TÉCNICAS DE ANÁLISIS AUTOMÁTICO DE TEXTOS CON RUIDO

APLICACIÓN A PRENSA HISTÓRICA ESPAÑOLA DURANTE LA TRANSICIÓN DEMOCRÁTICA (1977-1982)

Carlos G. Figuerola [figue@usal.es]

Universidad de Salamanca, 2019

MINERÍA DE TEXTOS Y PRENSA HISTÓRICA

- *Topic Modeling* y otras técnicas de minería de textos se utilizan con frecuencia con textos históricos
- Esto tiene sentido si se trabaja con grandes cantidades de texto
- Periódicos digitalizados son una fuente histórica importante

ALGUNOS PROBLEMAS EN LA DIGITALIZACIÓN DE PERIÓDICOS

- Generalmente el proceso de digitalizacón consiste en pasar por un *scanner* las páginas de los periódicos
- Esto requiere un programa de OCR que nos permita acceder al texto
- El OCR tiene que trabajar con tipografías antiguas, papel degradado, plantillas de páginas …
- El objetivo de este trabajo es comprobar funcionan bien con textos con *ruido*

EL ABC

- se eligió el periódico ABC para experimentar con ténicas de *topic modeling*
- ABC se edita desde 1903
- todos los ejemplares desde entonces han sido digitalizados
  - la digitalización consiste en pasar por un escáner, luego un OCR, etc.
- en ABC encontramos plantillas en columnas, diversas tipografías, etc.

Los otros dos ex carcelados son Diego Elorrieta Derio, de cincuenta y cinco años de edad, casado, con una hija. Fue condenado en septiembre de 1973 a doce años de prisión, y Pedro Aulestiaondarroa, natural de Ondárroa, de veinte años de edad, estudiante, que fue condenado el 18 de julio de 1975. Este detenido había cumplido a lo largo de su estancia en prisión ciento cuarenta días en celda de castigo.

Continúan saliendo reclusos

Con la aplicación de las nuevas medidas, en la noche de ayer eran ya 296 los presentencias.

Siempre según estas mismas fuentes, próximas a los abogados defensores, en el caso de que la amnistía no sea aplicada de oficio por los Tribunales Militares, los interesados pedirán a la Audiencia Nacional que las aplique.

Con relación a los que se encuentran en prisión por estar relacionados con los secuestros, no está claro si se podrá aplicar la amnistía, al haber comenzado éstos antes de la fecha del referéndum; pero, de no ser estimado así, los abogados recurrirán al Tribunal Supremo.

EL DOMINGO

"IZQUIERDA DEMOCRATICA" Y LA "F. P. D." PUEDEN FIRMAR UN PACTO DE FEDERACION

Madrid. (De nuestra Redacción.) La «Federación Popular Democrática» e «Izquierda Democrática» pueden llegar a la constitución de una federación de partidos, con presidencia, secretariado general y ejecutiva comunes. Esta federación tendría carácter transitorio, en tanto los respectivos congresos, y posteriormente un Congreso Constituyente, decidan sobre la unidad de ambos partidos demócrata cristianos.

La confirmación de la unidad se retrasará hasta después de las elecciones, por considerar «Izquierda Democrática» que no queda tiempo material para realizar todo el proceso de integración con anterioridad a los comicios. Sin embargo, desde el momento mismo en que se apruebe el acuerdo de federación entrarán en funcionamiento los órganos comunes y de hecho funcionará como un único partido.

Por el momento, según fuentes democristianas, ha sido elaborado ya un documento base en el que se regula el funcionamiento de los órganos comunes transitorios.

EL «P. P. D. C.».—La operación de unidad podría extenderse igualmente hacia el «Partido Popular Demócrata Cristiano» y la «Unión Democrática Española». De hecho, el primero de estos partidos acudió a la reunión conjunta de las ejecutivas de «Izquierda Democrática» y «Federación Popular Democrática», pero su representante en la reunión, Iñigo Cavero, no tenía facultad decisoria y asistió como observador.

El «Partido Popular Demócrata Cristiano» mantiene paralelamente conversaciones orientadas a la fusión con «Unión Democrática Española». Por el momento se han redactado varios borradores de trabajo, pero no parece haberse concretado nada. Fernando Alvarez de Miranda, consultado por A B C, se remitió al comité político que se reúne el próximo domingo en Oviedo, y en el que se informará de las gestiones realizadas hasta el momento.

