TextDigester

resumen de texto ubicuo multilingüe

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

I Hackathon de Tecnologías del Lenguaje @ 4YFN – 27/2/2017

Equipo

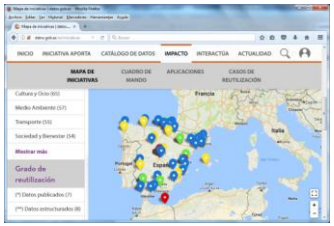
Francesco Ronzano / @francescopiu

Horacio Saggion / @h_saggion

Pablo Accuosto / @PabloAccuosto

Francesco Barbieri / @fvancesco

Sobrecarga de información



Generalitat de Catalunya

GOBIERNO DE ESPAÑA

BOLETÍN OFICIAL DEL ESTADO

Comunidad de Madrid

ISI Web of KNOWLEDGE
Transforming Research

WEB OF SCIENCE™

Scopus

PubMed

2,500,000 artículos científicos al año (uno cada 13 segundos)

> 5,000,000 tweet al día (6,000 tweet por segundo)

Google

WIKIPEDIA
The Free Encyclopedia

5,346,028 páginas (800 páginas nuevas al día)

upf. Universitat Pompeu Fabra Barcelona

taln

TextDigester:
resumen de texto ubicuo multilingüe

- 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

Análisis de documentos

PubMed

Text Importer

Language Identification

Tokenization, Sentence Splitting,
POS tagging, Lemmatization

Statistics (tf*idf / doc. embeddings)

Vector Representation

Feature Computation

Summarization

Visualization

FreeLing
Hooked on a FreeLing

general architecture
GATE
for text engineering

DEEPLEARNING4J

SUMMA

IDFs

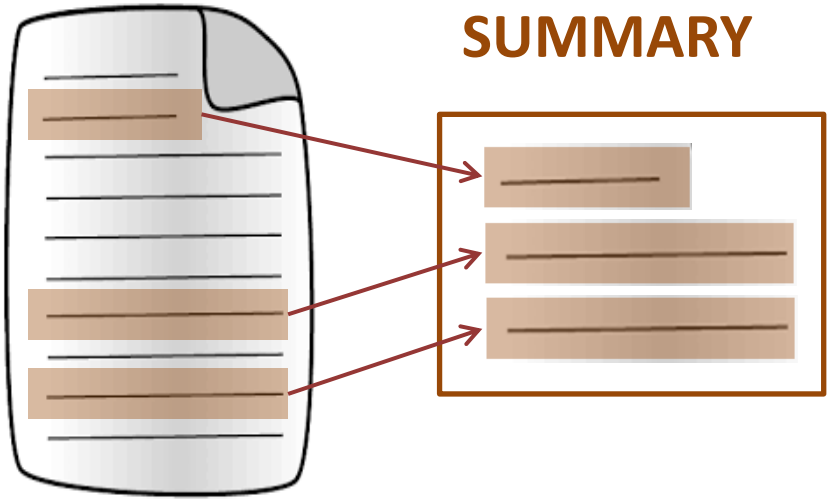
WORD
EMBEDDINGS

upf. Universitat Pompeu Fabra Barcelona

taln

Métodos de resumen

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



Código

- **TextDigester**: self-contained Java library
  - Maven project working with Java 1.8
  - Open-source code: https://github.com/fra82/textdigester

- *Basada en*:
  - **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/

Destacados

- Resumen de uno o varios documentos
- Datos anotados y framework Java para desarrollar y entrenar tu algoritmo
- Enseñanza de PLN
- Reproducibilidad de experimentos de resumen automático

TextDigester

resumen de texto ubicuo multilingüe

Thanks for your attention!

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

I Hackathon de Tecnologías del Lenguaje @ 4YFN – 27/2/2017