Introduction to NLP with *Sentiment Analysis*

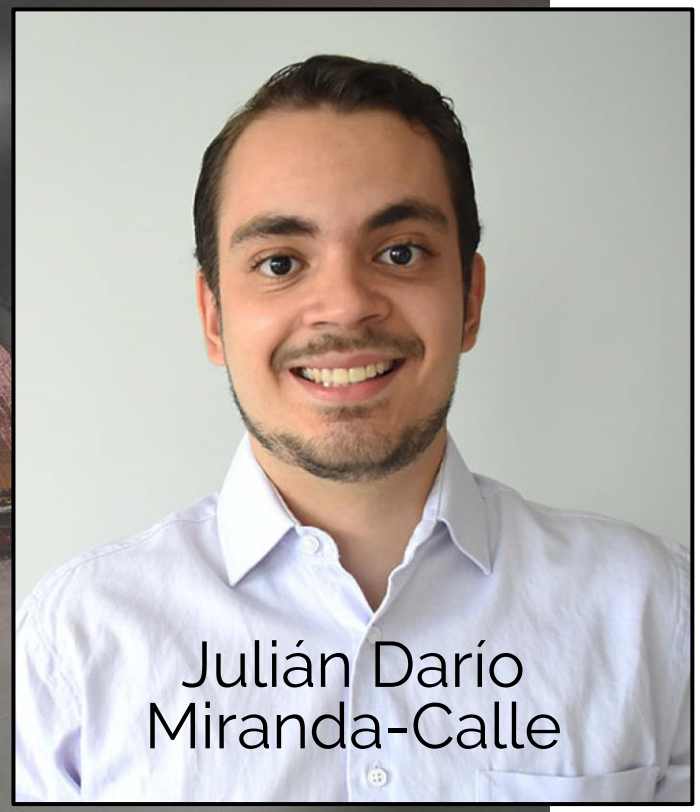
Julián Darío Miranda-Calle

Universidad Pontificia Bolivariana
SECCIONAL BUCARAMANGA

DPhi Data Science Bootcamp

DPhi Data Science Bootcamp

About me



Internal Auditor ISO 27001:2013, Cybersecurity Specialist, Computer Science Engineer, and Electronics Engineer. Currently, graduate and undergraduate professor at the Faculty of Computer Science Engineering.

Scrum Master, developer and researcher in cryptography, steganography, and steganalysis using Data Science, Machine Learning and Deep Learning techniques. Leader and Coach of the teams that will participate in the XXXIV ACIS/REDIS National Programming Contest 2020.

linkedin.com/in/juliandariomiranda/ 0000-0002-7580-2361
https://www.researchgate.net/profile/Julian_Miranda2

Agenda

01 Introduction
- Natural Language processing
- NLP Challenges
- Cyberattack practices

02 Fundamentals
- NLP Pipeline
- Data preparation

03 Applied Data Science
NLP applications to different contexts

04 Use case
Exploring Amazon customer reviews

05 Conclusions
- Conclusions
- Future work

DPhi Data Science Bootcamp

Julián Darío Miranda-Calle
Introduction to NLP

Introduction

Natural Language Processing description

Language is a communication system in which there is a **context** and combined principles that support it

Language can be present in different ways. Among them: **text**, speech, signs, and fingerprint-detectable characters, among others

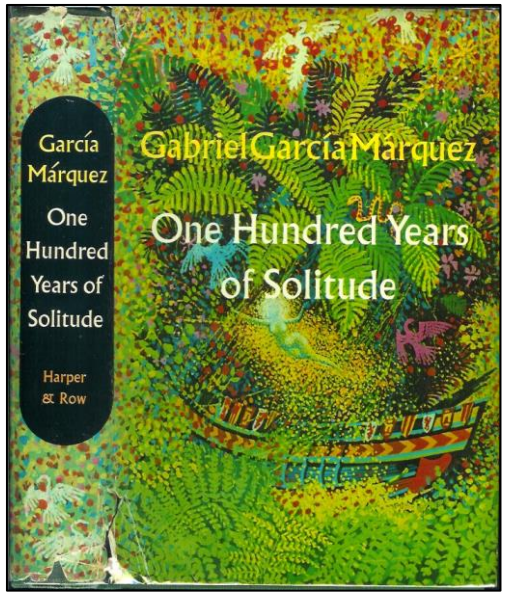
Natural language processing allows human beings to have an **understandable and interpretable** communication with the machine

