Maastricht University

Institute of Data Science

PGGM

Text Analytics Bootcamp
PGGM 2020

UMIO expand *your perspective*

Maastricht University

BISS

Bootcamp Lecturer

Pedro V Hernandez Serrano

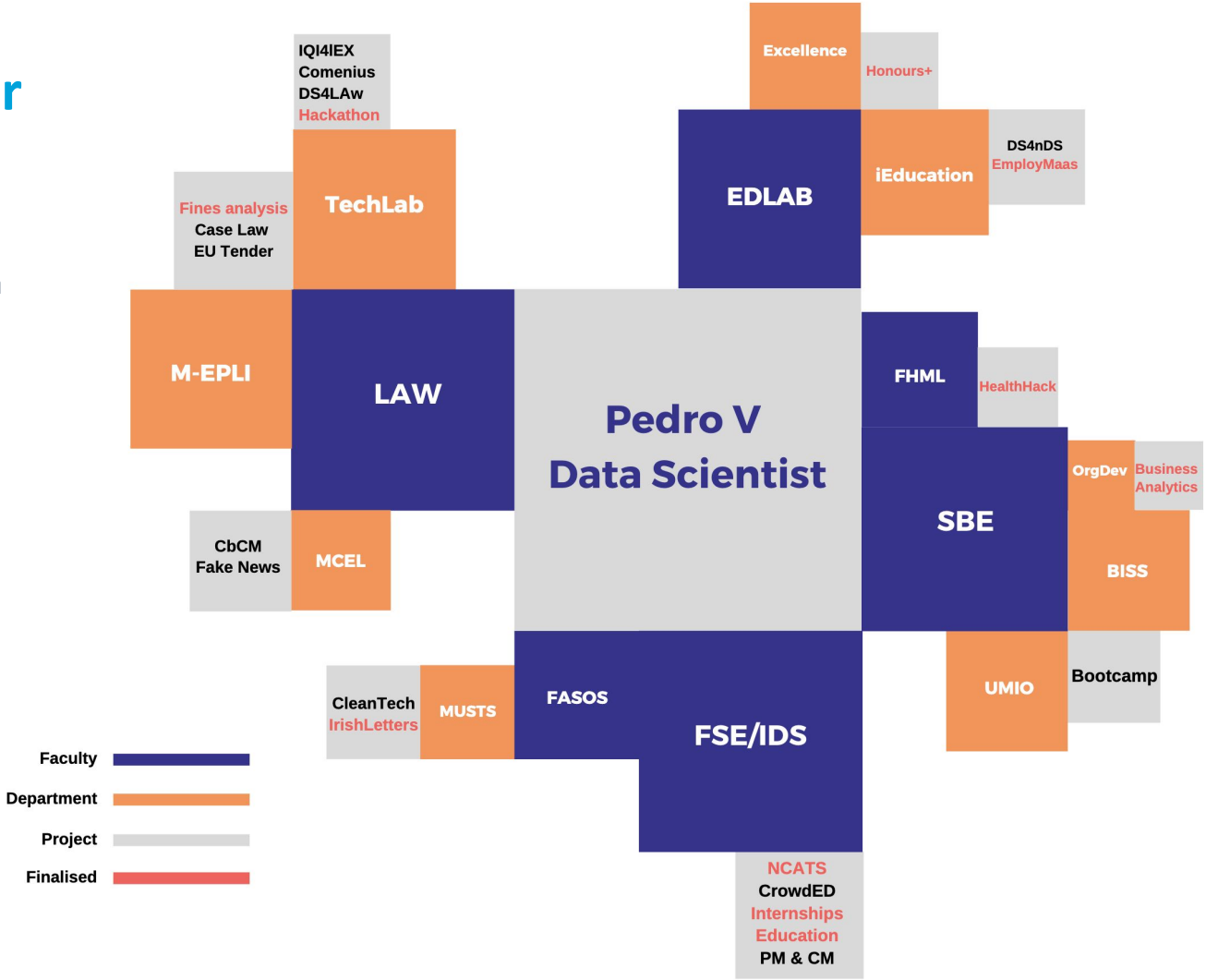
Data Scientist at Institute of Data Science, Maastricht University

As Data Scientist at UM:

- Teaching the fundamentals of Data Science
- Implementing Data Science for research
- Enabling cross-discipline collaborations

Research focus:

- Law & Policy Data Analytics
- Statistical Crowdsourcing
- Reproducible Data Science



Maastricht University

Outline

Text Analytics Bootcamp Outline

Text Analytics Bootcamp Outline

- Intro to Text Analytics Presentation

Text Analytics Bootcamp Outline

- 1.1 Basic functions
- 1.2 Text extraction
- 1.3 Text processing
- 1.4 Data analysis

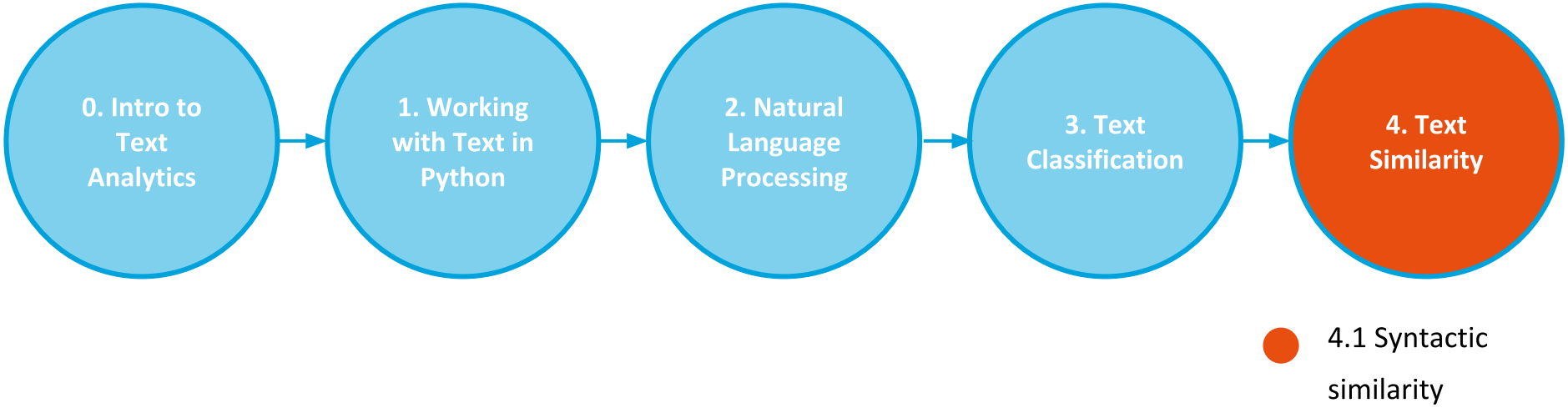
Text Analytics Bootcamp Outline

- 2.1 NLP tasks with Python
- 2.2 Text Features Construction

Text Analytics Bootcamp Outline

- 3.1 Financial sentiment analysis
- 3.2 Supervised ML Classifiers for text

Text Analytics Bootcamp Outline



Text Analytics Bootcamp Outline

Day 1

Part 1
- Bootcamp introduction
- Intro to Text Analytics
- *1.1 Basic functions.ipynb*
- *1.2 Text extraction.ipynb*

Lunch Break

Part 2
- *1.3 Text processing.ipynb*
- Assignment 1

Day 2

Part 1
- *1.4 Data analysis.ipynb*
- *2.1 NLP tasks with Python.ipynb*
- *2.2 Features Construction.ipynb*
- *3.1 Financial sentiment analysis*

Lunch Break

Part 2
- *3.2 Supervised ML Classifiers*
- Assignment 2

Day 3

Part 1
- *4.1 Syntactic similarity.ipynb*

Part 2
- Use case discussion
- Use case hands-on

Lunch Break

- Use case output: Recommender system for responsible investment

**.ipynb* (Jupyter Notebooks)

Goals

- To study the nature of text as a data source for knowledge discovery and identify its relevance to the information needs of diverse individuals.
- To study some of the techniques by which text is automatically processed.
- To demonstrate the types of information which can be extracted from text.
- To discuss the various ways in which text can be analyzed, brainstorm
- We will use open source text analytic tools
- We will conduct text analysis tasks
- We will learn through doing.

