DIGIT.B4 – Big Data PoC

DG GROW

D03.02.Dictionary

everis Spain S.L.U

Table of contents

List of figures

1 INTRODUCTION

1.1 Context of the project

The objective of the proof of concept showcasing the use of big data in the procurement domain, in cooperation with DG GROW, is to prove the usefulness and policy benefits that big data can bring.

This proof of concept shall also demonstrate the use of natural language analysis techniques to check the compliance of the transpositions sent by the European Member States related to some directives. In the context of the PoC, one directive and its respective national transpositions will be analysed, with the objective of supporting the manual checks currently done by European Commission staff.

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

2 DICTIONARY

The creation of a term dictionary is one of the most important activities in text mining techniques. It is even more necessary for making a segmentation of the documents, and once we identify 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 obtain the categories we are looking for. The whole treatment process has been used in directive and transposition documents for each language: German, English, French and Spanish.

There are four key points in this process:

- Keywords translation;
- Keywords synonyms;
- Ngrams dictionaries;
- Unification of terms.

More than 8.000 terms in the directive and transposition documents have been analysed to create the final dictionary for each language, with approximately 200 keywords needed to make the segmentation.

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

2.1 Keywords translation

There are several keywords that belong to each directive article and are common in the four languages. Therefore, the received keywords were translated from English to German, French and Spanish.

The following figure shows some examples:

English	Spanish	French	German
Statutory interest 8 percentage points reference rate	acta de notoriedad Ocho por ciento tipo de referencia	intérêt légal Huit pourcentage taux directeur	gesetzlich zins Achtknoten Prozentpunkt Leitzinssatz

Figure 1 - Example of keywords

2.2 Keywords synonyms

The keyword dictionary also includes some synonyms of the main keywords in order to identify more terms and improve the segmentation of the articles and classification of the paragraphs.

Originals	Synonyms
Días de calendario Desviación en bruto Título ejecutivo	Días naturales Desviación grave Título habilitante

Figure 2 - Example of keywords synonyms

2.3 Ngrams dictionaries

There are terms which are formed by multiple words and they should be treated as a single word in order to increase the information fed to the algorithms. For example, Scottish Minister, European Parliament, Secretary of State, Department for Business Innovation and Skills, etc.

2.4 Unification of terms

All relevant terms and keywords from an article are transformed to a unique term. This transformation groups different terms, giving higher frequency to the keywords and aiming to solve the problem of low term frequency.

2.4.1 English

The dictionary of keywords below shows the transformation of some terms, applied in English directive and transposition documents.

Terms	Translation
statutory_interest	art_2 art_4
eight_percentage	art_2
reference_rate	art_2
amount_due	art_2
between_undertakings	art_3
without_necessity_reminder	art_3 art_6
grossly_unfair_creditor	art_3 art_4
max_60_calendar_days	art_3 art_4
six_months	art_3
expressly	art_3 art_4
public_authority	art_4
30_calendar_days	art_4
objectively_justified_light_particular_nature_features_contract	art_4
schedule	art_5
eur_forty	art_6
reasonable	art_6
gross_deviation	art_7
good_faith	art_7
exclude_interest	art_7
exclude_compensation_recovery_cost	art_7
enforceable_title	art_10
90_calendar_days	art_10
not_disputed	art_10

Figure 3 - Unification of English terms

All the terms around article 7 are transformed into the term ‘art_7’. The algorithm has no intelligence and does not know that the terms ‘gross_deviation’, ‘good_faith’ or ‘exclude_interest' belong to the same article. These actions bring an intelligence to the segmentation process that a machine cannot have by itself.

2.4.2 Spanish

The dictionary of keywords below shows the transformation of some terms, applied to the Spanish version of the directive and the Spanish transposition documents.

