BigData o NoData

Text Analytics (Análisis de Texto)

José M. Castaño

Grupo GALLI: http://www.galli.dc.uba.ar
Departamento de Computación
FCEyN, UBA

Big Players: Big Data

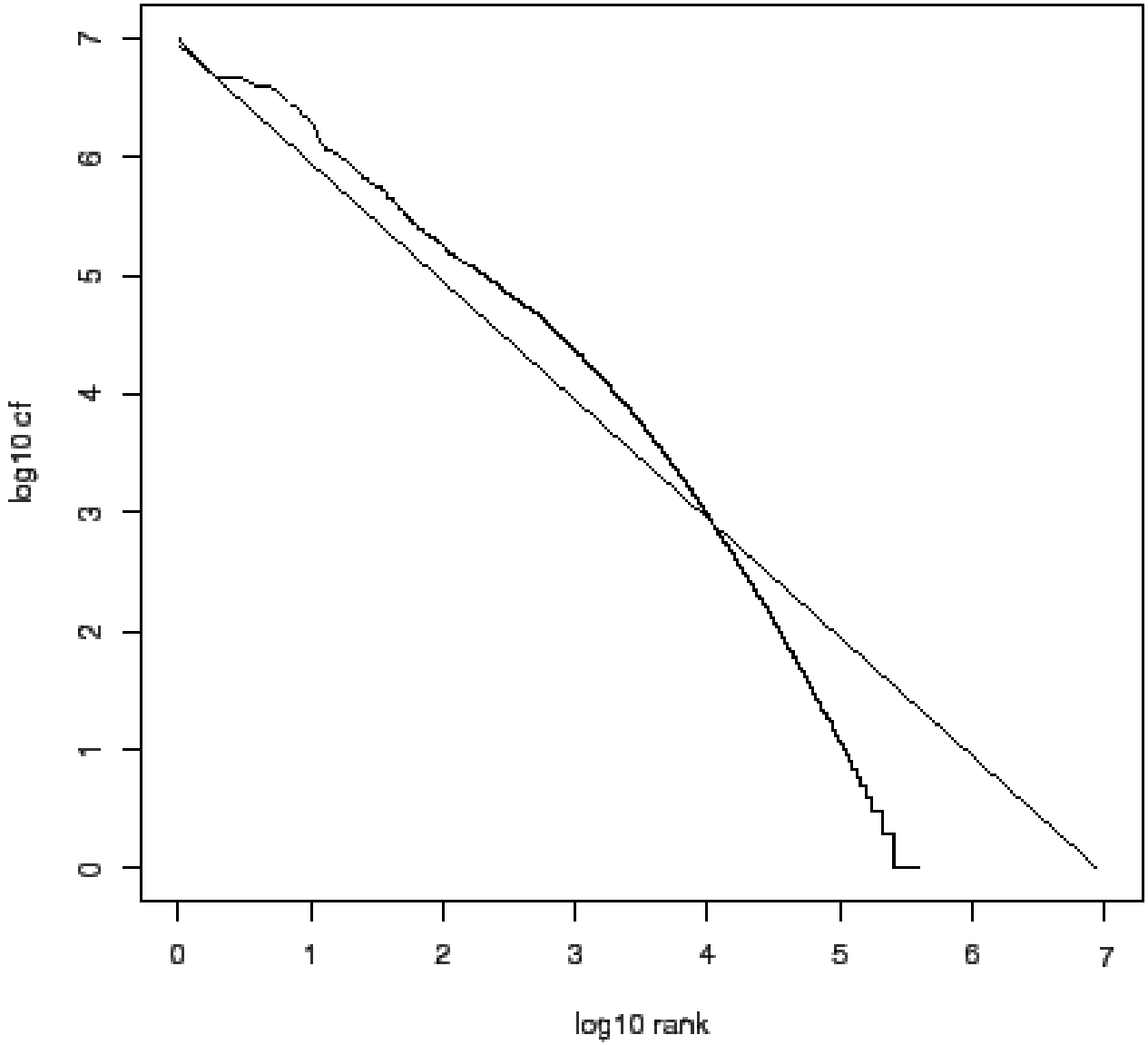
- Buscadores: Una compañia 70% del Mercado
- Social Media: Una compañia 70% del Mercado
- Necesidad de Información: un país determina la agenda tecnológica
- Históricamente: la comunidad de inteligencia (NSA, CIA, DARPA, ARDA, otros).
- También aparecen nichos: BioNLP