Literature

• Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media, Inc. [link]
• Bayley, R., Cameron, R., & Lucas, C. (2013). The Oxford handbook of sociolinguistics. Oxford University Press.
• Gaskell, M. G. (2007). The Oxford handbook of psycholinguistics. Oxford University Press.
• Huang, R. (2017). The Oxford handbook of pragmatics. Oxford University Press.
• Manning, C. D., Raghavan, P., & H. Schutze. (2008). Introduction to information retrieval. Cambridge University Press.
• Taylor, J. R. (Ed.). (2015). The Oxford handbook of the word. Oxford University Press.
• Mining Text Data. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.
• Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Ragha- van, and Hinrich Schuetze, Cambridge University Press, 2007.

Text Analytics

Analytics Defined

Mark Ryan M. Talabis, ... D. Kaye, in Information Security Analytics, 2015

Text Mining

Text mining is based on a variety of advance techniques stemming from statistics, machine learning and linguistics. Text mining utilizes interdisciplinary techniques to find patterns and trends in "unstructured data," and is more commonly attributed but not limited to textual information. The goal of text mining is to be able to process large textual data to extract "high quality" information, which will be helpful for providing insights into the specific scenario to which the text mining is being applied. Text mining has a large number of uses to include text clustering, concept extraction, sentiment analysis, and summarization.

Text Analytics is:
a collection of methods to extract knowledge from text to support decision making.

This field includes:
natural language processing, databases, text mining, deep learning, computational linguistics, etc.

Source: https://www.sciencedirect.com/topics/computer-science/text-mining

Text Analytics

- And this text data is growing really fast. It grows exponentially and continues to grow so. And it's estimated to be about 2.5 Exabytes, that is 2.5 million TB, a day. [1]
- And over 1.5 trillion queries on Google in a year.[2]
- 80 percent of the info is unstructured text
- In the “AI in Banking Vendor Landscape and Capability Map report” we find more vendors selling NLP-based products to banks than any other single AI approach, making up 28.1% of the total AI Approaches
- NLP (information retrieval) 47.4%, NLP (internet pasing) 28.9%, Only

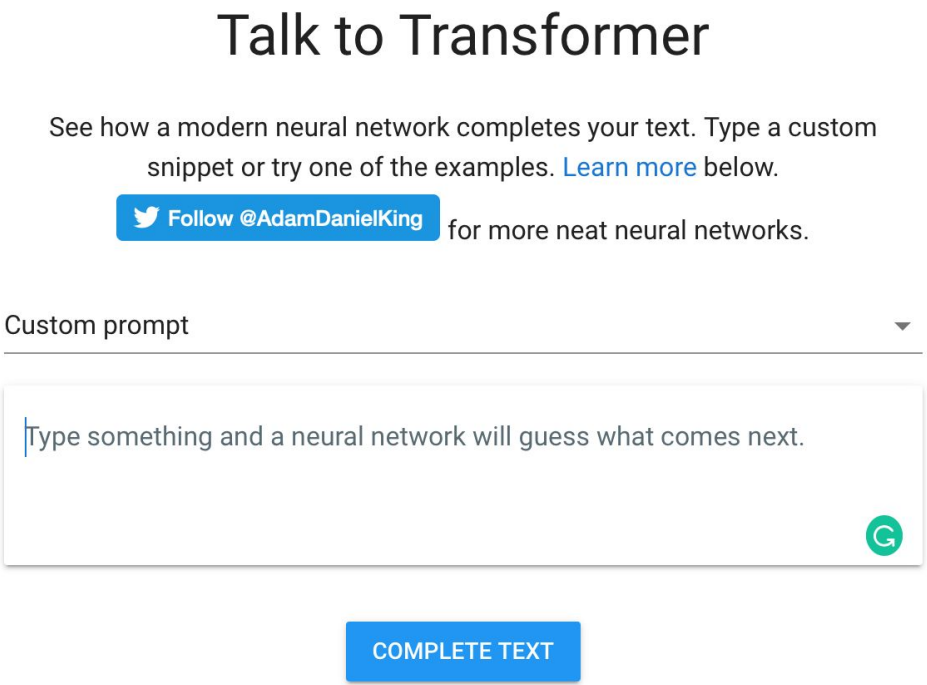
[1] https://ourworldindata.org
[2]https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#521af12960ba

