Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

PLAN DE IMPULSO DE LAS TECNOLOGÍAS DEL LENGUAJE

Corpus Viewer, una herramienta para el análisis de políticas públicas en I+D y la asistencia a la evaluación de la innovación

Caso de Estudio: Inteligencia Artificial

Secretaría de Estado para el Avance Digital

26 de junio de 2019

Índice

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. ¿Qué es Corpus Viewer?

- Corpus Viewer:
  - Plataforma en producción en la Secretaría de Estado para el Avance Digital (SEAD)
  - Desarrollo desde 2016, ahora en uso por SEAD, FECYT, SEUIDI
  - Incorpora desarrollos en colaboración con grupos de investigación universitarios (UC3M, UPM, UPF, UPV, UAM)
- Plataforma de propósito general. Actualmente contiene fundamentalmente corpus de datos relacionados con la I+D

Fuentes de Datos

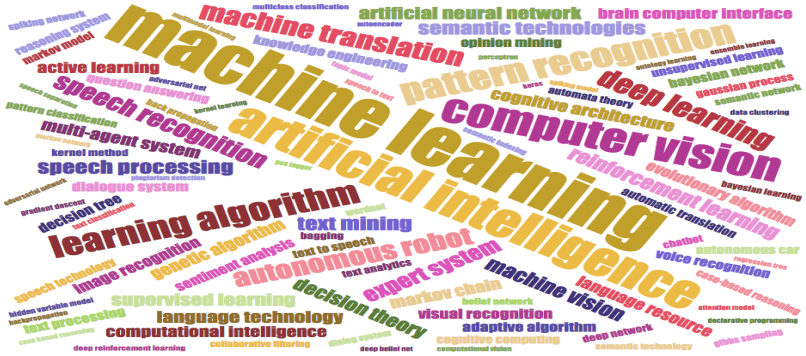
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Corpus	Docs en corpus	Horizonte Temporal
Proyectos del Plan Estatal de I+D+i	110 K	2004 -2016
Proyectos europeos de I+D (CORDIS)	78 K	1984 - 2018
Proyectos estadounidenses de I+D (NSF)	150 K	1985 - 2017
Proyectos estadounidenses de I+D en salud (NIH)	1,8 M	1983 - 2017
Solicitudes de patentes (PATSTAT)	90 M	1898 - 2017
Artículos científicos con contribuciones de autores afiliados a una institución española (SCOPUS)	680 K	2006 - 2018

Fuentes de Datos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- Próximas fuentes a incorporar
  - Artículos científicos: Semantic Scholar (incluye PubMed)
  - Software (Github)
  - Corpus de Ofertas de Empleo
  - Contratación Pública
- Integración con Recursos Open Science
- Generación de subcorpus temáticos
  - Inteligencia Artificial (+ENE, +BIO)
  - Importancia de unificar metodologías



Usuarios

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

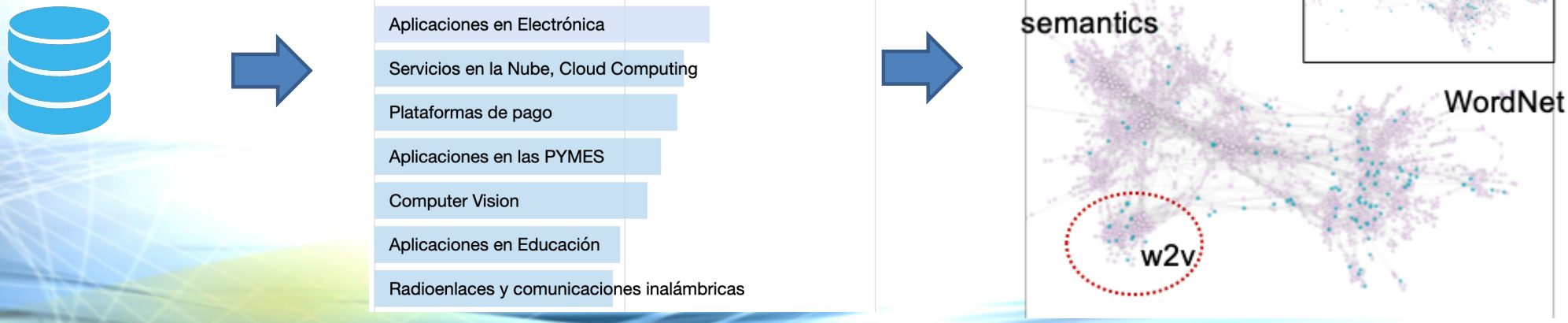
Los perfiles de usuarios de la plataforma incluyen:

- Responsables de políticas de I+D
- Gestores y coordinadores de programas de I+D (implementación de políticas)
- Evaluadores de subvenciones y ayudas
- Investigadores, organismos públicos de investigación, empresas

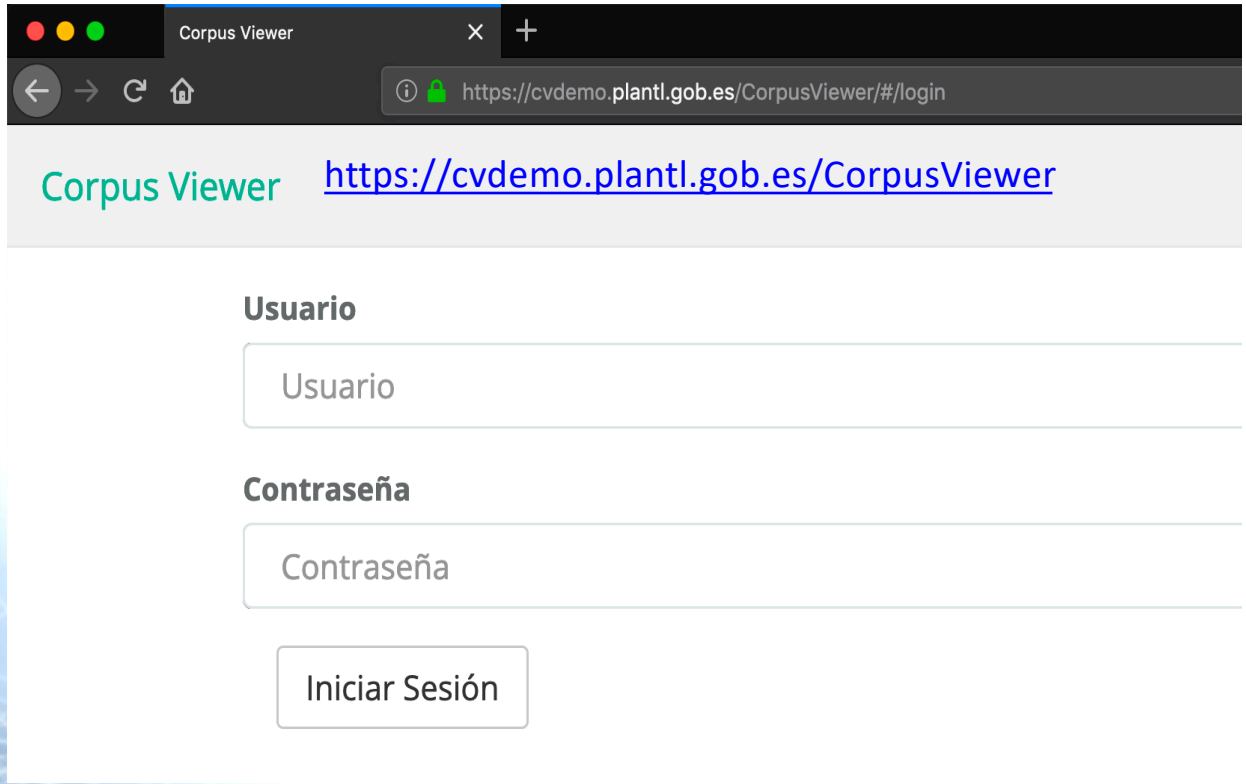
CASO DE USO	Perfil de usuario				
	Decisor	Gestor Ayudas	Evaluador	Investigador	Empresa
DISEÑO POLÍTICAS PÚBLICAS (seguimiento, prospectiva, planificación)	X	X	X	X	X
Herramientas de soporte a EVALUACIÓN (evaluadores, similitud de doc., clasificació taxonomías, estimación innovación)		X	X	X	X
SISTEMA DE ALARMAS (plagio, patrones de fraude)		X	X		X
SISTEMA DE RECOMENDACIÓN (evaluadores,licitación,formación, cruce, conocimiento implícito)	X	X		X	X

2. Técnicas empleadas: Modelado de Tópicos

- Permite detectar las temáticas presentes en un corpus de datos a partir del texto completo de las ayudas, de los artículos científicos, solicitudes de patentes …
- Un tópico puede definirse como “un conjunto de palabras que coocurren” en varios documentos
- Cada documento queda representado como perteneciente a uno o varios tópicos
- Permite hacer seguimiento temporal de dichos tópicos: detectar emergencias, hibridaciones, temáticas en decadencia
- Permite construir una distancia para medir similitudes semánticas entre textos:



