SPARK NLP IN ACTION:

INTELLIGENT, HIGH-ACCURACY FACT EXTRACTION

FROM LONG FINANCIAL DOCUMENTS

Saif Addin Ellafi

Paul Parau

CONTENTS

1.

The Challenge

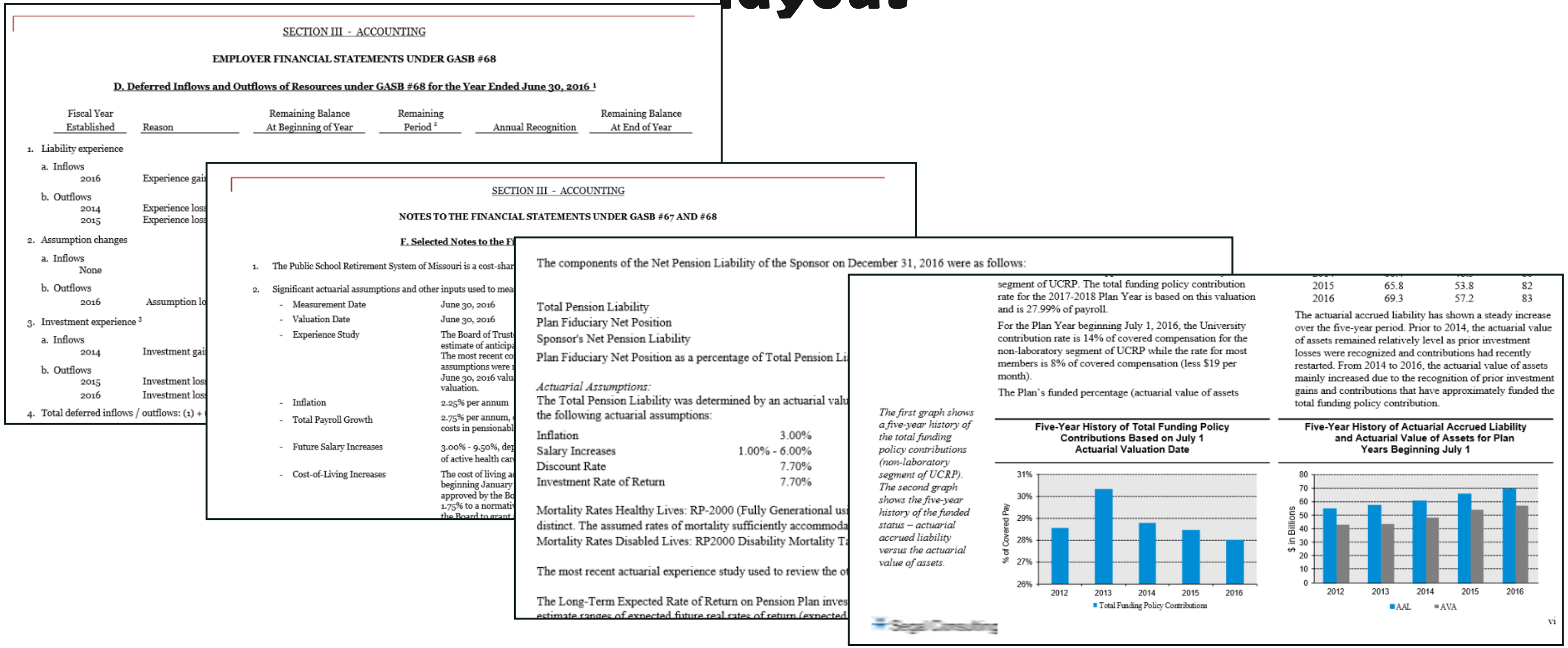
a.

Automate information extraction from pension fund documents and legal forms

Data vs. Information

b.

Extract details from hundreds of raw unstructured documents in PDF format – including tables, pictures and layout



c.

Domain specific natural language with numbers, currency, magnitude & firm name identification (Pensions, social plans, retirement system)

A historical perspective of how the participant population has changed over the past ten valuations can be seen in this chart.

CHART 1

Member Population: 2007 – 2016

Year Beginning July 1	Active Members	Terminated Vested Members[1]	Retired Members, Disabled Members and Beneficiaries[2]	Ratio of Retirees to Actives
2007	118,885	59,056	47,682	0.40
2008	114,242	64,566	50,171	0.44
2009	115,745	54,883		
2010	114,928	55,037		
2011	115,568	60,903		
2012	116,888	67,318		
2013	118,321	73,589		
2014	120,568	78,229		
2015	123,768	75,165		
2016	128,513	81,595		

[1] *Includes terminated nonvested members due a refund of member contributions LLNS defined benefit plans who will be entitled to a CAP balance payment fro*

[2] *Excludes deferred retirees and deferred beneficiaries who are entitled to futur*

CHART 6

Determination of Actuarial Value of Assets for Year Ended June 30, 2016 ($ in 000s)

From	To	Total Actual Market Return (net)	Expected Market Return (net)	Investment Gain/(Loss)[1]	Deferred Factor	Unrecognized Return[2]
7/2011	6/2012	$115,864	$3,133,623	$(3,017,759)	0.0	$0
7/2012	6/2013	4,833,339	3,086,770	1,746,569	0.2	349,314
7/2013	6/2014	8,009,979	3,379,298	4,630,681	0.4	1,852,272
7/2014	6/2015	1,993,802	3,969,206	(1,975,404)	0.6	(1,185,243)
7/2015	6/2016	(1,104,655)	3,995,788	(5,100,443)	0.8	(4,080,354)

1. Total Unrecognized Return[3]	$(3,064,011)
2. Market Value of Assets	54,164,531
3. Actuarial Value of Assets (Item 2 – Item 1)	$57,228,542
4. Actuarial Value of Assets as a Percentage of Market Value (Item 3 ÷ Item 2)	105.7%

[1] *Total return minus expected return, both on a market value basis.*

[2] *Recognition at 20% per year over 5 years.*

[3] *Deferred return as of June 30, 2016 recognized in each of the next four years:*

(a) Amount Recognized during 2016/2017	$(139,720)
(b) Amount Recognized during 2017/2018	(489,033)
(c) Amount Recognized during 2018/2019	(1,415,169)

Example content to be extracted

Actuarial Firm Name	*NLP - Named entity recognition*
Fiscal End Year	*Learning rule-based extraction*
Actuarial Value of Assets	*Currency numbers from tables in different magnitudes (hundreds, millions, etc)*

WHAT MAKES READING DOCUMENTS HARD?