Text Analytics

Applications

- Text classification
- Document similarity
- Topic modeling
- Natural language text generation
- Data augmentation

Applications



Source: https://talktotransformer.com

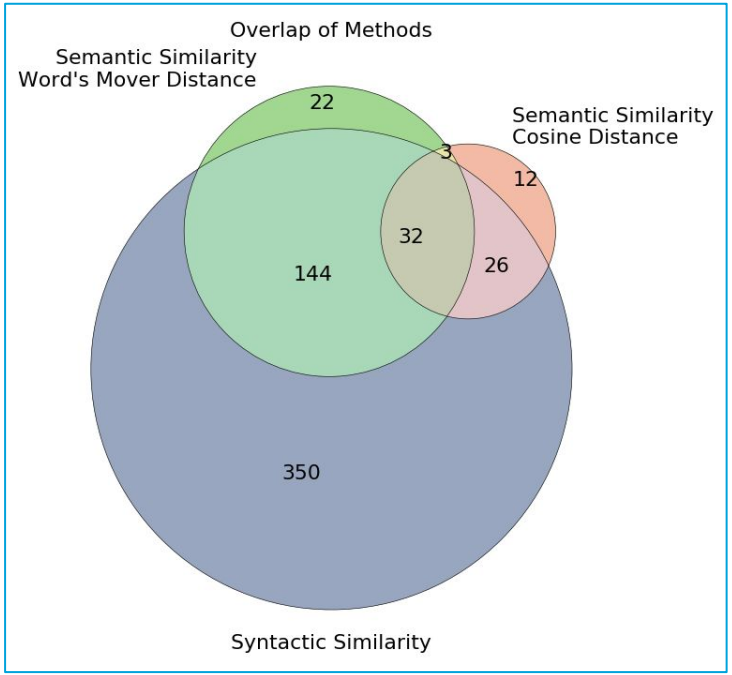
Applications

Similarity and Relevance of Court Decisions: A Computational Study on CJEU Cases

Authors	Kody Moodley, Pedro V. Hernandez Serrano, Gijs van Dijck, Michel Dumontier
Pages	63 - 72
DOI	10.3233/FAIA190307
Category	Research Article
Series	Frontiers in Artificial Intelligence and Applications
Ebook	Volume 322: Legal Knowledge and Information Systems

Abstract

Identification of relevant or similar court decisions is a core activity in legal decision making for case law researchers and practitioners. With an ever increasing body of case law, a manual analysis of court decisions can become practically impossible. As a result, some decisions are inevitably overlooked. Alternatively, network analysis may be applied to detect relevant precedents and landmark cases. Previous research suggests that citation networks of court decisions frequently provide relevant precedents and landmark cases. The advent of text similarity measures (both syntactic and semantic) has meant that potentially relevant cases can be identified



Maastricht University

Similarity and Relevance of Court Decisions: a Computational Study on CJEU Cases
Moodley, K., Hernández Serrano, P., van Dijck, G. & Dumontier, M., Dec 2019, Legal Knowledge and Information Systems. Araszkiewicz, M. & Rodríguez-Doncel, V. (eds.). p. 63-72 10 p. (Frontiers in Artificial Intelligence and Applications; vol. 322).

Applications

			Total Overlap in Percentage of Sampled Cases			
Similarity Type	Similarity Method	Vectorization Method	Data Protection	Public Health	Social Policy	3 Topics Together
Syntactic	Cosine Similarity	N-grams (N=5)	38,1%	40,4%	39,9%	39,6%
Syntactic	Jaccard Distance	N/A	38,1%	27,2%	37,4%	35,0%
Syntactic	Cosine Similarity	TF-IDF	37,2%	22,3%	38,5%	34,2%
Semantic	Word Mover's Distance	Law2Vec Embeddings	6,5%	9,8%	20,1%	14,6%
Semantic	Word Mover's Distance	GoogleNews Embeddings	7,8%	7,9%	20,1%	14,4%
Semantic	Word Mover's Distance	CJEU Embeddings	3,0%	7,5%	15,6%	10,9%
Semantic	Cosine Similarity	CJEU Embeddings	3,9%	3,8%	4,7%	4,3%
Semantic	Cosine Similarity	Law2Vec Embeddings	1,7%	3,8%	3,0%	2,9%
Semantic	Cosine Similarity	GoogleNews Embeddings	2,2%	2,3%	3,1%	2,7%

Similarity and Relevance of Court Decisions: a Computational Study on CJEU Cases
Moodley, K., Hernández Serrano, P., van Dijck, G. & Dumontier, M., Dec 2019, Legal Knowledge and Information Systems. Araszkiewicz, M. & Rodríguez-Doncel, V. (eds.). p. 63-72 10 p. (Frontiers in Artificial Intelligence and Applications; vol. 322).