3. Corpus Viewer: Instancia abierta



- Corpus incorporados:
  - ACL (publicaciones)
  - CORDIS
  - CORDIS_AI
- Incorporación progresiva de nuevos corpus
- Requiere solicitud de acceso: plantecnologiaslenguaje@mineco.es

4. Caso de Estudio: Inteligencia Artificial

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Vamos a hacer un repaso de las funcionalidades de Corpus Viewer, con el objetivo de ilustrar cómo dar respuesta a las siguientes cuestiones sobre IA:

1. Breve descripción metodológica: Selección de subcorpus IA
2. ¿Qué porcentaje de proyectos financiados emplean IA?

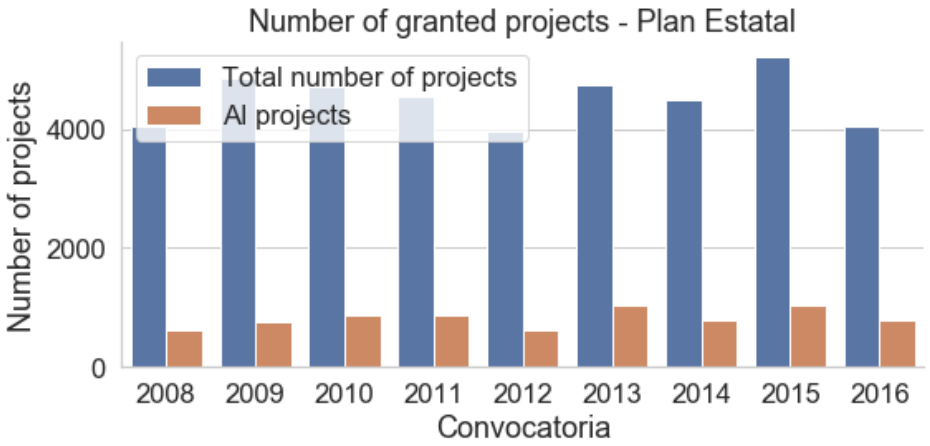
   ¿Existen diferentes evoluciones temporales en los diversos corpus?
3. ¿Cuáles son las principales subáreas temáticas de los proyectos IA?
4. ¿Cuál es la penetración de la IA en otras áreas temáticas?
5. ¿Existen diferencias en cuanto a las áreas temáticas asociadas al IA según áreas geográficas?