- **Domain specific, Context specific**
  - *Your models & rules must be specific to what you are looking for*
  - *Reading currency amounts, company names, people, locations, units, and other facts depends heavily on context*
- **Tables & images**
  - *Ensure your OCR extraction is well capable of reading data from tables and pictures without breaking content coherence*
- **Heterogeneous**
  - Even if all documents contain the same content, it's not always in the same place, order or format

A SOLUTION FRAMEWORK

	What it does?	Why is it useful?
UiPath	OCR Parsing, fact extraction, learning rules	It converts unstructured data into ready to process text
Spark-NLP	Extremely fast NLP & NLU with machine learning algorithms at scale	Scalable batch or streaming NLP pipelines with applied ML and DL models
Akka http	Fast communication across services	Integrates Spark NLP and UI Path in an asynchronous manner
John Snow Labs NLP models	Pre-trained NLP models for Spark-NLP	Highly accurate extraction of domain specific information

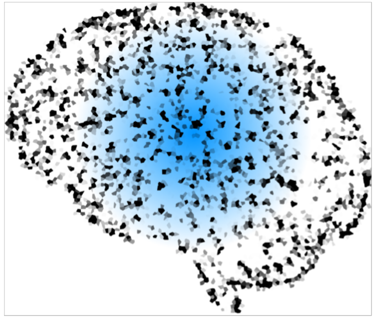
2.

Introducing UiPath

DATA EXTRACTION CHALLENGES

Where do I get my documents from?

How do I validate data?

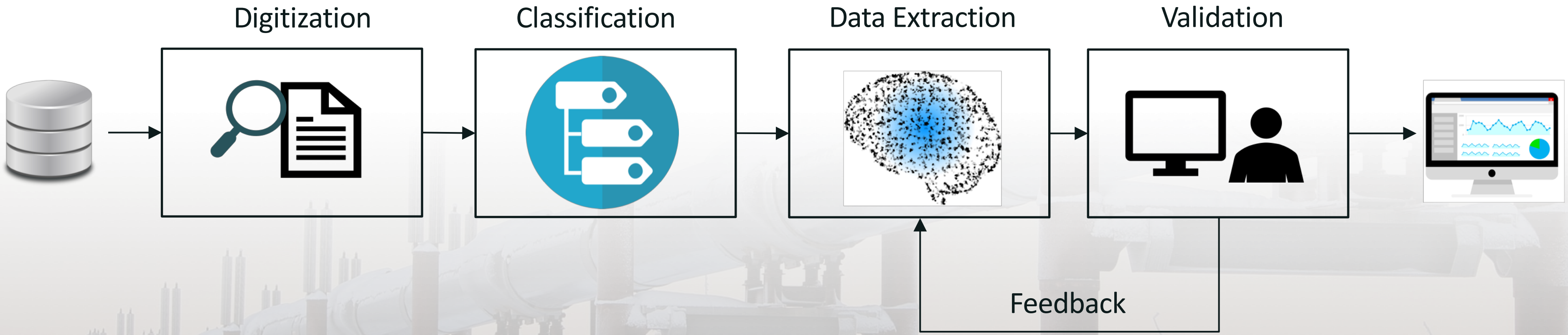


Extraction
algorithms
ML
NLP
[...]

How do I prepare the documents for extraction?

What do I do with the results?

UIPATH DOCUMENT PROCESSING PIPELINE



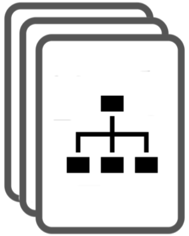
Infrastructure for:

- Document retrieval
- Document processing
- Manual validation
- Learning feedback

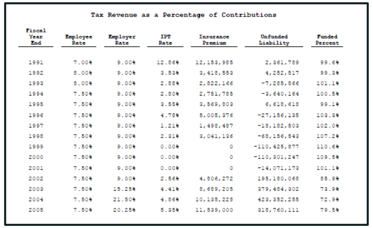
DIGITIZATION: RETRIEVE TEXT, PRESERVE ORIGINAL LAYOUT

Input documents: PDFs, with text content or scanned files, images

Text retrieval: OCR, PDF

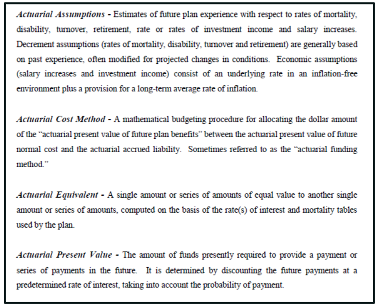


Document Layout Analysis



Tax Revenue as a Percentage of Contributions

Fiscal Year End	Employee Rate	Employer Rate	IPT Rate	Insurance Premium	Unfunded Liability	Funded Percent
1991	7.00%	9.00%	12.86%	12,153,985	2,361,789	99.6%
1992	8.00%	9.00%	3.53%	3,418,553	4,252,517	99.3%
1993	8.00%	9.00%	2.88%	2,822,166	-7,285,866	101.1%
1994	7.50%	9.00%	2.80%	2,751,785	-3,640,164	100.5%
1995	7.50%	9.00%	3.55%	3,569,803	6,618,618	99.1%
1996	7.50%	9.00%	4.78%	5,005,376	-27,156,135	103.3%
1997	7.50%	9.00%	1.21%	1,498,487	-18,182,503	102.0%
1998	7.50%	9.00%	2.31%	3,041,136	-68,156,543	107.2%
1999	7.50%	9.00%	0.00%	0	-110,425,877	110.6%
2000	7.50%	9.00%	0.00%	0	-110,301,247	109.5%
2001	7.50%	9.00%	0.00%	0	-14,071,173	101.1%
2002	7.50%	9.00%	2.56%	4,806,272	195,180,068	85.9%
2003	7.50%	15.25%	4.41%	8,689,205	379,484,302	73.9%
2004	7.50%	21.50%	4.86%	10,135,228	423,352,255	72.9%
2005	7.50%	20.25%	5.35%	11,539,000	318,760,111	79.5%