Terms	Translation
acta notoriedad	art_2
ocho ciento	art_2
ocho puntos porcentuales	art_2
tipo referencia	art_2
deuda aduanera	art_2
cooperación empresarial	art_3
necesidad recordatorio	art_3
manifiestamente abusivo acreedor	art_3
abusivo perjuicio acreedor	art_3
primero enero	art_3
primero julio	art_3
seis meses	art_3
medio año	art_3
exceda sesenta días naturales	art_3 art_4
excedan sesenta días naturales	art_3 art_4
máximo sesenta días naturales	art_3 art_4
ninguno caso poder acordar plazo superior sesenta día natural	art_3 art_4
expresamente	art_3 art_4
poder público	art_4
poderes públicos	art_4
intereses legales	art_4
interés legal	art_4
legal interés	art_4
treinta días naturales	art_4
treinta días siguientes	art_4
mínimo cuarenta euros	art_6
cuarenta euros	art_6
razonable	art_6
desviación grave	art_7
buena fe	art_7
interés excluido	art_7
compensación costes cobro	art_7
indemnización costes cobro	art_7
título ejecutivo	art_10
título habilitante	art_10
títulos habilitantes	art_10
noventa días	art_10
deudas discuten	art_10

Figure 4 - Unification of Spanish terms

2.4.3 French

The dictionary of keywords below shows the transformation of some terms, applied to the French version of the directive and the French transposition documents.

Terms	Translation
intérêt légal	art_2 art_4
intérêts légaux	art_2 art_4
huit point pourcentage	art_2
huit points pourcentage	art_2
taux directeur	art_2
montant dû	art_2
coopération interentreprises	art_3
rappel nécessaire	art_3
prèmiere janvier	art_3
prèmiere juillet	art_3
soixante jours civils	art_3 art_4
soixante jours	art_3 art_4
trente jours civils	art_4
trente jours	art_4
nommément	art_3
abus manifeste égard créancier	art_3 art_4
autorité publique	art_4
excède soixante jours civils	art_3 art_4
excède aucun soixante jours civils	art_3 art_4
maximum soixante jours civils	art_4
objectivement justifié nature particulière certains éléments contrat	art_4
échéancier	art_5
quarante euros	art_6
rappel nécessaire	art_6
raisonnable	art_6
écart brut	art_7
bonne foi	art_7
excluant versement intérêts	art_7
excluant indemnisation frais recouvrement	art_7
titre exécutoire	art_10
quatrevingt dix jours	art_10

Figure 5 - Unification of French terms

2.4.4 German

The dictionary of keywords below shows the transformation of some terms, applied to the German version of the directive and the German transposition documents.

Terms	Translation
gesetzlich zins	art_2 art_4
gesetzlichen zins	art_2 art_4
gesetzlicher zins	art_2 art_4
achtknoten prozentpunkt	art_2
acht prozentpunkten	art_2
leitzinssatz	art_2
zinssatz	art_2
bezugszinssatz	art_2
zus zahlender betrag	art_2
fälliger betrag	art_2
fälligen betrag	art_2
zusammenarbeit zwischenn unternehmen	art_3
zusammenarbeit vooon unternehmen	art_3
ohneee dieee notwendigkeit eineeer mahnung	art_3
ohneee daaass eees eineeer mahnung bedarf	art_3
ohneee mahnung	art_3
ersten januar	art_3
ersten juli	art_3
halbjahr	art_3
naaach meeehr aaals sechzig tage	art_3 art_4
keineeem fall sechzig kalendertage überschreitet	art_3 art_4
zahlungsfrist sechzig kalendertage nichta überschreite	art_3 art_4
sechzig kalendertage	art_3
sechzig tage	art_3
ausdrücklich	art_3 art_4
grob nachteilig füra deeen gläubiger	art_3 art_4
füra deeen gläubiger nichta grob nachteilig	art_3 art_4
gläubigers nichta grob unbillig	art_3 art_4
staatliche behörde	art_4
zuständigen behörde	art_4
zuständige behörde	art_4
dreißig kalendertage	art_4
dreißig tage	art_4
iiin anbetracht deeer besonderen natur odeeer merkr	art_4
aufgrund deeer besonderen natur odeeer merkmale c	art_4
aufgrund deeer besonderen natur odeeer deeer merk	art_4
fälligkeitsplan	art_5
ratenzahlung	art_5
mindestens vierzig euros	art_6
höhe vooon vierzig euros	art_6
sinnvoll	art_6
angemessen	art_6
brutto abweichung	art_7
grobe abweichung	art_7
guter glaube	art_7
guten glaubens	art_7
ohneee zins	art_7
ausgeschlossen entschädigung füra beitreibungskoste	art_7
entschädigung füra beitreibungskosten ausgeschlosse	art_7
vollstreckungstitel	art_10
vollstreckbarer titel	art_10
neunzig tage	art_10
neunzig kalendertage	art_10
schulden nichta bestritten	art_10

Figure 6 - Unification of German terms

3 EXAMPLES OF CORPUS

After applying all the rules previously mentioned that are related to the creation of the dictionary, it is easier to identify the category of the article. 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 dictionaries.

