms with no transformation:

Term	Frequency
cancer	7.666
tumor	4.517
blood	2.860
brain	1.965
hiv	1.942
injury	1.880
lesion	1.758
syndrome	1.703
bone	1.699
liver	1.697
dna	1.656
breast	1.654
lung	1.621
diabetes	1.573
genetic	1.532
heart	1.442
inflammatory	1.408
muscle	1.291
renal	1.278

Figure 4 - 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 category of the documents. These actions also allow the clustering 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
myocardial blood_flow coronary circulation mbf focal coronary microvascular coronary atherosclerosis focal percutaneous coronary coronary circulation	heart blood_flow cardiovascular circulation mbf focal cardiovascular cardiovascular cardiovascular stroke focal skin cardiovascular cardiovascular circulation
diabetic hypertensive fatigue renal renal nephritis renal nephritis nephritis inflammatory	metabolic blood_disease respiratory_disease renal renal inflammatory renal inflammatory inflammatory inflammatory
melanoma skin cancer metastatic melanoma metastatic melanoma prognosis systemic metastatic melanoma metastatic melanoma angiogenesis metastatic melanoma melanoma metastatic melanoma	cancer skin cancer cancer cancer cancer cancer prognosis cancer cancer cancer cancer angiogenesis cancer cancer cancer cancer cancer
obstetric hysterectomy obstetric ile obstetric hysterectomies obstetric hysterectomy ile uterine sepsis vaginal fistula renal maternal fetal obstetric hysterectomy ile uterine obstetric hysterectomy	obstetric reproduction obstetric ile obstetric hysterectomies obstetric reproduction ile reproduction infection reproduction oral_gastro_disease renal reproduction reproduction obstetric reproduction ile reproduction obstetric reproduction
enthesitis spondyloarthritis lesion lesion enthesitis knee_joint inflammatory tendon ligament lesion lesion knee_joint enthesitis enthesitis lesion histological joint enthesitis	inflammatory musculoskeletal_disease injuries injuries inflammatory knee_joint inflammatory tendon musculoskeletal injuries injuries knee_joint inflammatory inflammatory injuries histological joint inflammatory
metabolism cholesterol triglyceride woman ship waist alcohol cholesterol cholesterol sex triglyceride triglyceride woman metabolism metabolism	metabolic metabolic metabolic ship musculoskeletal drug metabolic metabolic metabolic metabolic metabolic metabolic
cerebellar hemorrhage rch spinal rch spinal cerebrospinal hypovolemia rch spinal rch rch hemorrhagic venous cerebellar vein cerebellar cerebellar rch rch rch rch spinal rch	brain stroke rch spinal rch spinal stroke blood_disease rch spinal rch rch stroke cardiovascular brain cardiovascular brain brain rch rch rch rch spinal rch
physiologic anxiety anxiety skin anxiety anxiety anxiety anxiety physiologic anxiety anxiety	mental_health mental_health mental_health skin mental_health mental_health mental_health mental_health mental_health mental_health mental_health

Figure 5 - Example of corpus tranformed

4 CONCLUSIONS

To split the documents following the HRCS categorization we need to translate specifics terms to global terms and give intelligence to the algoritihm.

A dictionary with more than 3.000 relevant and specific terms have been created out of 5000 individual terms.

The data preparation process execution allows to obtain the input for the segmentation optimized and guided in order to classify the documents categories.

5 ANNEX

5.1 Dictionary

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 abstract. It is the file used in '2.2 meaningless terms'

D03.02.Dictionaries_ v1.0.xlsx