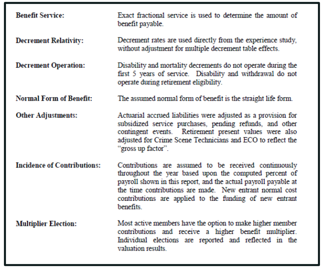


***Actuarial Assumptions* -** Estimates of future plan experience with respect to rates of mortality, disability, turnover, retirement, rate or rates of investment income and salary increases. Decrement assumptions (rates of mortality, disability, turnover and retirement) are generally based on past experience, often modified for projected changes in conditions. Economic assumptions (salary increases and investment income) consist of an underlying rate in an inflation-free environment plus a provision for a long-term average rate of inflation.

***Actuarial Cost Method* -** A mathematical budgeting procedure for allocating the dollar amount of the "actuarial present value of future plan benefits" between the actuarial present value of future normal cost and the actuarial accrued liability. Sometimes referred to as the "actuarial funding method."

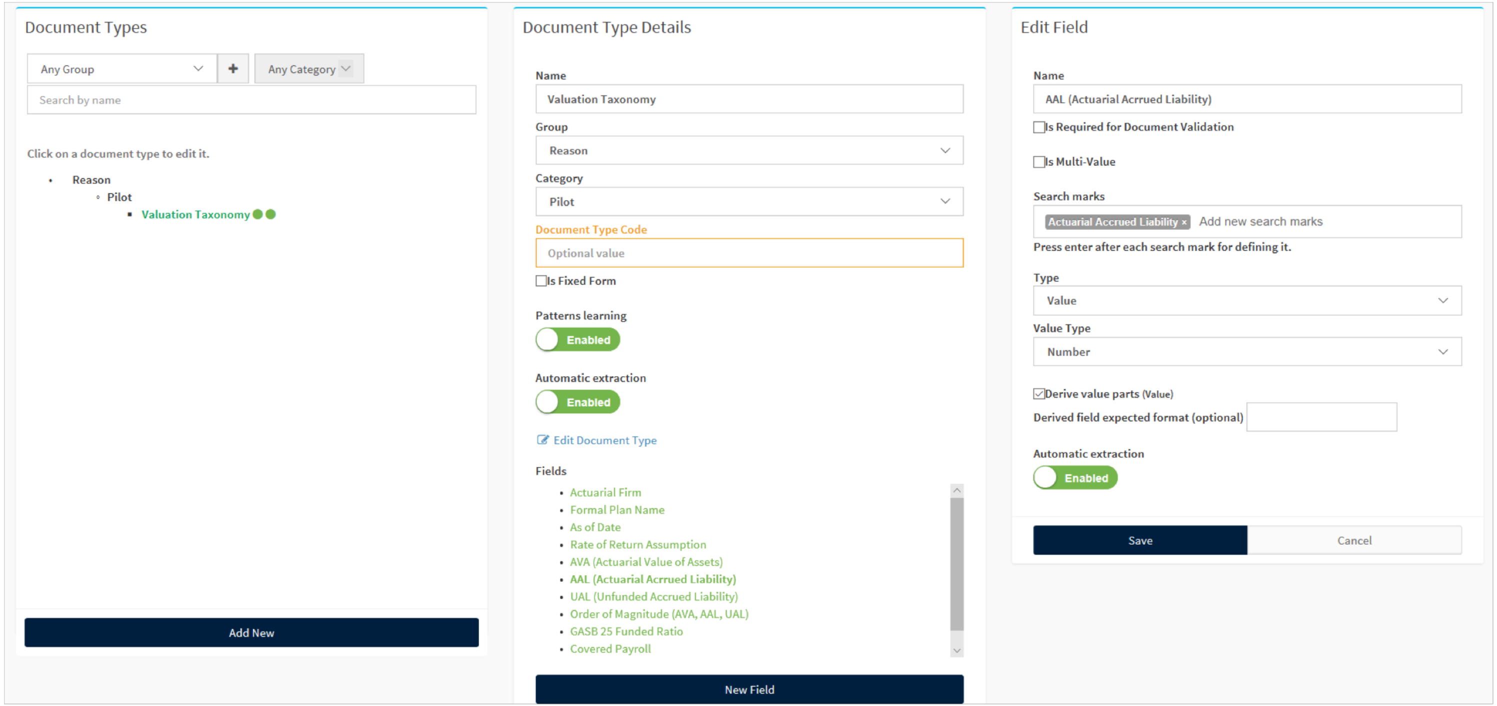
***Actuarial Equivalent* -** A single amount or series of amounts of equal value to another single amount or series of amounts, computed on the basis of the rate(s) of interest and mortality tables used by the plan.

***Actuarial Present Value* -** The amount of funds presently required to provide a payment or series of payments in the future. It is determined by discounting the future payments at a predetermined rate of interest, taking into account the probability of payment.



Benefit Service:	Exact fractional service is used to determine the amount of benefit payable.
Decrement Relativity:	Decrement rates are used directly from the experience study, without adjustment for multiple decrement table effects.
Decrement Operation:	Disability and mortality decrements do not operate during the first 5 years of service. Disability and withdrawal do not operate during retirement eligibility.
Normal Form of Benefit:	The assumed normal form of benefit is the straight life form.
Other Adjustments:	Actuarial accrued liabilities were adjusted as a provision for subsidized service purchases, pending refunds, and other contingent events. Retirement present values were also adjusted for Crime Scene Technicians and ECO to reflect the "gross up factor".
Incidence of Contributions:	Contributions are assumed to be received continuously throughout the year based upon the computed percent of payroll shown in this report, and the actual payroll payable at the time contributions are made. New entrant normal cost contributions are applied to the funding of new entrant benefits.
Multiplier Election:	Most active members have the option to make higher member contributions and receive a higher benefit multiplier. Individual elections are reported and reflected in the valuation results.