Corpus	Corpus after dictionary
For the purposes of this Directive, the following definitions shall apply: (1) 'commercial transactions' means transactions between undertakings or between undertakings and public authorities which lead to the delivery of goods or the provision of services for remuneration; (2) 'public authority' means any contracting authority, as defined in point (a) of Article 2(1) of Directive 2004/17/EC and in Article 1(9) of Directive 2004/18/EC, regardless of the subject or value of the contract; (3) 'undertaking' means any organisation, other than a public authority, acting in the course of its independent economic or professional activity, even where that activity is carried out by a single person; (4) 'late payment' means payment not made within the contractual or statutory period of payment and where the conditions laid down in Article 3(1) or Article 4(1) are satisfied; (5) 'interest for late payment' means statutory interest for late payment or interest at a rate agreed upon between undertakings, subject to Article 7; (6) 'statutory interest for late payment' means simple interest for late payment at a rate which is equal to the sum of the reference rate and at least eight percentage points; (7) 'reference rate' means either of the following: (a) for a Member State whose currency is the euro, either: (i) the interest rate applied by the European Central Bank to its most recent main refinancing operations; or (ii) the marginal interest rate resulting from variablerate tender procedures for the most recent main refinancing operations of the European Central Bank; (b) for a Member State whose currency is not the euro, the equivalent rate set by its national central bank; (8) 'amount due' means the principal sum which should have been paid within the contractual or statutory period of payment, including the applicable taxes, duties, levies or charges specified in the invoice or the equivalent request for payment; (9) 'retention of title' means the contractual agreement according to which the seller retains title to the goods in question until the price has been paid in full; (10) 'enforceable title' means any decision, judgment or order for payment issued by a court or other competent authority, including those that are provisionally enforceable, whether for immediate payment or payment by instalments, which permits the creditor to have his claim against the debtor collected by means of forced execution.	purpose directive definition commercial_transaction transaction art_3 art_3 public_authorities lead delivery good provision service remuneration art_4 contract authority define point article directive article directive regardless subject value contract undertaking organisation art_4 act course independent economic professional activity even activity carry single person payment contractual statutory period payment condition lay article article satisfy interest art_2 art_4 interest rate agree upon art_3 subject article art_2 art_4 simple interest rate equal sum art_2 little art_2 point art_2 either whose currency euro either interest rate european_central_bank recent main refinance operation marginal_interest_rate result variablerate tender procedure recent main refinance operation european_central_bank whose currency euro equivalent_rate set national central bank art_2 principal sum pay contractual statutory period payment include applicable tax duty levy charge specify invoice equivalent_request payment retention title contractual agreement accord seller retain title good question price pay full art_10 decision judgment order payment issue court competent authority include provisionally enforceable whether immediate payment payment_schedule permit creditor claim debtor collect force execution