FIRMA DEL PACTO.—El próximo domingo, en una reunión que mantendrán en el Colegio Mayor San Pablo el Comité federal del «F. P. D.» y la ejecutiva de «Izquierda Democrática», se espera que se firme el acuerdo de federación y el compromiso de convocar sus respectivos Congresos para aprobar la integración.

Anteriormente, el sábado, se reunirá el Comité federal de la «Federación Popular Democrática» para debatir y aprobar el pacto de federación.

ABC, RESULTADO DE UN OCR

SANCHO D AVILA, 19 - T«!s, 245 80 43 - 246 20 9 8
El «Partido Popular Demócrata CristiaMADRID-28
no» mantiene paralelamente conversaciones orientadas a la fusión con
«unión Democrática Española». Por el momento se
han redactado varios borradores de trabajo, pero no parece haberse
concretado líada. Fernando Alvarez de Miranda, consultado por ABC, se
remitió al comité político que se reúne el próximo domingo en
Oviedo, y en el que se Informará de las
gestiones idealizadas hasta el momento.
FIRMA DEL PACTÓ.—El próximo doa diez minutos de la nieve, edificable
mingo, en una reunión que mantendrán en
Precio total: 747.000 pesetas
el Colegio Mayor San Pablo el Comité federal del «P.P.p.» y la ejecutiva
de «IzFACILIDADES
quierda Democrática», se espera que se firme el acuerdo de federación y
el comproTeléfonos 253 56 64 y 254 58 86
miso de convocar sus respectivos Congresos para aprebar la Integración.
Anteriormente, el sábado, se reunirá el P I S O E N F U B N T E R R A B
1 A
GóMt! ÍNléral de la «Badéraclóa Popular ©áÜício Mlramar Torreón, todo
exterior.
232 m2 útiles. Info»na<íloa en Madrid,
D^b&ráliél^ Q$m debatir y aprobar tí
Teiéfoiio 21$ 28 77. Tardes.
pací» dé fgáeráácl6n.

EL PAÍS COMO LÍNEA DE COMPARACIÓN

- El Pais comenzó 1976
- tiene una hemeroteca digital
  - realizada transcribiendo o convirtiendo cada noticia individual
  - no se ha utilizado escáner ni OCR
  - es navegable por internet

LAS NOTICIAS EN LA HEMEROTECA DE EL PAÍS

Paro total en Renfe

EL PAÍS

22 SEP 1977

A partir de las ocho de la mañana de hoy quedarán paralizados todos los trenes del país, durante dos horas, por huelga de los trabajadores de Renfe. Los trenes en circulación se detendrán, según las previsiones de los huelguistas, en la estación inmediata al punto donde se encuentren a la hora de inicio del paro.El Pleno General de Ferroviarios, organismo de representación unitaria de los trabajadores que ha convocado el paro, ha constituido durante los últimos días distintos comités de huelga, para que el desarrollo de la acción de protesta sea ordenado y repercuta en la menor medida posible sobre la cobertura de seguridad de Renfe.

La empresa, que a última hora de la mañana de ayer aumentó su oferta a los trabajadores con la intención de que se desconvocara el paro, resporisabilizar a los comités de huelga de la seguridad de la circulación ferroviaria, y califica de *totalmente desproporcionada* con as divergencias existentes la protesta

EL PERÍODO DE LA TRANSICIÓN

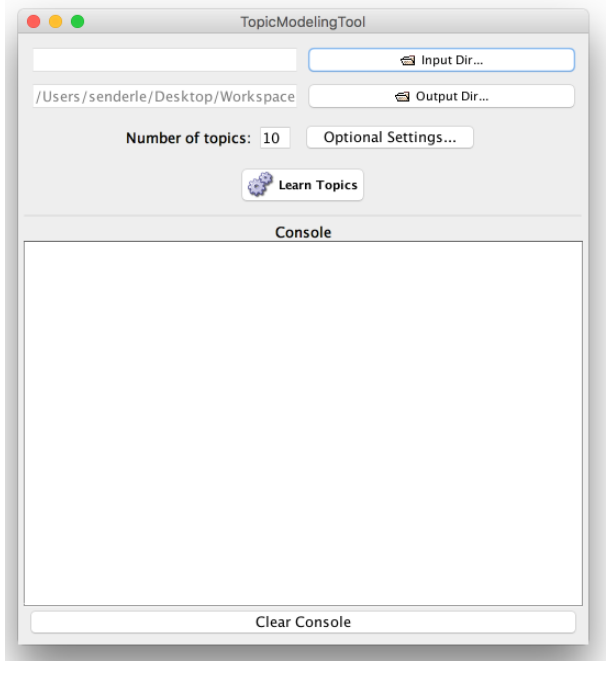
- podemos considerar que va de 1977 a 1982
- están disponibles las colecciones completas de ABC y El País para este período
- la comparación entre ambos periódicos puede darnos una idea de las posibilidades de estás técnicas en textos degradados (ABC)