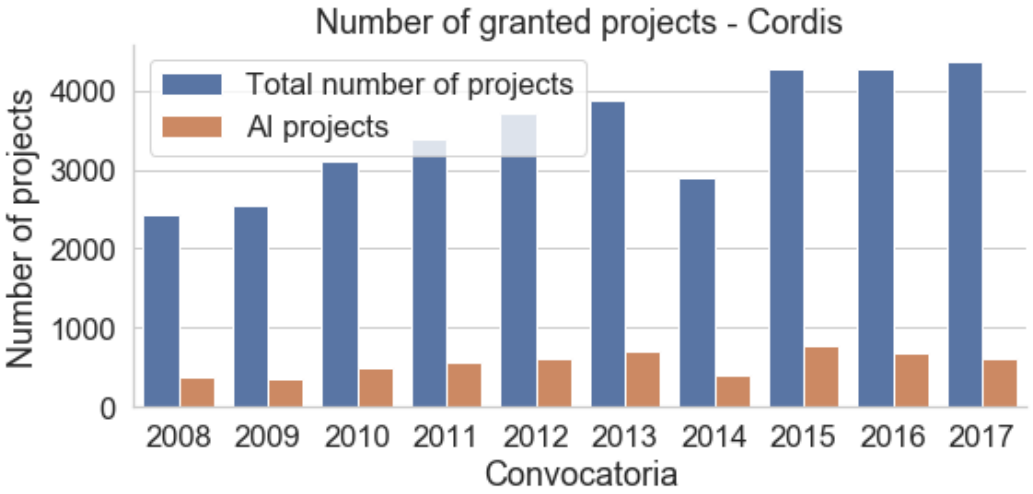
4. Selección del Universo IA

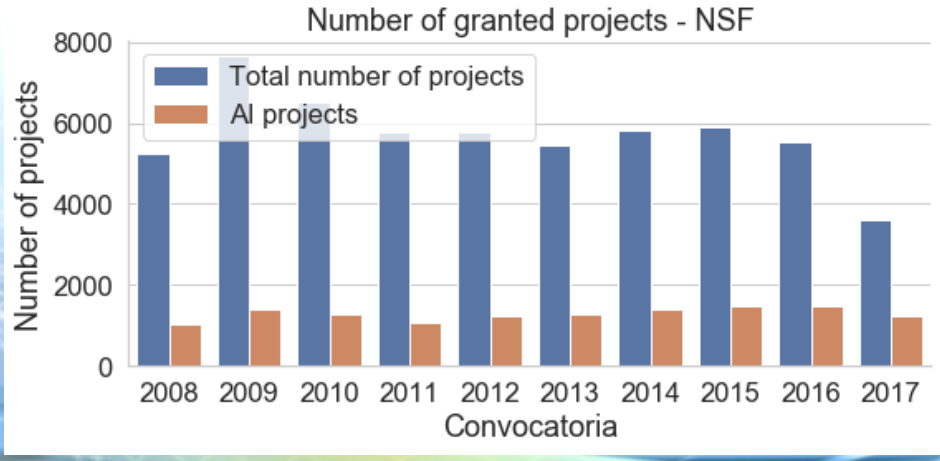
- Preselección en base a palabras clave o metadatos
- Ampliación del Universo con Técnicas de Machine Learning
- Incorporación de expertos en el proceso
- Importancia de unificar metodologías

¿Qué porcentaje de proyectos financiados emplean IA?
¿Existen diferentes evoluciones temporales en los diversos corpus?

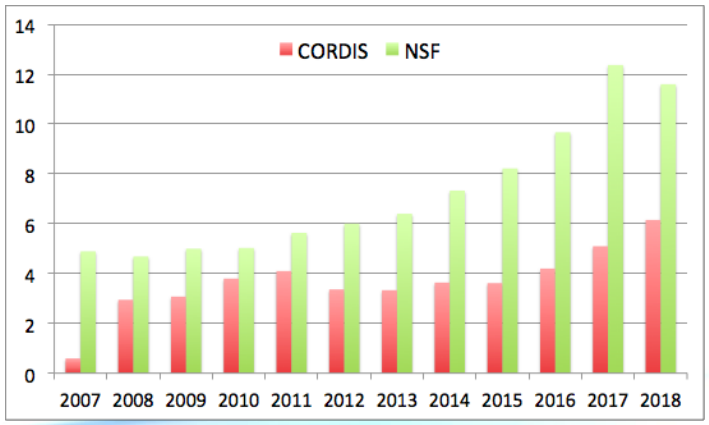
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje







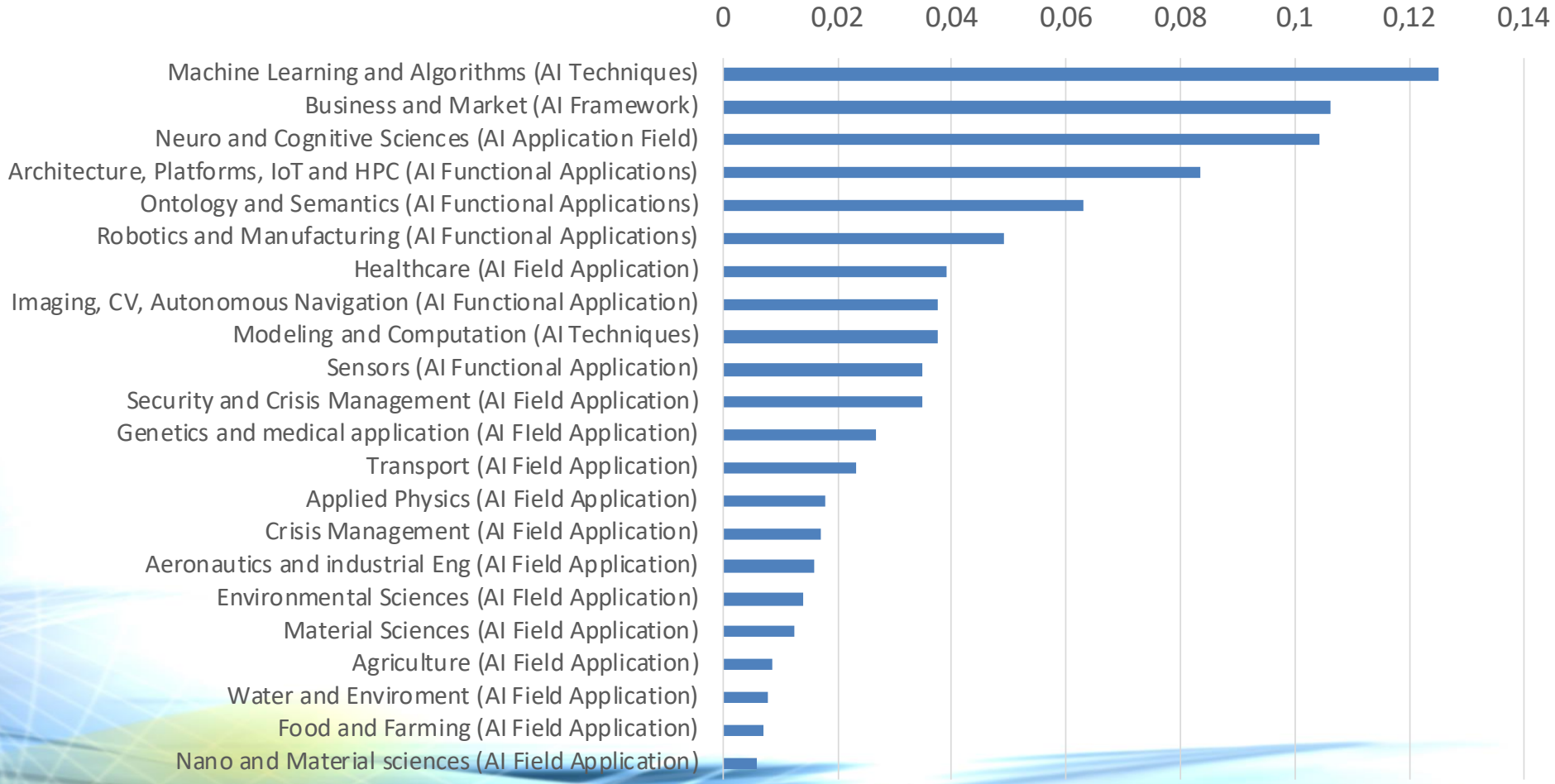
Porcentaje de proyectos con palabras clave



