5641 - Clasificación automática de textos y explotación BI

Javier Buill Vilches

Estructura

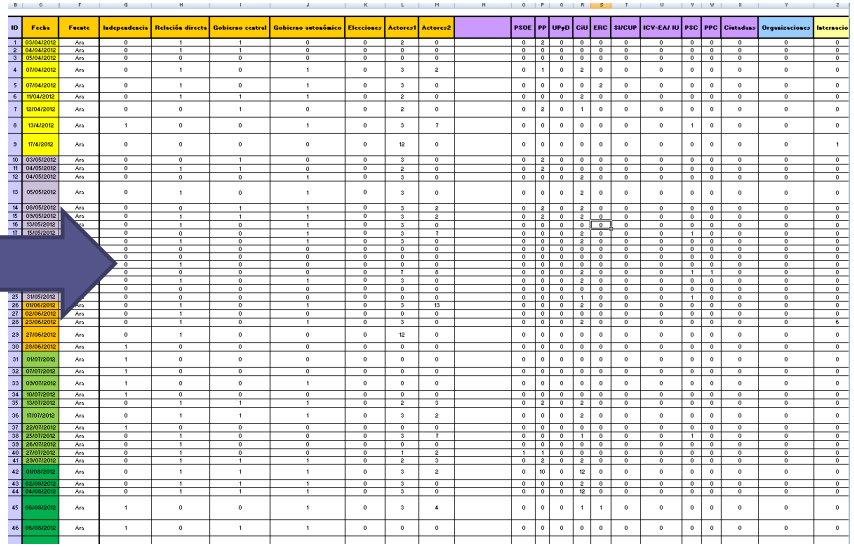
- Motivación
  - Problemática
  - Metodología actual
  - Propuesta de solución
- Clasificación automática
  - Método de Bayes
- Business Intelligence (BI)
  - Qlikview
- Resultados
  - Clasificación
  - Visualización
- Conclusiones
  - Mejoras