CLASSIFICATION: FIND DOCUMENT TYPE

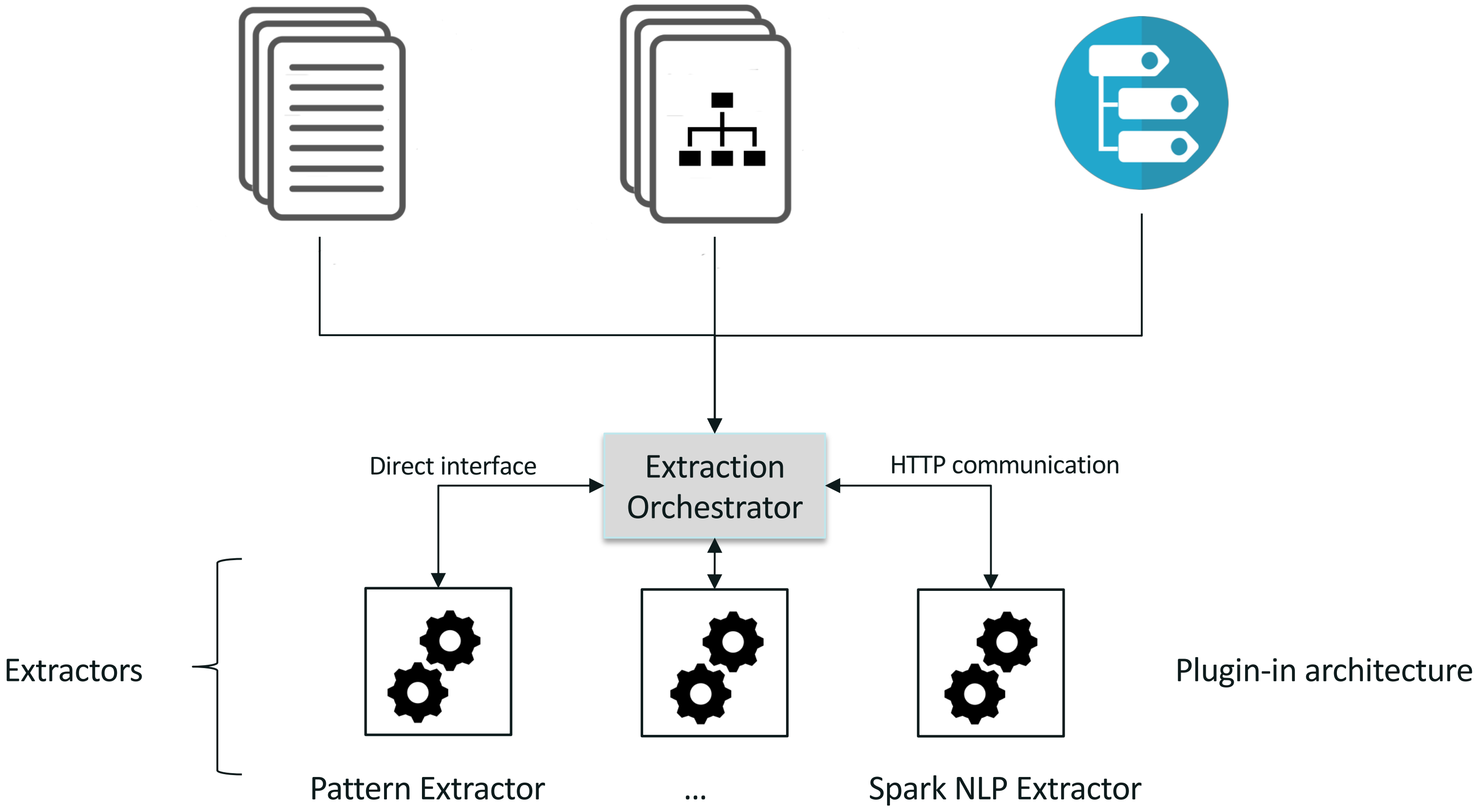


Document Taxonomy: collection of document types

Document type: collection of fields

Field: data point type and properties

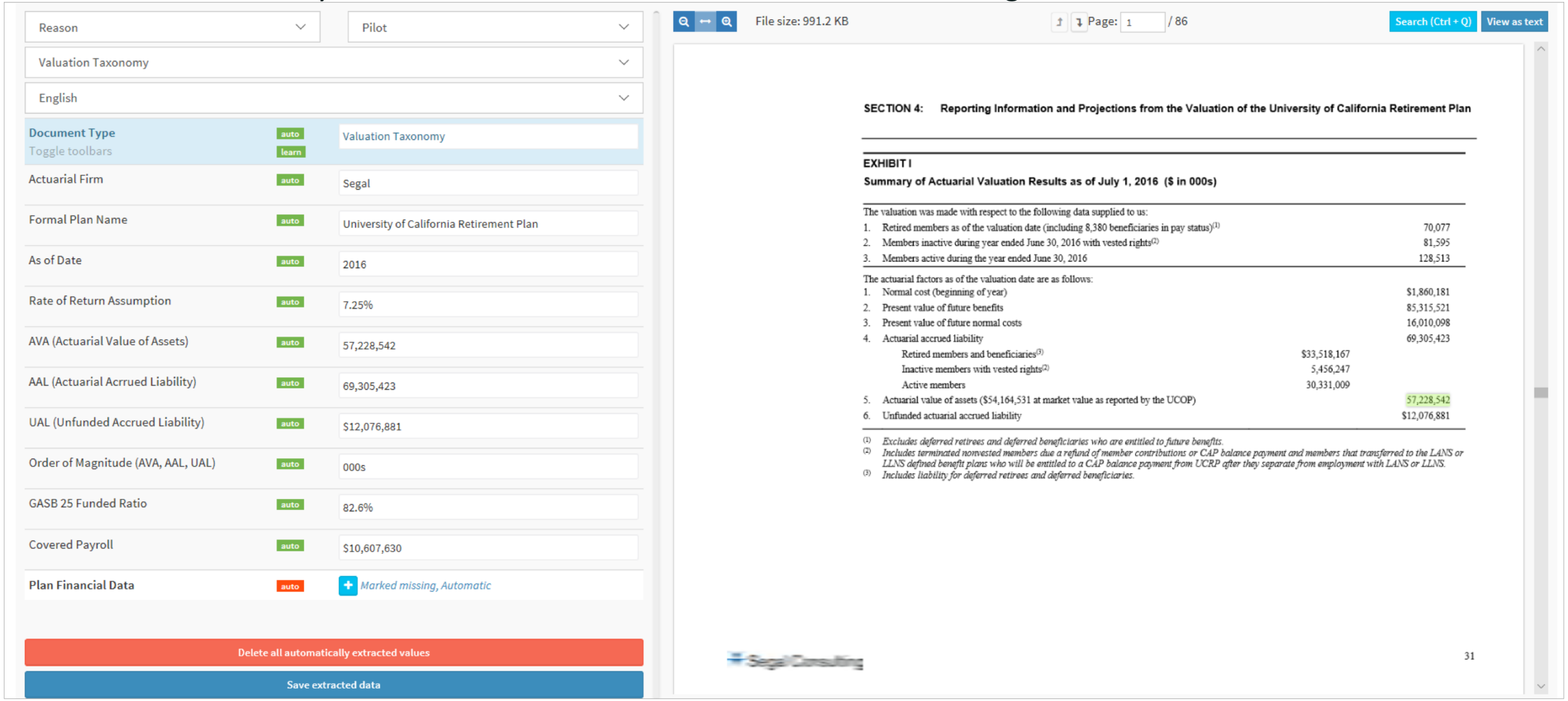
EXTRACTION: FIND DATA POINTS SPECIFIC TO DOCUMENT TYPE



MANUAL VALIDATION & FEEDBACK LOOP

Automatically extracted data

Original document



Validated document → Extractor: learning and model adjustment

PATTERN-BASED EXTRACTION STRATEGY

- Learns value context rules, based on manual extractions

Field	Type	Description
Fiscal End Year	Date	The year for which the valuation report is filed
Actuarial Value of Assets	Number	The value of pension plan investments and other property
Funded Ratio	Number	Ratio of a pension or annuity's assets to its liabilities

The results of the ***45th Annual Actuarial Valuation*** of the City Retirement System are presented in this report. The purpose of measure the System's funding progress and to determine the Ci the ensuing fiscal year in accordance with the established fundi the valuation may not be applicable for other purposes.

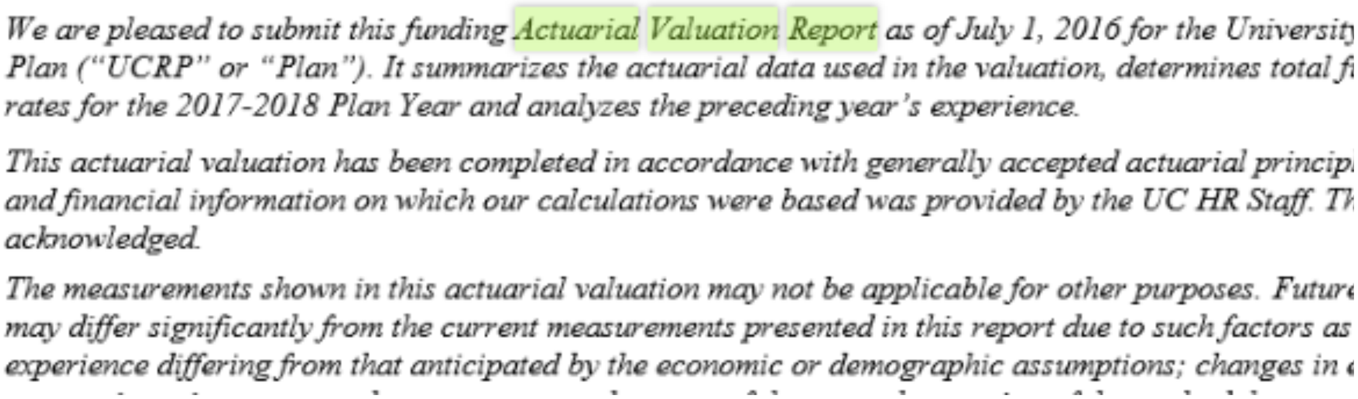
The date of the valuation was ***June 30, 2012***.

This report should not be relied on for any purpose other than tho prepared at the request of the Board and is intended for use by the those designated or approved by the Board. This report may be pr the System only in its entirety and only with the permission of the