¿Cuáles son las principales subáreas temáticas de los proyectos IA?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

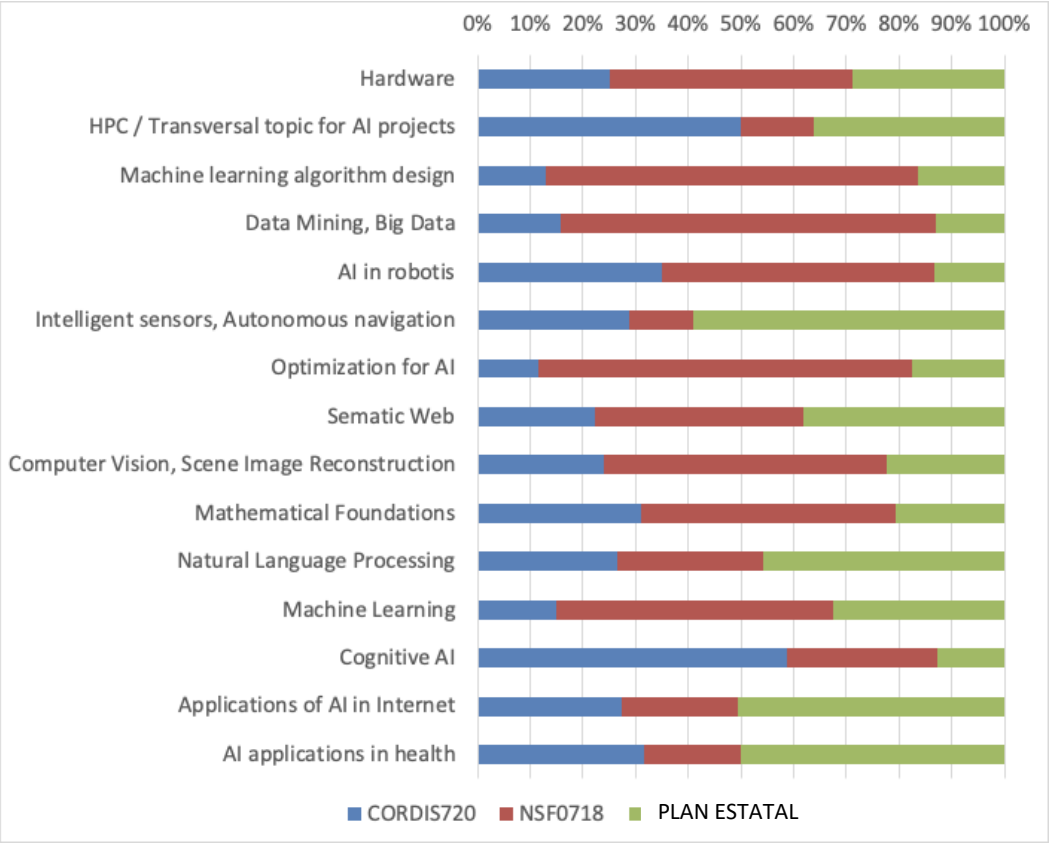
Topic Model for AI selected projects in CORDIS dataset (25 topics)

