

Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI

Document classification:

- Document processing
- Web scraping
- GDPR and other laws
- Poor quality
- Handwritten text
- Different templates

From: Amre g@alcatel.com <amre.g@perfileslogisticas.com>
Sent: Monday, June 13, 2017 12:18 PM
To: Erika B. Lucas @ al
Cc: lucas.g@perfileslogisticas.com; lewis.arsenio@perfileslogisticas.com
Subject: RE: NESTIT 18631287
Attachments: 2017_A 18631287.pdf

Dear Sr/Madam,

AIRWAY BILL NO. 0302586420

CUSTOMER: PLASTIC INJECTION ROTARY MOULD

CUSTOMER REF: SA30800

SUPPLIER: VITE PRO PARTS S.R.L.S

PO: E1706599325

ENAVY: 2017/06/13 09:01:04

TAKE: 18631287

CREDIT USED PAY ADVANCE REMUNING OF USD 13,000

TOTAL FOR ~~30000~~ PARABLE TO SPURIN LOGISTICS LIMITED

Request
Reference

--- Original Message ---
From: Erika B. Lucas @ al <erika.b@alcatel.com>
Sent: Tuesday, June 13, 2017 5:42 AM
To: Amre g@perfileslogisticas.com
Subject: Forwarding: 2017/06/13 09:01:04

Air Way Bill

NO.	DESCRIPTION	WEIGHT
1.2	X 18631287	X 2.00
1.3	X 18631287	X 2.00
1.4	X 18631287	X 2.00
1.5	X 18631287	X 2.00
1.6	X 18631287	X 2.00
1.7	X 18631287	X 2.00
1.8	X 18631287	X 2.00
1.9	X 18631287	X 2.00
1.10	X 18631287	X 2.00
1.11	X 18631287	X 2.00
1.12	X 18631287	X 2.00
1.13	X 18631287	X 2.00
1.14	X 18631287	X 2.00
1.15	X 18631287	X 2.00
1.16	X 18631287	X 2.00
1.17	X 18631287	X 2.00
1.18	X 18631287	X 2.00
1.19	X 18631287	X 2.00
1.20	X 18631287	X 2.00
1.21	X 18631287	X 2.00
1.22	X 18631287	X 2.00
1.23	X 18631287	X 2.00
1.24	X 18631287	X 2.00
1.25	X 18631287	X 2.00
1.26	X 18631287	X 2.00
1.27	X 18631287	X 2.00
1.28	X 18631287	X 2.00
1.29	X 18631287	X 2.00
1.30	X 18631287	X 2.00
1.31	X 18631287	X 2.00
1.32	X 18631287	X 2.00
1.33	X 18631287	X 2.00
1.34	X 18631287	X 2.00
1.35	X 18631287	X 2.00
1.36	X 18631287	X 2.00
1.37	X 18631287	X 2.00
1.38	X 18631287	X 2.00
1.39	X 18631287	X 2.00
1.40	X 18631287	X 2.00
1.41	X 18631287	X 2.00
1.42	X 18631287	X 2.00
1.43	X 18631287	X 2.00
1.44	X 18631287	X 2.00
1.45	X 18631287	X 2.00
1.46	X 18631287	X 2.00
1.47	X 18631287	X 2.00
1.48	X 18631287	X 2.00
1.49	X 18631287	X 2.00
1.50	X 18631287	X 2.00

PERSONAL DATA OF THE EMPLOYEE
Last Name: ...
First Name: ...
Middle Name: ...
Date of Birth: ...
Sex: ...
Nationality: ...
Identification Number: ...

EMPLOYEE'S CURRENT CONTRACT INFORMATION
DATE OF EMPLOYMENT: ...
NAME: ...
PERSONAL DETAILS: ...
DATE OF BIRTH: ...
EMPLOYMENT: ...
REFERENCE: ...
TERMS OF PAYMENT: ...