Applications

Cross-border Corporate Mobility in the EU: Empirical Findings 2019 (Vol. 1)

85 Pages • Posted: 20 Jun 2019

Thomas Biermeyer
Maastricht University - Faculty of Law

Marcus Meyer
Maastricht University - Faculty of Law

Date Written: June 12, 2019

Bekanntmachungstext

In ( ) gesetzte Angaben der Anschrift und des Geschäftszweiges erfolgen ohne Gewähr: Vorgänge ohne Eintragung 11.01.2019 HRB 27380: BVO Ortner GmbH, Bernau a.Chiemsee, Geigelsteinstr. 8, 83233 Bernau a.Chiemsee. Beim Handelsregister des Amtsgerichts Traunstein wurde am 17.12.2018 der Entwurf des Verschmelzungsplans vom 22. November 2018 (Verschmelzungsplan) über die grenzüberschreitende Verschmelzung der BVO Vertrieb Limited mit dem Sitz in 69 Great Hampton Street, Birmingham, West Midlands, B18 6EW, Großbritannien als übertragende Gesellschaft auf die BVO Ortner GmbH mit Sitz in Bernau am Chiemsee, Deutschland, als aufnehmende Gesellschaft eingereicht. Bei der übertragenden BVO Vertrieb Limited handelt es sich um eine Gesellschaft mit beschränkter Haftung nach dem Recht von England und Wales. Die bekannt zu machenden Angaben gem. § 122d Satz 3 UmwG haben folgenden Inhalt: 1. An der Verschmelzung sind beteiligt als übertragende Gesellschaft die BVO Vertrieb Limited, eine Gesellschaft mit beschränkter Haftung nach dem Recht von England und Wales mit dem Sitz in 69 Great Hampton Street, Birmingham, West Midlands, B18 6EW, Großbritannien und als aufnehmende Gesellschaft die BVO Ortner GmbH, eine Gesellschaft mit beschränkter Haftung deutschen Rechts mit dem Sitz in Bernau am Chiemsee, Deutschland. Die übertragende Gesellschaft ist eingetragen im Gesellschaftsregister für England und Wales unter der Nummer 05293246. Die aufnehmende Gesellschaft ist eingetragen im Handelsregister des Amtsgerichts Traunstein, Deutschland, unter HRB 27380. 2.a) Die Rechte der Gläubiger der übernehmenden deutschen GmbH ergeben sich aus § 122a Abs. 2 UmwG i. V. m. § 22 UmwG. Danach ist den Gläubigern

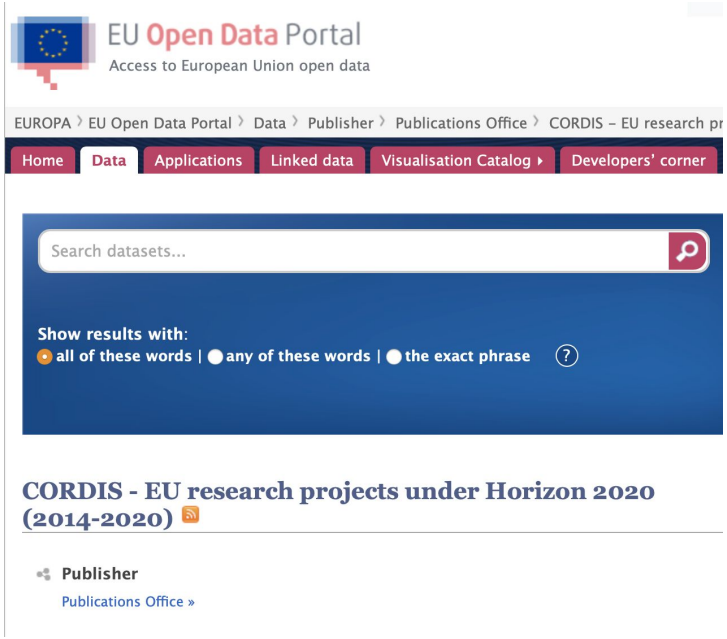
Abstract

This report on cross-border mobility in the European Union focuses, in its third edition, particularly on cross-border mergers and cross-border seat transfers between 2013 and 2019.