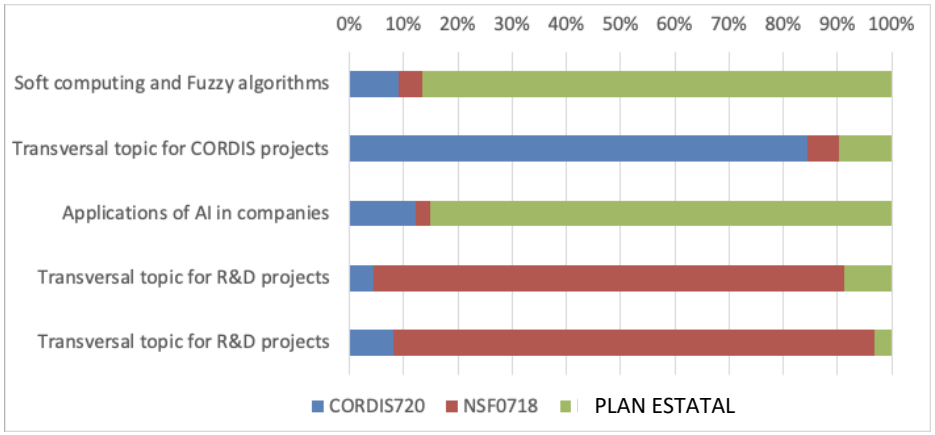


¿Cuáles son las principales subáreas temáticas de los proyectos IA?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- Análisis comparativo entre distintas entidades financiadoras
- Selección conservadora de proyectos (AI core, no campos de aplicación)



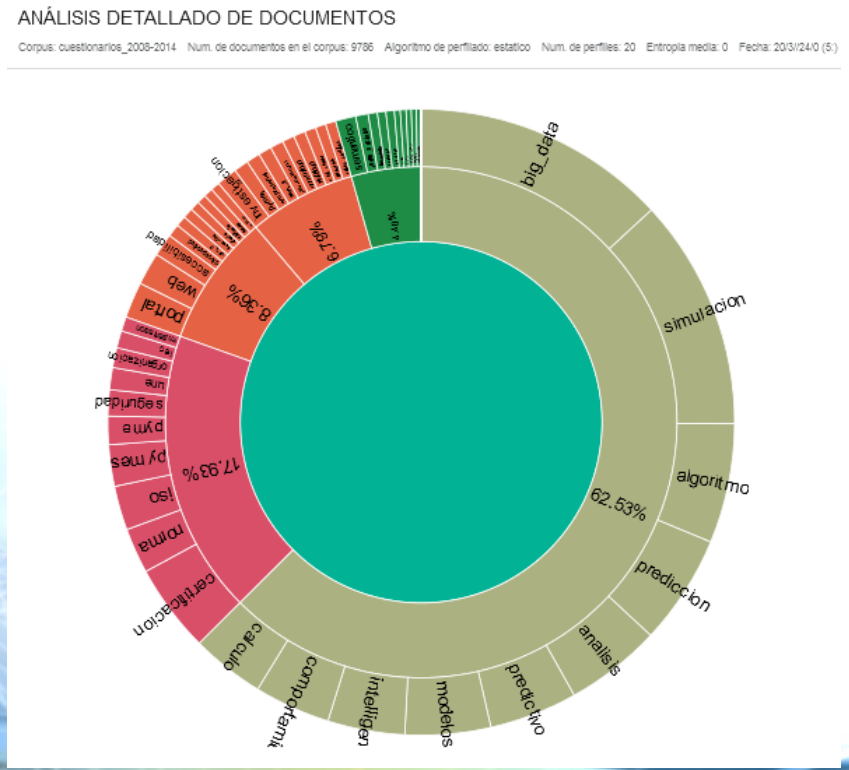


- Muestramos número de proyectos, no importe
- Los tópicos “transversales” suelen ser específicos de corpus
- Corpus Viewer admite modelos multi-corpus

  (requiere selección cuidadosa de datos a incorporar)