1. Los Estados miembros se asegurarán de que, en los casos en que resulte exigible el interés de demora en las operaciones comerciales con arreglo a los artículos 3 o 4, el acreedor tenga derecho a cobrar al deudor, como mínimo, una cantidad fija de 40 EUR. 2. Los Estados miembros se asegurarán de que la cantidad fija mencionada en el apartado 1 sea pagadera sin necesidad de recordatorio como compensación por los costes de cobro en que haya incurrido el acreedor. 3. Además de la cantidad fija establecida en el apartado 1, el acreedor tendrá derecho a obtener del deudor una compensación razonable por todos los demás costes de cobro que superen la cantidad fija y que haya sufrido a causa de la morosidad de este. Esta podría incluir, entre otros, los gastos que el acreedor haya debido sufragar para la contratación de un abogado o una agencia de gestión de cobro.	asegurar caso resultar exigible interés_demora operación_comercial arreglo artículo acreedor derecho cobrar deudor mínimo cantidad fijo art_6 asegurar cantidad fijo mencionar pagadero art_3 art_7 incurrir acreedor además cantidad fijo establecer acreedor derecho obtener deudor compensación art_6 demás coste cobro superar cantidad fijo sufrir causa morosidad poder incluir gasto acreedor debido sufragar contratación abogado agencia gestión cobro
1. Les États membres veillent à ce qu'un titre exécutoire, quel que soit le montant de la dette, puisse être obtenu, y compris au moyen d'une procédure accélérée, normalement dans les quatrevingt-dix jours civils après que le créancier a formé un recours ou introduit une demande auprès d'une juridiction ou d'une autre autorité compétente, lorsqu'il n'y a pas de contestation portant sur la dette ou des points de procédure. Les États membres s'acquittent de cette obligation en conformité avec leurs dispositions législatives, réglementaires et administratives nationales respectives. 2. Les dispositions législatives, réglementaires et administratives nationales s'appliquent dans les mêmes conditions à tous les créanciers qui sont établis dans l'Union. 3. Pour calculer le délai visé au paragraphe 1, il n'est pas tenu compte des périodes suivantes: a) les délais requis pour la signification et la notification des documents; b) tout retard causé par le créancier, tel que les délais nécessaires à la rectification de demandes. 4. Le présent article s'applique sans préjudice des dispositions du règlement (CE) no 1896/2006.	veiller art_10 monter dette pouvoir être obtenir comprendre moyen procédure accélérer normalement art_10 civil créancier former recours introduire demander auprès juridiction autre autorité compétent lorsqu contestation portant dette point procédure acquitter obligation conformité disposition législatif réglementaire administratif national respectif disposition législatif réglementaire administratif national appliquer même condition tout créancier établir union calculer délai viser tenir compter période suivant délai requérir signification notification document tout retard causer créancier délai nécessaire rectification demander présent applique préjudice disposition règlement
(1) Die Mitgliedstaaten stellen sicher, dass im Geschäftsverkehr zwischen Unternehmen der Gläubiger Anspruch auf Verzugszinsen hat, ohne dass es einer Mahnung bedarf, wenn folgende Bedingungen erfüllt sind: a) Der Gläubiger hat seine vertraglichen und gesetzlichen Verpflichtungen erfüllt, und b) der Gläubiger hat den fälligen Betrag nicht rechtzeitig erhalten, es sei denn, dass der Schuldner für den Zahlungsverzug nicht verantwortlich ist. (2) Die Mitgliedstaaten stellen sicher, dass folgender Bezugszinssatz angewendet wird: a) für das erste Halbjahr des betreffenden Jahres der am 1. Januar dieses Jahres geltende Zinssatz; b) für das zweite Halbjahr des betreffenden Jahres der am 1. Juli dieses Jahres geltende Zinssatz. (3) Für die Fälle, in denen die in Absatz 1 genannten Bedingungen erfüllt sind, stellen die Mitgliedstaaten Folgendes sicher: a) Der Gläubiger hat Anspruch auf Verzugszinsen ab dem Tag, der auf den vertraglich festgelegten Zahlungstermin oder das vertraglich festgelegte Ende der Zahlungsfrist folgt. b) Ist der Zahlungstermin oder die Zahlungsfrist nicht vertraglich festgelegt, so hat der Gläubiger Anspruch auf Verzugszinsen nach Ablauf einer der folgenden Fristen: i) 30 Kalendertage nach dem Zeitpunkt