TOPIC MODELING

- técnicas que tratan de identificar (y cuantificar) los temas tratados en una colección de documentos
- Latent Dirichlet Allocation es una de las más frecuentes
  - en una colección de documentos existe un conjunto de temas
  - cada documento tiene una cantidad (%) concreta de cada tema
  - LDA produce una serie de términos que definen cada tema, también el porcentaje de cada tema en cada documento

TOPIC MODELING

existen programas fáciles de usar que aplican estas técnicas (Mallet y otros)



MALLET
machine learning for language toolkit

gensim
topic modelling for humans

TOPICS DE EL PAÍS

agua aceite productos agricultura agricultores medio enfermedad zonas
bank chemical valores índice mercado steel ltd bolsa
política economía precios crisis crecimiento empresas mercado inflación
soviética unidos urss israel moscú guerra china paz
millones pesetas crédito empresa economia bancos dinero valor
internacional presidente unidos argentina militar carter reagan británico
trabajadores huelga empresa economia convenio sindicatos paro acuerdo
francia gobierno europa relaciones internacional exteriores otan francés
petróleo energía nuclear dólares producción mercado industria precio

CANTIDAD DE CADA TOPIC EN CADA NOTICIA

filename	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic
pais19770104-064	0,01 %	0,02 %	0,03 %	1,34 %	0,01 %	0,01 %	4,59 %	
pais19770105-060	0,11 %	31,53 %	0,20 %	0,10 %	0,11 %	0,08 %	0,48 %	
pais19770107-027	0,10 %	0,16 %	0,18 %	0,09 %	0,10 %	0,07 %	2,22 %	
pais19770111-047	23,04 %	0,10 %	0,12 %	0,06 %	0,06 %	0,05 %	2,58 %	
pais19770111-079	10,55 %	20,61 %	0,05 %	0,02 %	0,03 %	0,02 %	14,47 %	
pais19770112-132	10,13 %	0,03 %	0,59 %	0,01 %	0,30 %	0,01 %	1,75 %	
pais19770113-029	0,04 %	20,06 %	2,39 %	0,04 %	0,04 %	0,03 %	9,42 %	
pais19770114-021	0,04 %	0,07 %	0,08 %	0,04 %	0,04 %	0,03 %	0,19 %	
pais19770115-062	0,03 %	12,86 %	0,06 %	0,03 %	0,03 %	0,03 %	0,76 %	
pais19770118-112	0,03 %	24,80 %	0,05 %	0,02 %	0,03 %	0,02 %	0,12 %	
pais19770202-025	0,05 %	1,05 %	0,10 %	0,05 %	0,05 %	0,04 %	1,19 %	
pais19770206-137	0,01 %	1,53 %	7,34 %	0,01 %	0,01 %	0,01 %	0,03 %	
pais19770208-123	0,06 %	0,09 %	0,10 %	0,05 %	0,05 %	0,04 %	0,24 %	
pais19770209-035	0,07 %	0,11 %	0,13 %	0,06 %	2,53 %	0,05 %	26,22 %	
pais19770209-149	0,01 %	0,02 %	0,02 %	5,88 %	0,01 %	0,01 %	6,79 %	
pais19770222-052	1,28 %	9,43 %	0,06 %	0,03 %	0,03 %	0,03 %	0,15 %	
pais19770226-158	4,67 %	5,79 %	0,02 %	0,01 %	0,01 %	0,01 %	0,23 %	
pais19770227-157	0,12 %	0,19 %	0,21 %	0,11 %	0,11 %	0,09 %	17,16 %	
pais19770302-109	15,43 %	0,08 %	0,09 %	0,04 %	0,05 %	0,04 %	0,21 %	