Caracterización de documentos

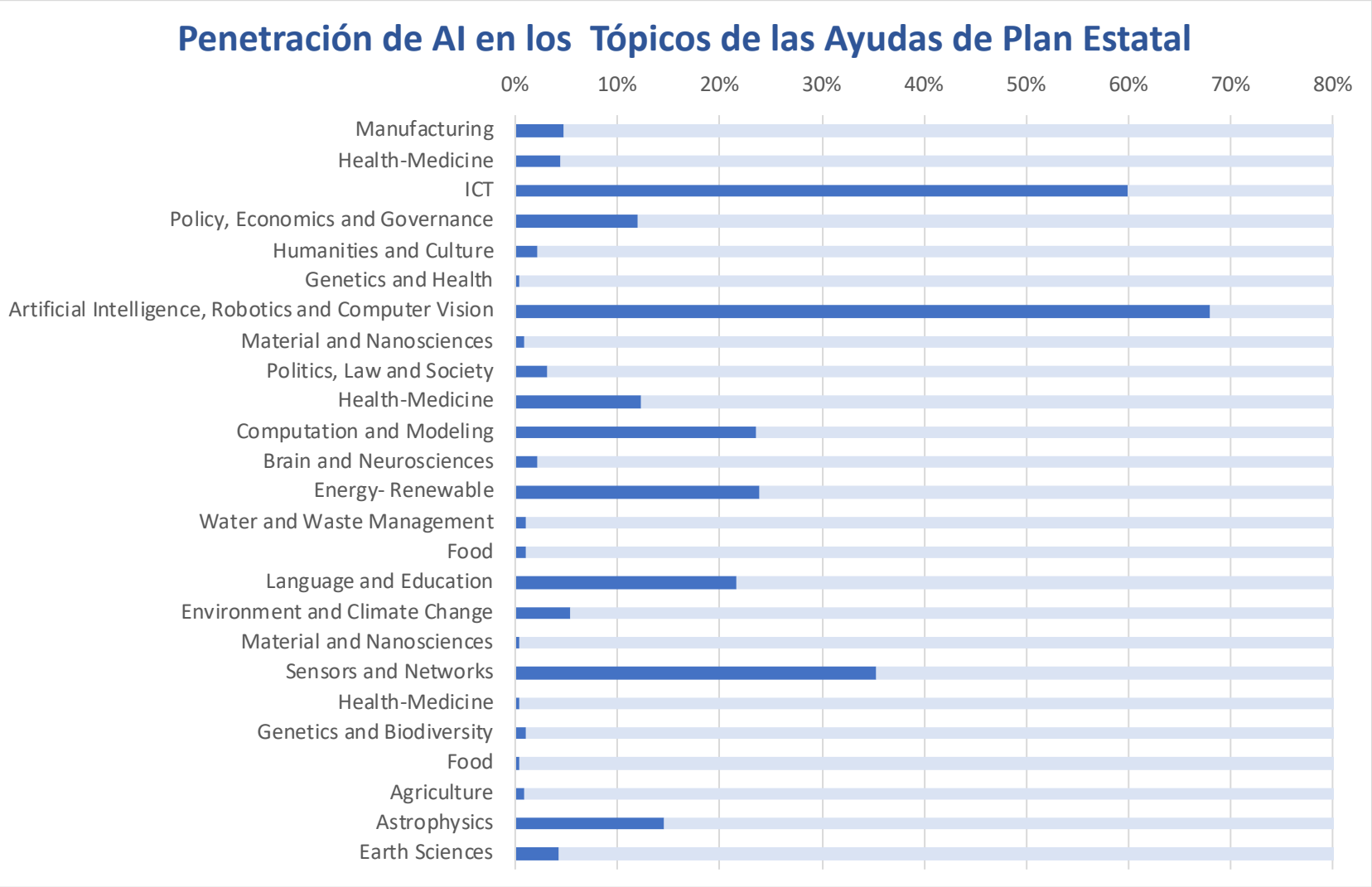
- Sistema de recuperación de información basado en similitud de documentos (semántica y BM25)
- Caracterización de documentos (publicaciones, ayudas)



- Porcentajes de texto de las secciones técnicas de la solicitud referidos a cada tópico:
  - 62% al tópico “Big data”
  - 17% al tópico “Certificaciones”
  - 8% al tópico “Portales web”
  - 6% al tópico “Investigación”
  - 4% al tópico “Semántica”
- Más práctico con modelos pequeños (< 25 tópicos)