Otros Países

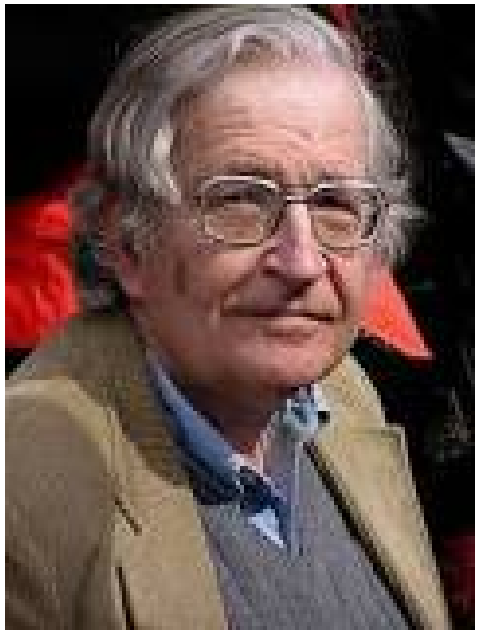
- Dependencia tecnológica y cultural
- Parsing determinístico sobre BigData textual.
- Todo el mundo suministra data
- El control del flujo de esos data está en pocas manos.
- Una porción significativa está abierta.
- Hay respuestas para la soberanía?
- Siempre en una posición desventajosa?

Data Driven: Zipf por todos lados

Zipf (texto) en escala log



AI: Racionalismo vs. Empirismo



Racionalismo vs. Empirismo

- Tiempo y espacio infinito?
- Si procesamos petabytes de texto.
- Podemos obtener algún conocimiento?
- Aprendizaje No supervisado: No alcanza.
- Definir problemas concretos y obtener Data:
  - Salud
  - Justicia
  - Educación

Cómo pescar en un mar de palabras?

- Extraer información relevante de una secuencia de texto
  - Blah blah blah **relevante** blah blah blah
- Específica para un Dominio
  - ej. Negocios
- El significado de *relevantes* predefinido
- No es un problema de BigData
- Problema: Data de calidad
- Data anotado.

El conocimiento ("Oro") codificado en texto

Más del 80% de la información: formato no estructurado.

- Correo electrónico
- Reclamos (seguros/servicios)
- Noticias
- Páginas Web
- Patentes
- Artículos científicos o técnicos,
- Salud
- Educación
- Organismos del Estado
- ...
- Muchas perspectivas sobre los datos.

Problema Acceso

- Hay áreas que no tienen
- No hay un acceso uniforme a las diversas fuentes
- Cada fuente tiene su propio formato y almacenamiento
- Cada fuente tiene su propia sintaxis

Qué se Necesita

- Medidor de Bolsas (y bolsitas) de Palabras
- Autómatas de Estados Finitos
- Métodos Estadísticos
- Aprendizaje Automático
- Ingeniería del Conocimiento
- Corpus Anotado
- Preprocesamiento Sintáctico (Tagging/Chunking)