TOPICS EXTRAÑOS EN ABC

interna necesita niños casa sueldo informes cocina teléfono llamar madrid playa

mercedes estrenar extras bmw particular completo nuevo automóviles

metros dormitorios piso garaje piscina facilidades zona particular

ANUNCIOS POR PALABRAS EN ABC

GENERALISIMO, vacío, amueblado estrenar. 2503348.
ARGÜELLES, amueblado vacío, confortabilísimo. 2504295.
OFICINAS con teléfono. 2500942.
SALAMANCA, amueblado, desamueblado, elegante. 2595840.
¡¡¡**AGUAYO**!!! English Française, apartamentos chalet, oficinas, pisos. 2500942.
DIRECTAMENTE. Apartamentos Rueda, lujosísimos. Zona Castellana. Servicios incluidos. Teléf. 4469777.
ALMAGRO, 10. Apartamentos «Andrómeda». Estrenar, climatizados, alquileres semanales-mensuales, todos servicios. 4470900.
PRECIOSOS apartamentos climatizados, zona glorieta Bilbao. Servicio. Alquileres semanales - mensuales. Velarde, 9. 2328920.
CASTELLANA, 56. Importante piso, 90.000, dos plazas garaje, 400 metros. 2223133.
ROSALES, 28, oficinas, 50.000, superconfortabilísimas. 2224833.
RESIDENCIAL, chalet, 85.000. Aravaca. 2319290.
AMUEBLADO, 40.000, superconfortable. 2320644.
RESIDENCIAL, vacío, 40.000. 2223133.
HABANA, 140.000. Oficinas, quince habitaciones, 470 metros. 2500942.
ARAPILES, 40.000, amueblado, cuatro dormitorios. 2503348.
CIUDAD Periodistas, 35.000, vacío, 220 metros, [illegible] frente a plaza de España. Facilidad de aparcamiento. Directamente propietario. Teléfonos 2478572 y 2486826.
MONTERA. Pisos, apartamentos oficinas residenciales, vacíos amueblados. 2315690.
ALQUILAMOS oficinas zona Cibeles-Colón. Teléfono, aire acondicionado. Información. 2759190.
SE alquila oficina 330 metros, planta primera, edificio moderno, con teléfono. Calle General Martínez Campos, 9. Información: Teléfono 4552102.
APARTAMENTOS alquiler, amueblados, estrenar, garaje, piscina, aire acondicionado. Alberto Alcocer, 43.
APARTAMENTOS amueblados zona Castellana, servicios parking, sauna, gimnasio, cafetería Edificio Cónsul. Alquilo directamente. Señor Carro. 4191650.
OFICINA amueblada nueva, 4 despachos. María de Molina, 26. Mañanas.
DISPONEMOS oficinas todas zonas, superficies, precios. 2312324.
DIVERSIDAD pisos, chalets, oficinas disponemos. Inmediata ocupación. 2326553.
SOL. Mayor, 1, segundo. Oficinas y locales.
APARTAMENTOS semanales desde 4.200, mensuales desde 16.800, servicios incluidos. Vargas, 3 (esquina General Sanjurjo) 4412500.
OFICINAS. Salesas. Ocho **PUERTA** Hierro, alquilo piso amueblado, 4 dormitorios, 3 baños, salón, comedor separado, despacho, hermosa vista, octava planta, Arroyo Fresno, 26, cerca hotel Monterreal, razón portería. 2793552.
APARTAMENTOS amueblados de lujo, zona plaza España, espaciosos, luminosos, climatizados, TV., servicios completos incluidos. Véalos: Onésimo Redondo dieciséis. 2425900.
OFICINAS, teléfono garaje propio, laborables Cea Bermúdez, 14.
OFICINAS, inmejorables zonas. 2623726. Baikal.
GENERALISIMO, vacío, lujo, cuatro dormitorios, garaje, orientación Mediodía. 60.000. 6501498.
PEQUEÑO, amueblado lujosísimo, gran terraza, chimenea. 2599521.
¡¡¡**OFICINAS**!!! Desde 150 m2. hasta edificios de 15.000 m2. Exclusivamente zonas residenciales, Uve y Eme «Boutique Inmobiliaria». Castelló, 41. Teléfonos 2755511 - 62.
APARTAMENTO junto Puerta del Sol, habitación, salón amplísimo, cocina, baño, dos terrazas más invernadero, de 10 m2 propio pintores, escultores artesanos. 2667705.
APARTAMENTO con garaje. 31.000. 4197101.
ATICO. Salamanca, cuatro dormitorios. 4488174.
PISO Salamanca. 4022838.
¡**OFICINA**! General Mola, 100 metros, 22.000 cuadrados, 27.000 pesetas. Teléfono 248 11 81. De cuatro a seis.
OFICINA céntrica moderna, teléfono, aparcamiento. Cuesta Santo Domingo, 11.
OFICINA, 80 metros, mejor tramo paseo Habana, dos amplios despachos, calefacción, servicio. 25.000. 4469801.
APARTAMENTO amueblado, estrenar, próximo paseo Habana, calefacción central. 34.000. 4469802.
CHAMARTIN, vacío, lujo, dos dormitorios, calefacción, agua centrales, teléfono, garaje, 28.000. 4469801.
EDIFICIO oficinas, avenida América, teléfono, 90 metros, 30.000. 2423616.
FLORIDA, chalet amueblado estilo inglés, 140.000. 2422542.
PUERTA Hierro, amueblado, estrenar, muy lujoso, 70.000. 2422542.
MEDICO, consulta lujo, totalmente equipada. Barrio Salamanca. 55.000. 2152342.
OFICINA Viso, 200 m. Renta interesante. 2419957.
VACIOS amueblados Argüelles, Generalísimo, Salamanca, todas rentas. 2419957.
CHALETS Chamartín, Aravaca, Somosaguas. Todas rentas. 2419957.
OFICINAS chalet Chamartín, 400 m., buena renta. 2419957.
OFICINAS Generalísimo, Argüelles, Salamanca. Todas superficies, todas rentas. 2419957.