TERMS AND CONDITIONS OF EMPLOYMENT
WORKER'S OBLIGATION: ...
EMPLOYEE'S OBLIGATION: ...
NOTE: CONTRACT IS SUBJECT TO PROVISIONS OF THE FOLLOWING DOCUMENTS:
1. ...
2. ...
3. ...
4. ...
5. ...
6. ...
7. ...
8. ...
9. ...
10. ...
11. ...
12. ...
13. ...
14. ...
15. ...
16. ...
17. ...
18. ...
19. ...
20. ...
21. ...
22. ...
23. ...
24. ...
25. ...
26. ...
27. ...
28. ...
29. ...
30. ...
31. ...
32. ...
33. ...
34. ...
35. ...
36. ...
37. ...
38. ...
39. ...
40. ...
41. ...
42. ...
43. ...
44. ...
45. ...
46. ...
47. ...
48. ...
49. ...
50. ...
51. ...
52. ...
53. ...
54. ...
55. ...
56. ...
57. ...
58. ...
59. ...
60. ...
61. ...
62. ...
63. ...
64. ...
65. ...
66. ...
67. ...
68. ...
69. ...
70. ...
71. ...
72. ...
73. ...
74. ...
75. ...
76. ...
77. ...
78. ...
79. ...
80. ...
81. ...
82. ...
83. ...
84. ...
85. ...
86. ...
87. ...
88. ...
89. ...
90. ...
91. ...
92. ...
93. ...
94. ...
95. ...
96. ...
97. ...
98. ...
99. ...
100. ...

NER or named entities recognition:

- OCR engines
- Document formats
- General or document specific (e.g. date)
- Columns and other formatting
- Abbreviations

Invoice No.	106566	
Tax Date	14/11/201	
Our Reference	Terms	Due Dat
51521	30 Days...	14/12/2016

Amount	VAT AMT	VAT%
1,000.00	160.00	16.00%

Already Forwarded Sent

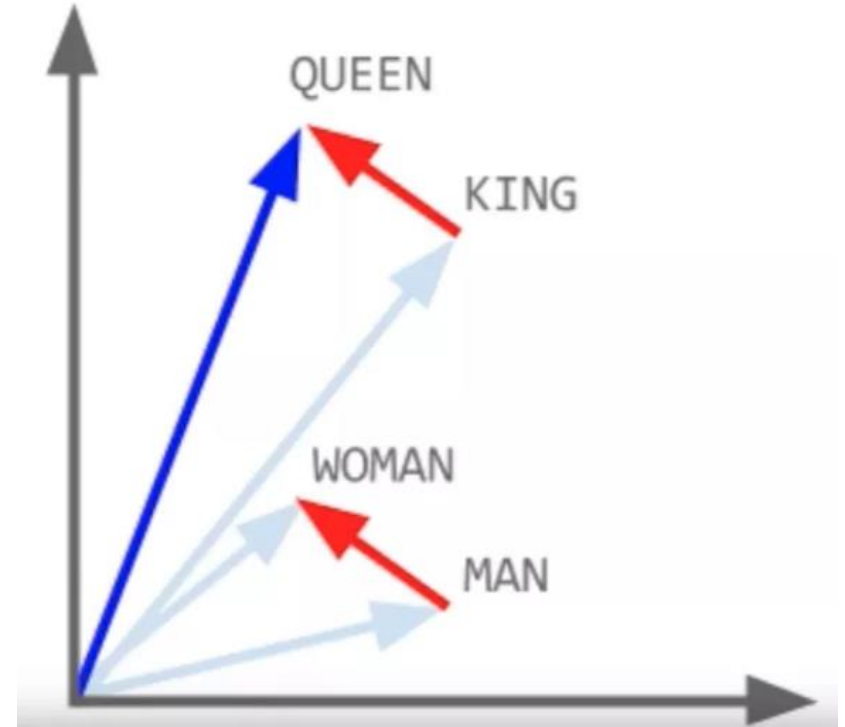
Marks	Nu. of Articles	
AD24 4811323	1x ⁵ / ₁₀	STC
EAL: 01227	CNER	THE
DM: 122444		PAP