The signing actuaries are independent of the plan sponsor.

EXTRACTED FIELD EXAMPLES

- Generic, domain-independent
- Not suited for all scenarios
- Different fields => different extraction strategies


We are pleased to submit this funding Actuarial Valuation Report as of July 1, 2016 for the Universit
Plan ("UCRP" or "Plan"). It summarizes the actuarial data used in the valuation, determines total f
rates for the 2017-2018 Plan Year and analyzes the preceding year's experience.

This actuarial valuation has been completed in accordance with generally accepted actuarial princip
and financial information on which our calculations were based was provided by the UC HR Staff. Th
acknowledged.

The measurements shown in this actuarial valuation may not be applicable for other purposes. Futur
may differ significantly from the current measurements presented in this report due to such factors as
experience differing from that anticipated by the economic or demographic assumptions; changes in

3.

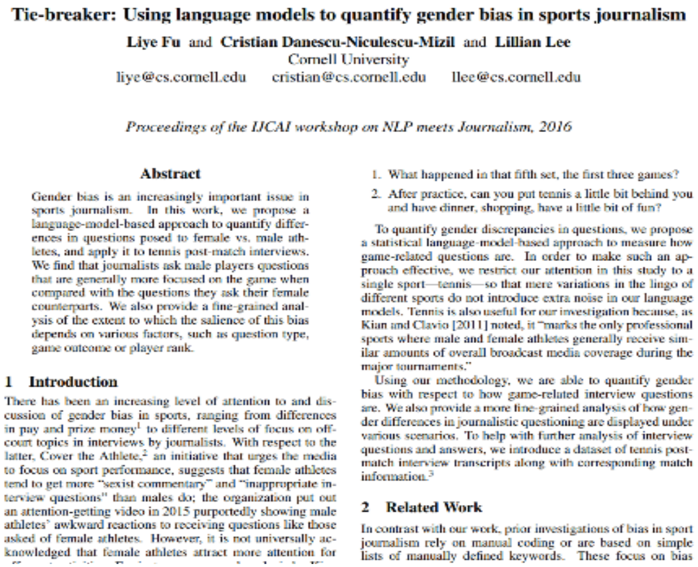
Introducing Spark NLP

WHAT IF THERE'S NO GRAMMATICAL RULE OR PATTERN?

states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting

Since yeatreday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back

Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue.