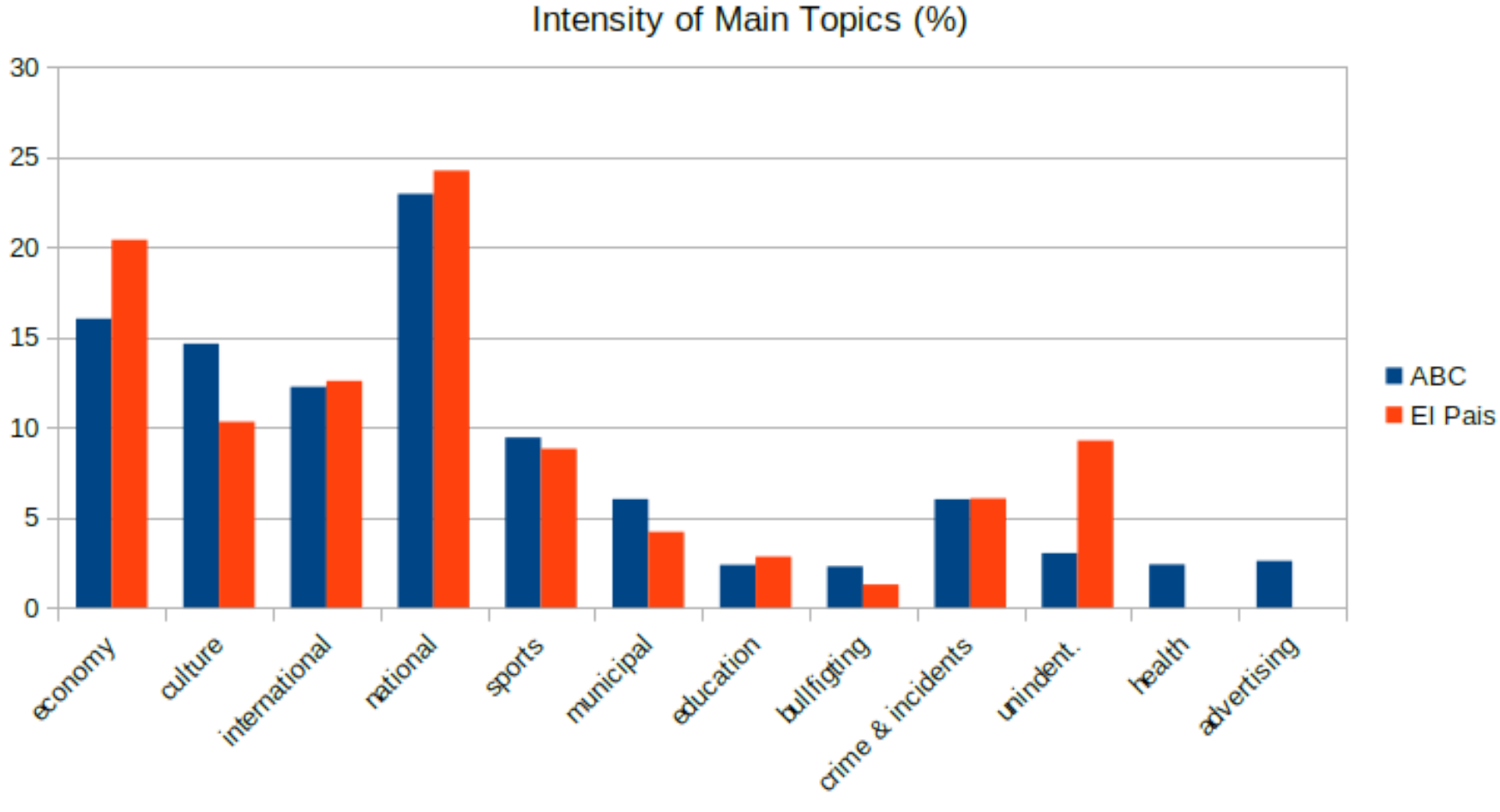
TOPICS EN ABC (2ª VUELTA)