INDUSTRIAL & DOMESTIC BRG
MANILLA TWINE, POLYTHENE S
PLASTIC & PAPER STATIONERY PR
STRAWS, HOUSEHOLD PLASTIC PR

... No.	Way Sr. No.	
	N 2	
DESCRIPTION	UNIT	QTY.
STRAPPING 16MM BASKET.	PCS	110

Typical NLP process:

- OCR
- Cleaning (symbols, stop-words, etc)
- Tokenization (sentence and word)
- Embeddings (BOW or word2vec)



Issues with the typical process:

- Missed information (color, style, template)
- Tokenization of German-like languages
- Arabic and Hebrew languages
- Table columns
- Contextual labeling is expensive



The image shows a screenshot of a financial spreadsheet titled 'ABWASSERLOGG' Forecast in Millions (Planned Budget Spreadsheet) for the year 2010. The spreadsheet has multiple columns with headers like 'Actual Budget', 'Forecast', 'Variance', 'Budget', 'Actual', 'Variance', and 'Budget'. The rows are organized into sections with red headers, such as 'Operating Expenses', 'Capital Expenses', and 'Total Expenses'. The data is presented in a grid format with various numerical values.



בְּרוּךְ אַתָּה יְהוָה אֱלֹהֵינוּ
מֶלֶךְ-הָעוֹלָם
הַמוֹצִיא לָהֶם מִן-הָאָרֶץ
אָמֵן

polyglot:

- 40 models
- Trained on Wiki
- Semi-supervised learning
- Embeddings plus Shallow Neural Network

Language	Sentence	Translation
English	Simien was traded from the Heat along with Antoine Walker and Michael Doleac to the Minnesota Timberwolves on October 24, 2007, for Ricky Davis and Mark Blount .	-
Hungarian	Dimitri beszélt egy utat Rómába .	Dimitri talked about a trip to Rome .
Spanish	Pešek nació en Praga y estudió dirección de orquesta piano en la Academia de Artes allí, con Václav Smetacek	Pešek born in Prague and studied orchestra direction piano at the Academy of Arts there, Vaclav Smetacek .
Russian	Уроженец Рио-де-Жанейро, Хосе Родригес Триндади использовал сокращенную форму как своим сценическим псевдонимом.	A native of Rio de Janeiro , Jose Rodriguez Trindade used as a shortened form of his stage name.
Korean	도널드 스미스 는 1976 년에 자이드 압둘 아지즈 에 그의 이름을 바꿨다 .	Donald Smith in 1976 changed his name to Zaid Abdul Aziz .
French	La France veut satisfaire à ses engagements envers l' Union européenne .	France wants to meet its commitments to the European Union .
Turkish	Erdoğan, Türkiye 'de twitter yasaklandı.	Erdogan banned twitter in Turkey .
Arabic	قال غاريث بيل ، نجم فريق ريال مدريد ، إن الفوز بدوري أبطال أوروبا سيظل معه الى الأبد، وذلك بعدما ساعد هدفه النادي الملكي على هزيمة نادي اتليتيكو مدريد في Spain .	Said Gareth Bell , the star of Real Madrid , said that winning the Champions League will remain with him "forever", and that after his goal helped the club to defeat Atletico Madrid in Spain .
Indonesian	Rendjambe meninggal dalam keadaan tidak jelas , yang mengakibatkan kerusuhan oleh pendukung oposisi marah di Port Gentil - dan Libreville .	Rendjambe died in unclear circumstances, which resulted in riots by angry opposition supporters in Port Gentil - and Libreville .
Chinese	新华社 25 日报道 援引 新疆 公安厅 。	Xinhua News Agency report quoted the 25th Xinjiang Public Security Department .
Greek	Ἐνῆλ ἔφερε τα Γκούτιους κάτω από το λόφους ανατολικά του Τίγρη, να φέρει το θάνατο σε ολόκληρη τη Μεσοποταμία.	Enlil brought the Gutians down from the hills east of the Tigris , to bring death throughout Mesopotamia .