Tie-breaker: Using language models to quantify gender bias in sports journalism

Liye Fu and **Cristian Danescu-Niculescu-Mizil** and **Lillian Lee**
Cornell University
liye@cs.cornell.edu cristian@cs.cornell.edu llee@cs.cornell.edu

Proceedings of the IJCAI workshop on NLP meets Journalism, 2016

Abstract

Gender bias is an increasingly important issue in sports journalism. In this work, we propose a language-model-based approach to quantify differences in questions posed to female vs. male athletes, and apply it to tennis post-match interviews. We find that journalists ask male players questions that are generally more focused on the game when compared with the questions they ask their female counterparts. We also provide a fine-grained analysis of the extent to which the salience of this bias depends on various factors, such as question type, game outcome or player rank.

1 Introduction

There has been an increasing level of attention to and discussion of gender bias in sports, ranging from differences in pay and prize money[1] to different levels of focus on off-court topics in interviews by journalists. With respect to the latter, Cover the Athlete,[2] an initiative that urges the media to focus on sport performance, suggests that female athletes tend to get more "sexist commentary" and "inappropriate interview questions" than males do; the organization put out an attention-getting video in 2015 purportedly showing male athletes' awkward reactions to receiving questions like those asked of female athletes. However, it is not universally acknowledged that female athletes attract more attention for

1. What happened in that fifth set, the first three games?
2. After practice, can you put tennis a little bit behind you and have dinner, shopping, have a little bit of fun?

To quantify gender discrepancies in questions, we propose a statistical language-model-based approach to measure how game-related questions are. In order to make such an approach effective, we restrict our attention in this study to a single sport—tennis—so that mere variations in the lingo of different sports do not introduce extra noise in our language models. Tennis is also useful for our investigation because, as Kian and Clavio [2011] noted, it "marks the only professional sports where male and female athletes generally receive similar amounts of overall broadcast media coverage during the major tournaments."

Using our methodology, we are able to quantify gender bias with respect to how game-related interview questions are. We also provide a more fine-grained analysis of how gender differences in journalistic questioning are displayed under various scenarios. To help with further analysis of interview questions and answers, we introduce a dataset of tennis post-match interview transcripts along with corresponding match information.[3]

2 Related Work

In contrast with our work, prior investigations of bias in sport journalism rely on manual coding or are based on simple lists of manually defined keywords. These focus on bias

MEDICAL RECORDS: Facts to extract	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy

ACADEMIC PAPERS: Facts to extract	
Summarize Main Result	Double Blind?
Theory or Experiment?	Sample Size?
Benchmark Used	All Results Published?

THIS HAPPENS VERY OFTEN IN PRACTICE

Company	Outcome
Winstar Invesments	Set up tour
Winstar Invesments	Wants to see three more locations.
Red Cloud	Looking for 5MM in the six cap range.
Red Cloud	Here is the five year cash flow.
Winstar Invesments	Property tour: His client wants to put in an offer
Champion Partners LLC	I spoke to George about the deal. He wants to make an offer. We agreed to meet on Tuesday morning.
Red Cloud	Looking for 20MM in the six cap range.
Red Cloud	Looking for properties in the sw.
Diversified Investment Assoc.,	Looking for 4,000
Winstar Invesments	Looking for 5MM
	I'm so excited about this one
	The only thing that would make it a lot of fun
ASB Capital Management LLC	LOI To: rbellinger@asbc345m.com, ashley@345ecooper.com Hi Robert, Here are the new changes to the LOI. Best regards, John Dawson Managing Director Taylor Commercial Real Estate 123 Main Street St. Louis, MO 63131 314-526-5555

The State has no further witnesses. Judge, at this time we would offer, file and introduce S-1 -- for motion purposes only -- S-1, which is a copy of the crime lab in this matter showing that the evidence confiscated from the defendant, both the hand-rolled cigar and the nine plastic baggies were each positive for marijuana; and S-2, a copy of the defendant's prior conviction.

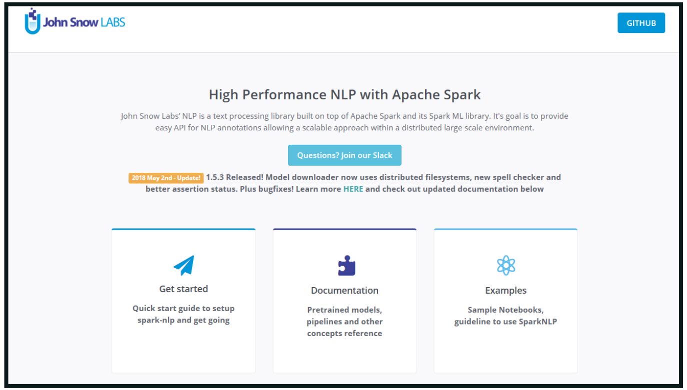
THE COURT:
Any objection for motion purposes?
MS. JANE:
Not for motion purposes.
MS. SMITH:
With that, Judge, the State submits.
THE COURT:
Anything by the Defense?
MS. JANE:
No.
THE COURT:
Submitted?
MS. JANE:
I would submit.
MS. SMITH:
The State submits.
THE COURT:
The Court finds probable cause as charged. The Court denies the Motion to Suppress Evidence. I will note an objection on behalf of the Defense to the Court's ruling.

CRM NOTES: Facts to extract	
Budget	Timeframe
Authority	Level of Urgency
Need	Need executive support?