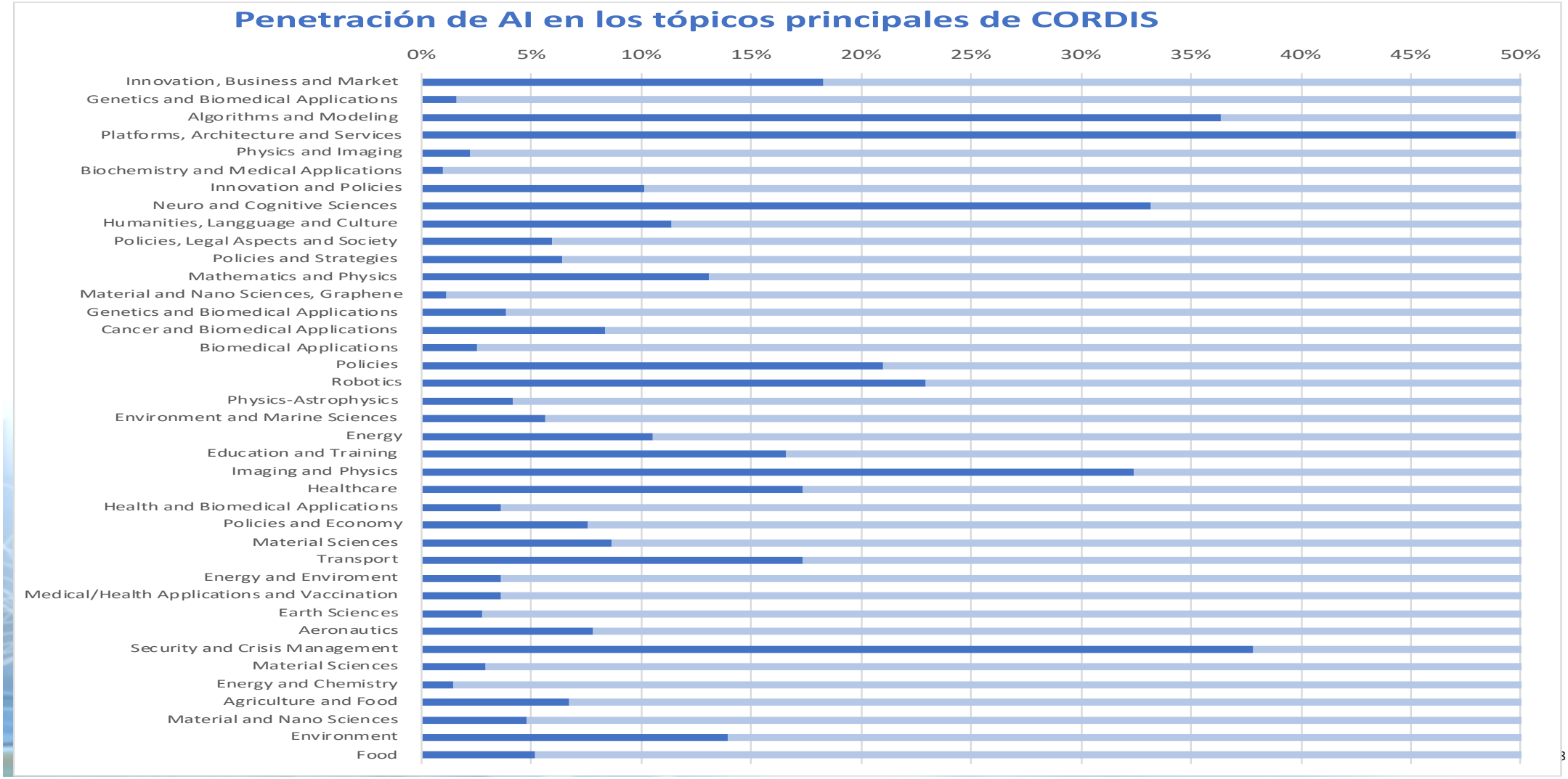
¿Cuál es la penetración del AI en otras áras temáticas?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje



¿Cuál es la penetración del AI en otras áras temáticas?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje



¿Existen diferencias en cuanto a las áreas temáticas asociadas al IA según áreas geográficas?

Conclusiones

- Corpus Viewer es una herramienta en producción, principal resultado del proyecto Faro de Inteligencia Competitiva del Plan TL
  - Incluye técnicas “estado del arte” de machine learning e inteligencia artificial
  - Está alineada con otras iniciativas similares a nivel europeo e internacional para aplicar técnicas de análisis de datos al análisis de la I+D+i
- Se ha creado una instancia en abierto que estará operativa a partir de la semana que viene
  - El número de corpus disponibles se incrementará progresivamente en las próximas semanas
  - Solicitud de sugerencias y recomendaciones para mejorar la experiencia de uso
- Corpus Viewer es un proyecto en rápido desarrollo
  - A lo largo de 2019 se integrarán nuevos desarrollos ya concluidos
  - Se está trabajando ya con la navegación empleando mapas basados en grafos semánticos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

¡¡Gracias!!

www.PlanTL.gob.es

plantecnologiaslenguaje@mineco.es

Plan TL - Tecnologías Lenguaje