spaCy:

- Subword embeddings
- Affixes
- Convolutions with residuals
- No Bi-LSTM
- Custom models
- Active Learning

displaCy Named Entity Visualizer

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun, now the co-founder and CEO

Model

English - en_core_web_sm (v2.0.0)

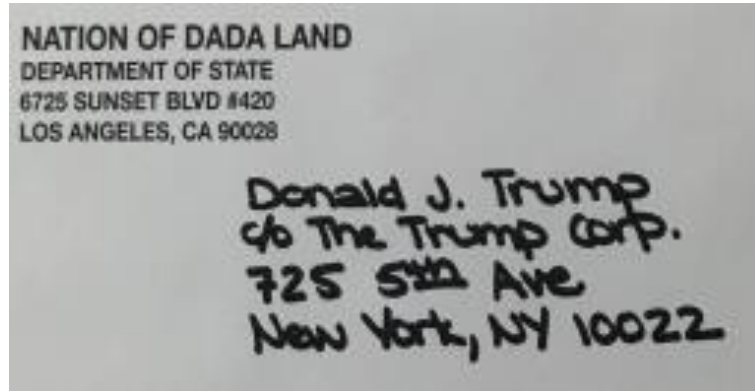
Entity labels (select all)

- | | | | |
|--|--|---|---|
| <input checked="" type="checkbox"/> PERSON | <input checked="" type="checkbox"/> NORP | <input type="checkbox"/> FACILITY | <input checked="" type="checkbox"/> ORG |
| <input checked="" type="checkbox"/> GPE | <input checked="" type="checkbox"/> LOC | <input checked="" type="checkbox"/> PRODUCT | <input type="checkbox"/> EVENT |
| <input type="checkbox"/> WORK OF ART | <input type="checkbox"/> LANGUAGE | <input checked="" type="checkbox"/> DATE | <input type="checkbox"/> TIME |
| <input type="checkbox"/> PERCENT | <input type="checkbox"/> MONEY | <input type="checkbox"/> QUANTITY | <input type="checkbox"/> ORDINAL |
| <input type="checkbox"/> CARDINAL | | | |

When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

Ready-Solution problems:

- Missed capital letters info
- Disambiguations (Bank or Bank, numbers)
- Work for generic terms, not for document specific
- Only contextual labels
- Only basic ENs (PER, ORG, LOC)
- Only few languages working good