LEGAL TRANSCRIPTS: Features	
Kind of hearing	Evidence presented
Kind of motion	Witnesses presented
Who filed it	Decision

SPARK-NLP

- Industrial Grade NLP for Apache Spark ecosystem
- Design Goals
    1. **Performance & Scale**
    2. **Frictionless Reuse**
    3. **Enterprise Grade**
- Built on top of Spark ML API's
- Open Source Apache 2.0 licensed
- Active development & support



NATIVE SPARK EXTENSION

High Performance Natural Language Understanding at Scale

John Snow LABS

Part of Speech Tagger
Named Entity Recognition
Sentiment Analysis
Spell Checker
Tokenizer
Stemmer
Lemmatizer
Entity Extraction

Spark MLlib

Topic Modeling
Word2Vec
TF-IDF
String distance calculation
N-grams calculation
Stop word removal
Train/Test & Cross-Validate
Ensembles

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

Data Sources API

FRICTIONLESS REUSE

```
pipeline = pyspark.ml.Pipeline(stages=[
              document_assembler,
              tokenizer,
              stemmer,
              normalizer,
              stopword_remover,
              tf-idf,
              lda])

topic_model = pipeline.fit(df)
```

Spark NLP annotators

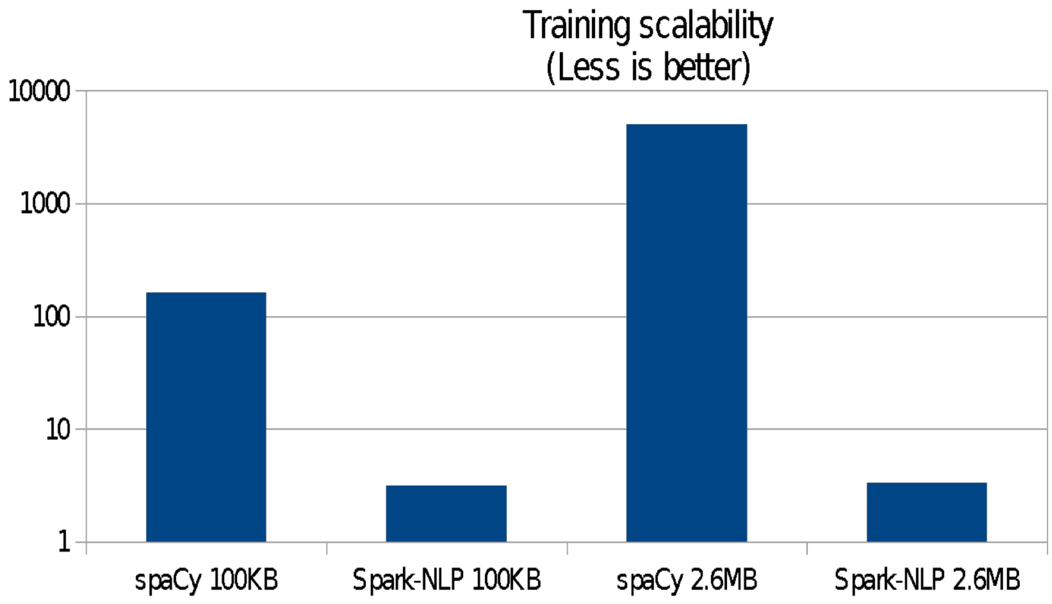
Spark ML featurizers

Spark ML LDA implementation

Single execution plan for the given data frame

BENCHMARK: TRAINING

- Run on a desktop PC, Linux Mint with 16GB RAM, local SSD drives, & Intel core i5-6600K processor running 4 cores at 3.5GHz
- Data has been taken from the National American Corpus (http://www.anc.org), utilizing the MASC 3.0.2 written corpora from the newspaper section.
- Pipeline has Sentence Boundary, Tokenization & Part of Speech

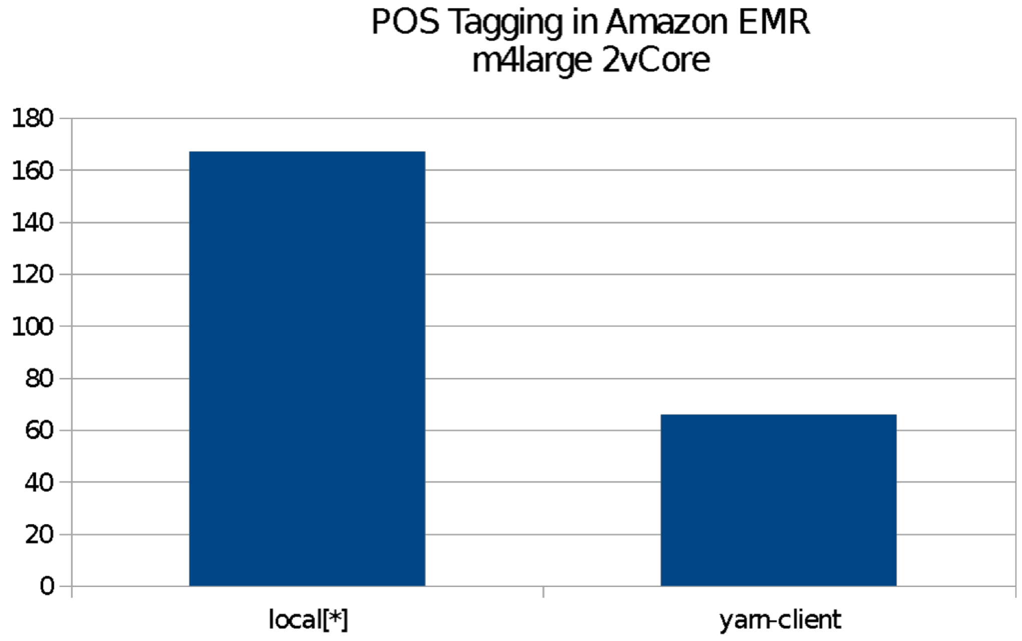


Spark-NLP was 38 times faster to train on 100kb of data

Spark-NLP was 80 times faster to train on 2.6mb of data

BENCHMARK: SCALING

- Spark-NLP against itself
- 2.5x speedup with a 4-node cluster
- Zero code changes
- Spark scales as Spark does:
  1 to 3 orders of magnitude faster
  depending on cluster setup
- Not compares to spaCy, since it cannot leverage a cluster