des Eingangs der Rechnung oder einer gleichwertigen Zahlungsaufforderung beim Schuldner; . ii) wenn der Zeitpunkt des Eingangs der Rechnung oder einer gleichwertigen Zahlungsaufforderung unsicher ist, 30 Kalendertage nach dem Zeitpunkt des Empfangs der Waren oder Dienstleistungen; iii) wenn der Schuldner die Rechnung oder die gleichwertige Zahlungsaufforderung vor dem Empfang der Waren oder Dienstleistungen erhält, 30 Kalendertage nach dem Zeitpunkt des Empfangs der Waren oder Dienstleistungen; iv) wenn ein Abnahme- oder Überprüfungsverfahren, durch das die Übereinstimmung der Waren oder Dienstleistungen mit dem Vertrag festgestellt werden soll, gesetzlich oder vertraglich vorgesehen ist und wenn der Schuldner die Rechnung oder eine gleichwertige Zahlungsaufforderung vor oder zu dem Zeitpunkt, zu dem die Abnahme oder Überprüfung erfolgt, erhält, 30 Kalendertage nach letzterem Zeitpunkt. (4) Ist ein Abnahme- oder Überprüfungsverfahren vorgesehen, durch das die Übereinstimmung der Waren oder Dienstleistungen mit dem Vertrag festgestellt werden soll, so stellen die Mitgliedstaaten sicher, dass die Höchstdauer dieses Verfahrens nicht mehr als 30 Kalendertage ab dem Zeitpunkt des Empfangs der Waren oder Dienstleistungen beträgt, es sei denn im Vertrag wurde ausdrücklich etwas anderes vereinbart und vorausgesetzt, dass dies für den Gläubiger nicht grob nachteilig im Sinne von Artikel 7 ist. (5) Die Mitgliedstaaten stellen sicher, dass die vertraglich festgelegte Zahlungsfrist 60 Kalendertage nicht überschreitet, es sei denn im Vertrag wurde ausdrücklich etwas anderes vereinbart und vorausgesetzt, dass dies für den Gläubiger nicht grob nachteilig im Sinne von Artikel 7 ist.	geschäftsverkehr unternehmen gläubiger anspruch verzugszinsen art_3 folgende bedingungen erfüllt gläubiger vertraglichen gesetzlichen verpflichtungen erfüllt gläubiger art_2 rechtzeitig erhalten schuldner zahlungsverzug verantwortlich folgender art_2 angewendet erste art_3 betreffenden jahres art_3 jahres geltende art_2 zweite art_3 betreffenden jahres art_3 jahres geltende art_2 fälle genannten bedingungen erfüllt folgendes gläubiger anspruch verzugszinsen tag vertraglich festgelegten zahlungstermin vertraglich festgelegte ende zahlungsfrist zahlungstermin zahlungsfrist vertraglich festgelegt gläubiger anspruch verzugszinsen ablauf folgenden fristen art_4 eingangs rechnung gleichwertigen zahlungsaufforderung schuldner eingangs rechnung gleichwertigen zahlungsaufforderung unsicher art_4 empfangs dienstleistungen schuldner rechnung gleichwertige zahlungsaufforderung empfang dienstleistungen erhält art_4 empfangs dienstleistungen abnahme überprüfungsverfahren übereinstimmung dienstleistungen vertrag festgestellt gesetzlich vertraglich vorgesehen schuldner rechnung gleichwertige zahlungsaufforderung abnahme überprüfung erfolgt erhält art_4 letzterem abnahme überprüfungsverfahren vorgesehen übereinstimmung dienstleistungen vertrag festgestellt höchstdauer verfahrens art_4 empfangs dienstleistungen beträgt vertrag wurde ausdrücklich vereinbart art_3 art_4 sinne vertraglich festgelegte art_3 art_4 vertrag wurde ausdrücklich vereinbart art_3 art_4 sinne

Figure 7 - Example corpus articles in the different languages

4 CONCLUSIONS

Different techniques have been applied to the directive in order to solve the problems identified in the data understanding phase (low frequency of terms and keywords):

- Ngrams unification;
- Unification of terms;
- Keywords synonyms.

With this process, the term/document matrix (input for segmentation and classification models) will have more relevant terms and better results will be obtained in the model phase.

5 ANNEX

5.1 Dictionary

The following Excel includes three sheets:

- Sheet 1: Dictionary with the bigrams in each language;
- Sheet 2: Dictionary with the keywords and its global term in the right column, called translation. It is the file with all the transformations used in '2.1 Unification of terms';
- Sheet 3: All terms that have been removed from the original abstract.

D03.02.Dictionaries_ v1.0.xlsx