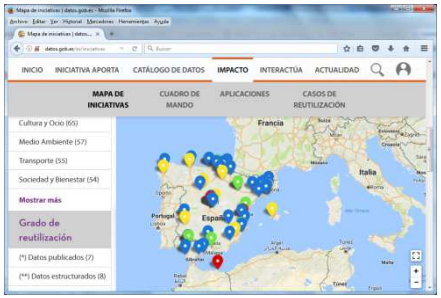
TextDigester:
resumen de texto ubicuo multilingüe

Francesco Ronzano & Horacio Saggion
Large Scale Text Understanding Systems Lab
TALN Group
Universitat Pompeu Fabra

Equipo

- Francesco Ronzano @francescopiu 
- Horacio Saggion @h_saggion 
- Pablo Accuosto @PabloAccuosto 
- Francesco Barbieri @fvancesco 

Problema



Generalitat de Catalunya

GOBIERNO DE ESPAÑA

ISI Web of KNOWLEDGE
Transforming Research

> 90M records

BOLETÍN OFICIAL DEL ESTADO

Comunidad de Madrid

> 1M pages (en)

Google

PubMed

> 24M records

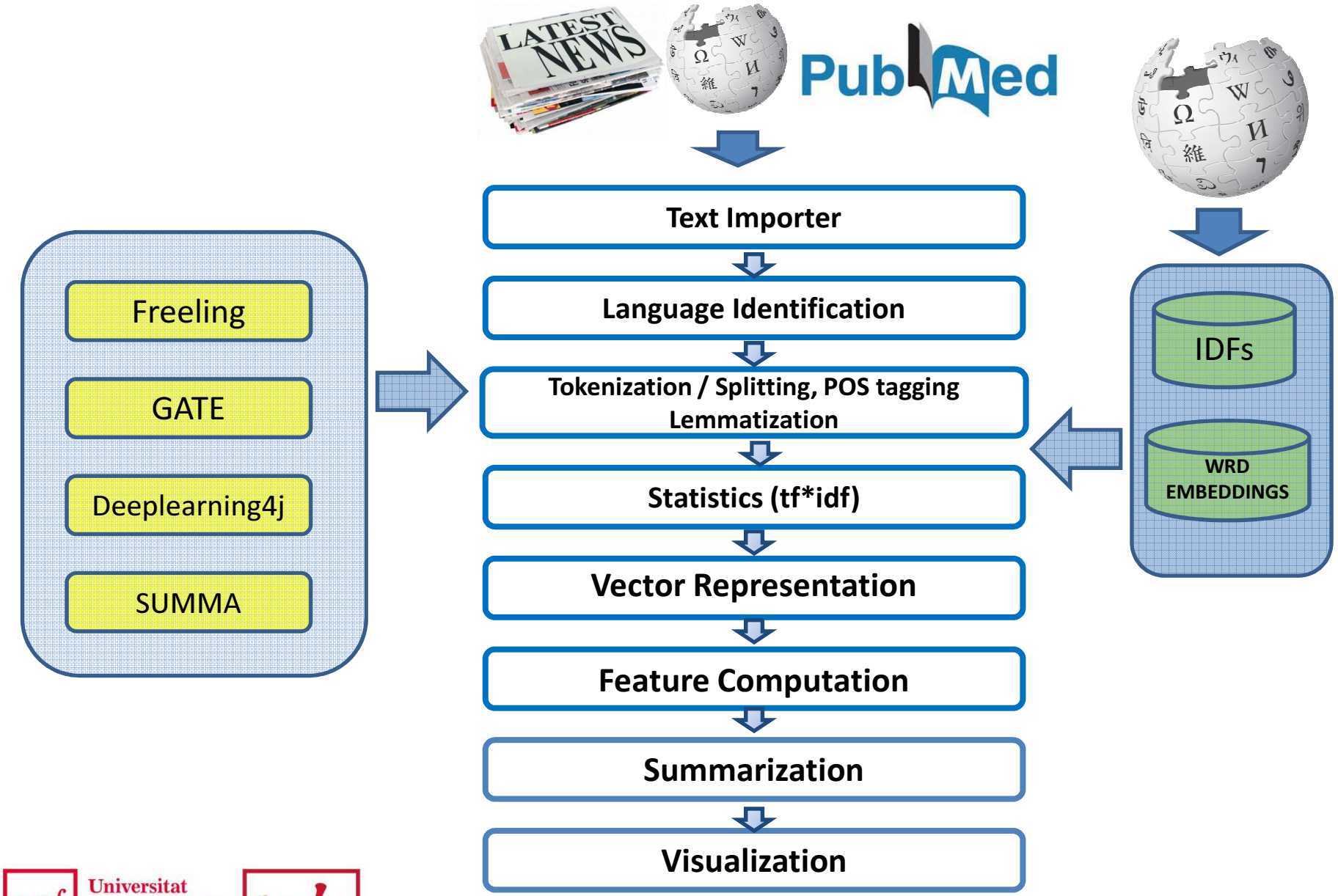
>500M / day

upf. Universitat Pompeu Fabra Barcelona

taln

Solución

- Una librería para generar resúmenes de uno o varios documentos en inglés, castellano, y catalán (extensible a otros idiomas)
  - Desarrollada p/ #HackathonPLN !!!
- TextDigester utiliza datos abiertos
  - Wikipedia en Español, Catalán, Inglés
  - Creación de recursos (word embeddings; tablas de frecuencias)
- Modelo único de documento para los distintos idiomas
- Algoritmos computan *valores de relevancia* de oraciones y anotan los documentos



upf. Universitat Pompeu Fabra Barcelona

taln

Análisis de textos

- Tokenization, sentence splitting, lemmatization, POS tagging
- *if*idf* and word embeddings computation
  - basado en corpora
- Vector representation
  - tf*idf & word embeddings
- Centroide computation
  - tf*idf & word embeddings

Métodos de resumen

- LexRank
  - tf*idf
  - Word embeddings
- Centroide
  - tf*idf
  - Word embeddings
- First similarity
- Document similarity
- Semantic relevance
- Position
- Term Frequency

Código

- TextDigester: self-contained Java library:
  - https://github.com/fra82/textdigester
- Based on:
  - **Freeling** (v 4.0): http://nlp.cs.upc.edu/freeling/
  - **GATE** (v 8.3): https://gate.ac.uk/
  - **Deeplearning4j** (v 0.7.2): https://deeplearning4j.org/
  - **SUMMA**: http://www.taln.upf.edu/pages/summa.upf/
- TextDigester is structured as a Maven project working with Java 1.8.

upf. Universitat Pompeu Fabra Barcelona taln GitHub Java general architecture GATE for text engineering

Destacados

- Resumen de uno o varios documentos
- Datos anotados para entrenar tu algoritmo
- Enseñanza de PLN
- Reproducibilidad

TextDigester:
resumen de texto ubicuo multilingüe

Francesco Ronzano & Horacio Saggion
Large Scale Text Understanding Systems Lab
TALN Group
Universitat Pompeu Fabra