Introduction

Natural Language Processing Challenges

01 The text is **unstructured** and **Highly Dimensional** in nature, making it difficult to interpret.

For example, consider the book *One Hundred Years of Solitude* by the Nobel Prize *Gabriel García-Márquez*:



Taken from: https://www.chanticleerbooks.com/pages/books/22817/

If we analyze the text in the book as:

- **Characters:** the book has more than 1 million characters, hence a 1M dimensional representing vector.
- **Words:** the book has more than 120,000 unique word expressions, hence a +120K dimensional representing vector.
- **Sections:** the book has around 20 sections with extremely complex language, which would be uninterpretable for a machine.

DPhi Data Science Bootcamp

Introduction

Natural Language Processing Challenges

The text totally depends on the **context**, and hence its processing stages

02

For example, consider the term ***compact.*** This term has several uses in English (homograph terms):

A synonym of small

To compress something

A verb that denotes firmness

A small case for makeup

A compact car

A cassette

Fundamentals

Julián Darío Miranda-Calle
Introduction to NLP

Natural Language Processing Pipeline

Business context

01
Acquisition
Collection of data for processing for decision-making from the real-world measures.

02
Data preparation
Structuring, cleaning and enrichment of the data.

03
Data exploration
Identify visible patterns in the data and test hypotheses.

04
Modeling
Train Machine Learning and Deep Learning models to predict variables of interest.

05
Validation
Validate that predictive and prescriptive models execute tasks properly.

06
Data visualization
Communicate the findings with stakeholders through interactive data visualization.

Decision making

Fundamentals

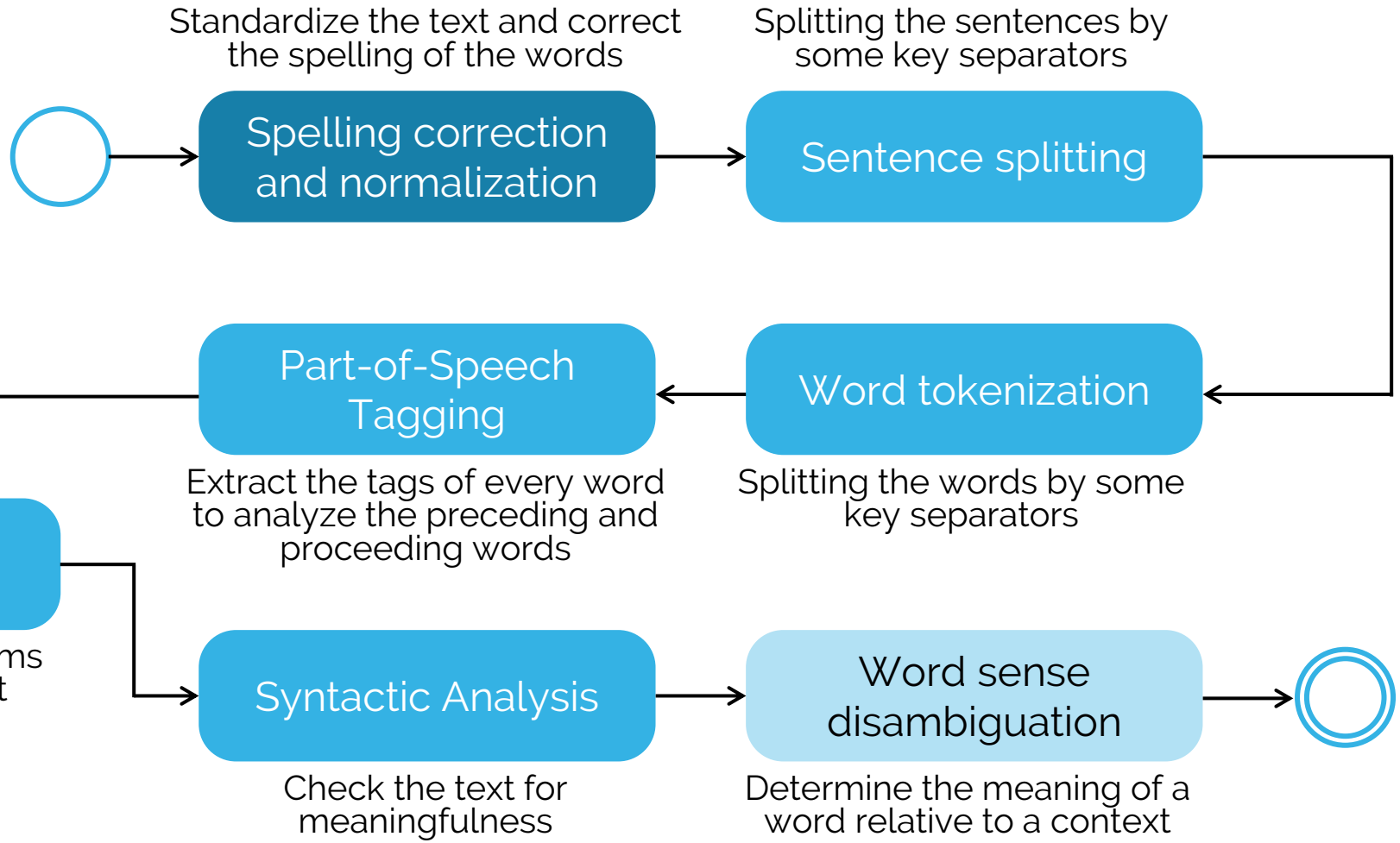
Julián Darío Miranda-Calle
Introduction to NLP

