Project Proposal

Large Scale Data Collection and Preprocessing in Spark

Divya Tyagi
dxt180002@utdallas.edu

Harsh Verma
hxv180001@utdallas.edu

Nirbhay Sibal
nxs180002@utdallas.edu

Prachi Rajendra Hagwane
pxh180000@utdallas.edu

Swathy Priya Sathishbarani
sxs175832@utdallas.edu

Related Literature:

In the current era, big data systems collect complex data streams and give rise to 6 Vs of big data, namely – Volume, Velocity, Value, Variety, Variability and Veracity. The reduced and relevant data streams are more useful than collecting raw, redundant, inconsistent and noisy data. Another reason for big data reduction is that million variables cause curse of dimensionality which requires unbounded computational resources to uncover actionable knowledge patterns.

The above said issues can be resolved by the process of deduplication which is one of the essential tasks in data preprocessing. This process results in data cleaning and replica-free repositories which allow retrieving increased high quality information. The project aims at running such deduplication algorithm at content level and comparing two articles from different URLs to find out whether they cover the same story.

Goal:

Comparing contents of various Spanish news websites and ascertain whether the articles cover the same stories

Input:

Raw unstructured text from Spanish news sources.

Output:

Data from the articles, including raw text and processed information – in MongoDB.

Project Timeline:

Project Phases	Tech Stack	Dates
Creation of web crawler to collect data from Spanish news websites	Scrapy (python)	28th March - 5th April
Processing the extracted content; segment the sentences within the content; generate universal dependency parse for each sentence within the content; store processed data in MongoDB	*Apache Spark *For Text processing: 1. Python NLTK 2. Spacy 3. Textacy * MongoDB	6th April - 15th April
Deduplication Algorithm development	* Dedupe (Python) * NLTK (python)	16th April - 22nd April
Model Deployment and Testing phase		23rd April - 30th April

References:

1. News-please: https://github.com/fhamborg/news-please
2. Scrapy: https://scrapy.org/
3. Apache Kafka: https://kafka.apache.org/
4. SPEC paper: https://ieeexplore.ieee.org/document/7474330
5. Universal Dependency: https://universaldependencies.org/
6. ufal-udpipe python package