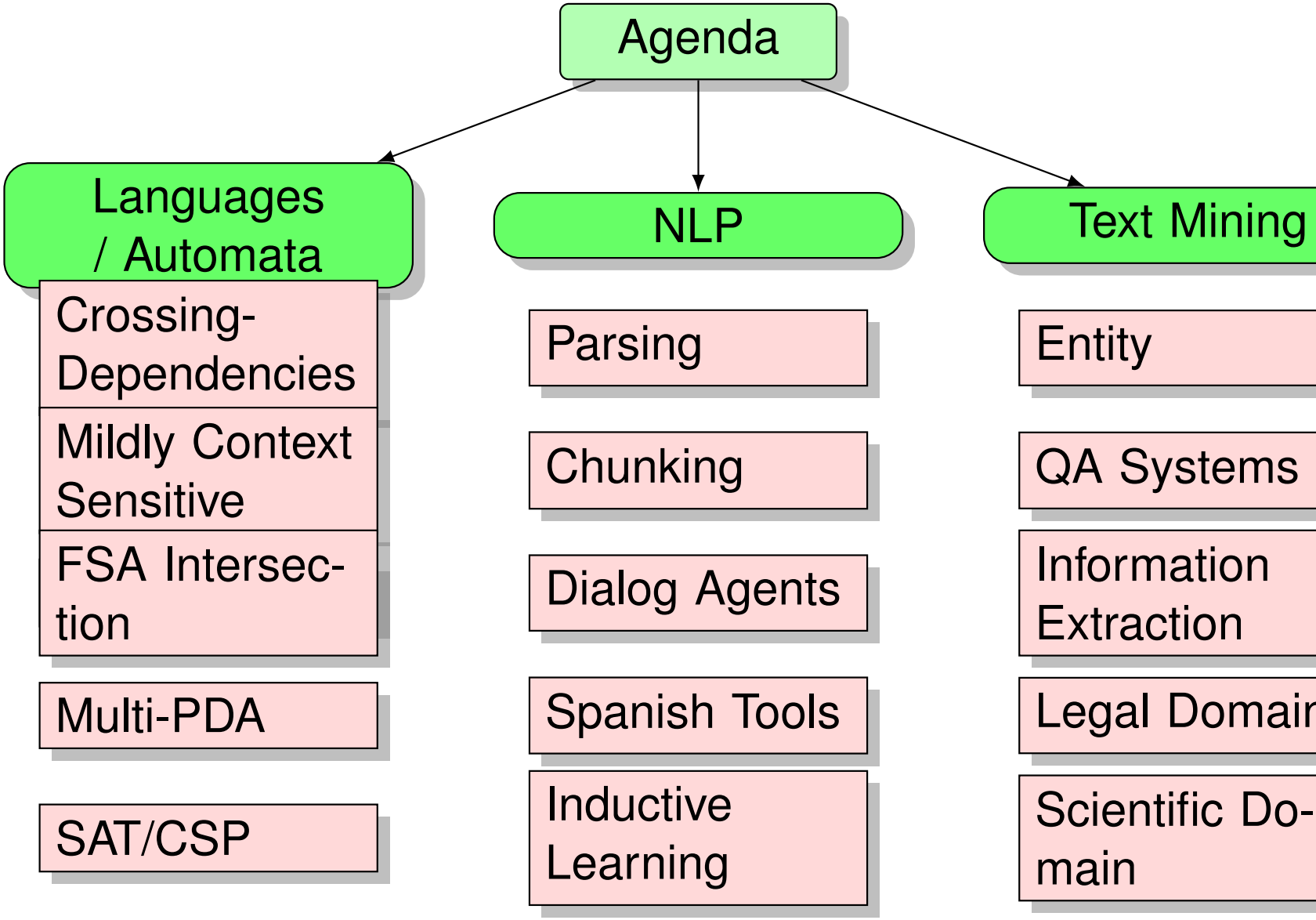
Historia de IE: PRISM

- MUC: Message Understanding Conferences (1987)
  - DARPA (en este siglo ARDA)
  - Estandard
  - Evaluación
  - Diseminación
  - Programa TIPSTER de DARPA : hasta 1998
    - Detección de Documentos
    - Resumen y Extracción de Información
  - TREC (Text Retrieval Conferences)
  - TAC (Text Analysis Conferences)

Named Entity Recognition (NER)

- Jerarquías de NE
  - Persona
  - Organización
  - Locaciones
- Pero también:
  - Artefacto
  - Facilidad
  - Entidad Geopolítica
  - Vehículo
  - Arma
  - Etc.
- Muchos tipos dependientes del dominio (SEKINE & NOBATA '04)

Grupo GALLI


Agenda
Languages / Automata
NLP
Text Mining
Crossing-Dependencies
Mildly Context Sensitive
FSA Intersec-tion
Multi-PDA
SAT/CSP
Parsing
Chunking
Dialog Agents
Spanish Tools
Inductive Learning
Entity
QA Systems
Information Extraction
Legal Domain
Scientific Do-main