Motivación

- Extracción de datos y clasificación

Información de los medios

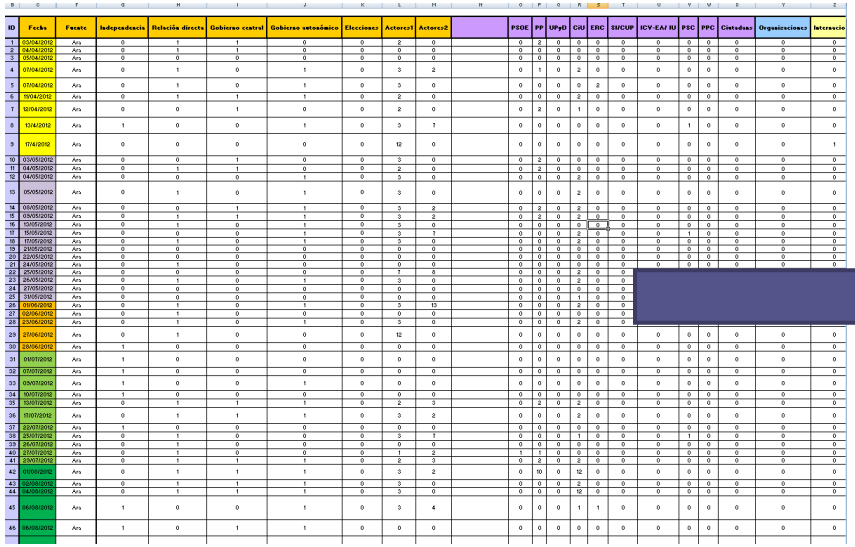
Clasificación manual



Motivación

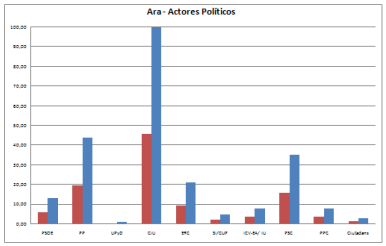
- Análisis de la información

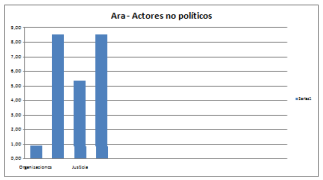
Clasificación manual

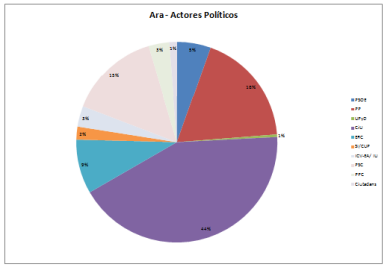


ID	Fecha	Fecatc	Independencia	Relación directa	Gobierno central	Gobierno autonómico	Eleccionesc	Actorcs1	Actorcs2		PSOE	PP	UPyD	CiU	ERC	SI/CUP	ICV-EA/ IU	PSC	PPC	Ciutadans	Organizacionesc	Interesio
1	03/04/2012	Ara	0	1	1	0	0	2	0		0	2	0	0	0	0	0	0	0	0	0	0
2	04/04/2012	Ara	0	1	1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
3	05/04/2012	Ara	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
4	07/04/2012	Ara	0	1	0	1	0	3	2		0	1	0	2	0	0	0	0	0	0	0	0
5	07/04/2012	Ara	0	1	0	1	0	3	0		0	0	0	0	2	0	0	0	0	0	0	0
6	10/04/2012	Ara	0	1	1	1	0	2	0		0	0	0	2	0	0	0	0	0	0	0	0
7	12/04/2012	Ara	0	0	1	0	0	2	0		0	2	0	1	0	0	0	0	0	0	0	0
8	13/4/2012	Ara	1	0	0	1	0	3	7		0	0	0	0	0	0	0	1	0	0	0	0
9	17/4/2012	Ara	0	0	0	0	0	12	0		0	0	0	0	0	0	0	0	0	0	0	1
10	03/05/2012	Ara	0	0	1	0	0	3	0		0	2	0	0	0	0	0	0	0	0	0	0
11	04/05/2012	Ara	0	1	1	0	0	2	0		0	2	0	0	0	0	0	0	0	0	0	0
12	04/05/2012	Ara	0	0	0	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
13	05/05/2012	Ara	0	1	0	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
14	08/05/2012	Ara	0	1	1	1	0	3	2		0	2	0	2	0	0	0	0	0	0	0	0
15	09/05/2012	Ara	0	1	1	1	0	3	2		0	2	0	2	0	0	0	0	0	0	0	0
16	10/05/2012	Ara	0	1	0	1	0	3	0		0	0	0	0	0	0	0	0	0	0	0	0
17	15/05/2012	Ara	0	0	1	1	0	3	1		0	0	0	2	0	0	0	1	0	0	0	0
18	17/05/2012	Ara	0	1	0	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
19	21/05/2012	Ara	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
20	22/05/2012	Ara	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
21	24/05/2012	Ara	0	1	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
22	25/05/2012	Ara	0	0	0	0	0	7	8		0	0	0	2	0	0	0	0	0	0	0	0
23	26/05/2012	Ara	0	1	0	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
24	27/05/2012	Ara	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
25	31/05/2012	Ara	0	0	0	0	0	0	0		0	0	0	1	0	0	0	0	0	0	0	0
26	01/06/2012	Ara	0	1	0	1	0	3	13		0	0	0	2	0	0	0	0	0	0	0	0
27	02/06/2012	Ara	0	1	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
28	23/06/2012	Ara	0	1	0	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
29	27/06/2012	Ara	0	1	0	0	0	12	0		0	0	0	0	0	0	0	0	0	0	0	0
30	28/06/2012	Ara	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
31	01/07/2012	Ara	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
32	07/07/2012	Ara	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
33	09/07/2012	Ara	1	0	0	1	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
34	10/07/2012	Ara	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
35	13/07/2012	Ara	0	1	1	1	0	2	3		0	2	0	2	0	0	0	0	0	0	0	0
36	17/07/2012	Ara	0	1	1	1	0	3	2		0	0	0	2	0	0	0	0	0	0	0	0
37	23/07/2012	Ara	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
38	25/07/2012	Ara	0	1	0	0	0	3	1		0	0	0	1	0	0	0	1	0	0	0	0
39	26/07/2012	Ara	0	1	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
40	27/07/2012	Ara	0	1	0	0	0	1	2		1	1	0	0	0	0	0	0	0	0	0	0
41	29/07/2012	Ara	0	1	1	1	0	2	3		0	2	0	2	0	0	0	0	0	0	0	0
42	01/08/2012	Ara	0	1	1	1	0	3	2		0	10	0	12	0	0	0	0	0	0	0	0
43	02/08/2012	Ara	0	1	1	1	0	3	0		0	0	0	2	0	0	0	0	0	0	0	0
44	14/08/2012	Ara	0	1	1	1	0	3	0		0	0	0	12	0	0	0	0	0	0	0	0
45	06/08/2012	Ara	1	0	0	1	0	3	4		0	0	0	1	1	0	0	0	0	0	0	0
46	09/08/2012	Ara	1	0	1	1	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0

Análisis de los datos clasificados







Problemática

¿Problema? TIEMPO

Para un texto de aproximadamente 4000 frases (~200 páginas) el tiempo estimado de codificación manual ronda los 2 meses en el trabajo de una persona a media jornada dedicada exclusivamente a esta tarea

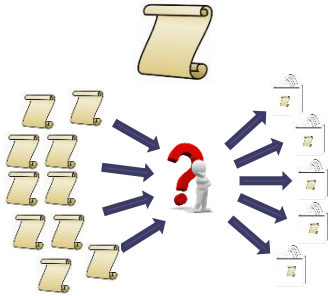
Metodología actual

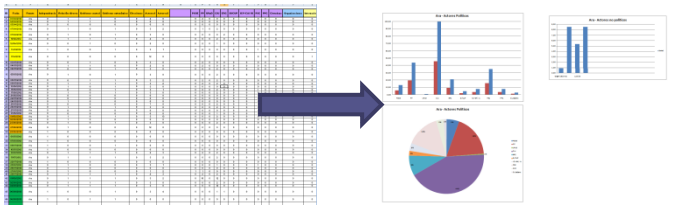
- Pasos de la actual metodología:
  - Formación de codificador
  - Lectura de texto (1)
  - Separación en cuasi-frases
  - Lectura de texto (2)
  - Clasificación manual
  - Recogida de datos
  - Estudio y análisis para los datos clasificados (sólo útil para éstos)

Hemos podido, en este contexto, desarrollar políticas de vertebración territorial para combatir los desequilibrios demográficos internos y luchar contra el despoblamiento rural,