Keywords: cross-border mobility, cross-border mergers, cross-border seat transfers, cross-border divisions, empirical findings

Biermeyer, Thomas and Meyer, Marcus, Cross-border Corporate Mobility in the EU: Empirical Findings 2019 (Vol. 1) (June 12, 2019).https://ssrn.com/abstract=3403791

Applications



Applications

Project Title	Project Description	Project Coordinator	
Molecular mechanisms underlying selective neuronal death in motor neuron diseases	The mechanisms behind neuronal death in different motor neuron diseases (MND) remain unknown. These MNDs include the devastating spinal muscular atrophy (SMA) and amyotrophic lateral sclerosis (ALS). A fascinating question in neurodegeneration research is why mutations in ubiquitously expressed genes result in the selective death of a specific neuronal subtype. The ubiquitously expressed and conserved survival of motor neuron (SMN) protein receives its name because its deficit results in MN degeneration. However, SMN known functions -spliceosome assembly and axonal mRNA transport- do not explain the selective MN vulnerability. Accumulation of intracellular aggregates in neurons is a hallmark of most neurodegenerative diseases. The lysosome-autophagy system is the main catabolic pathway for recycling of protein aggregates and	AGENCIA ESTATAL CONSEJO SUPERIOR DEINVESTIGACIONES CIENTIFICAS	 False Hit
Carbon 2 Butanol, a breakthrough technology in eco-innovation that cuts GHG emissions by converting industrial waste gases into chemicals and biofuel.	C2B is a novel microbially-mediated process that captures large amounts of CO2 from industrial plants and converts it into biofuels and chemicals. C2B process is based on Oakbio™ proprietary microbial strain, which uses CO2 from any flue gas and hydrogen (H2) as a feedstock to produce n-butanol, a valuable drop-in biofuel. In fact, n-butanol is primarily used to make durable acrylic plastics, but is also a superior biofuel which addresses a massive market as a potential gasoline replacement. Oakbio™ microbial strain grows in a standard fermenter that can be located next to the flue stack of the factories (cement plants, power plants and refineries). It can capture flue gas directly at the point source with minimal retrofitting. This will allow such factories to cut 70% of their direct GHG emissions, while the n-butanol production adds a significant revenue stream to their bottom line: Oakbio estimates a return of	OAKBIO LTD	 Correct Tag

Opportunities

- Which process you think a summary is useful?
- Is there any repetitive reading task?
- You want to learn more from you clients, comments and feedback?
- Financial sentiment analysis
- ...