Natural Language Processing Pipeline – Data Preparation

02

Data preparation

Structuring, cleaning and enrichment of the data.



Julián Darío Miranda-Calle
Introduction to NLP

Applied NLP

How and where are NLP practices applied?

Information retrieval

Retrieve the relevant results for customer segments

Sentiment Analysis

Analyze sentiments included in customer reviews for decision making

Spam filtering

Using NLP techniques to detect language components in spam messages.

Phishing emails

Using NLP and Machine Learning techniques to unveil phishing attempts.

Machine translation

Translate documents from one language to another

Natural Language Processing applications

Most of the techniques are build on top of EDA, Regression, Classification, supervised and unsupervised learning tasks

Web Mining

Discover web patterns through data mining and NLP

Use case

Julián Darío Miranda-Calle
Introduction to NLP

Let's analyze a dataset of **real observations of Amazon's customer reviews** to propose a sentiment analysis approach

Source code and datasets available **here**

Taken from: https://annerallen.com/2015/11/disappearing-amazon-reviews-facts/

	Id	ProductId	UserId	Score	Time	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	5	1303862400	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	1	1346976000	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	4	1219017600	This is a confection that has been around a fe...

Conclusions

Julián Darío Miranda-Calle
Introduction to NLP

Although there are different types of pre-processing involved in textual data, **not everything has to be applied in each case**. The procedure explained in the use case can be **extrapolated for all types of NLP studies** in which text data records are kept, which can provide great **information on customer patterns** and business analysis.

The analysis of the language components through Natural Language Processing **allows to make decisions** in various spectra, among which we can highlight marketing, advertising, communications and social networks.

Future work

Current research

Sentiment Analysis on customer reviews of digital products of the MercadoLibre.com (Colombia) site
Discrete Maths II (UPB). 2020

Sentiment Analysis Algorithm for Political Control on Twitter of Bucaramanga Politicians.
J. Flórez, G. Bohórquez, J. Miranda. 2020

A study of the state of the art of NLP techniques for Digital Forensics Analysis.
Y. Reddy, J. Miranda. 2020

Development of an algorithm for PQRS processing applied to a given context
Discrete Maths II (UPB). 2019

Julián Darío Miranda

juliandariomiranda@gmail.com

/juliandariomiranda

0000-0002-7580-2361

www.researchgate.net/profile/Julian_Miranda2