puerto mar avión zona aguas aeropuerto costa pesca acuerdo
millones pesetas economía mercado empresas banco dólares productos
trabajadores huelga empresa acuerdo reunión sindicales comisión gobierno
energía pesetas precios producción consumo toneladas mercado nuclear
banco junta consejo pesetas accionistas administración sociedad acciones
unidos soviética presidente moscú gobierno carter internacional militar
israel presidente ministro gobierno irán paz árabes acuerdo
gobierno partido ministro presidente francia parís elecciones socialista europa

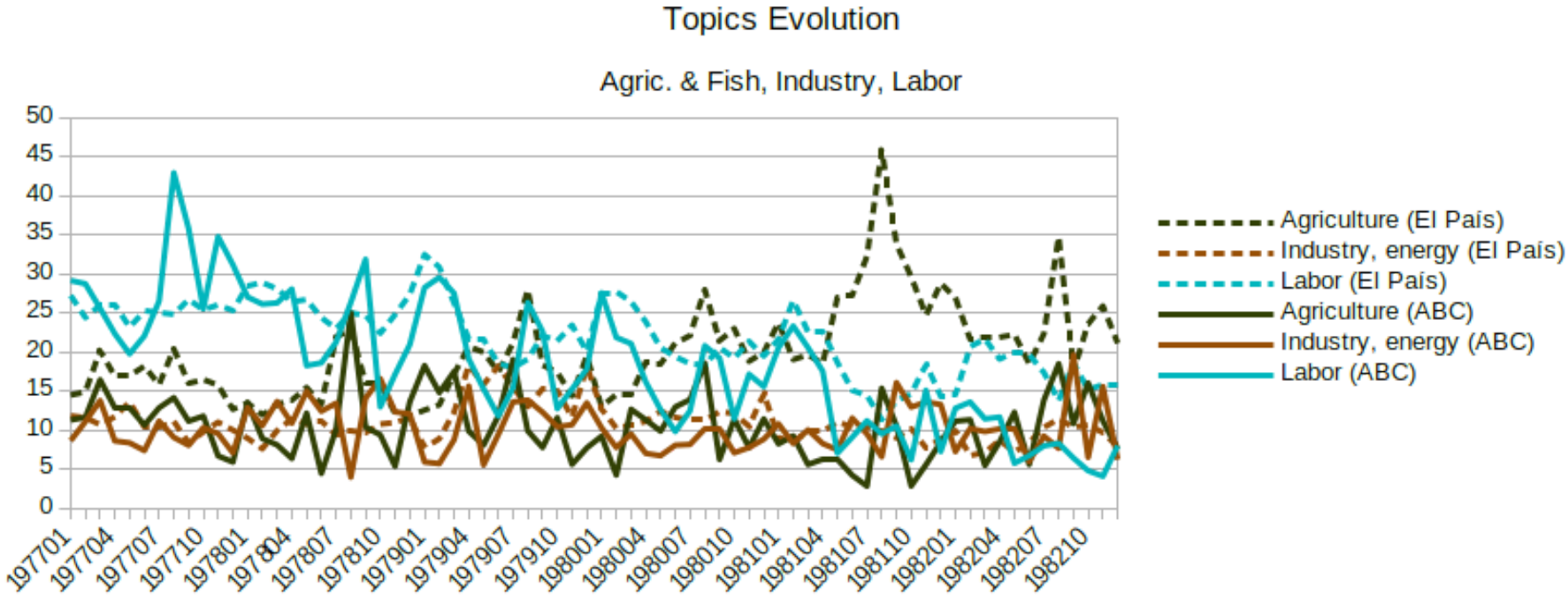
ALGUNAS CIFRAS

	El País	ABC	ABC (2ª vuelta)
1977	41359	33963	15967
1978	40105	32831	14574
1979	41396	33084	14419
1980	42150	33491	14479
1981	40090	33091	14248
1982	40224	37096	16303
Total	**245324**	**203556**	**89990**