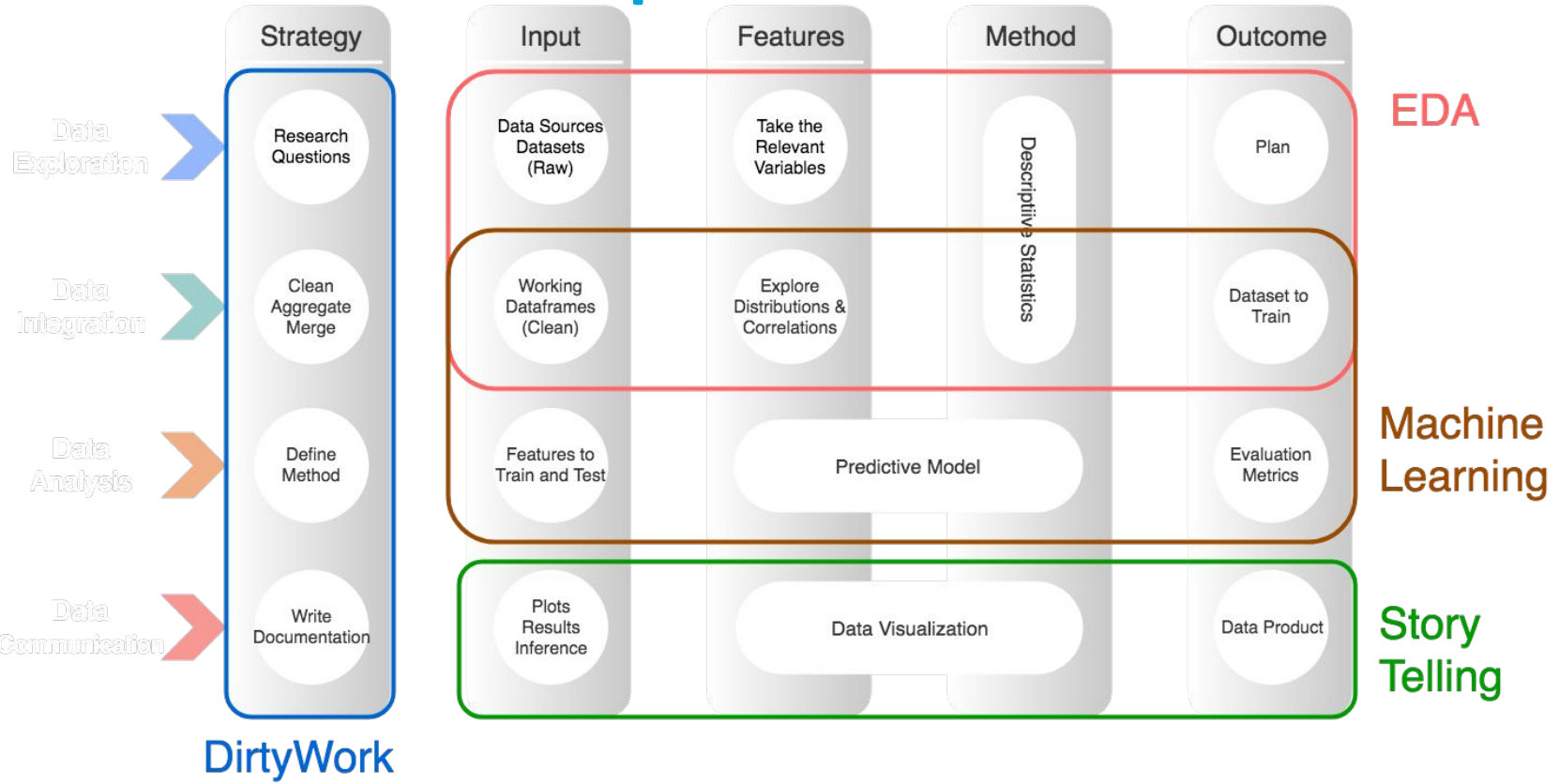
Anaconda + Python

- Python is an , open-source high-level , general purpose, interpreted, programming language.
- A data science/text analytics project may include everything from scraping data from the web, analyzing a mixture or text and numerical data, computing features, training a model, creating high-quality graphs, and then hosting a webapp with results.
- It is explicitly and user-friendly by design.
- It has a massive user community, who contribute to a large number of high-quality, well maintained open-source tools.
- Widely used in industry.

Technologies and Methods

- NLTK: Most popular library for NLP in Python
- Pickle-mixin: Binary objects manager
- Regex: Pattern search implemented in Python by default
- Textract: Library for text extraction from “difficult” formats
- Beautifulsoup: Library for parsing internet-friendly formats
- Textdistance: Library for similarity algorithms
- Textblob: Precomputed sentiment algorithms

The Data Science Pipeline



Text Analytics Concepts

- Corpus: dataset or collection of textual data/files/reports to analyze
- Dictionary: classification of important words into sentiment categories (positive, negative, risk, etc.), aka bag of words, sentiment lexicon, word categorization
- Syntax: the way words are grouped together into larger constituents and phrases and the way these phrases can be ordered
- Morphology: the way words are formed
- Entity/Aspect /Features: textual content map with related entities and their attributes, components and characteristics
- Token: Minimal textual object in a corpus

Github repository:

github.com/pedrohserrano/text-analytics-bootcamp-pggm

Notebook help:

bit.ly/2RkqKMa