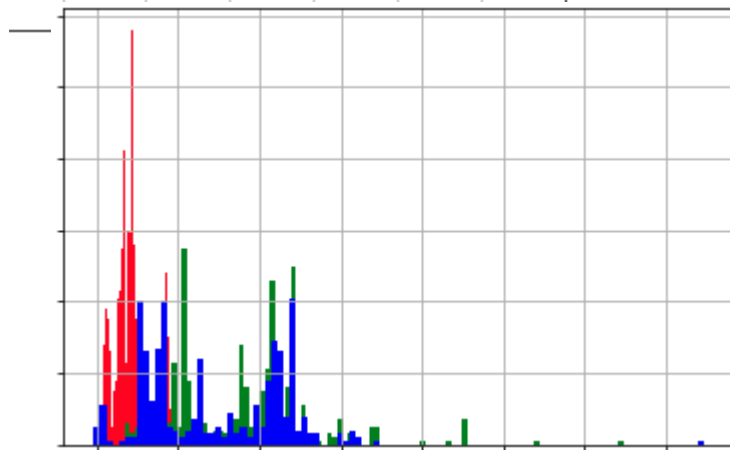
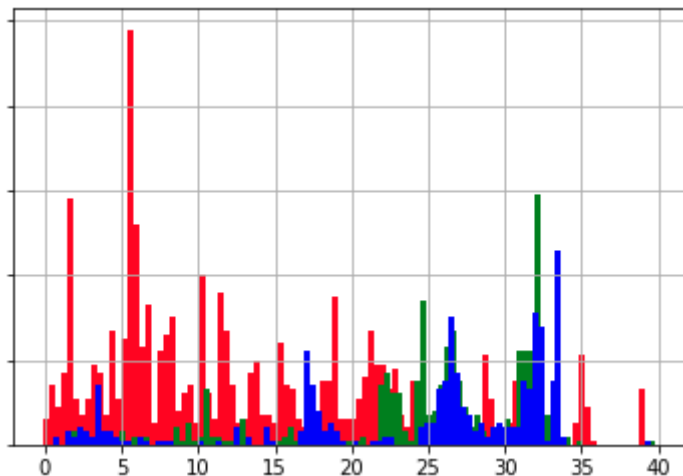
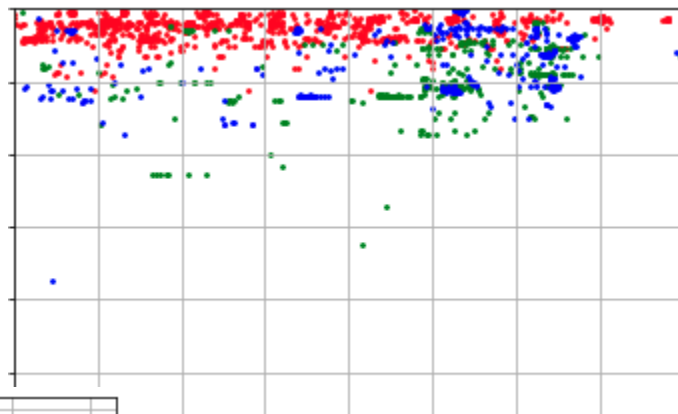
How we do it - Document classification:

- Typical BOW model
- CV analysis using CNN or Shallow NN
- 70 + 70 = 93



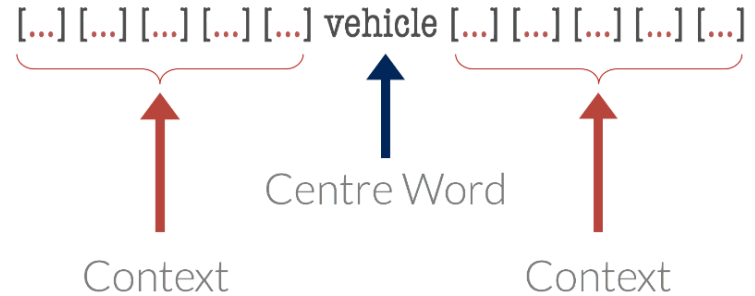
How we do it - Location model:

- statistical model
for every EN location on the doc
- 62% accuracy out of 22 docs



How we do it - Word model:

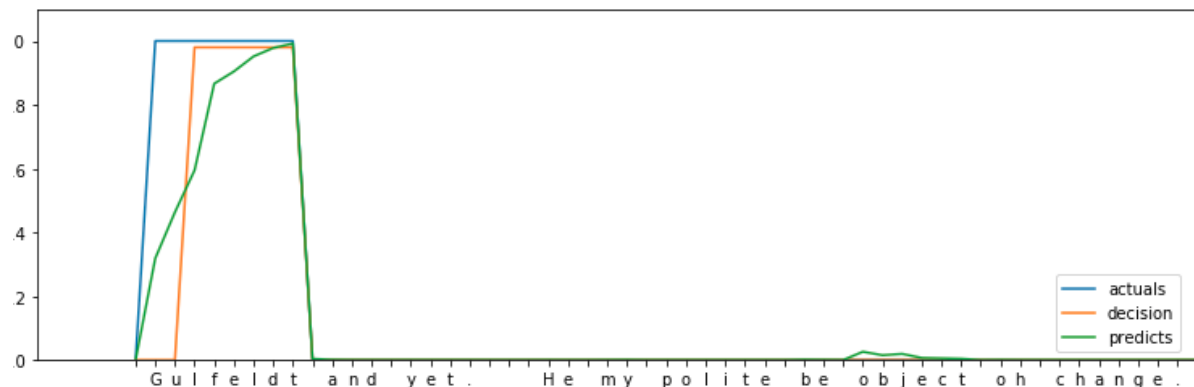
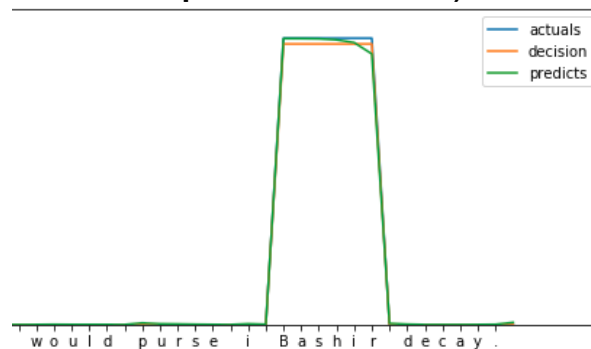
- word2vec embeddings
- Horizontal and vertical tokenizer
- 1D CNN + Bi-directional LSTM
- 80% accuracy out of 12 ENs