Creer en una economía sostenible,

diversificada, equilibrada y tecnológicamente avanzada,

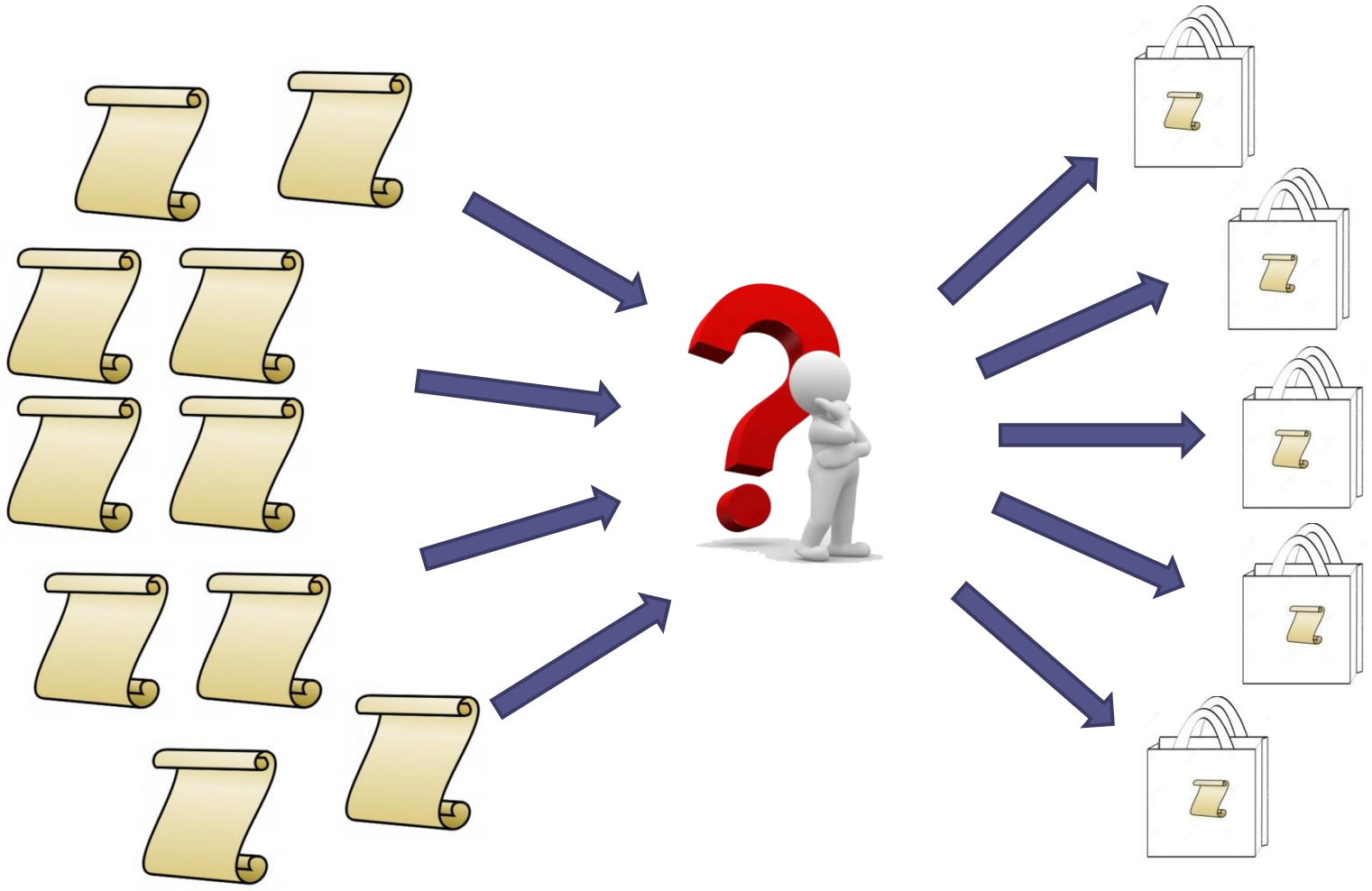




Propuesta de solución - Objetivos

- Clasificador automático
  - Rápido
  - Alto índice de acierto
  - Fácil manejo
- Herramienta *Business Intelligence* (BI)
  - Cuadro de mando preparado para presentar esta información para análisis de datos

Clasificación automática



Algoritmo propuesto

Bayes

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)}$$

Donde:

$P(A_i)$ → Probabilidad a priori

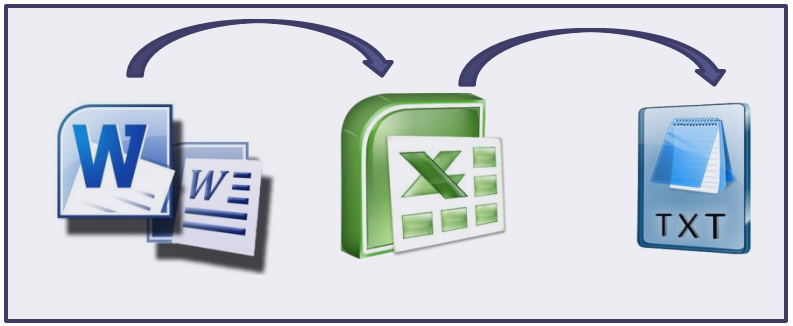
$P(B|A_i)$ → Probabilidad condicional

$P(B)$ → Probabilidad total

$P(A_i|B)$ → Probabilidad a posteriori

Metodología desarrollada (1)

Preparación y transformación de los datos

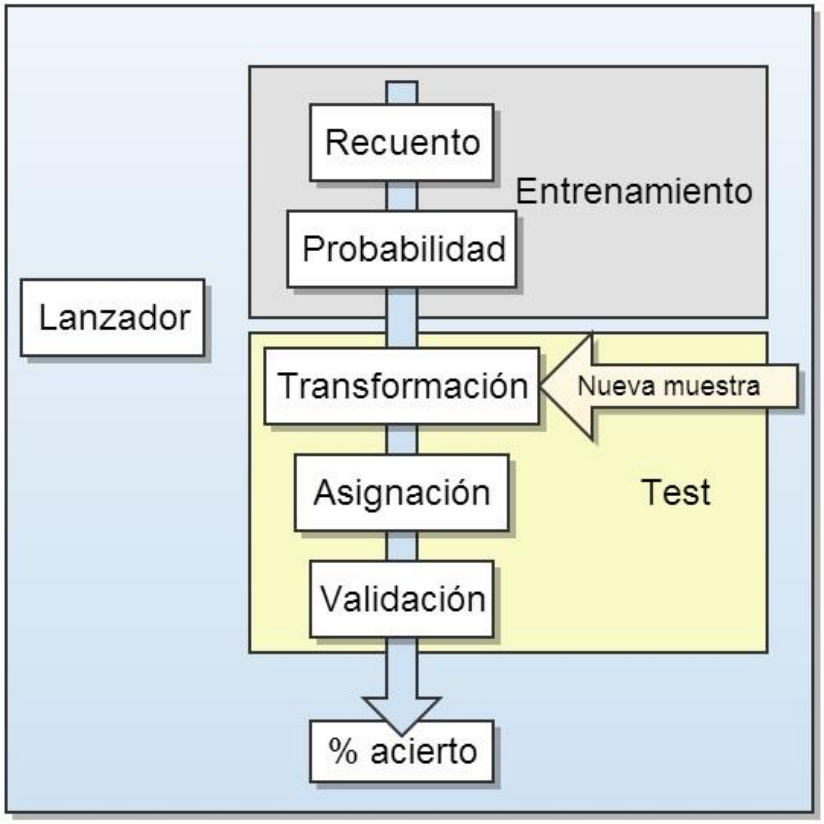


5	En el contexto económico	5	20_408
6	Desde el Partido Popular	6	00_408
7	Abordaremos una reforma :	7	00_408
8	• Reduciremos progresiv	8	00_408
9	• Rebajaremos la fiscal:	9	00_408
10	• Mejoraremos las deduc	10	90_410
11	• Reduciremos el Impues	11	00_408
12	acompañado además de nu	12	00_408
13	• Abogaremos para que e	13	00_408
14	extenderemos el ámbito	14	20_408

- Uso de formatos sencillos para usuarios ajenos a la informática
- Datos provistos en documentos de texto (.doc/.docx)
- Transformados con Excel
- Exportados a 2 archivos de texto .txt separados por:
  - Texto
  - Categoría

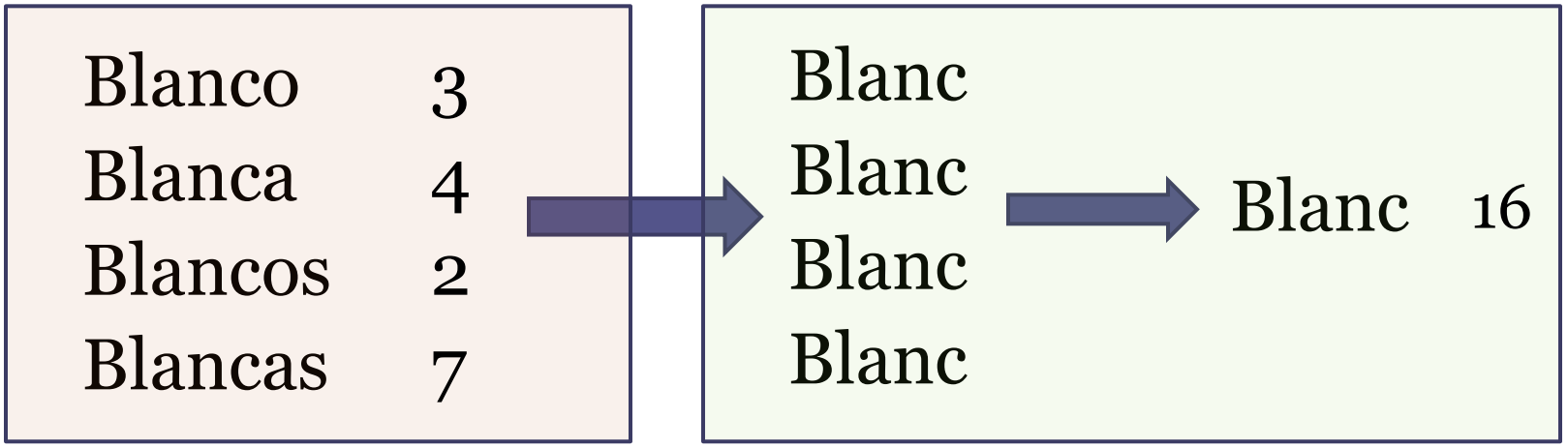
Metodología desarrollada (2)

- Flujo de ejecución:
  - 2 Fases:
    - Entrenamiento
    - Test
- Entrenamiento con 32 configuraciones para cada texto combinando:
  - *Stemmed words*
  - Stop words
  - *Laplace Smoothing*
  - Prod(prob)/sum(log)
  - Probabilidad a priori



Metodología desarrollada (3)

Stemmed words



Metodología desarrollada (3)

Stop words

Listado de palabras "sin significado" que no dan peso a ninguna categoría en particular.

Son las palabras que más aparecen

ahí, tal, de, aquí, allí, allá, la, que, el, en, y, a, los, del, se, las, por, un, para, con, no, una, su, al, lo, como, más, pero, sus, le, ya, o, este, sí, pues, decir, entonces, vez, porque, esta, entre, cuando, muy, sin, sobre, también, me, hasta, hay, donde, quien, desde, todo, nos, durante, todos, uno, les, ni, contra, otros, ese, eso, ante, ellos, e, esto, mí, antes, algunos, qué, unos, yo, otro, otras, otra, él, tanto, esa, estos, mucho, quienes, nada, muchos, cual, poco, ella, estar, estás, algunas, algo, nosotros, mi, mis, tú, te, ti, tu, tus, ellas, nosotras, vosotros, vosotras, os, mío, mía, míos, mías, tuyo, tuya, tuyos, tuyas, suyo, suya, suyos, suyas, nuestro, nuestra, nuestros, nuestras, vuestro, vuestra, vuestros, vuestras, esos, esas, ser, haber, tener, hacer, estar.

Metodología desarrollada (3)

Laplace Smoothing

$$P(X_i = x_{ij} \mid Y = y_k) = \frac{\#D\{X_y = x_{ij} \wedge Y = y_k\} + 1}{\#D\{Y = y_k\} lM}$$

Corrección cuando el valor de recuento = 0

palabra	cat1	cat2	cat3	cat4	cat5	cat6
a	10	9	10	7	5	8
b	3	2	4	2	1	2
c	1	8	2	10	2	6
d	0	6	2	4	4	6
e	6	3	8	1	7	4
f	4	2	10	2	10	7
g	5	7	7	2	1	8
h	8	7	8	8	9	2
i	6	2	0	5	8	5
h	2	5	6	4	10	6
k	3	7	4	9	5	5
l	8	8	8	2	3	3
m	3	4	7	6	2	5
n	8	6	3	10	2	5
o	6	1	9	7	7	8
p	8	2	0	7	4	0
q	5	3	6	9	9	1
r	10	8	9	3	8	9
s	8	2	7	0	8	3
t	8	3	10	0	9	5
u	0	9	1	5	2	4
v	0	2	1	2	4	6
w	0	1	9	3	0	5
x	0	0	2	8	3	1
y	1	0	6	3	5	2
z	6	2	5	1	3	6

Metodología desarrollada (3)

Prod(prob) – Sum(log)

Al multiplicar muchos valores cercanos a 0, puede causar imprecisión numérica, generando ceros de manera similar al caso anterior.

Se puede cambiar el cálculo y en vez de hacer la multiplicación de probabilidades, calcular su logaritmo, y por tanto, queda como el sumatorio de los logaritmos

$$v = argmax(v_{j \in V}(P(v_j) \prod_{ai \in X} (P(a_i|v_j))))$$

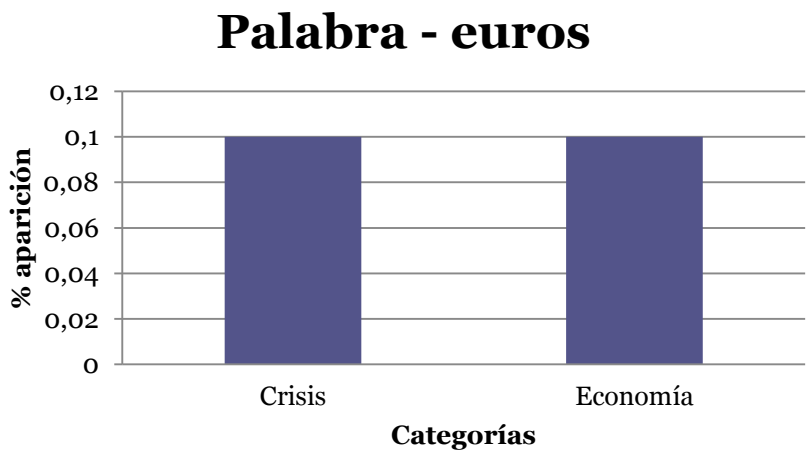
$$v = argmax(v_{j \in V}(logP(v_j) + \sum_{a_i \in X} logP(a_i|v_j)))$$

Metodología desarrollada (3)

Probabilidad a priori

$$v = argmax v_{j \in V}(P(v_j) \prod_{ai \in X}(P(a_i|v_j)))$$

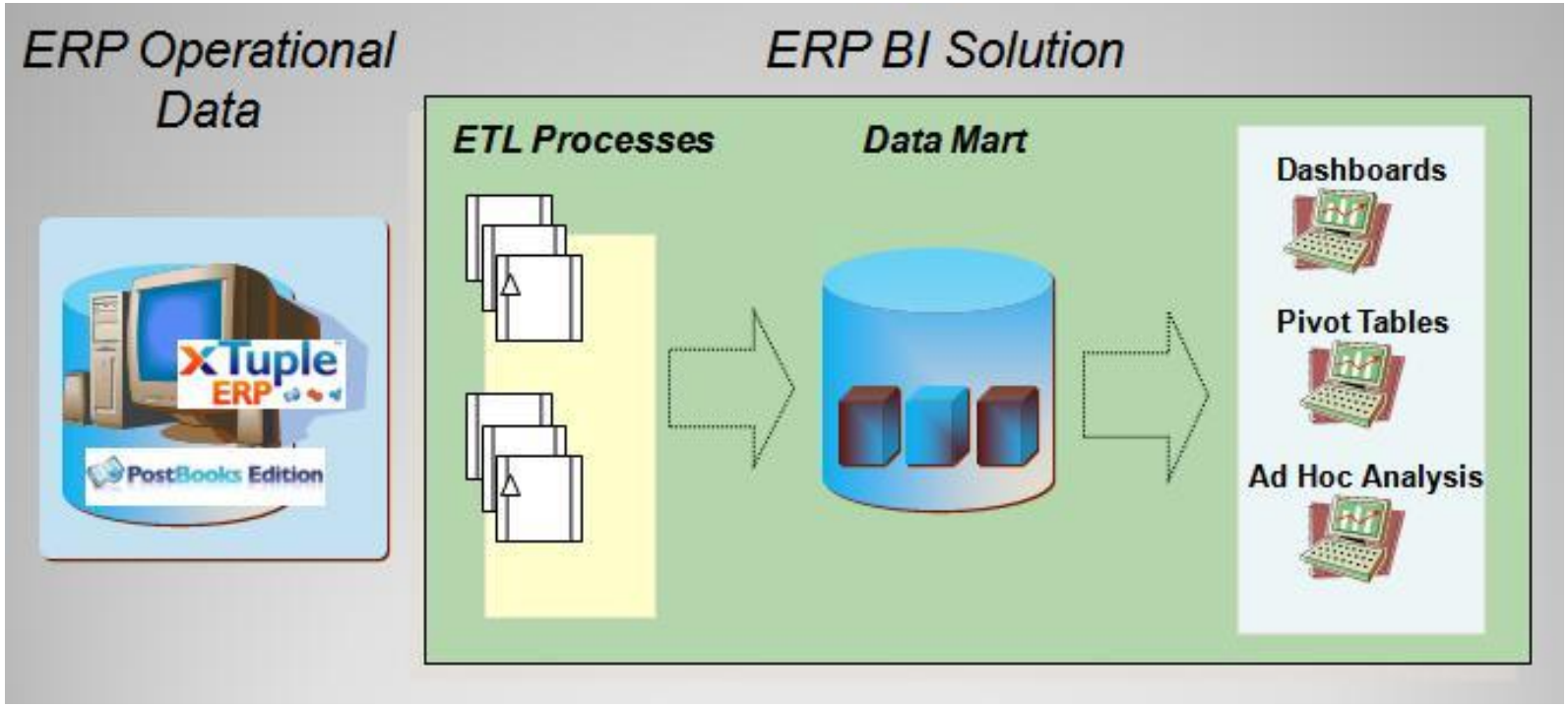
$$v = argmax v_{j \in V}(P(v_j) \prod_{ai \in X}(P(a_i|v_j)))$$



Categorías	# Palabras
Crisis	500
Economía	2500

Business Intelligence (BI)

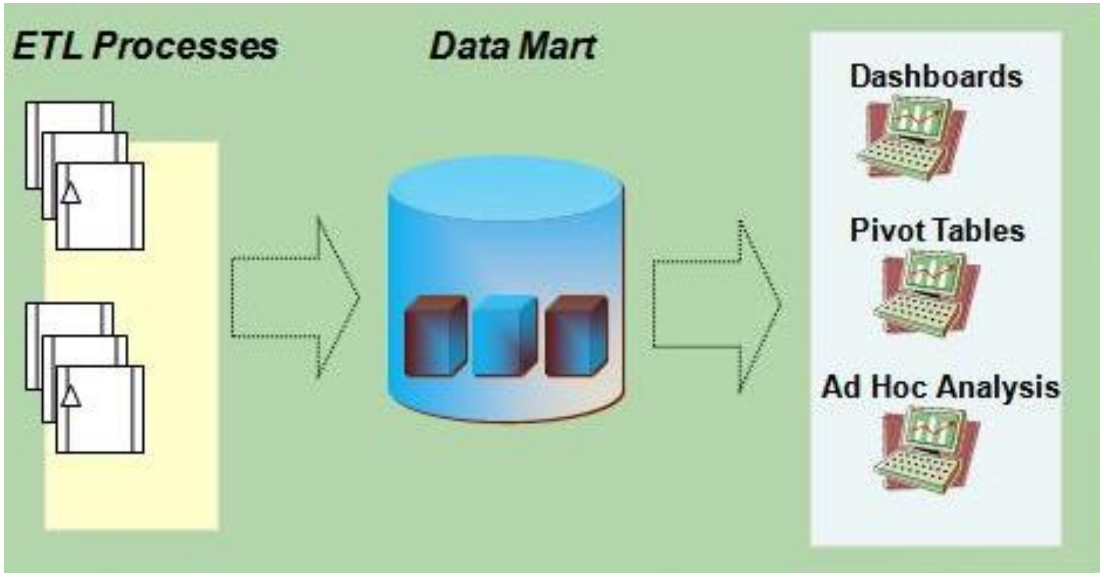
Business Intelligence (BI)



Transformación de gran cantidad de datos provenientes del ERP a tablas y gráficos fácilmente interpretables

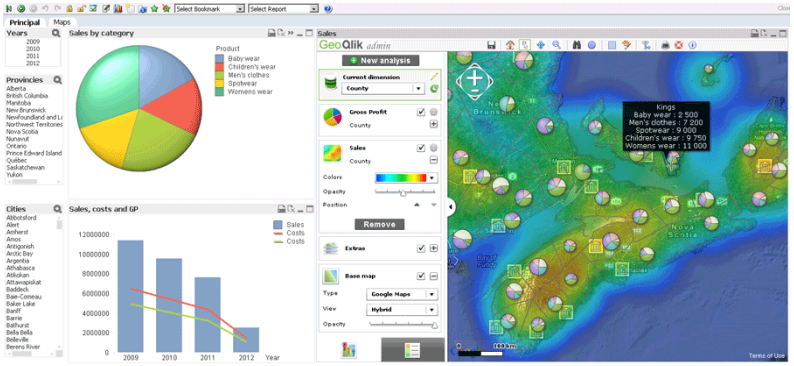
Business Intelligence (BI)

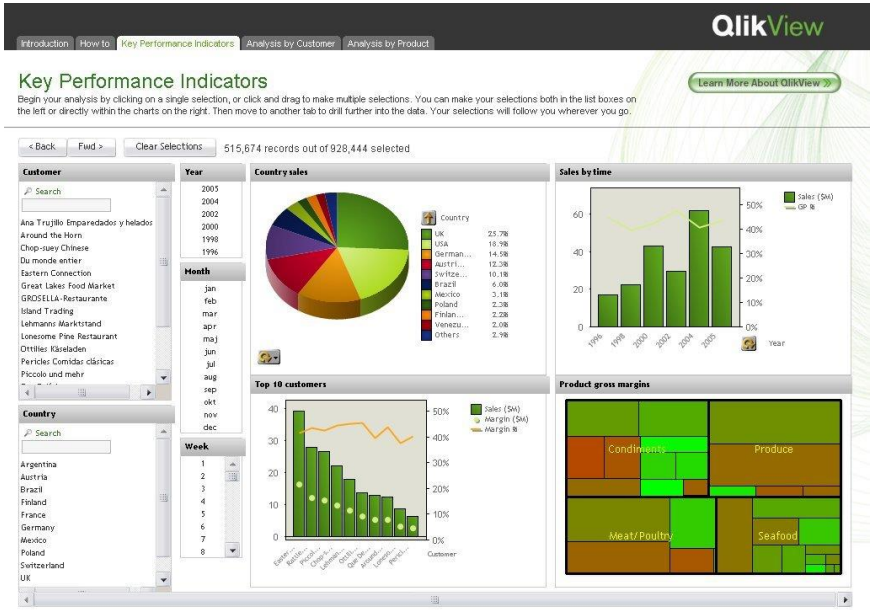
- Proceso de analizar datos de una empresa y extraer conocimiento de ello
- Uso de almacén de información (*Datawarehouse*) como herramienta estratégica
- Habilidad de explorar y analizar datos para revelar la existencia de tendencias



Qlikview

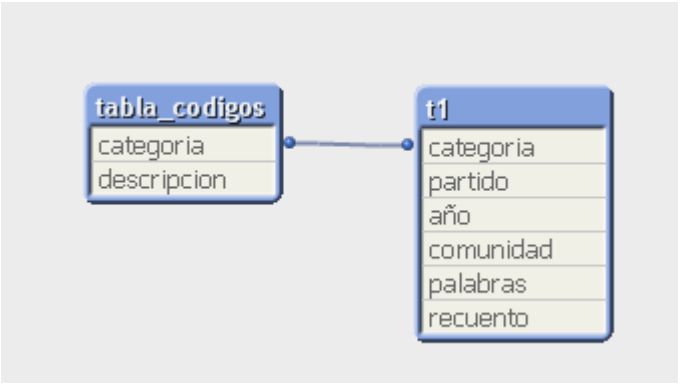
- Es una de las principales plataformas para el Business Discovery
- Aporta un lenguaje propio de modelado de datos
- Utiliza un modelo de datos asociativo que se carga en memoria



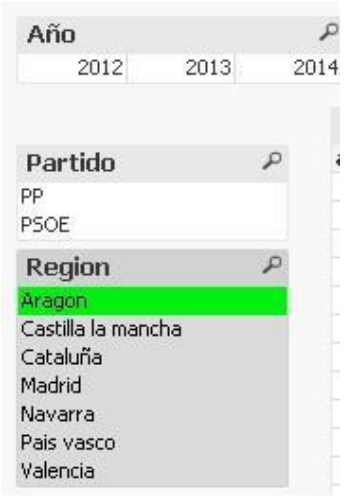


Qlikview (2)

- ETL

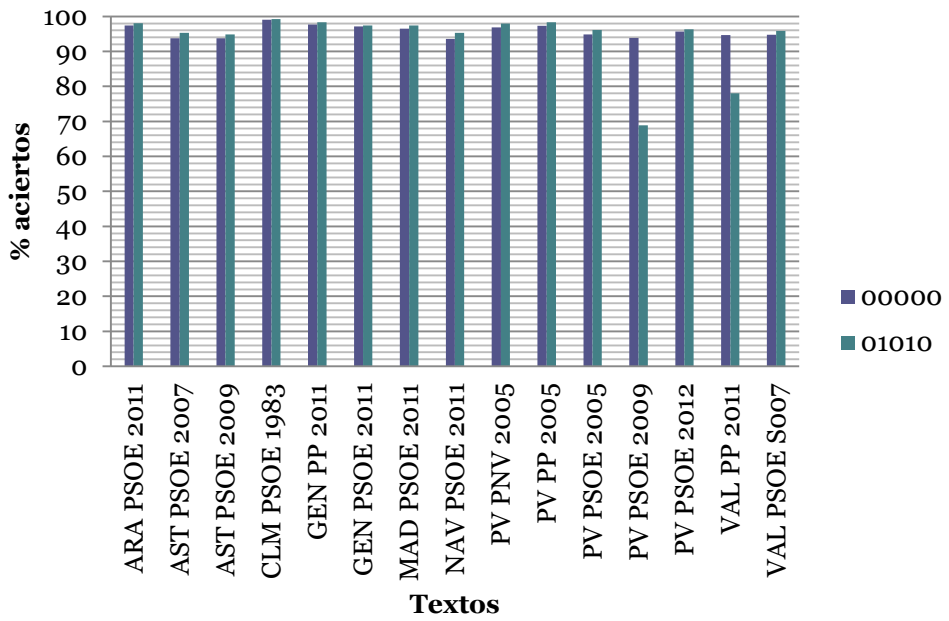


- Selección



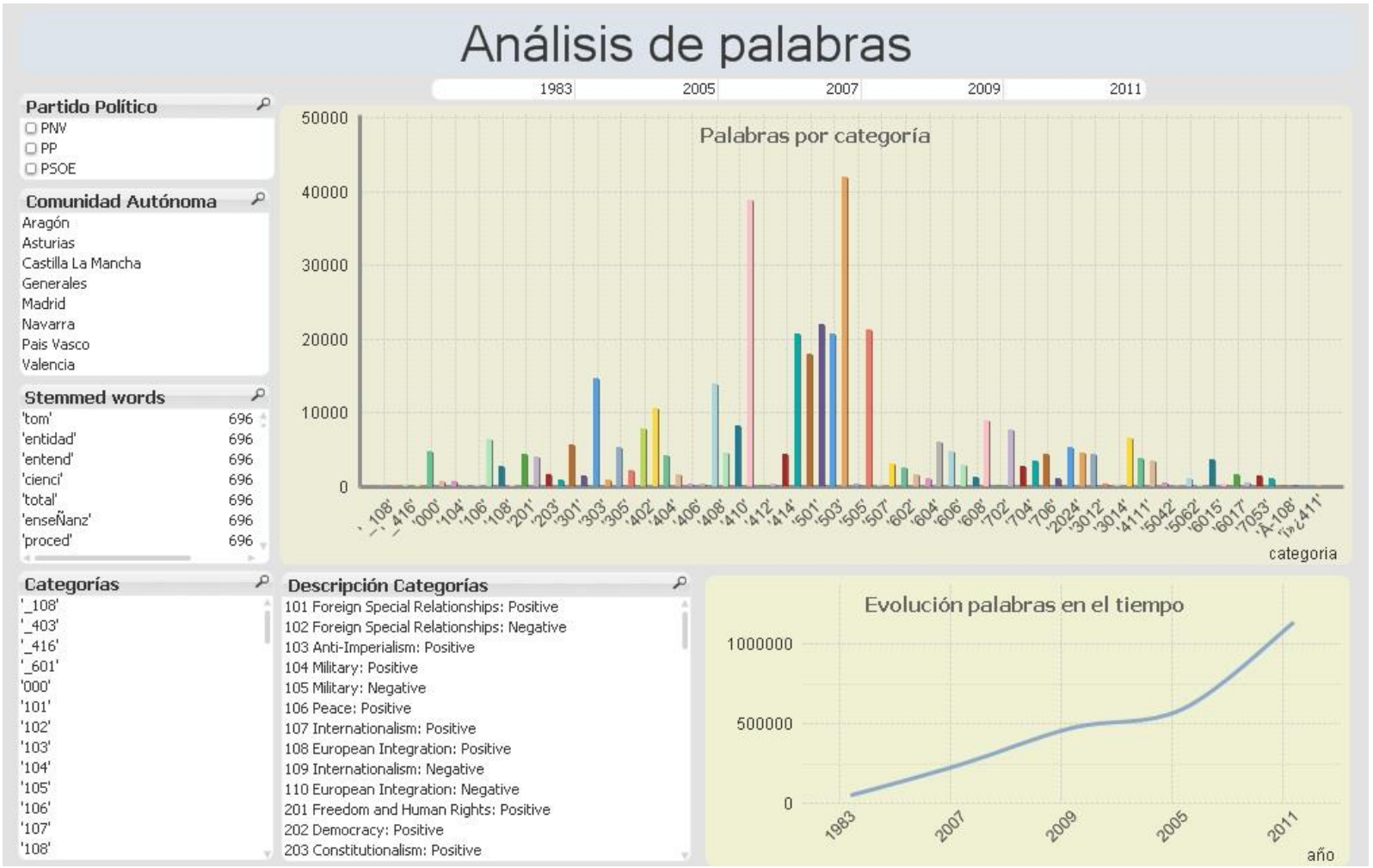
Resultados clasificación

Mejores resultados para la configuración 01010 stop words + sum(log)



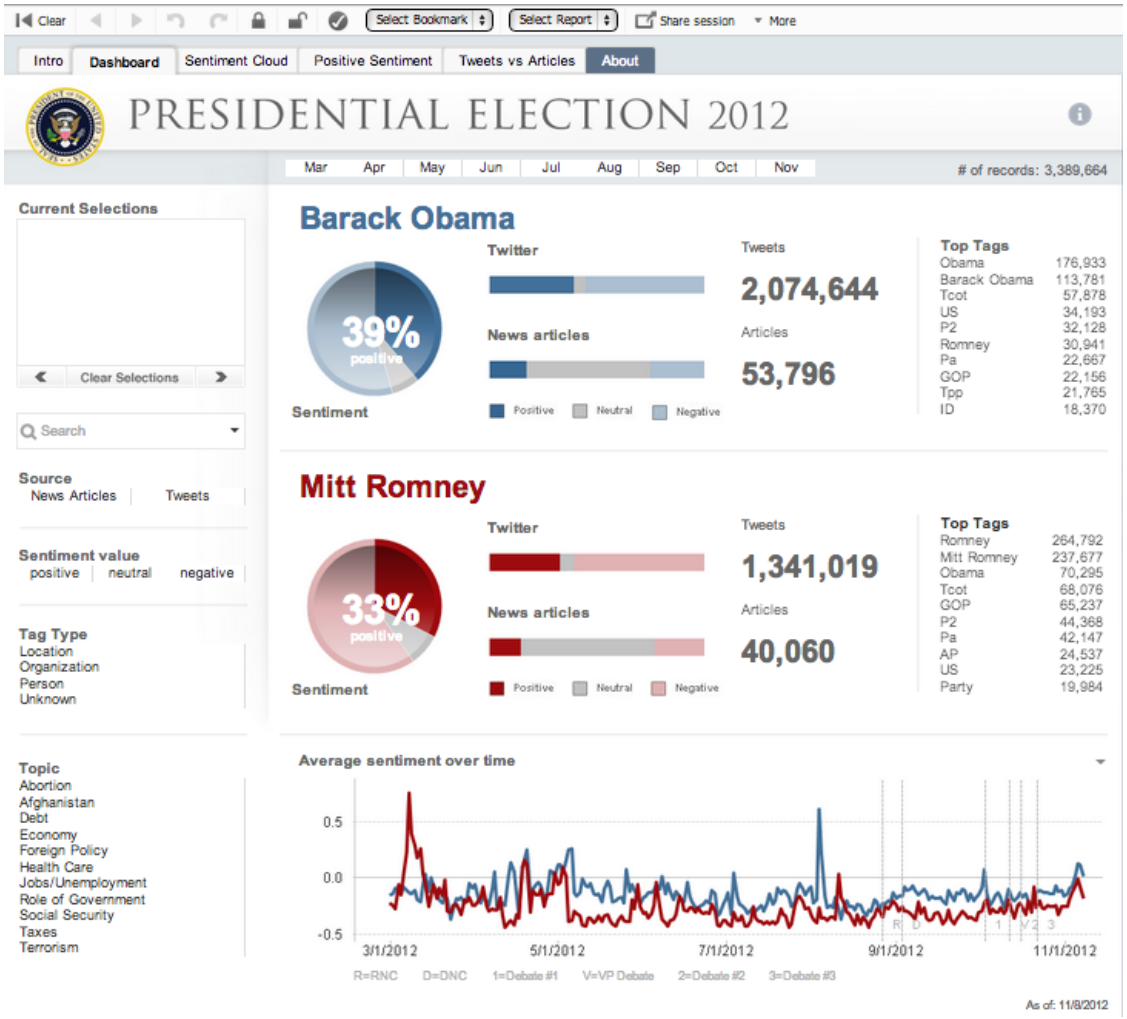
Texto	% acierto 01010
ARA PSOE 2011	98,06
AST PSOE 2007	95,35
AST PSOE 2009	94,87
CLM PSOE 1983	99,28
GEN PP 2011	98,28
GEN PSOE 2011	97,38
MAD PSOE 2011	97,39
NAV PSOE 2011	95,31
PV PNV 2005	97,98
PV PP 2005	98,3
PV PSOE 2005	96,16
PV PSOE 2009	68,86
PV PSOE 2012	96,35
VAL PP 2011	78,03
VAL PSOE S007	95,86

Dashboard


Análisis de palabras
1983
2005
2007
2009
2011
Partido Político
PNV
PP
PSOE
Comunidad Autónoma
Aragón
Asturias
Castilla La Mancha
Generales
Madrid
Navarra
Pais Vasco
Valencia
Stemmed words
'tom' 696
'entidad' 696
'entend' 696
'cienci' 696
'total' 696
'enseÑanz' 696
'proced' 696
Palabras por categoría
50000
40000
30000
20000
10000
0
categoria
Categorías
'_108'
'_403'
'_416'
'_601'
'000'
'101'
'102'
'103'
'104'
'105'
'106'
'107'
'108'
Descripción Categorías
101 Foreign Special Relationships: Positive
102 Foreign Special Relationships: Negative
103 Anti-Imperialism: Positive
104 Military: Positive
105 Military: Negative
106 Peace: Positive
107 Internationalism: Positive
108 European Integration: Positive
109 Internationalism: Negative
110 European Integration: Negative
201 Freedom and Human Rights: Positive
202 Democracy: Positive
203 Constitutionalism: Positive
Evolución palabras en el tiempo
1000000
500000
0
1983
2007
2009
2005
2011
año

Planificación

Propuesta continuación



Conclusiones

- El tiempo invertido en la clasificación automática es prácticamente nulo y se debe considerar como la principal ventaja.
- Altos índices de acierto en la clasificación, llegando hasta 99.2%
- Cuadro de mando preparado para datos formateados, permite estandarizar. Simplemente al cargar nuevos datos se dispone de información ya procesada para analizar con eficacia.

Mejoras

- Mejorar el porcentaje de acierto (a través de muchos más datos de entrenamiento)
- Multi-idioma (actualmente sólo castellano)
- Mejorar el método de stemming
- Añadir palabras a la lista de stop words
- Mejorar el tiempo de clasificación
- Realizar un estudio a medida de las necesidad para mostrar en Qlikview para explotar al máximo la información proveniente de los datos

Gracias