THE TWO SPARK NLP PIPELINE TYPES

Real time processing

↓

Light Pipelines

10x speedup for 'small data'
(<= 40k single-row documents)

Batch processing

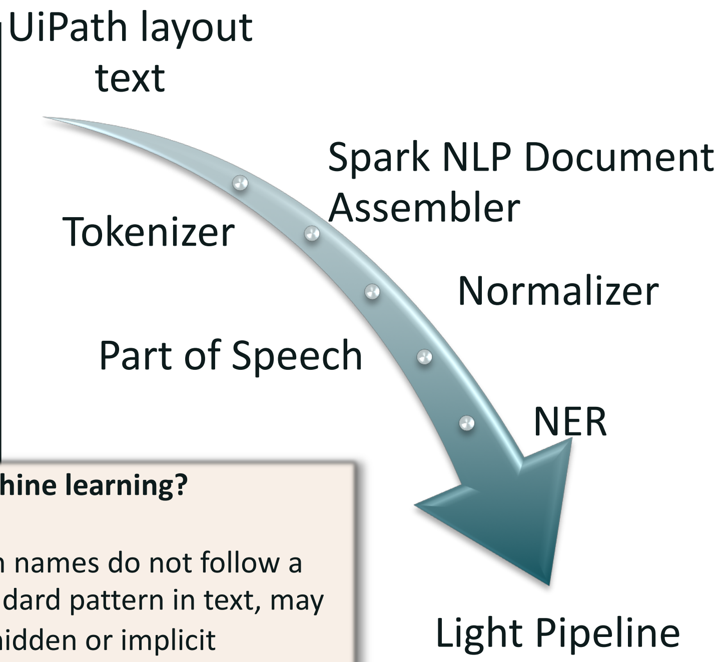
↓

Spark Pipelines

Only open source distributed NLP library, for large batches or very large documents

OUR USE CASE: SPARK NLP COMPANY NAME PIPELINE

```
def props(): Props = Props(new FirmNameExtractor)

val document: DocumentAssembler = new DocumentAssembler()
  .setInputCol("text")
  .setOutputCol("sentence")

val token: Tokenizer = new Tokenizer()
  .setInputCols("sentence")
  .setOutputCol("token")

val normalizer: Normalizer = new Normalizer()
  .setInputCols("token")
  .setOutputCol("normal")

val pos: PerceptronModel = PerceptronModel.pretrained()
  .setInputCols(Array("sentence", "normal"))
  .setOutputCol("pos")

val ner: NerCrfModel = NerCrfModel.pretrained()
```

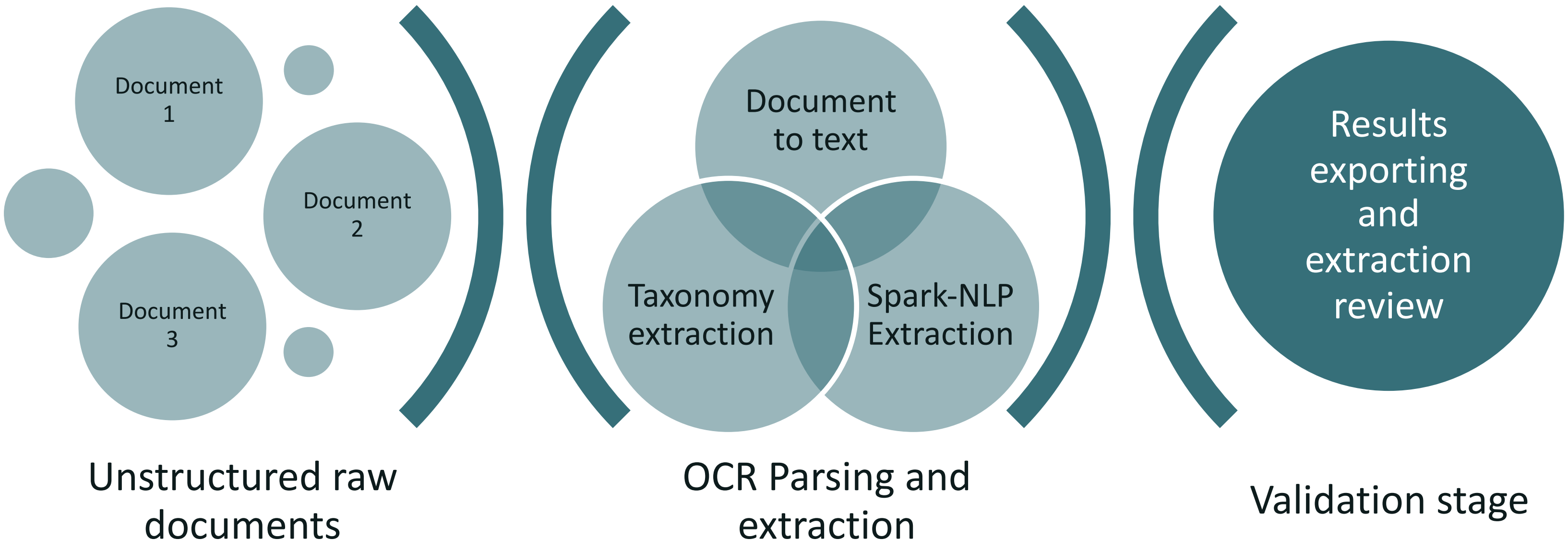


Why machine learning?

1. Firm names do not follow a standard pattern in text, may be hidden or implicit

2. Trained on domain specific language allows accurate and in-scope identification

SOLUTION OVERVIEW



USING SPARK NLP

- Homepage: https://nlp.johnsnowlabs.com
  – Getting Started, Documentation, Examples, Videos, Blogs
  – Join the Slack Community
- GitHub: https://github.com/johnsnowlabs/spark-nlp
  – Open Issues & Feature Requests
  – Contribute!
- The library has Scala and Python 2 & 3 API's
- Get directly from maven-central or spark-packages
- Tested on all Spark 2.x versions

THANK YOU!

Saif Addin Ellafi
saif@johnsnowlabs.com

Paul Parau
paul.parau@uipath.com