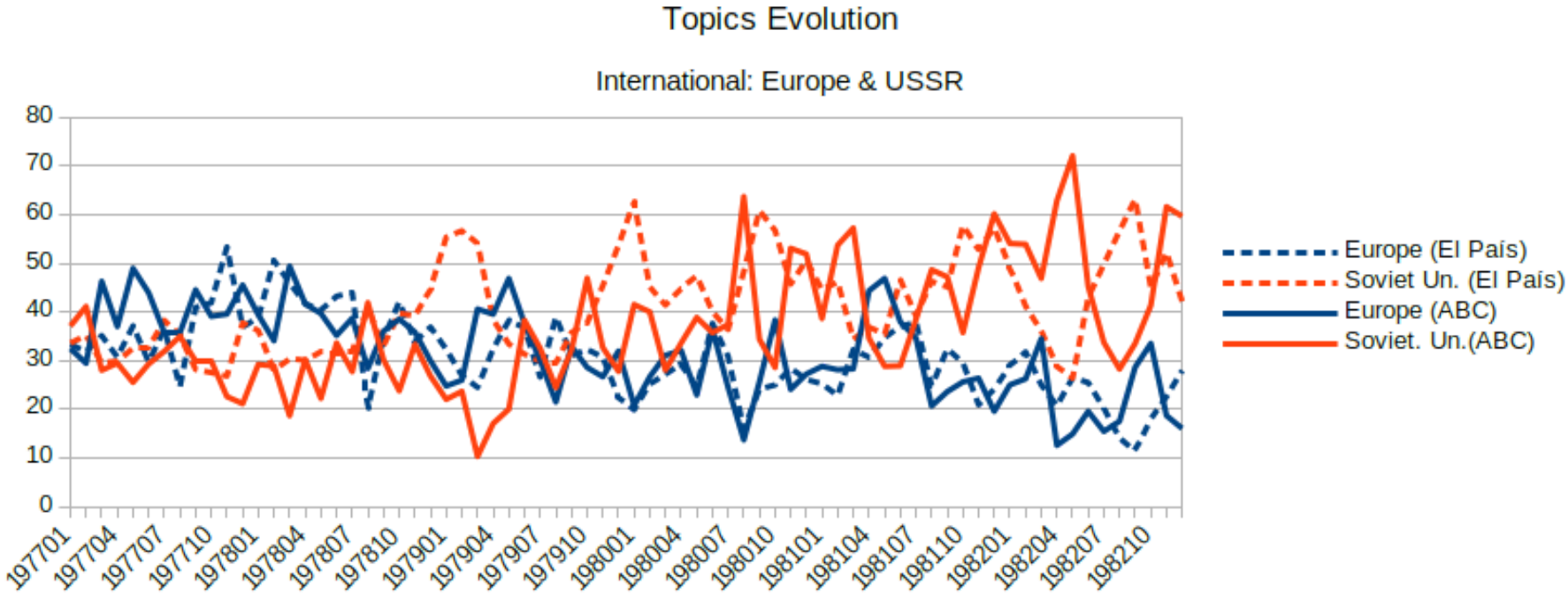
INTENSIDAD DE LOS *TOPICS*



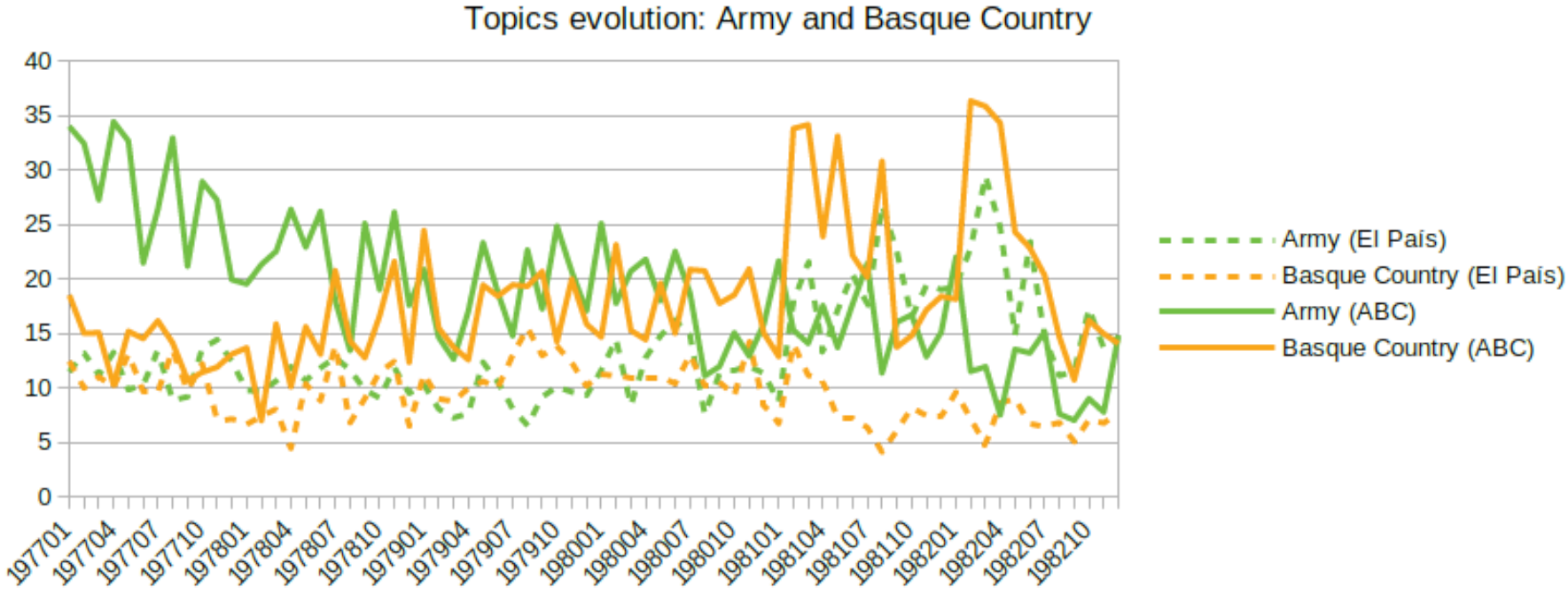
ECONOMÍA


Topics Evolution
Agric. & Fish, Industry, Labor
50
45
40
35
30
25
20
15
10
5
0
197701
197704
197707
197710
197801
197804
197807
197810
197901
197904
197907
197910
198001
198004
198007
198010
198101
198104
198107
198110
198201
198204
198207
198210
Agriculture (El País)
Industry, energy (El País)
Labor (El País)
Agriculture (ABC)
Industry, energy (ABC)
Labor (ABC)

INTERNACIONAL



EJÉRCITO Y PAÍS VASCO



CONCLUSIONES

- hemos aplicado LDA a texto con ruido procedente de páginas de periódico escaneadas
  - la unidad de análisis es la página
  - plantillas irregulares (columnas, anchuras diferentes, etc.)
  - tipografías antiguas, anuncios …
- se han aplicado las mismas técnicas a texto procedente de otro periódico
  - texto plano y limpio
  - la unidad de análisis es la noticia individual
  - no hay publicidad

CONCLUSIONES

- los *topics* son los mismos para ambos periódicos
  - las palabras más significativas son diferentes
- la evolución de la intensidad de los *topics* muestra la parecidas tendencias
  - en algunos casos específicos las tendencias son opuestas
- esas diferencias pueden explicarse por las diferencias ideológicas entre ambos periódicos
- LDA parece una herramienta útil con texto en malas condiciones (utilizando la página como unidad de análisis, en lugar de las noticias individuales)

TÉCNICAS DE ANÁLISIS AUTOMÁTICO DE TEXTOS CON RUIDO

Aplicación a prensa histórica española durante la Transición Democrática (1977-1982)

Carlos G. Figuerola, Universidad de Salamanca

figue@usal.es