Order-Date	Region	Repr	Item	Units	Unit-Cost	Total
1/23/10	Ontario	Kivell	Binder	50	19.99	999.50
2/9/10	Ontario	Jardine	Pencil	36	4.99	179.64
2/26/10	Ontario	Gill	Pen	27	19.99	539.73
3/15/10	Alberta	Sorvino	Pencil	56	2.99	167.44
4/1/10	Quebec	Jones	Binder	60	4.99	299.40
4/18/10	Ontario	Andrews	Pencil	75	1.99	149.25
5/5/10	Ontario	Jardine	Pencil	90	4.99	449.10
5/22/10	Alberta	Thompson	Pencil	32	1.99	63.68

How we do it - Character model:

- Feature engineering (style, capitalization, color, punctuations)
- Bi-directional LSTM
- Hysteresis
- Cheap automatic labeling
- Data Anonymization



How we do it - stacking models:

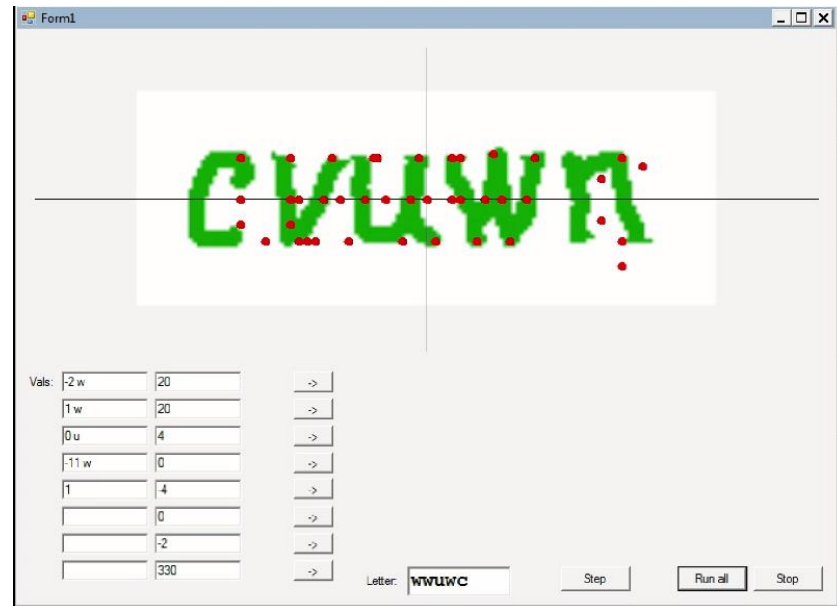
- RE filters
- Joining contexts (street and number)
- Entity linking (removing disambiguation)
- Stacking horizontally as features
- Use Bayesian rule
- Test accuracy over 95%

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$



30 years ago: ML model to play checkers (thanks to my father and Martin Gardner)

(picture from www.instructables.com)



in .NET

Jh

that has tr-like operators, optimizer and

even automatic hyper-param tuner

Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI